CURSO PROGRAMADO DE ESTADÍSTICA

Asistido con EXCEL y POWER POINT

MANUEL PONTIGO ALVARADO.

Ficha de catalogación.
310
P816c Pontigo Alvarado, Manuel
Curso Programado de Estadística. – 1ª. ed. –
Cartago: M. Pontigo A., 2006. 250 p.
Contiene un disco compacto.
ISBN 978-9968 9634−3-5.
1. REGRESIÓN LINEAL 2. ESTADÍSTICA NO PARAMÉTRICA 3. MUESTREO. 4 ANDEVA. 5 REGRESIÓN.

Cartago Costa Rica. Teléfono 552-3618.
e-mail: mpontigo@itcr.ac.cr
ISBN:978- 9968-9634-3-5.

Impreso en Costa Rica, diciembre 2006.

Para:

Mi amada y paciente esposa Delfina.

Nuestros hijos: Manuel Esteban; Julio Alberto; Carlos Arturo; Marcelo.

Especialmente para nuestras nietas y nietos con la ilusión de que en cualquier lugar y momento tengan siempre presentes sus raíces.

PREFACIO.

El Curso Programado de Estadística se ha preparado para profesionales que requieren resolver problemas que se les presentan en su labor cotidiana como ocurre, por ejemplo, con:

- El ingeniero industrial que quiere probar la eficiencia de las cuatro líneas de producción de la fábrica que dirige;
- El encargado del marketing de una empresa dedicada a fabricar jaleas y mermeladas;
- El ingeniero agrónomo que quiere valorar el efecto de técnicas orgánicas en cultivos tradicionales;
- El médico de un hospital de la Seguridad Social que quiere conocer la incidencia del papiloma humano en una zona de la costa;
- En graduando de una universidad que requiere analizar los datos recopilados en su práctica de especialidad para optar al grado académico universitario.

Y otros más.

Es precisamente de esta fuente, que el autor ha recopilado información para preparar éste curso de estadística que utiliza las herramientas simples que ofrece la computadora: el editor de textos WORD, la hoja electrónica EXCEL y el carrusel de diapositivas POWER POINT (PP) operados de una manera simple e interactiva.

La estructura del administrador de diapositivas define la forma en que se preparan y refieren los temas pues cada una de ellas determina una etiqueta en el WORD y EXCEL manteniendo, por este medio, la concordancia.

El curso está preparado para que el estudiante, mediante el EXCEL con la Hoja *Generador* cree conjuntos de datos particularizado, mismo que deberá analizar siguiendo el secuenciado de las diapositivas o del documento impreso colocando los resultados en una presentación de PP abierta con instrucciones precisas en lo tocante a cuadros y respuestas.

El curso es integrado y va acompañado de Disquete Compacto (CD). El archivo de soporte programático que también puede encontrar y descargar desde la dirección:

http://www.itcr.ac.cr/escuelas/biblioteca

El curso puede ser utilizado como soporte al estudio de cursos de *Métodos Estadísticos* impartidos en las diferentes carreras universitarias.

Manuel Pontigo Alvarado 2006.

CONTENIDO:

1 Las Distribuciones en Estadística.

Los archivos para esta sección son:
E01_Distribuciones_W01.doc;
E01_Distribuciones_P01.ppt;
E01_Distribuciones_X01.xls.

1.1 Menú.

1.2 La era de la información.

Nunca la humanidad ha generado tantas *Noticias* y *Datos*.

La tecnología ha desarrollado aparatos que reciben, procesan y emiten señales de manera automatizada.

Tal es la magnitud de estas señales que se ha creado toda una *Teoría de la Información*:

Encargada de relacionar el medio, el canal y el código con los cuales se trasmite la información.

Los medios más usuales a la información son: La INTERNET (*Red Mundial de Computadoras*); La INTRANET (*Red Local de Computadoras*); El Radio y la Televisión; los medios gráficos como diarios, revistas y libros.

1.3 La Informática.

Es el conjunto de técnicas que permiten procesar datos dando resultados. Un proceso que es recomendable realizar mediante *ordenadores*.

Los *Ordenadores* también conocidos como *Computadoras* son las herramientas que han propiciado en gran medida el "*BUM*" *INFORMATIVO e INFORMÁTICO* que se está viviendo.

Con estos aparatos se captura, procesa y emite información con sentido, generalmente económico.

La pregunta que surge:

¿Por qué no se analizar la información? Si los mismos aparatos tienen incorporadas las herramientas.

1.4 Estadística: viene de estado.

Tiene dos acepciones:

- Sucesión numérica de datos sobre un tema con los que se pretende caracterizar a una población.
- Y, Ciencia cuyo propósito es la recopilación, agrupamiento y tratamiento de datos numéricos sobre fenómenos naturales o sociales, y el método que se usa.

Se puede agregar:

Con el objeto de facilitar al estudioso, el análisis, la síntesis y las recomendaciones que han sido el propósito de esa recopilación de datos.

Esto es, dar a la ***INFORMÁTICA*** un sentido analítico.

1.5 La recopilación y el almacenamiento de datos.

Cuando se investiga una población de individuos en una o más características, la lógica nos dice que al menos deben estudiarse algunos individuos que *representen* al grupo. A estos individuos que llamaremos *muestras,* se les toman una o varias medidas conocidas como *datos,* importantes para los fines de la investigación. El conjunto de datos define una variable. Al origen de una o más variables se les conoce como *Observación*.

La modernidad nos permite que las observaciones sean capturadas, almacenadas y tratadas en un *Ordenador*. En este curso se agregará:

Y analizarlas con las herramientas estadísticas de uso general que el mismo ordenador opera.

1.6 El propósito de la información.

En toda investigación, sea que se trate de recopilar información simple o muy complicada y costosa, el *Proyecto* debe establecerse de manera *clara, precisa y concisa*:

En o los propósitos que se persiguen con la investigación.

También llamados *Objetivos del Proyecto*. Estos son puntos en el horizonte del estudio a los que se llegará después de que la información se haya ordenado, procesado y analizado, esto es, *Informatizado.*

La claridad, precisión y concisión de los *Propósitos de la Investigación* propician recomendaciones cuyos resultados son previsibles con probabilidad conocida.

Con seguridad, estos propósitos aunque novedosos posiblemente no serán únicos, por tanto:

"No hay nada nuevo bajo el sol" simplemente una forma diferente de estudiarlos.

1.7 Análisis de la experiencia humana.

Frase del acervo popular que nos dice qué, por muy novedoso que suene "nuestro proyecto" alguien, en algún lugar, ya lo llevó a cabo o al menos hizo algo parecido.

Esto no debe quitarnos el ánimo, la ciencia avanza en un ciclo interminable de pruebas de acierto y error. Es posible que algunas circunstancias de "nuestro ensayo" provoquen diferencias o hagan evidentes errores cometidos en investigaciones similares.

Por esto, es indispensable enmarcar con precisión los *Propósitos de la Investigación*. Proceso que se conoce como *Marco Teórico* que incluye la *Revisión Bibliográfica.*

1.8 Y el método que se usará

Frase en la definición de la Ciencia Estadística de implicaciones trascendentales.

De poco sirve la experiencia humana, al menos en la investigación científica, si las recomendaciones de una experiencia no tienen bases creíbles, y mejor sí son ciertas.

Esta credibilidad es otorgada al proyecto sí los métodos de:

Aplicación de Estímulos; Manipulación de los Sujetos de la Experiencia; Obtención de Observaciones; Tratamiento Informático de los Datos; Método Estadístico de Análisis

Y cualquier otra manipulación directa o indirecta de los sujetos experimentados o explorados y sus observaciones están exhaustivamente descritos y correctamente aplicados.

1.9 Análisis de resultados.

Una vez que los datos se han recopilado se entra al proceso de:

Analizar los Resultados.

Cuando el proyecto ha considerado valorar los resultados mediante *Técnicas Estadísticas de Análisis*, sea mediante *Técnicas de Exploración* o mediante *Experimentos Planificados*, debe hacerse con base en el método que se ha determinado usar antes de que se iniciara la recopilación de la información o el tratamiento de las unidades experimentales.

Las *Técnicas de Análisis Estadísticos* ofrecen resultados objetivos y con *probabilidades totalmente determinadas* para que el investigador haga *recomendaciones* que le den la seguridad que él necesita, conociendo exactamente el *riesgo* que correrá el usuario de los resultados del proyecto.

1.10 Conclusión y recomendación.

Para que el proyecto sea útil, el *Análisis de los Resultados* debe derivar en:

Conclusiones y Recomendaciones.

En todo el proyecto de investigación se han considerado directa o indirectamente dos posiciones bien definidas:

- La del Investigador cuyos fines usualmente son prácticos y más de las veces económicos;
- Y las de los usuarios de los productos resultantes de la investigación.

Las Técnicas Estadísticas consideran estas posiciones mediante las probabilidades:

- De confianza que tiene el investigador de recomendar las conclusiones del proyecto;
- De riesgo que corre el "comprador" de los productos resultantes de la investigación.

También conocidos respectivamente como ***error del fabricante*** y ***error del consumidor***.

1.11 Puntualización.

El estudiante se habrá percatado que desde la diapositiva 1.6 se esquematizan los fundamentos de un proyecto de investigación en el ámbito del método científico:

- Introducción, que sirve para delinear el proyecto;
- Las Propósitos del Proyecto, en el que se establecen los objetivos que se persiguen con el proyecto;
- La Demarcación del Proyecto, una recopilación de lo que se ha hecho y se está haciendo sobre el proyecto.
- Determinación de los métodos de operación, inspección del material experimental y del método analítico de los resultados.
- Análisis de los resultados usando el método definido en el apartado anterior;
- Conclusiones y Recomendaciones, síntesis y prospectiva del proyecto.

En todo caso, esta normativa prevalecerá para todos los problemas que se traten en el curso.

1.12 Dos preguntas esenciales.

Puesto que se van a utilizar *Técnicas Estadísticas* en el análisis de proyectos, se estará hablando de conjuntos de individuos a los que se les toman datos numéricos. O sea, que en esencia se tratarán conjuntos de *observaciones* de números.

A la estadística interesan los conjuntos de datos, y más específicamente sus *Distribuciones*, respondiendo en todo momento a las preguntas:

- ¿De que *tipo* es la distribución de los datos?
- Y, ¿Con qué *distribución estadística* aproximo al conjunto de datos en cuestión?

Las dos interrogantes se deberán responder antes de dar inicio al *proyecto*.

1.13 Dos tipos de distribuciones.

La característica que define al tipo de dato, o considerada en su conjunto es *La Variable*, está, caracterizará a la distribución que el conjunto de datos determina.

- Si la distancia entre un dato con el teórico precedente es tan pequeña que resulta tan insignificante que puede graficarse con una línea, entonces *La Distribución de Datos será de tipo Continuo*. Formalizando, son *Variables* que pertenecen al menos al conjunto de los números *racionales* o de *razones* y más específicamente al conjunto de los números *reales*.
- Si la distancia entre un dato con el teórico precedente establece un espacio al menos de una unidad, entonces *La Distribución de Datos será de tipo Discreto*. Formalizando, son *Variables* que pertenecen al conjunto de los números Naturales.
- Si un dato es característico de una cualidad puede tomar dos valores, dígase 1 si el individuo posee la cualidad y 0 si no la posee, entonces *La Distribución de Datos será de tipo Cualitativa.*

1.14 Las distribuciones de tipo continuo.

Debe puntualizarse:

EL TIPO DE DISTRIBUCIÓN DE LOS DATOS ES INDISPENSABLE PARA APROXIMARSE A POBLACIÓN QUE LOS ORIGINA USANDO TÉCNICAS ESTADÍSTICAS.

Poblaciones a las que se les toman datos métricos —kilos, metros, litros, libras, onzas...— para ser caracterizadas deberán ser aproximadas mediante distribuciones de tipo *Continuo*.

Para los fines de este curso interesa en especial la denominada:

DISTRIBUCIÓN NORMAL

Y más específicamente:

LA DISTRIBUCIÓN NORMAL ESTÁNDAR

Que ha dado origen a poderosas herramientas de análisis y proyección.

1.15 Las distribuciones de tipo discreto.

Hay *variables* que se utilizan para valorar datos que por su naturaleza varían como mínimo de unidad en unidad, por ejemplo: la cantidad de huevos que pone una gallina en un año; el número de Gansos Canadienses que llegan a una laguna del centro de México a pasar el invierno austral; la cantidad de semillas que afloran de 1.000 semillas sembradas.

Estos valores provenientes de conteos son de tratamiento estadístico incómodo, la mayoría de las veces se aproximan mediante distribuciones continuas haciendo salvedades de continuidad.

En todo caso, este tipo de distribuciones establece condicionante que deben tomarse en cuenta para su tratamiento estadístico.

1.16 Las distribuciones de tipo cualitativo.

Estas distribuciones de datos reflejan pocos sucesos, la que más interesa al curso es la resultante de dos posibles resultados:

- Que el individuo estudiado cumpla una cualidad, entonces se el valor del dato será un uno (1);
- Que el individuo estudiado no cumpla esa cualidad, entonces el valor del dato será un cero (0).

De esta manera la distribución de datos —para fines prácticos individuos que poseen o no la cualidad— podrá representarse con mediante dos columnas.

En este curso interesa en especial la *Distribución Binomial.*

1.17 Las distribuciones relativas.

Los tres tipos de distribuciones de datos pueden llevarse a valores relativos, esto es, transformarlas a números puros que permitan concluir, independientemente de las unidades en que se mide la variable.

Las unidades relativas más utilizadas son los porcentajes, por esto, no es extraño escuchar qué tal o cuál individuo pertenece a ***X*** porcentaje de la población.

Esta facilidad matemática permite comparar distribuciones de números puros como son las distribuciones estadísticas con las distribuciones relativas de los datos que se estudian en los proyectos de investigación.

Otra distribución de importancia es el *orden estadístico*, esto es, la asociación de un valor de la variable con la posición ordenada ascendentemente que ocupa.

1.18 Las Distribuciones de probabilidad.

Puesto que se pueden comparar distribuciones de números puros, una parte muy importante de la *Teoría Estadística* se ha enfocado a estudiar las distribuciones de de datos para poder diseñar modelos estadísticos que emulen correctamente los resultados.

Las distribuciones estadísticas tienen cualidades simples pero determinantes:

- El área que cubre la distribución es exactamente la unidad;
- Si se toma una sección de esa área, los tamaños de las secciones está perfectamente determinado;
- Esas secciones de área representan probabilidades.

1.19 Problema 1.1.

Mediante un ejemplo se ilustrará el concepto de distribuciones de datos.

Un inversor ha decidido colocar su dinero en un proyecto agrícola consistente en la reproducción de una especie de gallina con muchas posibilidades en la producción de carne.

El inversor confía en su socio, un zootecnista dedicado a la crianza de aves, pues sabe poco de estas, pero sí conoce de análisis de la producción y sobre todo, del flujo de dinero encargándose del análisis de los datos.

Ha considerado tres variables:

- El número de huevos que una gallina pone en un año, dato de tipo discreto;
- El peso de los huevos, dato de tipo continuo;
- El número de machos y hembras que nacieron de los huevos, dato de tipo cualitativo.

Por facilidad de análisis se iniciará con la variable de tipo continuo.

1.20 La Hoja Electrónica.

Un motor del avance de la especie humana ha sido la consecución de instrumentos que hagan la vida del hombre más cómoda y que son universalmente aceptados, excepto por individuos reactivos al cambio, como aquél ingeniero que prefiere la regla de cálculo a la computadora. Al que sus compañeros de generación ven como un "bicho raro".

El criterio del curso es abordar el veloz autobús de la modernidad utilizando las herramientas modernas de uso general para el tratamiento informático de datos. Con los riesgos que esto implica para el estudiante poco dedicado que supone, por simple pachorra mental, que la herramienta lleva implícita la base teórica que soportará las conclusiones y recomendaciones de los proyectos.

1.21 Entrando a la Hoja Electrónica (HE).

A partir de este punto el estudiante podrá acceder al análisis de los datos que han sido recopilados durante un año de 280 gallinas adultas sujetas a la explotación de huevos para reproducción.

El objetivo del ejemplo es mostrar al estudiante en que consisten las distribuciones de datos y las diferencias entre estas.

El proyecto trata de una muestra trivariada, esto es, una observación que consiste en la recopilación de los tres datos en una gallina, a saber:

- ***X***, El peso medio de los huevos;
- ***Y***, El número de huevos viables ovopositados en una año por la gallina;
- ***Z***, El número de machos que resultaron de la incubación de esos huevos.

En este momento es pertinente que el estudiante abra el libro EXCEL y genere su conjunto de datos.

1.22 El Intervalo de Clases.

Se iniciará el trabajo con la variable continua, el peso promedio de los huevos. Al ser un promedio, el resultado de una división, la variable se transforma en una variable continua o muy aproximadamente continua.

La técnica que se ha desarrollado para observar las distribuciones de datos consiste en establecer un determinado *número de clases*, entre 5 y 15 considerando intervalos igualmente distanciados que incluyan a todos los individuos de la población.

Después, de acuerdo a su valor, cada individuo se asignará a la clase correspondiente llevando un conteo que se acomodará en una tabla especialmente diseñada para el caso.

Una regla empírica nos dice que un indicador del tamaño de las clases se puede obtenerse dividiendo la <u>*Desviación Estándar*</u> entre 2 y 4. Después dividir el <u>*Rango*</u> por este número y eligiendo el número de clases, usualmente entre 7 y 21.

Se irán utilizando formulas y estadísticos que serán definidos en el momento oportuno, por el momento, se mostrarán las instrucciones de la HE y la fórmula.

1.23 El número de clases.

La Desviación Estándar es un estadístico que ofrece una idea de la variación de la población. Se identifica con una *s* y se obtiene en la HE mediante:

s = DESVEST(B12 : B291) = 14,3017

Una idea del intervalo de clases o tamaño de clases se obtiene dividiendo por 2 y por 4.

$$IC_{>} = \frac{s}{2} = \frac{14{,}3017}{2} = 7{,}15; \quad IC_{<} = \frac{s}{4} = \frac{14{,}3017}{4} = 3{,}58$$

Dividiendo el *Rango = Máximo − Mínimo* entre ambos Intervalos de Clase se obtendrán los números de clase extremos.

$$NC_{<} = \frac{r}{IC_{>}} = \frac{\text{MAX(\$B\$12 : \$B\$291)} - \text{MIN(\$B\$12 : \$B\$291)}}{7{,}15} = \frac{76{,}2}{7{,}15} = 11$$

$$NC_{>} = \frac{r}{IC_{<}} = \frac{76{,}2}{3{,}58} = 21$$

Se puede elegir un IC de manera que se puedan conseguir de 13 a 15 clases. Por ejemplo *IC = 6* gramos, un número entero.

1.24 Los límites de las clases.

Para tener una perspectiva de las frecuencias de los pesos de los huevos, se acostumbra acomodar a los datos por su magnitud en un determinado número de clases. Usualmente se elige el valor mínimo para que sea el límite superior de la primera clase. Esto es:

$LS_1 = Mínimo = 13{,}5$

Que acomodará en la columna 3 (Columna D de la HE) titulada como *Límite Superior.* Después calculará el *Límite Inferior de la clase 1* restando el *intervalo de Clase:*

$LI_1 = LS_1 - IC = 13{,}5 - 6 = 7{,}5$

Después se calcula el promedio de la primera clase:

$$\bar{x}_1 = \frac{LI_1 + LS_1}{2} = \frac{7{,}5 + 13{,}5}{2} = 10{,}5$$

A cada uno de los límites se les suma el intervalo de clase hasta que el máximo caiga en la última clase.

1.25 Rango de las clases.

El *Rango de las Clases* es una lista de límites de clases que determinará cuáles individuos, de acuerdo a los valores que presentaros, pertenecen a qué clase.

El proceso siguiente es efectuar el conteo. Inspeccionar a cada dato y asignarlo a la clase correspondiente (se le dejará a la computadora).

Al llegar al final de los datos, cada valor habrá sido asignado a una clase y el número de individuos en una clase específica determinará la *Frecuencia de la Clase.* Y en conjunto, se habrá encontrado la distribución de frecuencias.

LÍMITES DE CLASES		
Inferior	Medio	Superior
7,5	10,5	13,5
13,5	16,5	19,5
19,5	22,5	25,5
25,5	28,5	31,5
31,5	34,5	37,5
37,5	40,5	43,5
43,5	46,5	49,5
49,5	52,5	55,5
55,5	58,5	61,5
61,5	64,5	67,5
67,5	70,5	73,5
73,5	76,5	79,5
79,5	82,5	85,5
85,5	88,5	91,5
91,5	94,5	97,5

1.26 Cuadro o tabla de frecuencias.

El conteo lo realizará la HE mediante algoritmos internos. Tradicionalmente el conteo se hacía manualmente preferiblemente entre dos personas:

Una canta las cantidades;

Y otra las ubica en una tabla que se denomina tabla de conteo haciendo una rayita.

También se acostumbraba a señalar diferencias (por ejemplo restarle 0,1 al cada límite superior) entre el límite superior de una clase y el inferior de la siguiente para evitar equivocaciones en el

conteo. Es Obvio que para el programa de la HE no es importante, pues basta señalar por ejemplo con < en lugar de ≤ para hacer la diferencia.

También se dice que el límite real de la clase, en una variable continua es igual para los límites superiores y el inferior de la siguiente. Por este concepto de continuidad, no se hacen diferencias entre los límites de las clases. La importancia se verá al momento de usar las probabilidades de las distribuciones estadísticas.

A la HE se le solicita el conteo de la primera clase con la instrucción de contar condicionada:

$$LS_1 = \text{CONTAR.SI}(\$B\$12:\$B\$291;"<13{,}5") = 0$$

Para la segunda clase se agrega la instrucción para que se resten los conteos de las clases ya consideradas:

$$LS_2 = \text{CONTAR.SI}(\$B\$12:\$B\$291;"<19{,}5") - \text{SUMA}(\$D\$314:D314) = 3$$

La instrucción se copia a las hileras siguientes y se arregla el criterio de búsqueda.

La HE ubica a cada gallina de la muestra en la clase correspondiente al peso promedio de sus huevos se obtiene la tabla de frecuencias en donde, bajo la columna *Frecuencias* se refiere el número de individuos de la clase.

Como era de esperarse la suma de frecuencias es igual al número de individuos medidos:

$$n = \sum_{i=1}^{15} f_i = 0+3+3+15+23+24+41+ \\ +50+42+36+19+15+7+2+0 = 280$$

LÍMITES DE CLASES			Frecuencias
Inferior	Medio	Superior	Observadas
7,5	10,5	13,5	0
13,5	16,5	19,5	3
19,5	22,5	25,5	3
25,5	28,5	31,5	15
31,5	34,5	37,5	23
37,5	40,5	43,5	24
43,5	46,5	49,5	41
49,5	52,5	55,5	50
55,5	58,5	61,5	42
61,5	64,5	67,5	36
67,5	70,5	73,5	19
73,5	76,5	79,5	15
79,5	82,5	85,5	7
85,5	88,5	91,5	2
91,5	94,5	97,5	0
		Suma	280

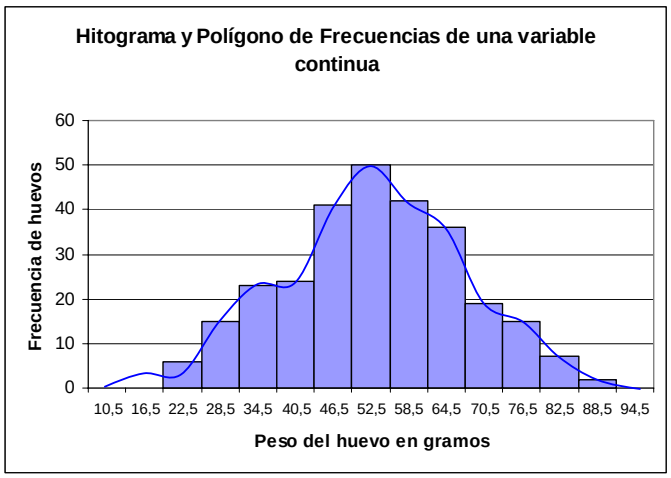
1.27 *Afinando el cuadro de frecuencias.*

Habrá notado que las columnas del límite inferior y el punto medio no se utilizaron. Sin embargo, debe considerar que cada clase forma un subconjunto acotado (limitado) con un punto central o punto medio o promedio que representa a todos los individuos de esa clase.

Tratándose de una variable continua, los valores mínimos y máximos son, apenas, un par de valores entre muchos posibles y para poder ser graficados se debe considerar el mínimo factible que sería cero y el máximo factible que sería un número desconocido.

Por otra parte, la manera de representar a una variable continua es mediante una línea sin interrupciones o por clases sin separaciones como se verá en los gráficos siguientes.

1.28 *Herramientas gráficas.*



En estadística se acostumbra usar estas figuras. *El Histograma* es el diagrama de barras, donde cada barra representa el peso relativo de la distribución. Entre más alta la barra más individuos hay en la clase y más peso relativo.

El polígono, aquí graficado como una línea suavizada, representa el área bajo una curva continua.

Más aún, se espera que las ondulaciones de suavicen aún más en una curva más uniforme, que se conseguiría si

se aumentará el número de observaciones.

1.29 Frecuencias relativas.

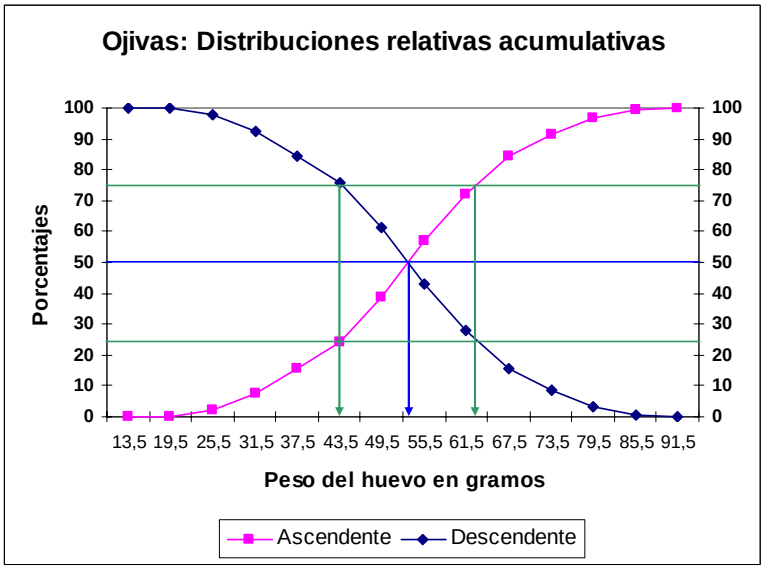
Dividiendo cada frecuencia entre el total de individuos se obtiene la proporción o porcentaje (si se multiplica por 100) de individuos en cada clase.

Si las frecuencias relativas se acumulan se obtienen las frecuencias acumulativas, útiles en procesos deductivos y para elaborar el siguiente gráfico.

LÍMITES DE CLASES			Fre. Relativas	Relativas Acumulativas	
Inferior	Medio	Superior	Observadas	Ascendente	Descendente
7,5	10,5	13,5	0,0	0,0	100,0
13,5	16,5	19,5	1,1	1,1	98,9
19,5	22,5	25,5	1,1	2,1	97,9
25,5	28,5	31,5	5,4	7,5	92,5
31,5	34,5	37,5	8,2	15,7	84,3
37,5	40,5	43,5	8,6	24,3	75,7
43,5	46,5	49,5	14,6	38,9	61,1
49,5	52,5	55,5	17,9	56,8	43,2
55,5	58,5	61,5	15,0	71,8	28,2
61,5	64,5	67,5	12,9	84,6	15,4
67,5	70,5	73,5	6,8	91,4	8,6
73,5	76,5	79,5	5,4	96,8	3,2
79,5	82,5	85,5	2,5	99,3	0,7
85,5	88,5	91,5	0,7	100,0	0,0

1.30 Las Ojivas o frecuencias acumulativas.

Las Ojivas son especialmente útiles para representar los *estadígrafos de orden*. Éstos son los que relacionan el número índice u ordinal con los valores de la variable.



El estadígrafo de orden que mejor se comprende es la *Medina*. Valor que divide a los datos en dos subconjuntos con los mismos elementos.

Está ubicada en la posición media de los estadígrafos de orden:

$$\frac{n+1}{2}=\frac{280+1}{2}=140{,}5$$

Esto es, el valor que presenta la observación 140. Sin la ayuda de la HE, los datos se debían ordenar a mano y ubicar la observación; a la HE se le solicita:

$\tilde{x}$ = MEDIANA(B12 : B291) = 53,2

1.31 Utilidad de las Ojivas.

Si la cantidad de observaciones es par, la mediana es el promedio del valor para el estadístico mediano x_{140} y el siguiente x_{141}. En el ejemplo

x_{140} = K.ESIMO.MENOR($B$12 : $B$291;140) = 53,1

Y

x_{141} = K.ESIMO.MENOR($B$12 : $B$291;141) = 53,3

Por tanto:

$$\tilde{x} = \frac{x_{140} + x_{141}}{2} = \frac{53{,}1 + 53{,}3}{2} = 53{,}2$$

La Mediana en la ojiva se identifica por ser la línea que parte de los valores de los márgenes en 50% y cae en el eje *x* sobre el valor 53,2.

De la misma manera se pueden obtener los cuartos o cuartiles y en general cualquier percentil mediante la fórmula, calculada para los cuartiles:

$$k_p = \frac{(n+1)P}{100};\ k_{25} = \frac{(280+1)25}{100} = 70{,}25;\ \ k_{75} = \frac{(280+1)75}{100} = 210{,}75$$

$\tilde{x}_{25}$ = CUARTIL($B$12 : $B$291;1) = 43,75;

$\tilde{x}_{75}$ = CUARTIL($B$12 : $B$291;3) = 63,025

Los valores para los cuartiles y en general para cualquier percentil se obtienen extrapolando pues, muchas veces, el número de orden no es entero. Por ejemplo, para el primer cuartil el número se encuentra en la posición k = 70,25, esto es, a un cuarto entre la variable x_{70} y la variable x_{71}. La fórmula para la extrapolación lineal es:

$$\tilde{x}_{25} = x_{70}\frac{70}{280} + x_{71}\left(1 - \frac{70}{280}\right) = 10{,}90 + 32{,}85 = 43{,}75$$

1.32 Variables estándar.

Una alternativa para obtener valores relativos es estandarizar las variables, esto es, dividir la diferencia entre un dato x_i con respecto al *Promedio* entre la *Desviación Estándar*.

$$z_i = \frac{x_i - \bar{x}}{s}$$

Esta variable **Z** posee unas características muy importantes en estadística, por el momento nos interesa saber que el promedio de las variables estandarizadas es 0 y que la desviación estándar es 1.

$$\bar{z} = \frac{\sum_{i=1}^{n} \frac{x_i - \bar{x}}{s}}{n} = 0;\ \ s^2 = \frac{\sum_{i=1}^{n} (z_i - \bar{z})^2}{n-1} = 1$$

1.33 La Distribución Normal Estándar.

Lo trascendente de esta variable **Z** es que existe una *Distribución de Probabilidad* ampliamente estudiada en el *Teoría Estadística* que posee media 0 y varianza 1. Que como de mencionó en la diapositiva 18, todas las probabilidades bajo el área bajo la curva están determinadas.

Entonces, si la distribución de datos estandarizada es similar a la distribución de probabilidad estadística, con esta se puede aproximar sin dificultad y efectuar estimaciones y proyecciones con probabilidades.

La *Distribución Normal Estándar* tiene forma de campana, tal que también se le conoce como campana de Gaus [Carl Friedrich Gauss (30 Abril 1777 – 23 Febrero 1855)]. Es simétrica y se

aproxima muy apropiadamente a variables biológicas, sociológicas, provenientes de procesos de fabricación y muchas otras de tipo continuo.

1.34 Los parámetros: La Media

Los *Parámetros* son valores que caracterizan de manera incompleta a las distribuciones de datos y por consiguiente a las poblaciones que les dieron origen.

Por el momento interesa la media, o valor medio definido por:

$$\bar{\bar{x}} = \frac{\sum_{i=1}^{m} f_i \bar{x}_i}{\sum_{i=1}^{m} f_i}$$

Fórmula para usar los datos de la tabla de frecuencias, y:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Para datos sin agrupar.

1.35 Los parámetros: La Varianza.

Valor que es un promedio ajustado de las desviaciones cuadráticas de las observaciones con respecto a la media, definida por:

$$s^2 = \frac{\sum_{i=1}^{m} f_i \left(\bar{x}_i - \bar{\bar{x}}\right)^2}{\left(\sum_{i=1}^{m} f_i\right) - 1}$$

Para datos agrupados en las tablas de frecuencias, y:

$$s^2 = \frac{\sum_{I=1}^{n} \left(x_i - \bar{x}\right)^2}{n-1} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]$$

Para datos individuales. Al momento se usarán las fórmulas para la tabla de frecuencias.

1.36 El Cálculo de media y varianza:

Suma de frecuencias:

$$n = \sum_{i=1}^{15} f_i = 0 + 3 + \ldots + 2 + 0 = 280$$

Suma Total:

$$T = \sum_{i=1}^{15} f_i \bar{x}_i = 0(10{,}5) + 3(16{,}5) + \ldots + 2(88{,}5) + 0(94{,}5) = 14.862{,}0$$

Promedio:

$$\bar{\bar{x}} = \frac{\sum_{i=1}^{15} f_i \bar{x}_i}{n} = \frac{14.862,0}{280} = 53,08$$

Suma de Cuadrados:

$$sc = \sum_{i=1}^{15} f_i (\bar{x}_i - \bar{\bar{x}})^2 =$$

$$= 0(10,5 - 53,08)^2 + 3(16,5 - 53,08)^2 + ... + 0(94,5 - 53,08)^2 = 57.902,27$$

La Varianza:

$$s^2 = \frac{SC}{n-1} = \frac{57.902,27}{280-1} = 207,54$$

La Desviación estándar:

$$s = \sqrt{s^2} = \sqrt{207,54} = 14,41$$

LÍMITES DE CLASES			Frecuencias		
Inferior	Medio	Superior	Observadas	f * xi	f(xi - xm)²
7,5	10,5	13,5	0	0,0	0,0
13,5	16,5	19,5	3	49,5	4.014,0
19,5	22,5	25,5	3	67,5	2.805,1
25,5	28,5	31,5	15	427,5	9.061,6
31,5	34,5	37,5	23	793,5	7.938,8
37,5	40,5	43,5	24	972,0	3.797,3
43,5	46,5	49,5	41	1906,5	1.774,4
49,5	52,5	55,5	50	2625,0	16,7
55,5	58,5	61,5	42	2457,0	1.234,5
61,5	64,5	67,5	36	2322,0	4.696,2
67,5	70,5	73,5	19	1339,5	5.766,6
73,5	76,5	79,5	15	1147,5	8.228,4
79,5	82,5	85,5	7	577,5	6.059,3
85,5	88,5	91,5	2	177,0	2.509,4
91,5	94,5	97,5	0	0,0	0,0

Estadísticos:			
n = suma frecuencias	280	Suma de cuadrados	57.902,27
Suma total	14.862,0	Varianza	207,54
Promedio	53,08	Desviación Estándar	14,41

1.37 Propiedades de la media.

La propiedad más importante del valor promedio es:

La suma de las desviaciones de las observaciones con respecto al promedio es cero;

$$D = \sum_{i=1}^{n} d_i = (x_1 - \bar{x}) + (x_2 - \bar{x}) + ... + (x_n - \bar{x}) = 0$$

Esta propiedad divide a la distribución de los datos en dos secciones con la misma probabilidad ,50% de valores inferiores al promedio y 50% superiores al promedio (la Mediana lo hace con las unidades de la muestra).

Otra implicación importante es que la suma de cuadrados de las desviaciones de las observaciones con respecto a la media en mínima.

$$D^2 = \sum_{i=1}^{n} d_i^2 = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_n - \bar{x})^2 = SC \downarrow$$

La *Media*, junto con la *Mediana* y la *Moda* son los tres parámetros de posición que se presentan al centro de las distribuciones.

1.38 Propiedades de la varianza.

El promedio ajustado de las desviaciones cuadráticas tiene la propiedad de ser la suma cuadrática mínima en una distribución. Al requerir del cálculo previo de la media, sus propiedades están sujetas a las propiedades de la media, por esto a la primera se le llama *Primer Momento* y a la segunda *Segundo Momento Muestrales*.

Por sí sola la varianza no indica valores útiles, al sacársele la raíz cuadrada se obtiene la *Desviación Estándar* que es un indicador de la variación de la población. Se espera que en el intervalo de más y menos una *Desviación Estándar* del promedio se ubiquen poco más o menos el 68% de los datos como se puede comprobar en la HE. Por tanto, cuando en trabajos de investigación se observe la expresión:

$$\bar{x} \pm s$$

Debe entenderse, para el caso del ejemplo:

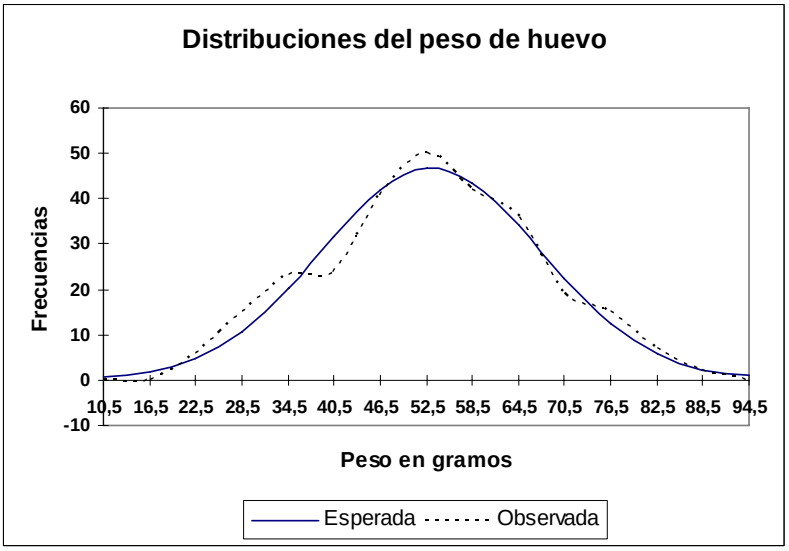
$$\text{Pr}obabilidad\{38{,}67 \geq \overline{X} \leq 67{,}48\} = 68\%$$

1.39 Ajustando la distribución esperada.

Como se apuntó, la importancia de las distribuciones de datos se centra en que puedan ser emuladas o aproximadas por alguna *Distribución Estadística de Probabilidad.*

En la HE se efectúa todo un proceso para crear el gráfico que compara las distribuciones *Observada y Teórica o Esperada.*

Se observa que hay semejanza entre las distribuciones de datos como lo confirma la prueba estadística utilizada de Chi-cuadrada que indica un 84,66% de que las frecuencias observadas y esperadas se parezcan. Estadísticamente suficiente para considerarlas iguales.



Como se ha mencionado, la teoría estadística utiliza las distribuciones esperadas de los datos para efectuar los análisis de los datos que permitirán sacar conclusiones y emitir recomendaciones con niveles de probabilidad conocidos. En este proceso, se utiliza la distribución de probabilidad que aproxime adecuadamente a la distribución de los datos.

En variables como la del problema, que es el resultado de una función fisiológica, se utiliza regularmente la distribución Normal de Probabilidad o la Distribución Normal Estándar.

El análisis estadístico se inicia probando la hipótesis nula:

Ho; X ~N(μ,σ)

Ha; No se distribuye Normal.

Que quiere decir: La variable aleatoria X se distribuye como una Norma, contra la hipótesis alternativa que niega que la variable se distribuya como una Normal.

Se inicia por definir la distribución de densidad Normal:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

La función acumulativa similar a la ascendente de la ojiva determina una probabilidad desde menos infinito hasta x. Por tanto, si x_{s1} es el valor del límite superior de la clase 1 se tendrá:

$$F(x_{s1}) = \frac{1}{14{,}41\sqrt{2\pi}} \int_{-\infty}^{13{,}5} e^{-\frac{(13{,}5-53{,}08)^2}{2(207{,}54)}} dx = 0{,}0030$$

La probabilidad desde menos infinito hasta 13,5 gramos del peso de huevos de gallina. Esta también será la probabilidad para el límite inferior de la clase 2. Ahora si x_{s2} es el límite superior de la clase 2:

$$F(x_{s2}) - F(x_{s1}) = \frac{1}{14{,}41\sqrt{2\pi}} \int_{-\infty}^{19{,}5} e^{-\frac{(19{,}5-53{,}08)^2}{2(207{,}54)}} dx - \frac{1}{14{,}41\sqrt{2\pi}} \int_{-\infty}^{13{,}5} e^{-\frac{(13{,}5-53{,}08)^2}{2(207{,}54)}} dx =$$

$$= 0{,}0099 - 0{,}0030 = 0{,}0069$$

Esto quiere decir que la clase que va de 13,5 a 19,5 gramos podría tener 0,69 huevos de cada 100. De la misma manera se obtienen las probabilidades para las clases restantes menos la última. Esta se calcula:

$$F(x_{15}) = 1 - \frac{1}{14{,}41\sqrt{2\pi}} \int_{-\infty}^{91{,}5} e^{-\frac{(91{,}5-53{,}08)^2}{2(207{,}54)}} dx = 1 - 0{,}9962 = 0{,}0038$$

La suma de las probabilidades de todas las clases deberá ser 1.

Al multiplicar la probabilidad de cada clase por el total de observaciones se obtiene la frecuencia esperada para cada clase.

LÍMITES DE CLASES			PROBABILIDADES			FRECUENCIAS		Chi-Cuadrada
Inferior	Medio	Superior	L. Inferior	L. Superior	La Clase	Esperada	Observada	Parcial
7,5	10,5	13,5	0,0000	0,0027	0,0027	0,8	0	0,759
13,5	16,5	19,5	0,0027	0,0091	0,0064	1,8	0	1,799
19,5	22,5	25,5	0,0091	0,0262	0,0171	4,8	6	0,306
25,5	28,5	31,5	0,0262	0,0645	0,0382	10,7	15	1,723
31,5	34,5	37,5	0,0645	0,1362	0,0718	20,1	23	0,420
37,5	40,5	43,5	0,1362	0,2494	0,1131	31,7	24	1,861
43,5	46,5	49,5	0,2494	0,3991	0,1498	41,9	41	0,021
49,5	52,5	55,5	0,3991	0,5657	0,1665	46,6	50	0,244
55,5	58,5	61,5	0,5657	0,7212	0,1555	43,5	42	0,054
61,5	64,5	67,5	0,7212	0,8431	0,1219	34,1	36	0,101
67,5	70,5	73,5	0,8431	0,9234	0,0803	22,5	19	0,539
73,5	76,5	79,5	0,9234	0,9678	0,0444	12,4	15	0,530
79,5	82,5	85,5	0,9678	0,9884	0,0206	5,8	7	0,260
85,5	88,5	91,5	0,9884	0,9964	0,0080	2,3	2	0,028
91,5	94,5	97,5	0,9964	1,0000	0,0036	1,0	0	0,998
				Sumas	1,0000	280,0	280	9,6440
							Probabilidad de Chi-Cuadrada	**0,7877**

El proceso para ajustar y valorar el ajuste que la distribución teórica tiene con respecto a la distribución de los datos lleva los pasos en la HE.

Definir los límites de las clases. Se copian de algún cuadro anterior: en la Columna A el Límite Inferior, en la columna B el punto medio de cada clase, en la columna C el límite superior de la clase;

Usar la función de probabilidad normal de la HE:

=DISTR.NORM(C452;$C$442;$F$442;1)

Para calcular las probabilidades de los límites inferiores en la columna D y superiores en la columna E. A la clase 1 se le digita 0 en la columna D a la clase 15, 1 en la columna F;

A las celdas de la columna E se le restan las celdas de la columna D y ubican en la columna F;

Se comprueba que la suma de probabilidades sea 1;

Se obtiene las frecuencias esperadas multiplicando las probabilidades de las celdas de la columna F por n = 280, valores que se ubican en la columna G;

Se arrastran de cuadros anteriores las frecuencias observadas ubicándolas en la columna H;

Se efectúa el cálculo de la variable que se conoce como χ^2 (Chi-Cuadrada) que proporcionarán los elementos para la prueba acomodándolos en la columna I;

Se suman las variables χ^2 para obtener la Chi-Cuadrada total;

Se valora mediante la función:

=DISTR.CHI(I467;CONTAR(I452:I466)-1) = 0,8466

la probabilidad de que las distribuciones observada y esperada se parezcan.
Al final obtendrá un cuadro como el de la HE que se reproduce a continuación:

1.40 La Importancia de que las distribuciones se consideren iguales.

Cuándo la distribuciones de datos se considera que es similar a una *Distribución Estadística de Probabilidad* la labor del investigador se facilita enormemente pues puede utilizar todo el acerbo de la *Ciencia Estadística* para *Aproximarse* a una población real conociendo perfectamente las probabilidades que respaldan las *Conclusiones y Recomendaciones*.

Cuando no es así, la *Teoría Estadística* proporciona herramientas para obtener *Conclusiones y Hacer Recomendaciones* con probabilidad conocida, sin embargo, bajo una serie de restricciones que pueden reducir de manera importante el ámbito de utilidad.

1.41 Conclusión para la variable, peso promedio del huevo.

- Debe tenerse presente que la variable es de tipo *continuo* y que deberá simularse con una *Distribución Estadística* de tipo continuo.
- La *Distribución del peso de los huevos* es de forma acampanada, similar a una distribución estadística que se conoce como *La Normal*;
- Gráficamente, las distribuciones de frecuencias observadas y las esperadas calculadas utilizando la distribución son muy similares;
- La prueba estadística de χ^2 indica una probabilidad similitud de 84,66%
- Se puede utilizar la *Distribución Normal Estándar o La Normal* para analizar los resultados del proyecto.

1.42 La variable cualitativa Sexo del producto.

Se tratará la variable cualitativa que dio origen al proyecto:

La alta proporción de machos para una raza de gallinas productora de carne. Se espera que al menos sea de 70% de nacimientos de machos viables. Esto significa que el 30% restante incluye nacimientos de hembras y productos no viables.

Es evidente que únicamente hay dos resultados posibles: 1 si el producto es un macho viable y 0 si el producto no es un macho viable, por esto se utilizará la distribución *Binomial* para aproximar los datos.

1.43 La Distribución Binomial.

La Distribución de Probabilidad Binomial está definida por:

$$F(x) = \sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x}$$

En donde p es la proporción de que un suceso ocurra —que el producto sea un macho— $q = (1 - p)$ la proporción de que el suceso no ocurra. Y ${}_rC_n$ que indica las r combinaciones en que pueden intercambiarse los sucesos en n muestras denominado *Coeficiente Binomial.*

Para el caso se toman los sucesos de 10 huevos acomodados en una charola de la incubadora. Así se presentarían los datos, recordando que 1 (uno) significa que el producto es un pollito que al menos llegará a las granjas de los avicultores.

La incubadora de pruebas es pequeña y las bandejas de reproducción tienen capacidad para 10 huevos.

Las x_i, i = 1, 2,…,10 son los pollitos machos que se envían a las granjas productoras. La suma de los unos hace referencia a los pollitos machos viables en una bandeja reproductora. Los ceros, representan a hembras y productos no viables.

1.44 Cuadro de frecuencias y estadísticos.

El cuadro de frecuencias proporciona una idea de la distribución y los estadísticos necesarios para valorar la hipótesis:

Ho; **X ~B(np; npq)**

La variable ***X*** se distribuye binomial, con media $np = 10 \times 0.7 = 7$ y varianza $npq = 10 \times 0{,}7 \times 0{,}3 = 2{,}1$.

Por tanto, la distribución que aproxime a los datos será una binomial con un muestra de tamaño n = 10, y una proporción de pollitos machos viables de 0,7 o 70%.

La Hoja Electrónica proporciona la función binomial muy simple de utilizar sí se conocen, como en el ejemplo, los parámetros.

Evento x machos	Frecuencia Observada	Sumas Parciales
0	0	0
1	0	0
2	0	0
3	0	0
4	1	4
5	3	15
6	4	24
7	6	42
8	6	48
9	7	63
10	1	10
Estadísticos		
Número de bandejas		28
Suma Total de pollitos machos		206
Promedio de pollitos por bandeja		7,36
Tamaño de la muestra n		10
Proporción de pollitos machos viables		0,7357
Proporción no viable		0,2643

1.45 Las probabilidades binomiales.

Las operaciones para obtener las probabilidades binomiales se detallan en la HE. Con estas, se elabora un cuadro que permitirá determinar si la distribución de frecuencias del evento que el producto sea un macho viable pueda aproximarse mediante la *Distribución de Probabilidades Binomial*.

Para esto se comparan las frecuencias *esperadas* que se obtienen multiplicando la probabilidad para cada evento *x* por el número de muestras de tamaño 10 —charolas de incubación— *observadas*, con las frecuencias observadas mediante la prueba de χ^2.

Se recuerda al estudiante que el capítulo está orientado a conocer las distribuciones de datos. Las pruebas debe utilizarlas como herramientas.

Para efectuar la prueba de ajuste mediante la HE se procede de la siguiente manera:

En la columna A se detallan los eventos posibles: x = 0, significa que en la bandeja de 10 huevos ninguno sea un machito viable; x = 1, que en la bandeja uno de los pollitos sea un macho viable; y así hasta x = 10 que significa que los 10 pollitos de la bandeja de incubación sean machitos viables;

Solicite a la HE la función binomial con proporción p y 10 ensayos: para cada evento x. Deberá indicar con un o que quiere la probabilidad del evento, con 1 aparece la probabilidad acumulativa. La operación se ubica en la columna B. Se ejemplifica no x = 2;

=DISTR.BINOM(A561;$B$555;$B$556;0)

Sobre la columna C se calcula la frecuencia esperada multiplicando cada probabilidad binomial por las 28 muestras (bandejas) consideradas en el estudio;

Sobre la Columna D se arrastra la frecuencia observada para cada evento;

Sobre la columna E se calcula la diferencia entre la frecuencia observada – frecuencia esperada de cada evento o clase;

Sobre la columna F se calculan las χ^2 parciales;

$$\chi_i^2 = \frac{(fo_i - fe_i)^2}{fe_i}$$

sin utilizar el corrector por continuidad que provocaría valores grandes que alterarían más la prueba que si no se usan;

Se suman las columnas de manera que: la de probabilidades (B) sume 1; las frecuencias esperadas (C) sume 28; la de las frecuencias observadas (D) sume 28; la de diferencias (E) sume 0;

La siguiente columna (F) será la

$$\chi^2_{(11-1)} = \sum_{i=0}^{10} \frac{(fo_i - fe_i)^2}{fe_i} = 4{,}9954$$

que deberá valuarse mediante la distribución de χ^2 teórica.

La valuación de esta variable

$$F_{[4,9954;11\text{-}1]} = Y_0 \int_0^{4,9954} (4{,}9954)^{\frac{1}{2}(11-1)} e^{-\frac{1}{2}4,9954} d\chi = 0{,}8915$$

indica que la probabilidad de que las distribuciones sean compatibles es de 89,15%. En un prueba al 5% de confianza, habrá que aceptar la hipótesis nula.

1.46 El Cuadro con la prueba de bondad de ajuste.

La prueba estadística de χ^2 indicó una probabilidad de 0,8915 o 89,15% de que los nacimientos de pollitos machos se distribuyan como una Binomial. En términos estadísticos, no hay evidencia para rechazar la hipótesis *Ho*; *X*~B(*np* = 7; *npq* = 2,1) con nivel de confianza del 5%. Notará que además de la distribución se ha valorado la proporción. Esto es, el p = 0,7353 puede considerarse como 0,7.

En esta prueba, se aprovechó para valorar también la aproximación de la proporción observada p = 0,7353 con la proporción esperada p = 0,7. El resultado es que las diferencias en las proporciones se pueden considerar debidas al azar y no a una condición o factor que esté propiciando un mayor nacimiento y viabilidad de pollitos machos.

Puede probar cambiar la proporción de p = 0,7 por p = 0,7353 y observará que la probabilidad de aproximación que indica la prueba de Chi-Cuadrada es de poco más o menos 98%.

Esto significa que se han valorado dos hipótesis:

Ho; X~B(np = 7; npq = 2,1)

y

Ho; P = 0,7

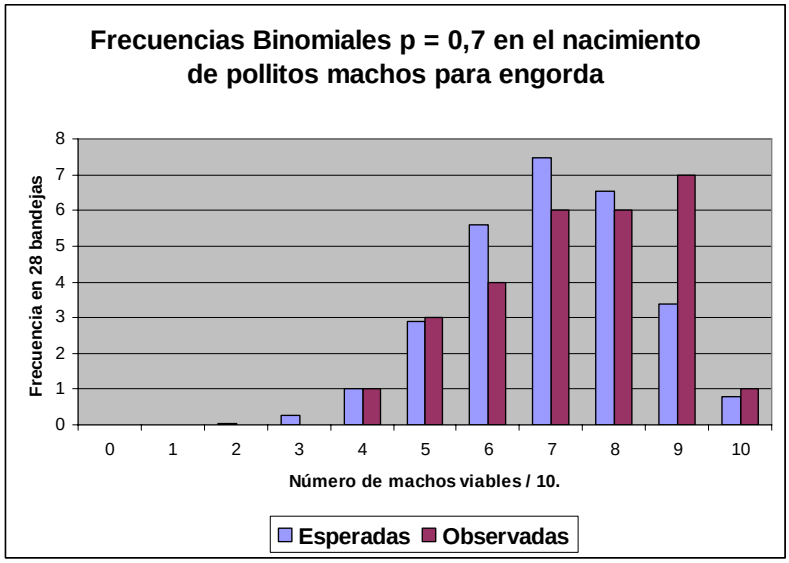
Realizado dos pruebas con una misma

$$F_{[4,9954;11\text{-}1]} = Y_0 \int_0^{4,9954} (4{,}9954)^{\frac{1}{2}(11-1)} e^{-\frac{1}{2}4,9954} d\chi = 0{,}8915$$

1.47 Un gráfico siempre es de ayuda.

Elaborando un Histograma con las frecuencias observadas en guinda y las esperadas en azul se aprecia una tendencia similar. Como en el caso de la distribución *Normal,* los resultados del proyecto pueden analizarse utilizando la distribución de probabilidad *Binomial*. Una conclusión que facilitará enormemente el análisis y la conclusión de proyecto en lo tocante al número de pollitos machos.

Es importante hacer notar al estudiante que los gráficos de conteo deben presentarse usando barras. Esto indicará al lector que se trata de una distribución de cualidades.



1.48 La variable discreta número de huevos.

Cuando se trabaja con variables cuya distribución brinca al menos por unidades debe tenerse cuidado. Casi siempre y sin mucho análisis, se trabajan como distribuciones continuas y más específicamente como distribuciones normales por la facilidad que esto implica.

El experimentador deberá tener, siempre en consideración, que está trabajando con una variable discreta que salta de unidad en unidad, pues las gallinas no ponen medios huevos. Aun cuando los estadísticos indiquen fracciones o sean elementos de los números racionales.

Para este ejemplo se iniciará solicitando a la HE el cómputo de las *Estadísticas Descriptivas.*

1.49 Estadísticas Descriptivas.

Los estadísticos importantes para determinar si la distribución de los datos puede aproximarse mediante una distribución normal son: *La Media, La Mediana, La Moda* como parámetros de tendencia central también llamados de posicionamiento.

El Coeficiente de Asimetría (valores críticos 0,230(5%) 0,360(1%)); y el Coeficiente de Curtosis con valores críticos de (-0,41 a +0,47 (5%) y -0,50 a +0,79 (1%).

Huevos	
Media	178,436
Error típico	3,581
Mediana	180
Moda	180
Desviación estándar	59,922
Varianza de la muestra	3.590,706
Curtosis	0,002
Coeficiente de asimetría	-0,075
Rango	324
Mínimo	18
Máximo	342
Suma	49.962
Cuenta	280

1.50 Las medias de posicionamiento.

Se presume que una distribución de datos se parece a una distribución normal cuando las medidas de posicionamiento están muy próximas:

Sí la *Media, Mediana y Moda* son iguales, al menos se presume que se tiene un distribución perfectamente centrada;

Sí el orden ascendente de los estadísticos es *Moda, Mediana y Media* se presume una cola a la derecha más larga:

Si el orden ascendente de los estadísticos *Media, Median y Moda* se presume una cola izquierda más larga.

En general en una distribución asimétrica, la *Media* con respecto a la *Moda* tiende a situarse al mismo lado que la cola más larga.

En el ejemplo con *Media* = 178,4 huevos / año, *Mediana* 180 huevos / año y *Moda* = 180 huevos año se puede considerar una distribución centrada. Siendo muye exigentes se puede presumir una distribución con sesgo negativo.

1.51 El Coeficiente de Curtosis.

El Coeficiente mide:

El alargamiento o estrechamiento de una distribución de datos con respecto a una distribución normal de los mismos datos.

Entre más se aproxime la distribución de los datos a una normal más próximo a 3 será el coeficiente. O a 0 cuando se corrige.

Según la tabla de para la valoración de la curtosis mediante los valores ajustados y para un nivel de confianza de 95% el coeficiente de curtosis debe mantenerse entre –0,41 y 0,47 para aceptar que la distribución se parece, por su estrechez a una normal. En el ejemplo se Acepta que la distribución es semejante a una normal.

1.52 El Coeficiente de Sesgo o Asimetría.

El Coeficiente mide:

La simetría de una distribución de datos con respecto a una normal.

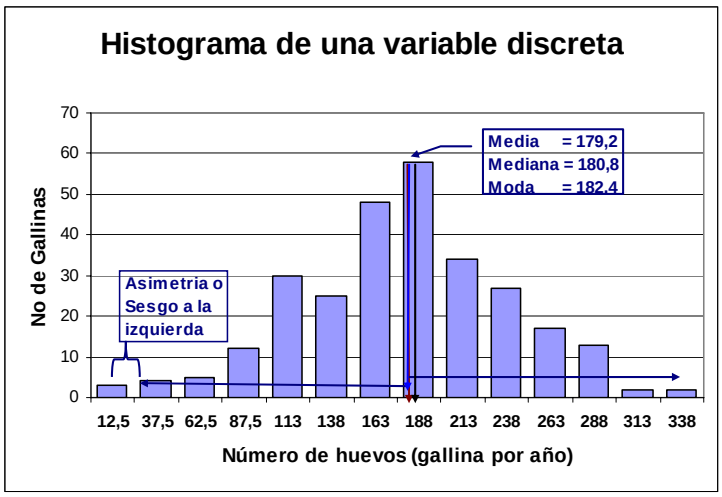
Este coeficiente siempre se valora con respecto a un valor cero en el que la distribución es, además de centrada simétrica.

En el ejemplo el coeficiente de asimetría o sesgo es de –0,0752 que para la valoración debe tomarse como valor absoluto. El límite teórico de la distribución del estadístico para n = 300 es de 0,23. Cómo 0,0752 es menor que 0,23, debe aceptarse que la distribución de datos es simétrica con respecto a la normal.

1.53 La Recomendación para las distribuciones discretas.

Tomando en cuenta el origen biológico de la variable ***Y*** = número de huevos en una año de 365 días por gallina y los estadísticos que se acaban de valorar puede concluirse que la variable sigue una distribución normal.

No obstante, es conveniente que el investigador obtenga una visión más directa mediante el gráfico de la distribución de datos y los estadísticos de posicionamiento y los coeficientes de sesgo y curtosis desde datos agrupados en



una distribución de frecuencias.

Este proceder es recomendable en análisis de resultados de una distribución absolutamente discreta.

1.54 El Histograma.

El gráfico simple muestra una distribución muy similar a una campana, característica de distribuciones de datos que se parecen a una distribución normal.

Es conveniente que las barras que representan el peso relativo de cada subclase no se unan, indicando con esto, que se trata de una distribución discreta.

Así mismo, no es conveniente unir las cúspides de las barras con la línea del polígono de frecuencias.

1.55 Estadísticos con datos agrupados.

LÍMITES DE CLASES			Frecuencia	$f_i\bar{x}_i$	$f_i(\bar{x}_i-\bar{\bar{x}})^2$	$f_i\left(\frac{\bar{x}_i-\bar{\bar{x}}}{s}\right)^3$	$f_i\left(\frac{\bar{x}_i-\bar{\bar{x}}}{s}\right)^4$
Inferior	Medio	Superior	Observada				
0	12,5	25	3	37,5	83.363,10	-63,56	175,87
25	37,5	50	4	150,0	80.311,51	-52,05	122,42
50	62,5	75	5	312,5	68.090,28	-36,34	70,40
75	87,5	100	12	1.050,0	100.898,82	-42,32	64,41
100	112,5	125	30	3.375,0	133.452,41	-40,71	45,07
125	137,5	150	25	3.437,5	43.464,80	-8,29	5,74
150	162,5	175	48	7.800,0	13.380,99	-1,02	0,28
175	187,5	200	58	10.875,0	3.999,06	0,15	0,02
200	212,5	225	34	7.225,0	37.710,35	5,74	3,18
225	237,5	250	27	6.412,5	91.781,27	24,47	23,69
250	262,5	275	17	4.462,5	117.971,25	44,95	62,15
275	287,5	300	13	3.737,5	152.485,63	75,53	135,79
300	312,5	325	2	625,0	35.539,68	21,67	47,95
325	337,5	350	2	675,0	50.120,04	36,29	95,36

Número de observaciones	280	Sumas de cuadrados	1.012.569,20
Suma Total	50.175,0	Varianza	3.629,28
Promedio de huevos	179,20	Desviación Estándar	60,24
Mediana	180,8	C. Asimetría	-0,128
Moda	182,4	C. Curtosis	0,089

Número de observaciones:

$$n=\sum_{i=1}^{14} f_i = 3+4+\ldots+2 = 280$$

Suma Total:

$$T=\sum_{i=1}^{14} f_i\bar{x}_i = 3(12{,}5)+4(37{,}5)+\ldots+2(337{,}5) = 50.175{,}0 \text{ Huevos.}$$

Promedio de huevos por gallina por año:

$$\bar{\bar{x}}=\frac{\sum_{i=1}^{14} f_i\bar{x}_i}{n}=\frac{50.175{,}0}{280}=179{,}20 \text{ Huevos/Gallina}$$

La Mediana:

$$\tilde{x}=LI_{me}+\left[\frac{\frac{n+1}{2}-S}{f_{me}}\right]IC=175+\left[\frac{140{,}5-127}{58}\right]25=180{,}8$$

En donde:

LI_{me} = Límite inferior de la clase mediana;

$$\frac{n+1}{2}$$

es el orden estadístico mediano;
S es la suma acumulativa hasta la clase anterior a la mediana;
f_{me} es la frecuencia de la clase mediana;
IC es el intervalo de clase.
La Moda:
En donde:
LI_{mo} es el límite inferior de la clase modal (la de mayor frecuencia);
.f_{mo} es la frecuencia de la clase modal;
.f_{mo-1} es la frecuencia de la clase anterior a la modal;
f_{mo+1} es la frecuencia de la clase posterior a la modal;
IC es el intervalo de clase.

$$\tilde{x} = LI_{mo} + \left[\frac{f_{mo} - f_{mo-1}}{(f_{mo} - f_{mo-1}) + (f_{mo} - f_{mo+1})}\right] IC =$$

$$= 175 + \left[\frac{58 - 48}{(58 - 48) + (58 - 34)}\right] 25 = 182{,}4$$

La Suma de Cuadrados:

$$SC = \sum_{i=1}^{14} f_i (\bar{x} - \bar{\bar{x}})^2 = 3(12{,}5 - 179{,}20)^2 + 4(37{,}5 - 179{,}20)^2 + \ldots + 2(337{,}5 - 179{,}20)^2 =$$

$$= 1.012.569{,}20$$

La Varianza:

$$s^2 = \frac{\sum_{i=1}^{14} f_i (\bar{x}_i - \bar{\bar{x}})^2}{n-1} = \frac{1.012.569{,}20}{279} = 3.629{,}28$$

La Desviación Estándar:

$$s = \sqrt{s^2} = \sqrt{3.629{,}28} = 60{,}24$$

Coeficiente de asimetría:

$$ca = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{14} f_i \left(\frac{\bar{x}_i - \bar{\bar{x}}}{s}\right)^3$$

$$= \frac{280}{(279)(278)} \left\{3\left(\frac{12{,}5 - 179{,}20}{60{,}24}\right)^3 + \ldots + 2\left(\frac{337{,}5 - 179{,}20}{60{,}24}\right)^3\right\} = -0{,}128$$

Coeficiente de curtosis:

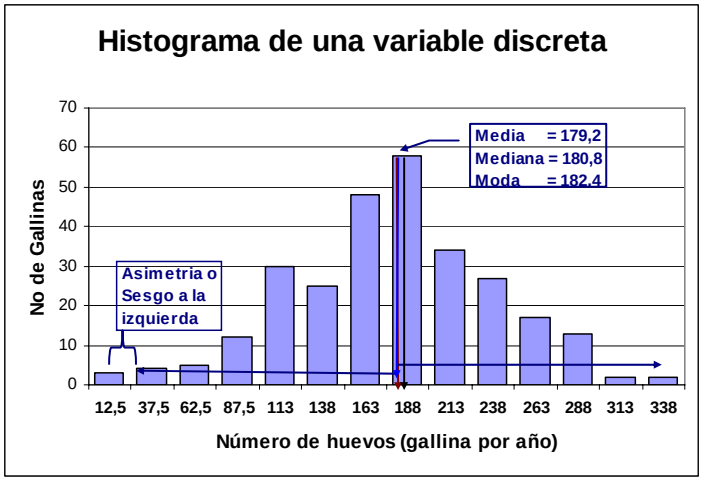
$$cc = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{14} f_i \left(\frac{\bar{x}_i - \bar{\bar{x}}}{s}\right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)} =$$

$$= \frac{280(281)}{(279)(278)(277)} \left\{3\left(\frac{12{,}5 - 179{,}20}{60{,}24}\right)^4 + \ldots + 2\left(\frac{337{,}5 - 179{,}20}{60{,}24}\right)^4\right\} - \frac{3(279)^2}{(278)(277)} = 0{,}089$$

1.56 *Interpretación.*

Se han señalado en el histograma los estadísticos de posición y dos líneas en la parte inferior del mismo tamaño para hacer evidente la asimetría también llamada sesgo señalada con una llave invertida. Recodaremos que no es significante.

La asimetría toma de referencia a la normal indicando una anormalidad de la distribución de los datos con respecto a la teórica. Los coeficientes de forma indican que las diferencias se deben al azar pudiendo, por tanto, utilizar a la distribución Normal Estándar en proceso de análisis, interpretación y predicción.



1.57 *La Prueba de Bondad de Ajuste.*

Si hubiera dudas se debe hacer la prueba de "*Bondad de Ajuste*" de las frecuencias esperadas y observadas mediante la Chi-Cuadrada. Para declarar que la distribución de datos no se asemeja a una normal la probabilidad de χ^2 debe ser inferior a 0,05 0 5%. Para todos los efectos la variable ***Y*** se opera como una variable continua con valores límites en el conjunto de los reales.

Dado que es más importante elaborar una prueba de Bondad de Ajuste cuando la variable es discreta y se va a aproximar mediante una variable continua, los pasos para elaborar la prueba sobre la HE se puntualizas:

1. ***Arrastre o elabore el cuadro de la distribución de frecuencias observadas. El la Columna A ubique el límite inferior de las clases, en la Columna B Limite Superior de la Clase (el punto medio no es necesario si se conoce la Media y la Desviación Estándar como es el caso);***
2. ***En la Columna C ubique la probabilidad normal acumulada que determina el valor del límite inferior de la clase. En la primera celda digite 0 (cero) para que considere la probabilidad desde menos infinito al límite superior de la primera clase. En la siguiente sería:*** $F(x_{LI2}) = \text{DISTR.NORM}(25;179{,}2;60{,}24;\text{Acumulado}) = 0{,}0052$ ***en la HE los valores se refieren a posiciones de celdas que almacenan esos valores;***
3. ***En la Columna D ubique la probabilidad normal acumulada que determina el valor del límite superior de la clase. En la última columna digite 1 para considerar la probabilidad del último limite inferior hasta infinito;***
4. ***En la Columna E se calcula la probabilidad de la clase, restando de la probabilidad acumulada determinada por los límites superiores de cada clase, la probabilidad acumulada de los límites inferiores de la misma clase. Se hace notar que el límite inferior de una clase es igual al límite superior de la clase anterior, recuerde que se está tratando de una distribución de probabilidad continua, de hecho hay una sobreestimación de media unidad. La suma de las probabilidades de cada clase debe ser la unidad;***
5. ***En la Columna F calcule la frecuencia esperada multiplicando la probabilidad de cada clase por el número de observaciones, esto es:*** $fe_i = n \times P(Clase_i);\ fe_1 = 280 \times 0{,}0052 = 1{,}5$***. La suma de las frecuencias esperadas debe ser exactamente el número de observaciones;***
6. ***En la Columna G arrastre u obtenga las frecuencias observadas para cada clase, evidentemente, la suma debe ser igual al número de unidades;***

LÍMITES DE CLASES		Probabilidad a los límites		Probabilidad	Frecuencias		Chi-Cuadradas
Inferior	Superior	Inferior	Superior	del intervalo	Esperadas	Observadas	parciales
0	25	**0,0000**	0,0052	0,0052	1,5	3	0,7268
25	50	0,0052	0,0160	0,0108	3,0	4	0,0795
50	75	0,0160	0,0419	0,0259	7,2	5	0,4183
75	100	0,0419	0,0943	0,0525	14,7	12	0,3269
100	125	0,0943	0,1842	0,0898	25,2	30	0,7505
125	150	0,1842	0,3140	0,1298	36,3	25	3,2363
150	175	0,3140	0,4722	0,1583	44,3	48	0,2289
175	200	0,4722	0,6351	0,1628	45,6	58	3,1080
200	225	0,6351	0,7765	0,1414	39,6	34	0,6542
225	250	0,7765	0,8801	0,1036	29,0	27	0,0783
250	275	0,8801	0,9441	0,0641	17,9	17	0,0105
275	300	0,9441	0,9775	0,0334	9,4	13	1,0555
300	325	0,9775	0,9922	0,0147	4,1	2	0,6367
325	350	0,9922	**1,0000**	0,0078	2,2	2	0,0497
			Sumas	1,0000	280,0	280	11,3602
					Probabilidad de Chi_Cuadrada		0,5807

7. En la Columna H se computa directamente las variables $\chi^2 = \frac{\left(\left|fo_i - fe_i\right| - 0{,}5\right)^2}{fe_i}$; $\chi_1^2 = \frac{\left(\left|3 - 1{,}5\right| - 0{,}5\right)^2}{1{,}5} = 0{,}7268$ ***parciales incluyendo el corrector por continuidad (-0,5), irrenunciable en las distribuciones de variables discretas. Esta condición le resta importancia a la diferencia;***

8. Sume las Chi-Cuadradas parciales;

9. Valore la suma de las Chi-Cuadradas parciales:

$$F_{[11,3612;14\text{-}1]} = Y_0 \int_0^{11,3612} (11,3612)^{\frac{1}{2}(14-1)} e^{-\frac{1}{2}11,3612} d\chi = \text{DISTR.CHI}(11,3602;13) = 0,5807$$

Puesto que la probabilidad es mucho mayor a 0,05 o 5%, se acepta que la distribución del número de huevos que pone una gallina en un año se puede aproximar mediante una distribución Normal o Normal Estándar.

1.58 Conclusión.

Éste capítulo hace referencia a las distribuciones de los datos.

Se han utilizado formulas, algunas muy complejas que requiere la teoría estadística para analizar resultados de pruebas y proyectos, pero fácilmente computables o obtenibles mediante funciones o algoritmos de la HE.

Se han abordado los tres tipos de datos: *continuos, discretos y cualitativos* asociando la *distribución de datos observadas* con las distribuciones estadísticas de mayor uso puntualizando criterios para determinar sí tal o cuál distribución estadística puede utilizarse para estudiar los resultados obtenidos a partir de conjuntos de datos de una población objetivo.

Se ha concluido con respecto a las implicaciones estadísticas de las tres variables ejemplificadas.

1.59 Recomendación.

Se recomienda al estudiante que entienda la notación matemática en las fórmulas para que pueda aplicarla en la HE sin reparar en la complejidad de la misma. Esto con el objeto de considerar a las fórmulas de cálculo de estadísticos como herramientas.

También se ha recomendado al estudiante que ponga atención en el significado de cada estadístico para que pueda interpretar y concluir desde los análisis de los resultados.

El estudiante habrá notado que la HE posee una gran cantidad de funciones y rutinas estadísticas y matemáticas que le facilitan el análisis de resultados de conjuntos de datos provenientes de exploraciones o de técnicas de experimentación: Utilícelas.

REFERENCIAS SELECTAS.

1. Hillier Frederick S., y Lieberman Gerard j., *Introducción a la Investigación de Operaciones*. Capítulo 19. Segunda edición en español traducida de la cuarta edición en ingles. McGraw-Hill Interamericana de México, S. A. De C. V., 1990.
2. Miller Irwin, Freund John E., Johnson Richard A: *Probabilidad y Estadística para Ingenieros*. Capítulo 11. Traducido de la cuarta edición en ingles; Prentice-Hall Hispanoamericana, S. A. 1992.
3. Murray R. Spiegel: *Serie de compendios Schaum, Teoría y Problemas de Estadística.* Capítulos 7. Primera edición en español, traducido de la primera edición en ingles; Libros McGraw-Hill de México, S. A. De C. V., 1973.
4. Ostle Bernard: *Estadística Aplicada*. Capítulos 3, 4 y 7. Primera edición en español traducida de la primera edición en ingles. Editorial Limusa, S. A., 1977.
5. Snedecor George W., y Cochran William G: *Statistical Methods*. Capítulos 2, 3, 8 y 9. Sexta edición; The Iowa State University, 1974.
6. Steel Robert G. D., Torrie James H: *Principles and Procedures of Statistics*. Capítulo 4. Primera edición; McGraw-Hill Book Company, Inc, 1960.

2 La Distribución Normal.

En esta sección se requieren los archivos:
E02_DNormal_W01.doc;
E02_DNormal_P01.pps;
E02_DNormal_X01.xls

2.1 Menú.

2.2 Introducción.

Los estudiosos de las ciencias biológicas se dieron cuenta que la distribución de las frecuencias de medidas de individuos de las poblaciones naturales presentaban una forma acampanada al dibujarla en un plano.

Por otro lado, los matemáticos buscaban modelos que pudieran emular las distribuciones que estas poblaciones presentaban.

Con el tiempo, se descubrió *El Promedio Aritmético* de un conjunto de datos. Casi inmediatamente se descubrió que la suma de las desviaciones de las medidas de las observaciones con respecto al promedio era cero.

$$d = \sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

Propiedad muy importante en la Teoría Estadística.

2.3 Los Parámetros.

Inmediatamente se descubrió la Varianza que no es otra cosa que las desviaciones elevadas al cuadrado y ponderadas por los grados de libertad.

A medida que se trabajaba con estos parámetros: *La Media*, *La Varianza* y su raíz cuadrada *La Desviación Estándar* se dieron cuenta que precisamente, la frecuencia de las desviaciones con respecto al promedio seguían una curva característica de campana, esto es, las medidas de los individuos se agrupaban hacia el centro, situación que se repetía constantemente.

Los matemáticos se dieron a la tarea de descubrir funciones que se pareciera a la ***DISTRIBUCIÓN DE FRECUENCIAS DE LAS DESVIACIONES.***

2.4 El modelo matemático.

GAUSS, Carl Friedrich (1.777 –1.855) matemático y astrónomo ideó la siguiente fórmula antes que la ciencia estadística se formalizara:

$$f(x;\mu;\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\left(e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right)$$

En donde μ es la media, σ la desviación estándar, π = 3,14159 y la base del logaritmo neperiano *e* = 2,71828. Que se conoce como Distribución de Densidad Normal o Campana de Gauss.

Ésta distribución tiene propiedades trascendentales para simular poblaciones reales, que se irán discutiendo a medida que se avanza en la lectura y usando un ejemplo.

Karl Friedrich Gauss. (Brunswick, actual Alemania, 1777 - Gotinga, id., 1855) Matemático, físico y astrónomo alemán. Nacido en el seno de una familia humilde, desde muy temprana edad Karl Friedrich Gauss dio muestras de una prodigiosa capacidad para las matemáticas (según la leyenda, a los tres años interrumpió a su padre cuando estaba ocupado en la contabilidad de su negocio para indicarle un error de cálculo), hasta el punto de ser recomendado al duque de Brunswick por sus profesores de la escuela primaria.

El duque le proporcionó asistencia financiera en sus estudios secundarios y universitarios, que efectuó en la Universidad de Gotinga entre 1795 y 1798. Su tesis doctoral (1799) versó sobre el teorema fundamental del álgebra (que establece que toda ecuación algebraica de coeficientes complejos tiene soluciones igualmente complejas), que Gauss demostró.

En 1801 Gauss publicó una obra destinada a influir de forma decisiva en la conformación de la matemática del resto del siglo, y particularmente en el ámbito de la teoría de números, las Disquisiciones aritméticas, entre cuyos numerosos hallazgos cabe destacar: la primera prueba de la ley de la reciprocidad cuadrática; una solución algebraica al problema de cómo determinar si un polígono regular de n lados puede ser construido de manera geométrica (sin resolver desde los tiempos de Euclides); un tratamiento exhaustivo de la teoría de los números congruentes; y numerosos resultados con números y funciones de variable compleja (que volvería a tratar en 1831, describiendo el modo exacto de desarrollar una teoría completa sobre los mismos a partir de sus representaciones en el plano x, y) que marcaron el punto de partida de la moderna teoría de los números algebraicos.

Su fama como matemático creció considerablemente ese mismo año, cuando fue capaz de predecir con exactitud el comportamiento orbital del asteroide Ceres, avistado por primera vez pocos meses antes, para lo cual empleó el método de los mínimos cuadrados, desarrollado por él mismo en 1794 y aún hoy día la base computacional de modernas herramientas de estimación astronómica.

En 1807 aceptó el puesto de profesor de astronomía en el Observatorio de Gotinga, cargo en el que permaneció toda su vida. Dos años más tarde, su primera esposa, con quien había contraído matrimonio en 1805, falleció al dar a luz a su tercer hijo; más tarde se casó en segundas nupcias y tuvo tres hijos más. En esos años Gauss maduró sus ideas sobre geometría no euclidiana, esto es, la construcción de una geometría lógicamente coherente que prescindiera del postulado de Euclides de las paralelas; aunque no publicó sus conclusiones, se adelantó en más de treinta años a los trabajos posteriores de Lobachewski y Bolyai.

http://www.biografiasyvidas.com/biografia/g/gauss.htm

2.5 El problema 2-1.

Un empresario dedicado a la acuacultura decidió hacer un estudio de truchas, para esto usó una poza con 253 peces a los que, con todo y el estrés que produce se les pesó en gramos. Los datos se ofrecen en la Hoja Electrónica.

El proyecto consiste en:

- Separar a los peces de bajo peso para probar tratamientos repositorios;

- Separar a los peces de mayor peso para reproducirlos;
- Separar a los peces promedio para la pesca deportiva, un atractivo de la empresa.
- Los análisis de resultados se valorarán con un nivel de confianza de 95%.

2.6 *Estadísticas descriptivas.*

Lo usual es solicitar a la HE las Estadísticas Descriptivas.

Son valores que definen estimadores si el conjunto de datos corresponde a una muestra o parámetros si se trata de una población.

La HE ofrece instrucciones para calcularlos independientemente, pues en ocasiones, no se requiere información de todos sobre todo, en distribuciones de datos cualitativos.

Peso gr.	
Media	811,94
Error típico	15,12
Mediana	823
Moda	762
Desviación estándar	240,51
Varianza de la muestra	57.846,23
Curtosis	0,09
Coeficiente de asimetría	-0,13
Rango	1.340
Mínimo	112
Máximo	1.452
Suma	205.421
Cuenta	253

2.7 *Interpretación.*

El interés es saber si la distribución de los pesos de las truchas se puede considerar normal para aprovechar las ventajas analíticas de ésta distribución.

- La *media* con 874,4 y la *mediana* con 870,0 gramos indican que podría tenerse una distribución equilibrada;
- El coeficiente de curtosis de 0,1315 se encuentra entre los límites permitidos a una normal de 250 datos entre –0,45 y 0,52 indicando, por el signo positivo un muy ligero alargamiento;
- El valor absoluto del coeficiente de Asimetría de –0,0827 es inferior a 0,251 de las tablas para 250 permitido a una normal. El signo nos indica una cola izquierda ligeramente mayor.
- La moda, con datos individuales es de poca utilidad.

2.8 *Preparando el Histograma: el intervalo de clase.*

Siempre es importante crear una figura que muestre la perspectiva de la distribución de los individuos. Para esto, se preparan un número apropiado de clases en las que se acomodarán los individuos de acuerdo a sus valores.

La primera interrogante a resolver es determinar el número apropiado de clases.

Una regla que se usa para determinar el tamaño del intervalo de cada clase cuando se requiere una alta precisión en las estimaciones de un cuadro de frecuencias es considerar intervalo de clase entre la mitad y un cuarto de la Desviación Estándar.

En el ejemplo irían de:

$$IC_{<} = \frac{s}{4} = \frac{240{,}51}{4} = 60; \quad IC_{>} = \frac{240{,}51}{2} = 120$$

El número de clases que se esperan con estos intervalos son:

$$NC_{>} = \frac{r}{IC_{<}} = \frac{1.340}{60} = 22; \quad NC_{<} = \frac{r}{IC_{>}} = \frac{1.340}{120} = 11$$

Se tiene la libertad de elegir un intervalo de clase entre 60 y 120. Por ejemplo 100 ó 110 gramos sería un intervalo de clase cómodo.

2.9 Preparando el Histograma: alternativa a la desviación estándar.

Usualmente no se conoce la *Desviación Estándar*. Ésta se puede estimar multiplicando mínimo + el máximo por 0,15

$$\hat{\sigma} = 0{,}15(Mín. + Máx.) = 0{,}15(112 + 1.452) = 234{,}6$$

Buscando un número apropiado entre 60 y 120 gramos se decidió 110.

El siguiente paso es determinar los límites de la primera clase. Para esto se considerará al mínimo como el límite superior de la primera clase LS_1 = *Mínimo* = 112. A este se le resta el intervalo de clase para obtener el límite inferior de la primera clase, $LI_1 = LS_1 - \text{IC} = 112 - 110 = 2$. Finalmente se obtiene el punto medio de la primera clase:

$$\bar{x}_1 = \frac{LI_1 + LS_1}{2} = \frac{2 + 112}{2} = 57$$

2.10 Preparando el Histograma: los límites de clase.

Puesto que el conteo lo hará la HE no es necesario diferenciar entre los límites superior de una clase y el inferior de la siguiente. Además, la variable es continua y se asume que no hay diferencia entre estos. Este concepto que se conoce como *Límites Reales* se utilizará más adelante.

Después de forma la tabla de conteos sumando a cala límite el *IC* hasta que el máximo quede contenido en la última.

Para los efectos del capítulo la primera clase y la última clase no tendrán elemento. Proceso necesario para indicar que se trata de una distribución continua. Por tanto, para elaborar el gráfico la tabla de frecuencias quedaría como se muestra en la siguiente diapositiva.

Recuerde que el límite superior de la primera clase corresponde al mínimo.

A este se le resta el intervalo de clase para obtener el límite inferior de la primera clase.

El punto medio se obtiene sumando los límites inferior y superior y dividiéndolos por 2.

Se suma el intervalo de clase a los datos anteriores hasta que el máximo quede en la última clase;

Agregue una clase más.

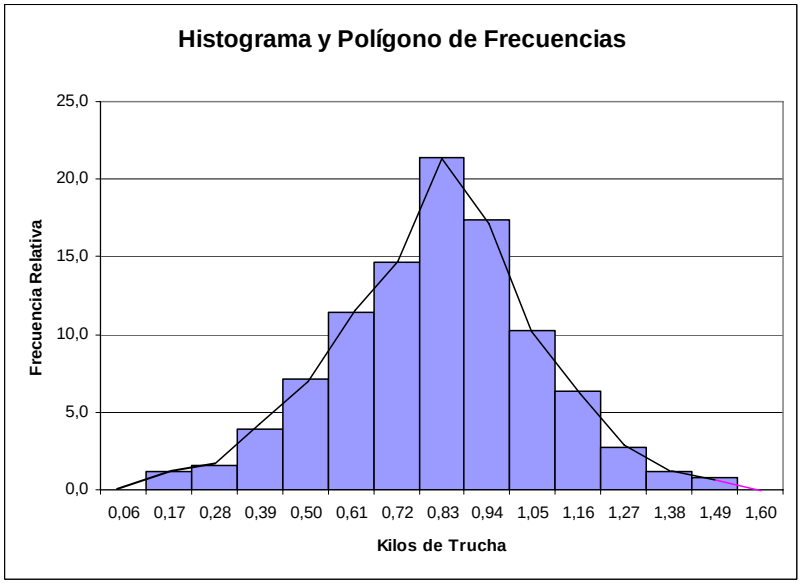
2.11 Cuadro de Frecuencias.

	Límites			Frecuencias	
Clase	Inferior	Punto Medio	Superior	Absolutas	Relativas
1	2	57	112	0	0,0
2	112	167	222	3	1,2
3	222	277	332	4	1,6
4	332	387	442	10	4,0
5	442	497	552	18	7,1
6	552	607	662	29	11,5
7	662	717	772	37	14,6
8	772	827	882	54	21,3
9	882	937	992	44	17,4
10	992	1047	1102	26	10,3
11	1102	1157	1212	16	6,3
12	1212	1267	1322	7	2,8
13	1322	1377	1432	3	1,2
14	1432	1487	1542	2	0,8
15	1542	1597	1652	0	0,0
			Sumas	253	100

2.12 El Histograma.

El Gráfico hace evidente que la distribución de los pesos presenta una forma de campana, en que los individuos con características más comunes en este caso el peso, se aglomeran hacia el centro, los más raros con pesos bajos o mayores a los extremos.

El polígono de frecuencias, la línea de color guinda que une las barras indica que la variable es de tipo continuo y no hace falta ningún ajuste por continuidad para utilizar en la aproximación que se efectúe a la *Distribución Normal* o *Normal Estándar*.



2.13 Selección por individuos.

Cuando se va a determinar un valor que servirá para elegir individuos y a utilizar *La Distribución Normal* o *La Normal Estándar* para establecer límites que involucran probabilidades, ***es indispensable que la distribución de la variable sea normal***.

También, para ciertos casos, es conveniente trabajar con datos agrupados en los cuadros de frecuencias, por representar una visión más apropiada de la distribución observada.

Por esta razón, se efectuarán los cálculos de las estadísticas descriptivas desde datos agrupados usando el cuadro de frecuencias.

2.14 Estadísticos con Datos Agrupados.

Clase	$\overline{x}_i$	Frecuencias Absolutas	$f_i \times \overline{x}_i$	$f_i(\overline{x}_i - \overline{\overline{x}})^2$	$f_i\left(\frac{\overline{x}_i - \overline{\overline{x}}}{s}\right)^3$	$f_i\left(\frac{\overline{x}_i - \overline{\overline{x}}}{s}\right)^4$
1	57	0	0,00	0,00	0,0000	0,0000
2	167	3	501,00	1.259.035,92	-56,7655	151,2652
3	277	4	1.108,00	1.157.027,60	-43,3085	95,8101
4	387	10	3.870,00	1.830.351,61	-54,4992	95,9076
5	497	18	8.946,00	1.818.241,59	-40,2188	52,5793
6	607	29	17.603,00	1.252.558,79	-18,1170	15,4875
7	717	37	26.529,00	354.087,90	-2,4108	0,9701
8	827	54	44.658,00	8.003,02	0,0068	0,0003
9	937	44	41.228,00	656.764,46	5,5844	2,8064
10	1.047	26	27.222,00	1.401.522,87	22,6465	21,6277
11	1.157	16	18.512,00	1.873.327,79	44,6117	62,7902
12	1.267	7	8.869,00	1.431.228,73	45,0404	83,7730
13	1.377	3	4.131,00	948.118,53	37,0955	85,7804
14	1.487	2	2.974,00	903.635,54	42,2730	116,8801
15	1.597	0	0,00	0,00	0,0000	0,0000
Número de Observaciones		253		Suma de Cuadrados		14.893.904,35
Suma Total		206.151		Varianza		59.102,80
Promedio		814,83		Desvío Estándar		243,11
Mediana		879,96		Coeficiente de asimetría		-0,0722
Moda		896,26		Curtosis		0,1568

Suma de Frecuencias = n;

$$n = \sum_{i=1}^{15} f_i = 0+3+4+10+18+29+37+54+44+26+16+7+3+2+0 = 253$$

Suma Total;

$$T = \sum_{i=1}^{15} f_i \bar{x}_i = 0(57)+3(167)+\ldots+2(1.487)+0(1.597) = 206.151$$

El promedio;

$$\bar{\bar{x}} = \frac{\sum_{i=1}^{15} f_i \bar{x}_i}{n} = \frac{206.151}{253} = 814,83$$

La Mediana;

$$\tilde{x} = LI_{me} + \left[\frac{\frac{n+1}{2} - S}{f_{me}}\right] IC = 772 + \left[\frac{127-101}{54}\right] 110 = 824,96$$

La Moda:

$$M_o = LI_{mo} + \left[\frac{f_m - f_{m-1}}{(f_m - f_{m-1}) + (f_m - f_{m+1})}\right] IC = 772 + \left[\frac{54-37}{(54-37)+(54-44)}\right] 110 = 841,26$$

Usualmente la moda se acomoda del lado de la cola más larga en la distribución de los datos y la mediana entre la media y la moda. La posición es correcta en los estadísticos obtenidos con los datos agrupados, no se presenta así en los estadísticos dato a dato.

La Suma de Cuadrados;

$$SC = \sum_{i=1}^{15} f_i \left(\bar{x}_i - \bar{\bar{x}}\right)^2$$

$$= 0(57-814,83)^2 + 7(167-814,83)^2 + \ldots + 0(1.597-814,83)^2 = 14.893.904,35$$

La varianza;

$$s^2 = \frac{\sum_{i=1}^{15} f_i \left(\bar{x}_i - \bar{\bar{x}}\right)^2}{n-1} = \frac{14.893.904,35}{253\text{-}1} = 59.102,80$$

La Desviación Estándar;

$$s = \sqrt{s^2} = \sqrt{59.102,80} = 243,11$$

El Coeficiente de asimetría;

$$ca = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{15} f_i \left(\frac{\bar{x}_i - \bar{\bar{x}}}{s}\right)^3$$

$$= \frac{253}{(252)(251)} \left[0\left(\frac{57-814,83}{243,11}\right)^3 + 7\left(\frac{167-814,83}{243,11}\right)^3 + \ldots + 0\left(\frac{1.597-814,83}{243,11}\right)^3\right] = \text{-}0,0722$$

El Coeficiente de Curtosis;

$$cc = \frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{15} f_i\left(\frac{\bar{x}_i - \bar{\bar{x}}}{s}\right)^4 - \frac{n(n-1)^2}{(n-2)(n-3)} =$$

$$= \frac{253(254)}{(252)(251)(250)}\left[0\left(\frac{057-814{,}83}{243{,}11}\right)^4 + \ldots + 0\left(\frac{1.516-814{,}83}{243{,}11}\right)^4\right] - \frac{3(252)^2}{(251)\lfloor 250\rfloor} = 0{,}1568$$

2.15 Diferencias con estadísticos directos.

Pocas son las diferencias que se tienen entre los cálculos de datos individuales con datos agrupados.

Media: 811,84 y 814,83. Mediana: 823 y 824,96. Moda 762 y 841,26.

El más significante se obtiene con la moda. Al ser el valor más frecuente, fue el que individualmente más se repite, pero no refleja con veracidad la verdadera tendencia modal, importante en muchos análisis de poblaciones.

La moda se acomoda al lado de la cola más corta de la distribución con la mediana en medio de ambos estadísticos de posición. Es evidente que los estadísticos individuales no presentan la secuencia indicada.

El coeficiente de asimetría muestra valores negativos en ambos casos -0,13 y -0,07 indicador de un sesgo a la izquierda, esto es, una cola izquierda ligeramente más larga.

2.16 La prueba de Bondad de Ajuste: la distribución de aproximación.

La prueba de denominada como *Bondad de Ajuste* requiere aproximar, esto es, usar un modelo estadístico que simule la distribución de datos de una población ideal. En este caso, se asume que la distribución de las frecuencias de los pesos de las truchas puede modularse mediante la distribución Normal Estándar, entonces, la prueba consistirá en probar que las frecuencias observadas se aproximan mucho a las frecuencias ideales o esperadas conseguidas con el modelo de probabilidad acumulativa normal definida por:

$$F(x;\mu;\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\,dx$$

La prueba requiere los límites de las clases mismos que se ubican en las Columnas B y C de la HE. Es insoslayable que los límites sean reales (el límite superior de una clase es igual al inferior de la siguiente).

2.17 La Prueba de Bondad de Ajuste. Continuación.

En la Columna E se obtiene la probabilidad de la primera clase usando la probabilidad acumulativa desde menos infinito al Límite Superior de la Clase 1 o $x = 112$;

$$F(x;\mu;\sigma) = \frac{1}{243{,}11\sqrt{2\pi}}\int_{-\infty}^{112} e^{-\frac{(112-814{,}83)^2}{2(243{,}11^2)}}\,dx = 0{,}00192$$

La probabilidad para la clase 2:

$$F(x_{LS2}) - F(x_{LI2}) = \frac{1}{243{,}11\sqrt{2\pi}}\int_{-\infty}^{222} e^{-\frac{(222-814{,}83)^2}{2(243{,}11^2)}}\,dx - \frac{1}{243{,}11\sqrt{2\pi}}\int_{-\infty}^{112} e^{-\frac{(112-814{,}83)^2}{2(243{,}11^2)}}\,dx =$$

$$= 0{,}00737 - 0{,}00192 = 0{,}00545$$

De manera similar se obtienen las probabilidades de las clases hasta la 14 o penúltima. La última se obtiene mediante:

$$F(x_{LI15}) = 1 - \frac{1}{243,11\sqrt{2\pi}} \int_{-\infty}^{1.542} e^{\frac{-(1.542-814,83)^2}{2(243,11^2)}} dx = 1 - 0,99861 = 0,00139$$

2.18 La Prueba de Bondad de Ajuste: la frecuencia esperada.

Pues debe considerar la probabilidad del límite inferior de la clase x_{15} = 1.542 hasta más infinito, por esto, la probabilidad acumulativa se vuelve complementaria y se resta de 1. Y

La suma de las probabilidades de todas las clases debe ser 1.

La probabilidad de la clase se obtiene restando a las probabilidades de la columna E de los límites superiores las de la D de los límites inferiores.

Una vez que se han obtenido las probabilidades esperadas, basta multiplicar cada probabilidad por el número de observaciones para obtener las frecuencias esperadas, sobre la Columna G, esto es;

$$fe_1 = 0,00192 \times 253 = 0,5; \quad fe_2 = 0,00545 \times 253 = 1,4; \quad fe_{15} = 0,00139 \times 253 = 0,4$$

Se hace notorio el sesgo de la distribución cuando las clases por arriba del promedio no tienen las mismas unidades esperadas que las que están bajo el promedio.

Se arrastran las frecuencias observadas a la Columna H y se lleva a cabo la prueba consiste en comparar el número de individuos que se esperarían de una distribución normal típica con los que realmente se contaron, para esto, la prueba aconsejada es la denominada de χ^2 definida por:

$$\chi^2_{(15-1)} = \sum_{i=1}^{15} \frac{(fo_i - fe_i)^2}{fe_i} = \frac{(0-0,5)^2}{0,5} + \frac{(3-1,4)^2}{1,4} + \ldots + \frac{(0-0,4)^2}{0,4} = 7,1662$$

No se deberá utilizar el corrector por continuidad puesto que el peso de las truchas es una variable continua.

2.19 La prueba completa.

Clase	Límites		Probabilidad Acumulada		Probabilidad	Frecuencias		Chi-Cuadrad
	Inferior	Superior	L. Inferior	L. Superior	Intervalo	Esperada	Observada	parciales
1	2	112	0,00000	0,00192	0,00192	0,5	0	0,4858
2	112	222	0,00192	0,00737	0,00545	1,4	3	1,9025
3	222	332	0,00737	0,02351	0,01614	4,1	4	0,0017
4	332	442	0,02351	0,06257	0,03905	9,9	10	0,0014
5	442	552	0,06257	0,13983	0,07726	19,5	18	0,1224
6	552	662	0,13983	0,26480	0,12497	31,6	29	0,2166
7	662	772	0,26480	0,43008	0,16529	41,8	37	0,5551
8	772	882	0,43008	0,60884	0,17876	45,2	54	1,7020
9	882	992	0,60884	0,76693	0,15809	40,0	44	0,4009
10	992	1102	0,76693	0,88125	0,11432	28,9	26	0,2952
11	1102	1212	0,88125	0,94884	0,06759	17,1	16	0,0709
12	1212	1322	0,94884	0,98152	0,03268	8,3	7	0,1944
13	1322	1432	0,98152	0,99444	0,01292	3,3	3	0,0220
14	1432	1542	0,99444	0,99861	0,00417	1,1	2	0,8437
15	1542	1652	0,99861	1,00000	0,00139	0,4	0	0,3516
				Sumas	1,00000	253,0	253	7,1662
							Probabilidad Chi-Cuadrada	**0,9281**

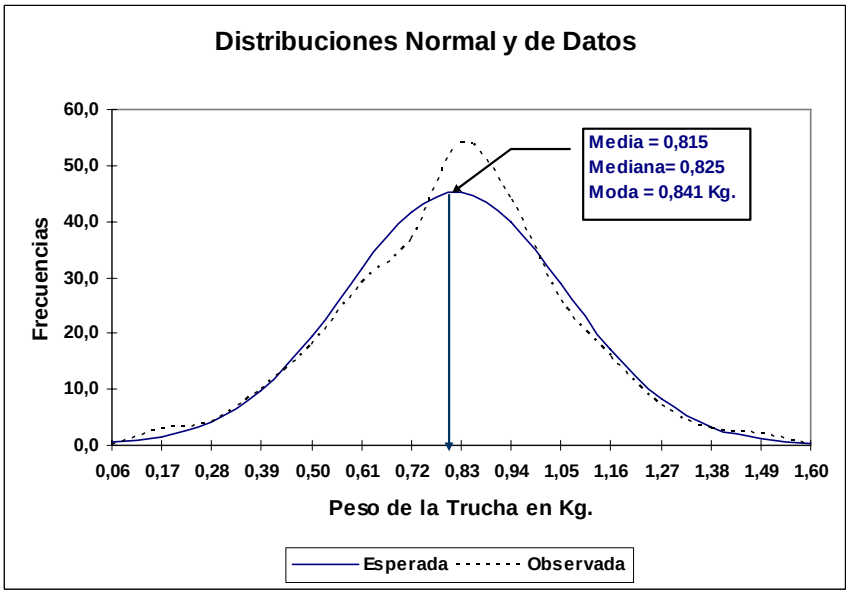
La probabilidad de la prueba:

$$F_{[7,1662;15\text{-}1]} = Y_0 \int_0^{7,1662} (7,1662)^{\frac{1}{2}(15-1)} e^{-\frac{1}{2}7,1662} d\chi = 0,9281$$

Indica que la aproximación que se haga mediante las Distribuciones Normales será 92,81% exacta. Dicho en términos estadísticos: no hay evidencia para rechazar la hipótesis Ho; $X \sim N(\mu;\sigma)$.

2.20 *El gráfico comparativo de las distribuciones.*

Las pruebas que se han realizado han indicado lo que el gráfico representa, que la distribución de los datos se puede aproximar con mucha seguridad por una distribución normal. Las pequeñas diferencias no provocarán conclusiones y recomendaciones riesgosas.



Siempre que los resultados estadísticos analizados a la luz de los resultados fisiológicos sean bien interpretados por el experimentador.

Para entender las diferencias entre la distribución de aproximación y la distribución teórica que se creará con la Distribución Normal Estándar.

2.21 *La distribución de datos estandarizados.*

Una variable estandarizada está definida, para datos individuales y agrupados por:

$$z = \frac{x_i - \bar{x}}{s};\quad z_a = f_i\left(\frac{\bar{x}_i - \bar{\bar{x}}}{s}\right)$$

La media de la variable estandarizada es 0 y la varianza es 1. La aseveración se probará para datos agrupados. Es promedio será:

$$\bar{z} = \frac{\sum_{i=1}^{15} f_i\left(\frac{\bar{x}_i - \bar{\bar{x}}}{s}\right)}{n} = \frac{0\left(\frac{57-814{,}83}{243{,}11}\right) + 3\left(\frac{167-814{,}83}{243{,}11}\right) + \ldots + 0\left(\frac{1.597-843{,}83}{243{,}11}\right)}{253} = 0$$

Y la varianza:

$$\bar{z}^2 = \frac{\sum_{i=1}^{15} f_i\left(\frac{\bar{x}_i - \bar{\bar{x}}}{s}\right)^2}{n-1}$$

$$= \frac{0\left(\frac{57-814{,}83}{243{,}11} - 0\right)^2 + 3\left(\frac{167-814{,}83}{243{,}11} - 0\right)^2 + \ldots + 0\left(\frac{1.597-843{,}83}{243{,}11} - 0\right)^2}{253-1} = \frac{252}{252} = 1$$

2.22 La distribución Normal Estándar.

Con los datos estandarizados se descubrió *La Distribución Normal Estándar*, con probabilidades idénticas a *La Distribución Normal*, pero con la ventaja de usar *Números Puros* basados en un sistema numérico definido por *Una Desviación Estándar*, cuya valor está definido por:

$$f(z;0;1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}$$

Y cuyas probabilidades acumulativas se encuentran resolviendo:

$$F(z;0;1) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2} dz$$

Supongamos una media ideal ubicada al centro de la distribución de promedios de los datos agrupados, esto es:

$$\overline{X} = \frac{\overline{x}_{\min} + \overline{x}_{\max}}{2} = \frac{167 + 1.487}{2} = 827$$

2.23 Preparando el gráfico de probabilidad estándar.

Y con la misma desviación estándar obtenemos las probabilidades del intervalo de manera similar a cuando se usó la Normal. Los límites estandarizados de la clase 1 serían;

$$z_{I1} = \frac{2-827}{243{,}11} = \text{-}3{,}3935; \quad z_{S1} = \frac{112-827}{243{,}11} = \text{-}2{,}9410$$

Las probabilidades respectivas (en las columnas D y E). Recuerde hacer 0 la probabilidad del límite inferior de la clase 1, y 1 el límite superior de la clase 15;

$$F(z_{S2};0;1) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\text{-}2{,}4886} e^{-\frac{1}{2}\left(\frac{2-827}{243{,}11}\right)^2} dz =$$

$$= \text{DISTR.NORM.ESTAND}(\text{-}2{,}4886) = 0{,}00641$$

La probabilidad del intervalo se obtiene restando de la probabilidad del límite superior, la inferior;

$$P(C_2) = F(z_{S2}) - F(z_{I2}) = 0{,}00641 - 0{,}00164 = 0{,}00478$$

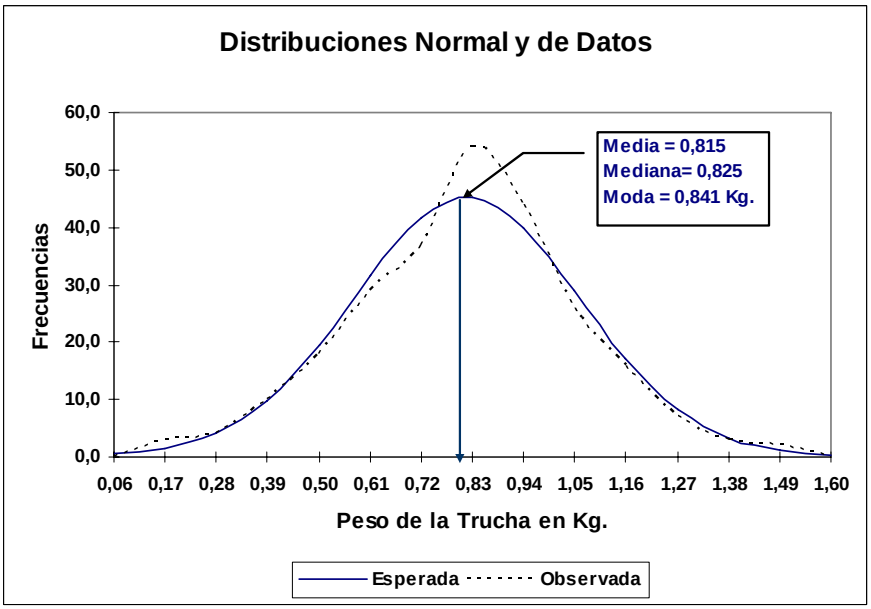
2.24 La distribución ideal y de aproximación.

Clase	Límites		Probabilidad Acumulada		Probabilidad	Frecuencias	
	Inferior	Superior	L. Inferior	L. Superior	Intervalo	Ideal	Aproximación
1	-3,3935	-2,9410	0,00000	0,00164	0,00164	0,4	0,5
2	-2,9410	-2,4886	0,00164	0,00641	0,00478	1,2	1,4
3	-2,4886	-2,0361	0,00641	0,02087	0,01446	3,7	4,1
4	-2,0361	-1,5836	0,02087	0,05664	0,03577	9,0	9,9
5	-1,5836	-1,1312	0,05664	0,12899	0,07235	18,3	19,5
6	-1,1312	-0,6787	0,12899	0,24866	0,11967	30,3	31,6
7	-0,6787	-0,2262	0,24866	0,41051	0,16185	40,9	41,8
8	-0,2262	0,2262	0,41051	0,58949	0,17898	45,3	45,2
9	0,2262	0,6787	0,58949	0,75134	0,16185	40,9	40,0
10	0,6787	1,1312	0,75134	0,87101	0,11967	30,3	28,9
11	1,1312	1,5836	0,87101	0,94336	0,07235	18,3	17,1
12	1,5836	2,0361	0,94336	0,97913	0,03577	9,0	8,3
13	2,0361	2,4886	0,97913	0,99359	0,01446	3,7	3,3
14	2,4886	2,9410	0,99359	0,99836	0,00478	1,2	1,1
15	2,9410	3,3935	0,99836	1,00000	0,00164	0,4	0,4
				Sumas	1,00000	253,0	253,0

Notará que la distribución ideal es equilibrada.

2.25 *Diferencias entre lo ideal y lo aproximado.*

En este problema en el que la distribución de Datos muestra pocas diferencias con respecto a la normal, hace que el sesgo negativo representado por el corrimiento de la curva de color guinda hacia la izquierda con respecto a la curva ideal representada en color azul.



Aun es más difícil por escaso, el alargamiento o curtosis de la curva de color guinda.

Sin embargo, no trabaja con la distribución *Ideal* si no con la de *Aproximación,* que reflejará las deformaciones de los datos de una manera suavizada por la *Esperanza* de un comportamiento poblacional uniforme.

2.26 *Consecuencias de que la distribución de datos sea normal.*

En muchos casos es necesario llevar algún sistema de selección del *Sujeto Estudiado*. Por ejemplo la selección genética que se hace en la agronomía y zootecnia; el control de la producción que se efectúa en la industria; el control de la calidad de productos manufacturados. En fin, un sinnúmero de situaciones en las que la unidad que proporciona los datos es sujeta a un proceso de selección.

CUALQUIER TIPO DE SELECCIÓN DE INDIVIDUOS (en particular) POR SUS CARACTERÍSTICAS FÍSCAS O CALIDADES PUEDE EFECTUARSE CON SEGURIDAD SÍ LA DISTRIBUCIÓN DE LA VARIABLE ES NORMAL (De las diferencias de la observación con respecto a la media).

De otra forma, los procesos de valoración que implique la selección se hace más complicada, al menos desde el punto de vista estadístico, teniendo que usar procesos menos confiables o más laboriosos.

2.27 *El Intervalo de Confianza.*

La Teoría Estadística ha desarrollado un postulado probabilístico fundamental en el desarrollo de la investigación planificada que involucra la responsabilidad del que investiga y la necesidad de quién creerá en los resultados de la investigación. Este se formaliza como:

$$\Pr\{\bar{x}_o - z(s) \geq \mu \leq \bar{x}_o + z(s)\} \leq 1 - \alpha$$

En donde:

El promedio de la población que se quiere estimar se indica como μ;

El límite inferior para μ lo establece el promedio de la muestra menos z veces la desviación estándar;

El límite superior para μ lo establece el promedio de la muestra más z veces la desviación estándar:

La probabilidad que determina a z se conoce como α o nivel de confianza.

2.28 Interpretación del Intervalo de Confianza.

El *Postulado Probabilístico* establece límites para el parámetro del promedio de la población μ usualmente desconocido considerando puntos importantes como:

El nivel de confianza α es la probabilidad que establece el investigador para no emitir una recomendación equivocada;

Esta probabilidad determina, en *La Distribución Normal Estándar* el escalador z;

Qué, multiplicado por *La Desviación Estándar* de la muestra establece *La Precisión de la Estimación*, que espera quién hará uso de las recomendaciones.

Una combinación de confiabilidad y precisión.

2.29 Problema de límite inferior: ¿Qué porcentaje de truchas pesan menos de 500 gramos?

Un poco de álgebra sobre el lado a la izquierda del postulado probabilístico dará el resultado resolviendo:

$$\Pr\{\bar{x}_o - z(s) \geq \mu \leq \bar{x}_o + z(s)\} \leq 1 - \alpha$$

En el cuadro de frecuencias los 500 gramos se ubican dentro de la clase 5. El número de individuos aproximado podría obtenerse interpolando. O bien, haciendo x_o del intervalo confiable igual a 500. Entonces;

$$\bar{\bar{x}} - z(s) = 500;$$

$$-z(s) = 500 + \bar{\bar{x}};$$

$$z = \frac{500 - \bar{\bar{x}}}{s} = \frac{500 - 814{,}83}{243{,}11} = -1{,}2950$$

Basta encontrar el área bajo la curva *Normal Estándar:*

$$F(z;0;1) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-1{,}2950} e^{-\frac{1}{2}\left(\frac{500-814{,}83}{243{,}11}\right)^2} dz =$$

$$= \text{DISTR.NORM.ESTAND(-1,2950)} = 0{,}0977$$

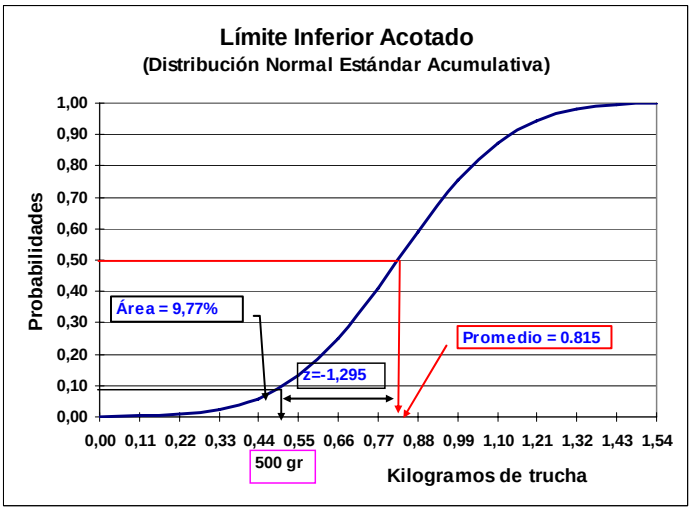
La respuesta es: el porcentaje esperado de truchas con peso inferior a 500 gramos es 9,77%. El conteo indica 28 truchas, lo que significa exactamente (28 × 100)/ 253 = 11,07%, una diferencia de poco más o menos 1, 30% debido, posiblemente, al sesgo a la izquierda ya reconocido.

2.30 Límite inferior acotado.

En el gráfico se representa la *Distribución Normal Estándar de aproximación* obtenida con los estimadores del problema.

No tenemos que calcular mucho para ubicar el promedio pues este parte la distribución de los datos en dos partes de igual probabilidad (50%)

De aquí hay que desplazarse (500-814,83)/243,11 veces a la izquierda, esto es, −1,295 veces. Este valor determina una probabilidad acumulativa de 0,0977 o 9,77%, representada por el área bajo la curva que está a la izquierda del límite de 500gr.



2.31 *Problema de límite superior: seleccionar el 20% de truchas más Pesadas.*

En este caso se requiere encontrar un valor que límite el 20% de los pesos más altos. Para esto se requiere encontrar el valor z que determina el 80% de probabilidad acumulativa en La Distribución Normal Estándar.

$$z_{(0,8)} = \int_0^z f(z;0,1)^{-1} dz = 0,8416$$

Sustituyendo el postulado probabilístico;

$$\bar{\bar{x}} + z(s) = x_o;$$

$$x_o = 814,83 + 0,8416(243,11);$$

$$= 1.019 \text{ gr.}$$

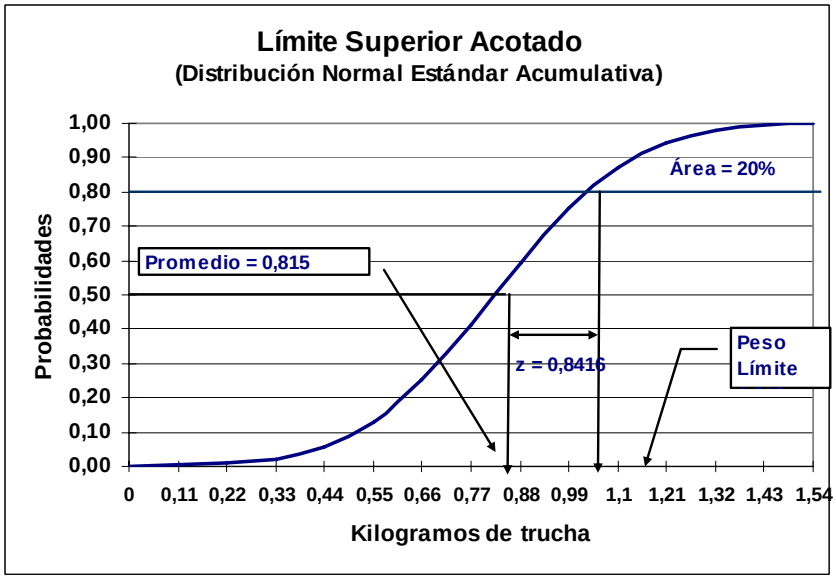
La comprobación se efectúa directamente con la HE pidiendo que cuente las truchas con un peso mayor o igual a 1.019 gramos. El resultado son 44 Truchas de 253 dan un 17,39% una diferencia de 2,61%.

$$P_{(20\%>)} = \frac{44 \times 100}{253} = 17,39\%$$

2.32 *El límite superior acotado.*

El 80% de truchas más pesadas determina un valor de z = 0,8416, este multiplicado por la Desviación Estándar y sumado al promedio indicó un peso de 1,019 Kg separa al 20% de Truchas más pesadas.

El conteo indicó 44 pesos iguales o superiores a 1,019 Kg., un 17,39% de los pesos. Nuevamente la proporción observada difiere de la esperada, debido a las diferencias con la normal teórica. No obstante, estos valores se espera que sean más exactos que otros estimados mediante otros métodos, aun cuando con los datos del ejemplo resulten más precisos.



2.33 *Trucha comercial en porcentajes: intervalo cerrado*

El peso comercial de la trucha de pesca en estanque está entre 750 y 1.000 gramos.

$$z_I = \frac{750 - 814{,}83}{243{,}11} = -0{,}2667; \quad z_S = \frac{1.000 - 814{,}83}{243{,}11} = 0{,}7617$$

$$F(z;0;1) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-0{,}2667} e^{-\frac{1}{2}\left(\frac{750-814{,}83}{243{,}11}\right)^2} dz = \text{DISTR.NORM.ESTAND}(-0{,}2663) = 0{,}3949$$

y

$$F(z;0;1) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+0{,}7617} e^{-\frac{1}{2}\left(\frac{1.000-814{,}83}{243{,}11}\right)^2} dz =$$

$$= \text{DISTR.NORM.ESTAND}(0{,}7617) = 0{,}7769$$

En este caso se procede a ubicar las desviaciones *z* para 750 que es de –0,2667 y para 1.000 que se calcula en 0,7617. Las probabilidades acumulativas que determinan son: 0,3949 y 0,7769. Efectuando la operación esquematizada:

$$F(z_S) - F(z_I) = 0{,}7769 - 0{,}3946 = 0{,}3820$$

Se estima la probabilidad del intervalo en 0,7769 – 0,3946 = 0,3820. Efectuando el conteo de la misma manera usando la herramienta que ofrece la HE indica: N(x < 750) = 86; para N(x > 1.000) = 50. Por tanto habrá 253 – 86 – 50 = 117 truchas con el peso en el intervalo, que corresponde a 46,25%.

2.34 *Intervalo interior.*

Varias conclusiones se pueden tener del intervalo interior determinado por el peso comercial de la trucha. Esto significa que entre más se encuentren en este rango mayor será la utilidad de le empresa.

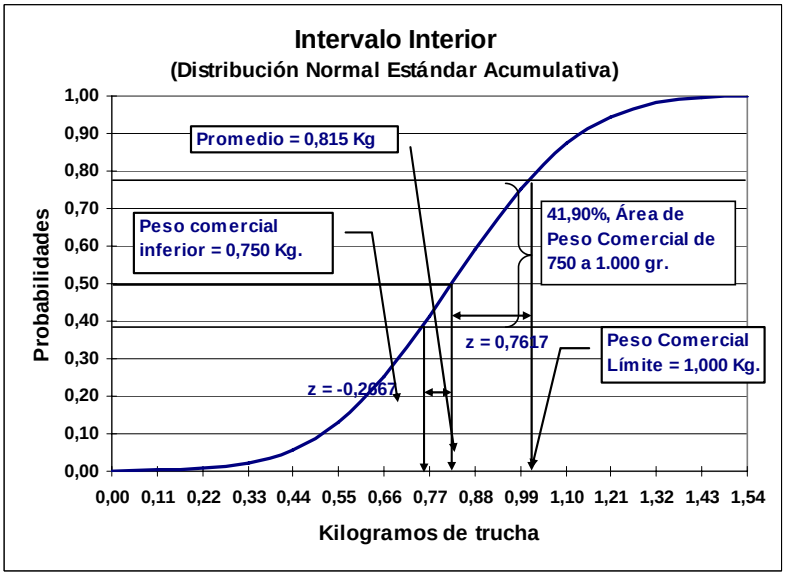
$$\hat{x}_I = \frac{86 \times 100}{253} = 33{,}99\%$$

No alcanzan el peso mínimo.

$$\hat{x}_S = \frac{50 \times 100}{253} = 19{,}76\%$$

Sobrepasan el máximo comercial.

Posiblemente, el impacto psicológico de pescar truchas grandes sea tanto más importante que vender al mínimo. Por tanto, es imperativo que la empresa encuentre la manera de reducir el porcentaje de peces de bajo peso.



2.35 ¿Qué sucede si la distribución de los datos no es Normal.

Se recalca que las respuestas estadísticas se pueden ofrecer cuando se involucran valores individuales, sí sólo sí la distribución de las diferencias con respecto al promedio es normal. O lo que se entiende regularmente como una *Distribución Normal*.

Cuando la distribución de las desviaciones no es normal, el uso de la aproximación mediante las Distribuciones Estadísticas Normales suele no ser la indicada. Para estos casos, se usa la distribución de orden estadístico o distribución libre que se estudia en el siguiente capítulo.

El hecho es que al experimentador siempre obtendrá una respuesta basada en métodos estadísticos a sus necesidades de información.

2.36 Estadísticos de estadísticos, una solución simple.

Supongamos que la empresa piscícola tiene una cantidad grande de pozas. A cada una de ellas se les toman datos y realizan estudios estadísticos más o menos por la misma época de alta demanda por pescadores "de fin de semana".

Si todos los resultados de los diferentes estudios estadísticos se arreglan de manera que cada columna corresponda a un estadístico, empezando con el promedio y terminando con el tamaño de la muestra. Al finalizar la época de alta demanda se podrán pedir las *Estadísticas Descriptivas* de los resultados de las diferentes pozas.

Esto es, una *Estadística Descriptivas de Estadísticas*. Y cada estadístico tendrá una muy particular distribución de las diferencias, muchas de ellas normal.

2.37 Una herramienta poderosa.

Centremos nuestro interés en los promedios. Estos, sin duda, mostrarán una *Distribución Normal* de las diferencias con respectos al promedio de todos los promedios. Esta conclusión práctica fue cristalizada en un teorema matemático que ha sido una herramienta poderosísima para el análisis científico de las poblaciones, *El Teorema central del Límite*. Que dice:

Una población definida por sus parámetros, media μ y varianza σ^2 finita, y siendo $\bar{x}_k$ la media de una muestra aleatoria de tamaño n de esa población. La variable;

$$\gamma_k = \frac{\bar{x}_k - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\sqrt{n}(\bar{x}_k - \mu)}{\sigma}$$

La distribución de la variable estandarizada γ_k se aproxima a la distribución Normal Estándar cuando n crece.

2.38 Interpretación del Teorema.

Lo trascendente del *Teorema Central del Límite*, es que asegura que independientemente de cuál sea la distribución de las diferencias de los datos, las diferencias de los promedios:

SE DISTRIBUIRÁ COMO UNA NORMAL ESTÁNDAR.

Esto significa, para el caso del piscicultor, que podrá utilizar todas las facilidades que ofrece la *Estadística de la Distribución Normal*, que involucra entre otras áreas, Las Técnicas de Muestreo, El Análisis de Experimentos, Los Métodos Cuantitativos, Las Técnicas de Simulación y Emulación. Una importante proporción de sistemas y métodos de análisis de resultados en diferentes ciencias.

2.39 Un ejemplo ayuda.

Estadístico	Poza 1	Poza 2	Poza 3	Poza 4	Poza 5	Poza 6	Poza 7	Poza 8	Poza 9	Poza 10
Media	874,40	871,73	909,01	870,31	905,80	719,16	868,23	877,52	880,03	894,65
Error típico	16,0068	13,7655	16,4283	18,8907	13,8652	15,0169	15,0511	18,8086	16,3861	18,8245
Mediana	870	867	903	864	899	714	870	873	872	892
Moda	813	867	890	856	899	714	870	873	872	892
Desviación estándar	254,6042	232,7952	274,4074	317,2283	232,0089	253,9584	254,0917	315,8509	274,1923	316,6763
Varianza de la muestra	64.823,2971	54.193,6080	75.299,4423	100.633,7728	53.828,1340	64.494,8723	64.562,6082	99.761,7807	75.181,4434	100.283,9091
Curtosis	0,1315	-0,3675	-0,2957	-0,3309	-0,2997	-0,3675	-0,3619	-0,3160	-0,29975	-0,3437
Coeficiente de asimetría	-0,0827	0,0072	0,1489	0,1036	0,1300	0,0072	0,0431	0,0835	0,13003	0,0222
Rango	1284	1089	1287	1485	1089	1188	1188	1485	1287	1485
Mínimo	216	328	292	151	382	126	282	153	261	157
Máximo	1500	1417	1579	1636	1471	1314	1470	1638	1548	1642
Suma	221.224	249.315	253.614	245.427	253.623	205.680	247.446	247.461	246409	253.186
Cuenta	253	286	279	282	280	286	285	282	280	283

Supongamos que la empresa tiene 10 pozas del mismo tamaño y los datos de peso de las capturas en la temporada de alta demanda turística se llevan pez a pez. En la HE se muestran los datos de las 10 pozas con lo siguientes estadísticos.

La empresa está buscando alternativas pues en ocasiones, el retraso que ocasiona el registro de peso, pez a pez incomoda al cliente.

2.40 El Desvío Típico.

Es fundamental para la utilización de las ventajas que ofrece el *Teorema Central del Límite,* interpretar correctamente él *Desvío Típico* o *Desvío Estándar de los promedios que está* definido por:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Para poblaciones grandes que son las apropiadas para la Distribución Normal Estándar. Para poblaciones pequeñas debe usarse:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}\sqrt{1 - \frac{1}{N}}$$

El agregado a la primera fórmula se conoce como *fracción muestral*, que se aproxima a 1 cuando *N* es grande. El término grande es muy relativo, usualmente 30, para el curso se considerará

una población como "grande" cuando tenga más de 120 elementos. La razón se debe al uso de la Distribución "t", adecuada a muestras pequeñas que se explicará más adelante.

2.41 La población y las muestras.

Es evidente que en el ejemplo tenemos 10 subpoblaciones de diferente tamaño de una población de 2.796 pesos de peces. Es posible obtener datos de dos fuentes sobre el mismo estadístico: Una obtenida dato a dato y otra que proviene de la información de los promedios de las muestras, como se muestra en el cuadro:

Nos interesan los promedios y las desviaciones típicas involucradas en el Teorema. Los promedios son muy parecidos, lo que se explica por otro teorema que dice: *La Esperanza Matemática de los Promedios es el Promedio Poblacional.*

Estadístico	Toda la Población	*Promedio de* Estadístico
Media	866,73	867,08
Error típico	5,2756	16,3044
Mediana	864	862
Moda	690	855
Desviación estándar	278,9574	272,5814
Varianza de la muestra	77.817,2120	75.306,2868
Curtosis	-0,1581	-0,2851
Coeficiente de asimetría	0,0713	0,0593
Rango	1.516	1286,70
Mínimo	126	234,80
Máximo	1.642	1521,50
Suma	2.423.385	242.338,50
Cuenta	2.796	279,60

2.42 Los Desvíos Típicos.

El *Desvío Típico* muestra una diferencia importante, el obtenido de la población es de:

$$S_{\mu} = \frac{S}{\sqrt{N}} = \frac{278{,}9574}{\sqrt{2.796}} = 5{,}2756$$

Mientras que el obtenido mediante el promedio de las 10 muestras es de 16,3044 gramos.

Usando la misma fórmula pero dividiendo la desviación estándar de la población por el promedio de unidades de las submuestras se obtiene:

$$s_{\bar{x}} = \frac{S}{\sqrt{n_o}} = \frac{278{,}9574}{\sqrt{279{,}60}} = 16{,}6828$$

Valor muy aproximado al obtenido con toda la población.

Es evidente que la llave para las estimaciones está en considerar el correcto valor de *n*.

2.43 El proceso de estimación.

Al empresario le interesa saber ¿qué tan aproximado estuvo el peso promedio de sus pozas al peso comercial?

La teoría estadística ha desarrollado mecanismos que permites obtener intervalos de confianza probabilística para los promedios. La ventaja es que estos no requieren que la población original de datos se distribuya normal. Para responder la pregunta, debe considerarse un límite probabilístico suficientemente preciso para ser útil al empresario; para el caso sería suficiente un 95%.

Los límites se obtienen usando *La Distribución Normal Estándar* aconsejada por el *Teorema Central del Límite*.

Algunos peces pueden pesar más de los límites comerciales, otro pueden pesar menos, por tanto debe establecerse una prueba para ambos lados de la distribución de diferencias o de dos colas.

2.44 Intervalo de Confianza para promedios.

Una modificación al postulado probabilístico ya conocido ofrece la solución:

$$\Pr\left\{\bar{x}_k - z\frac{s}{\sqrt{n}} \geq \overline{X} \leq \bar{x}_k + z\frac{s}{\sqrt{n}}\right\} \leq 1-\alpha$$

Partiendo de los datos de la población, como pudiera ser el caso, se pueden determinar los límites para los promedios definido un nivel confiable α para el ejemplo de 5%. El valor ***z*** que determina un 2,5% de probabilidad de menos infinito a **–*z*** y un 2,5% de ***z*** a más infinito es 1,96, por tanto, el área entre ambos límites será de 95%.

En este caso se habla de un *Intervalo Confiable de 95% de probabilidad*. Esto es, de cada 20 muestras el promedio de una muestra puede escapar al intervalo hacia arriba o hacia debajo de los límites.

2.45 Y ¿para una muestra promedio de 280 peces?

Para el límite inferior:

$$\Pr\left\{\overline{X} - z\frac{S}{\sqrt{n}}\right\} = \Pr\left\{\left(866{,}73 - 1{,}96\frac{278{,}9574}{\sqrt{280}}\right) < 834{,}06\right\} = 2{,}5\%$$

Para el límite superior:

$$\Pr\left\{\overline{X} + z\frac{S}{\sqrt{n}}\right\} = \Pr\left\{\left(866{,}73 + 1{,}96\frac{278{,}9574}{\sqrt{280}}\right) > 899{,}41\right\} = 2{,}5\%$$

La inspección de los promedios de las 10 pozas: 874,4; 871,73; 909,01; 870,31; 905,80; 719,16; 868,23; 877,52; 880,03 y 894,65 indican que 2 de los promedios sobrepasan el límite superior y 1 no alcanza el inferior.

Al ser más del 5% las muestras que salen de los rangos se debe sospechar que en las pozas 3 y 5 hay factores que hacen que las truchas ganen más peso y en la poza 6 lo pierdan. Con respecto al promedio de la población.

2.46 Lo usual es el proceso inverso.

Tomar el peso de las 2.796 truchas pescadas no es lo rentable, por esto, el empresario busca una alternativa. Lo Usual es tomar una única muestra para estimar lo que ocurre en la población, o repartir la muestra en las 10 pozas. Supongamos que se toma al azar una de las diez muestras. El generador indicó la poza 3. El intervalo confiable con los estadísticos de esta poza es:

$$\Pr\left\{909{,}01 - 1{,}96\frac{274{,}4074}{\sqrt{280}} \geq \overline{X} \leq 909{,}01 + 1{,}96\frac{274{,}4074}{\sqrt{280}}\right\} \leq 95\%$$

$$\Pr\left\{876{,}87 \geq \overline{X} \leq 941{,}15\right\}$$

Se asegura con 95% de confiabilidad que los promedios obtenidos con un poco más o menos 279 pesos fluctuarán entre 876,87 y 941,15 gramos. Cinco de los promedios están por debajo de estos límites.

2.47 ¿Qué cantidad de peces se deben medir? O tamaño de la muestra.

Tomar una poza al azar no le parece al empresario que sea el proceso correcto —y no lo es—, así se lo hace saber al ingeniero responsable. Operando con el mismo intervalo confiable y partiendo de la alternativa:

$$\Pr\left\{\mu = \bar{x}_k \pm z\frac{s}{\sqrt{n}}\right\} = 1-\alpha$$

Operando dentro del paréntesis y en una dirección se despeja n;

$$z\frac{s}{\sqrt{n}} = \bar{x}_k - \mu;$$

$$z^2\frac{s^2}{n} = (\bar{x}_k - \mu)^2;$$

$$n = \frac{z^2 s^2}{(\bar{x}_k - \mu)^2} = \frac{z^2 s^2}{d^2}$$

Se llega a la fórmula que determina el tamaño de muestra apropiado:

Donde d^2 es la precisión que conviene al proyecto y debe definir el investigador, puesto que μ es desconocida.

2.48 Definiendo confiabilidad y precisión en la estimación.

Puesto que el margen comercial del peso de la trucha de 750 a 1.000 gramos, esto es, un peso medio de 750 gramos, es amplio, no se requiere mucha precisión. Un 5% alrededor del promedio sería suficiente. Por tanto, tomando los estadísticos del muestreo de la poza 3:

$$d = 909{,}01 \times 0{,}05 = 45{,}45$$

Y la cantidad de peces que se deben muestrear serían:

$$n = \frac{1{,}96^2 \times 75.299{,}4423}{45{,}45^2} = 140$$

Dada la naturaleza de las instalaciones sería lógico pensar que deberían tomarse al azar el peso de 14 muestras de pesos de las truchas capturadas en cada una de las pozas, o pesar 14 truchas en cada poza.

Para continuar con el análisis es necesario obtener una muestra aleatoria de 14 peces en cada un de las posas. Para esto, el estudiante se puede auxiliar con una tabla de números aleatorios, crearla mediante la instrucción:

=ENTERO(ALEATORIO()*286)+1

U obtenerlos del generador de la HE.

Debe recordarse que debe pasar el conjunto de números aleatorios de la hoja Generador a la de Ejercicio con pegado especial valores pues los números aleatorios se generan constantemente.

El proceso de selección es simple:

1. ***Tome el primer número aleatorio;***
2. ***Marque la unidad a la que hace referencia;***
3. ***Copie esa unidad a un sitio aparte de la hoja en donde quedará la muestra;***
4. ***No puede seleccionar dos veces una unidad, si ya fue elegida, ignórela y escoja otro número aleatorio;***
5. ***Prosiga el ciclo hasta que complete 14 unidades en cada poza.***

2.49 La elección aleatoria de las muestras es absolutamente indispensable.

Muestra	Poza 1	Poza 2	Poza 3	Poza 4	Poza 5	Poza 6	Poza 7	Poza 8	Poza 9	Poza 10
1	905	1285	1085	601	1020	786	882	1503	1067	427
2	745	515	734	1051	668	402	678	588	352	427
3	1151	1186	526	766	910	810	666	1068	612	1117
4	437	911	877	901	514	126	918	558	443	817
5	925	1131	604	976	646	702	750	213	729	1567
6	831	1098	1085	766	855	558	786	1038	1548	547
7	940	922	864	1216	657	702	510	1218	807	517
8	911	1043	890	706	855	450	714	858	924	727
9	1244	680	955	1111	1339	942	1074	573	664	742
10	1039	592	422	1621	855	1002	714	648	378	442
11	682	1076	1488	616	1339	474	858	1128	443	1267
12	1214	889	1033	946	976	582	618	828	651	982
13	622	1395	396	1036	1009	678	1158	1413	872	727
14	738	966	1254	1201	965	1266	966	423	885	442

Cuando se tiene un marco de muestreo como el de los pesos de las truchas en las diez posas el problema es simple de solucionar pues basta elegir al azar 14 pesos en cada una de las pozas tal como se muestra en la HE. Aquí se presentan los datos finales.

2.50 Las estadísticas descriptivas de la muestra.

El primer análisis de los resultados se hace mediante las Estadísticas Descriptivas. Para el caso se presentan considerando toda la muestra y los promedios de las 10 pozas.

Los estadísticos indican que la distribución de los datos es normal en sus desviaciones. Resultado esperado pues sabemos que la distribución de las desviaciones de los pesos en la población es normal.

Estadístico	Toda la Muestra	Promedio por Poza
Media	845,43	845,43
Error típico	24,88	75,72
Mediana	855,00	836,40
Moda	855,00	713,14
Desviación estándar	294,36	283,33
Varianza de la muestra	86.649,2539	83.581,2967
Curtosis	-0,15	0,22
Coeficiente de asimetría	0,28	0,38
Rango	1495,00	1003,80
Mínimo	126,00	409,10
Máximo	1621,00	1412,90
Suma	118.360	11836,00
Cuenta	140	14,00

2.51 Intervalo de Confianza 95%.

Una de las primeras estimaciones que se realizan con los estimadores obtenidos de un muestreo es corroborar las condiciones del tamaño de muestra y ofrecer una predicción. Cómo la cantidad usual de las pozas anda sobre 285 peces, este es el tamaño que se usará para obtener el Desvío Típico.

$$\Pr\left\{845{,}43 - 1{,}96\frac{294{,}36}{\sqrt{280}} \geq \overline{X} \leq 845{,}43 + 1{,}96\frac{294{,}36}{\sqrt{280}}\right\} = 95\%$$

$$\Pr\left\{810{,}95 \geq \overline{X} \leq 879{,}91\right\} = 95\%$$

La precisión de la estima se calcula restando el límite inferior al superior y dividiendo por 2, en el ejemplo 34,48 gramos. Que implican un (34,42 * 100) / 845,43 = 4,1%, menos de lo fijado en el cálculo del tamaño de la muestra.

2.52 Interpretación.

El Intervalo de Confianza 95% para 280 peces indica los límites esperados para los pesos promedios en cualquiera de las pozas, y un *peso promedio* de cada 20 quedará fuera de los límites del intervalo en ausencia de factores que afecten el peso de las truchas.

Para poder establecer un *Intervalo de Confianza* para menos de 120 muestras:

- En este ejemplo se puede sospechar que hay factores en las pozas que afectan el peso de las truchas. Esto se puede corroborar elaborando un intervalo confiable para una muestra de 14 peces.
- El problema, es que la *Distribución Normal Estándar* se puede utilizar con muestras grandes (se fijó un límite de 120 individuos).

2.53 Las pequeñas muestras.

A principio del siglo XX *Gosset*, publicó bajo el pseudónimo de *Student* una función de probabilidad definida por:

$$Y = \frac{Y_o}{\left(1+\frac{t^2}{\upsilon}\right)^{\frac{\upsilon+1}{2}}}$$

En donde $\boldsymbol{Y}_0$ es una constante que depende de $\boldsymbol{N}$, de modo que el área bajo la curva sea uno, y donde la constante $\upsilon = (N-1)$ se llama *El número de grados de libertad* (υ es la letra griega *nu)*.

Y "t" está definida por la fórmula que aparece al lado derecho. Es evidente su parecido con la variable estandarizada **z**. La diferencia que la cantidad de muestras es pequeña.

$$t = \frac{\bar{x} - \bar{X}}{\frac{s}{\sqrt{n}}}$$

Cuando $n \geq 30$ la distribución de *t de Student* y la distribución *Normal Estándar* se aproximan estrechamente.

Guillermo Sealy Gosset (13 de junio de 1876 – 16 de octubre 1937) era químico y estadístico, mejora conocido como 'estudiante de la pluma'. Nace en Cantorbery, Inglaterra, sus padres a Agnes Sealy Vidal y coronel Frederic Gosset, Gosset estudió en la universidad de Winchester, la escuela privada famosa, de química y en la universidad. Se gradúa en Oxford en 1899, trabajando en la cervecería de Dublín de Arturo Guinness y hijo.

Otro investigador en Guinness había publicado previamente los secretos comerciales que contenían de papel de la cervecería de Guinness. Para prevenir acceso adicional de la información confidencial, Guinness prohibió a sus empleados publicar cualquier papel sin importar la información contenida. Esto significó que Gosset no podía publicar sus trabajos bajo su propio nombre. Por lo tanto él utilizó "Student" (estudiante) como seudónimo. Por lo tanto su logro más famoso ahora se refiere como la distribución "t" de Stúden que pudo haber sido la distribución de "t" de Gosset.

http://www-groups.dcs.st-and.ac.uk/~history/Biographies/Gosset.html

2.54 Cambios al Intervalo de Confianza.

La HE ofrece la distribución de las probabilidades de "t" y la distribución de los valores de "t" dadas la probabilidad y los grados de libertad. Aplicando la teoría del intervalo confiable, para pequeñas muestras queda definida por:

$$\Pr\left\{\bar{x} - t_{(gl;\alpha)}\frac{s}{\sqrt{n}}\sqrt{1-\frac{1}{N}} \geq \bar{X} \leq \bar{x} + t_{(gl;\alpha)}\frac{s}{\sqrt{n}}\sqrt{1-\frac{1}{N}}\right\} = 1-\alpha$$

Para el ejemplo, con n = 14 – 1 grados de libertad y nivel de significación $\alpha = 0{,}05$. El intervalo sería:

$$\Pr\left\{845{,}43 - 2{,}1604\frac{294{,}36}{\sqrt{14}}\sqrt{1-\frac{1}{285}} \geq \bar{X} \leq 845{,}43 + 2{,}1604\sqrt{1-\frac{1}{285}}\right\} = 95\%$$

$$\Pr\left\{769{,}39 \geq \bar{X} \leq 921{,}46\right\} = 95\%$$

Se espera que para una muestra de tamaño 14 en una población de 285 peces el promedio oscile entre 769,39 y 921,46 gramos, con un error de un promedio fuera de los límites cada 20 muestreos.

En algunos libros y reportes estadísticos la ecuación del desvío típico para poblaciones pequeñas se refiere como:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}\sqrt{1-\frac{n}{N}} = s_{\bar{x}}\sqrt{\frac{1}{n}-\frac{1}{N}}$$

2.55 Buscando las diferencias en las pozas.

Se contaron 3 muestras por abajo del límite inferior y 2 por arriba del límite superior. Esto nos lleva a suponer que en las pozas hay factores que afectan el peso de las truchas.

Dos promedios de una misma población pueden compararse mediante la siguiente prueba de "t":

$$t_c = \frac{\bar{x}_A - \bar{x}_B}{S_d\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}}$$

El resultado algebraico de restar dos intervalos confiables y en donde S_d es la varianza de las diferencias con $n_1 + n_2 - 2$ grados de libertad.

$$S_d = \sqrt{\frac{(n_A-1)S_A^2+(n_B-1)S_B^2}{n_A+n_B-2}}$$

Un promedio ponderado de las varianzas. Este principio requiere que los grupos provengan de la misma población.

2.56 La Hipótesis y la Prueba.

En la comparación se plantea prueba de hipótesis estadística sobre los promedios con una *Nula* y la *Alternativa:*

Ho; $\mu_A = \mu_B$

H*a;* $\mu_A \neq \mu_B$

A un nivel de significación α. En la HE la prueba se puede efectuar de dos maneras:

Calculando el estadístico "*t*" y comparándolo con el que corresponde al nivel de confianza, ejemplificando con las pozas 1 y 2.

$$t_{(26;2)} = \frac{|884{,}57 - 977{,}79|}{91{,}4894} = \frac{93{,}21}{91{,}4894} = 1{,}0189$$

Determinado para la prueba de $t_{(0{,}05;\ 26)} = 2{,}0555$ que se usa como criterio de manera que si el estadístico calculado es mayor o igual al que criterio de la prueba se rechaza la hipótesis nula. En este caso, 1,0189 es inferior a 2,0555 por tanto la hipótesis <u>*NO SE RECHAZA*</u>.

2.57 Contraste de hipótesis usando la probabilidad.

La otra manera que la HE ofrece un valor para la prueba, y sea quizá más ilustrativa es la probabilidad que señala el estadístico en la función de densidad con respecto a la zona de aceptación de la hipótesis, esto es, entre los límites de confianza:

$$F(1{,}0189;26) = Y_0\int_0^{\infty}\left(1+\frac{1{,}0189^2}{26}\right)^{\frac{26+1}{2}} dt = 0{,}3177$$

La función acumulativa desde 0, o sea el 50% de la distribución hasta el punto t = 1,0189 indicó una probabilidad de 0,3177 que está aun dentro del intervalo confiable definido de 0,95 o 95% también llamada zona de aceptación de la hipótesis. Mientras la probabilidad no sea inferior a 5% se tomará la decisión de aceptar la hipótesis nula, esto es: no hay diferencia entre los promedios de las truchas capturadas en la poza 1 y la poza 2.

La HE ofrece la opción de calcular la prueba de "t" para diferentes opciones de varianza, en este caso se elige la opción 2 = varianza común. El resultado será idéntico al obtenido manualmente.

2.58 Una varianza común.

Se puede comprobar que para el caso en que el número de observaciones de la muestra es igual, la varianza de las diferencias es, simplemente, el promedio aritmético de las varianzas. Considerando una varianza promedio común se puede calcular una única varianza de las diferencias usando el promedio de las varianzas.

El promedio aritmético de las varianzas de las 10 pozas puede usarse como un buen estimador de la varianza común con tantos grados de libertad como el producto 10 ×14 – 10 = 130

Esto es:

$$S_d^2 = \sum_{i=1}^{10} S_i^2 = \frac{53.269,1868 + 63.915,2582 + ... + 124.491,7582}{10} = 83.581,2967$$

2.59 Prueba de Contrastes.

Contraste	Media 1	Media 2	Diferencia	Estadístico "t"	Valoración de Ho: Probabilidad	Valoración de Ho: Criterio DMS
Poza 2 vs Poza 4	977,79	965,29	12,50	0,1144	0,9091	NS
Poza 2 vs Poza 5	977,79	900,57	77,21	0,7066	0,4811	NS
Poza 2 vs Poza 1	977,79	884,57	93,21	0,8531	0,3952	NS
Poza 2 vs Poza 3	977,79	872,36	105,43	0,9648	0,3364	NS
Poza 2 vs Poza 8	977,79	861,21	116,57	1,0668	0,2880	NS
Poza 2 vs Poza 7	977,79	806,57	171,21	1,5669	0,1196	NS
Poza 2 vs Poza 10	977,79	767,71	210,07	1,9225	0,0567	NS
Poza 2 vs Poza 9	**977,79**	**741,07**	**236,71**	**2,1663**	**0,0321**	**Significativo**
Poza 2 vs Poza 6	**977,79**	**677,14**	**300,64**	**2,7513**	**0,0068**	**Significativo**
Poza 4 vs Poza 5	965,29	900,57	64,71	0,5922	0,5547	NS
Poza 4 vs Poza 1	965,29	884,57	80,71	0,7387	0,4614	NS
Poza 4 vs Poza 3	965,29	872,36	92,93	0,8504	0,3966	NS
Poza 4 vs Poza 8	965,29	861,21	104,07	0,9524	0,3427	NS
Poza 4 vs Poza 7	965,29	806,57	158,71	1,4525	0,1488	NS
Poza 4 vs Poza 10	965,29	767,71	197,57	1,8081	0,0729	NS
Poza 4 vs Poza 9	**965,29**	**741,07**	**224,21**	**2,0519**	**0,0422**	**Significativo**
Poza 4 vs Poza 6	**965,29**	**677,14**	**288,14**	**2,6370**	**0,0094**	**Significativo**
Poza 5 vs Poza 1	900,57	884,57	16,00	0,1464	0,8838	NS
Poza 5 vs Poza 3	900,57	872,36	28,21	0,2582	0,7967	NS
Poza 5 vs Poza 8	900,57	861,21	39,36	0,3602	0,7193	NS
Poza 5 vs Poza 7	900,57	806,57	94,00	0,8602	0,3912	NS
Poza 5 vs Poza 10	900,57	767,71	132,86	1,2158	0,2262	NS
Poza 5 vs Poza 9	900,57	741,07	159,50	1,4597	0,1468	NS
Poza 5 vs Poza 6	**900,57**	**677,14**	**223,43**	**2,0447**	**0,0429**	**Significativo**
Poza 1 vs Poza 3	884,57	872,36	12,21	0,1118	0,9112	NS
Poza 1 vs Poza 8	884,57	861,21	23,36	0,2138	0,8311	NS
Poza 1 vs Poza 7	884,57	806,57	78,00	0,7138	0,4766	NS
Poza 1 vs Poza 10	884,57	767,71	116,86	1,0694	0,2869	NS
Poza 1 vs Poza 9	884,57	741,07	143,50	1,3132	0,1914	NS
Poza 1 vs Poza 6	884,57	677,14	207,43	1,8983	0,0599	NS
::::	::::	::::	::::	::::	::::	::::

Y la Desviación Estándar de las diferencias es:

$$Sd = \sqrt{S_d^2} = \sqrt{83.581,2967} = 289,1043$$

Y la Desviación Típica de la comparación:

$$S_{\bar{d}} = S_d \sqrt{\frac{1}{n_A} + \frac{1}{n_B}};\ S_{\bar{d}} = S_d \sqrt{\frac{2}{n_o}} = 289,1043 \sqrt{\frac{2}{14}} = 109,2712$$

$$DMS = t_{(gl;\alpha)}\left[S_d\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}\right]$$

En una secuencia de comparaciones pareadas e irrepetibles.

La prueba de "t" pareada puede dar resultados muy aproximados a los obtenidos con el método de la varianza común de todas las pozas.

2.60 Resultados de la prueba.

A la derecha se ofrece el cuadro de promedios arreglado desde el mayor al menor.

Los promedios con letras iguales no mostraron diferencias significativas. Las diferencias de la poza 2 a la 10 deben considerarse debidas al azar.

El las pozas 2, 4 y 5 hay factores que ayudan a que las truchas pesen más que las de las posas 9 y 6; o, en las pozas 9 y 6 hay factores que deprimen el peso de las truchas.

La consecuencia es: ¡Se encontró un motivo para profundizar en el análisis del peso diferencial de las truchas según la poza en que se alimentan hasta su edad comercial!

Poza	Promedio	Letras iguales no hay diferencias
Poza 2	977,79	***a***
Poza 4	965,29	***a***
Poza 5	900,57	***ab***
Poza 1	884,57	***abc***
Poza 3	872,36	***abc***
Poza 8	861,21	***abc***
Poza 7	806,57	***abc***
Poza 10	767,71	***abc***
Poza 9	741,07	***bc***
Poza 6	677,14	***c***

En la experimentación se usan diferentes maneras de presentar los resultados de contrastar promedios de tratamientos experimentados. Uno de ellos es hacer acompañar a los promedios que no muestran diferencias suficientes para rechazar la hipótesis nula con letras iguales. Todos los tratamientos acompañados con la letra a no mostraron diferencias significativas o lo que el la jerga del investigador quiere decir que el promedio se encontró dentro de los límites de un intervalo de confianza de 95% ó que no llegó a la zona de probabilidad fuera de los limites del intervalo de confianza.

En ocasiones es problemático poder tomar una decisión, en este caso, claramente en dos pozas, la 2 y la 4 hay uno o más factores que propicia un mayor desarrollo y que en las pozas 9 y 6 factores que deprimen el peso del pez.

El siguiente paso del experimentador sería emprender experimentos o exploraciones que hagan manifiestos los factores.

2.61 Dos distribuciones más de la familia de la Normal.

Es conveniente mencionar que hay dos distribuciones más que interesan al curso que son familia de la normal:

- La Distribución de χ^2 chi-cuadrada o chi-cuadrada o ji-cuadrada (por la letra griega χ = chi o ji);
- Y, La Distribución de *F*.

La primera se utiliza para valorar hipótesis relacionadas con variables cualitativas (se usó en la diapositiva 18).

Y la segunda, que es también muy cercana a la distribución de "***t***", en técnicas estadísticas más avanzadas. Esta distribución, llamada de ***F*** (por su descubridor R. A. Fisher) valora el cociente de dos varianzas. Cuando se comparan dos grupos, el estadístico "***t***" elevado al cuadrado es igual al estadístico ***F***.

La distribución de χ^2 se estudiará en el capítulo referente a la aproximación de la distribución Binomial mediante la Normal. La Distribución de F se utiliza exhaustivamente en el Curso de Diseño y Análisis de Experimentos.

Sir Ronald Aylmer Fisher, (17 de febrero de 1890 - el 29 de julio de 1962) británico. Era estadístico británico, biólogo evolucionista, y genetista. Lo han descrito como: un genio que creó casi sin ayuda los fundamentos para la ciencia estadística moderna.

http://en.wikipedia.org/wiki/R.A._Fisher

2.62 La Prueba de F.

Cuando se comparan dos o más poblaciones de la misma naturaleza se cera una nueva varianza que corresponde a los promedios de éstas poblaciones definida por:

$$S_B^2 = \frac{\sum_{i=1}^{b} n_i\left(\bar{x}_i - \bar{\bar{x}}\right)^2}{b-1} = \frac{14(884{,}57 - 931{,}18)^2 + 14(977{,}79 - 931{,}18)^2}{2-1} = 60.822{,}3214$$

En donde ***b*** es el número de grupos. La Distribución de *F* está definida por el teorema:

Si S_1^2 *y* S_2^2 son las varianzas de muestras aleatorias independientes de tamaño n_1 y n_2 respectivamente, tomadas de dos poblaciones normales que tienen la misma varianza, entonces;

$$F_c = \frac{S_1^2}{S_2^2}$$

Es un valor de una variable aleatoria que tiene distribución F con parámetros $\nu_1 = n_1 - 1$ y $\nu_2 = n_2 - 1$.

2.63 En el Ejemplo.

Las dos nuevas varianzas participan en la prueba de ***F***:

$$F_c = \frac{S_1^2}{S_2^2} = \frac{S_B^2}{S_P^2} = \frac{60.822{,}3214}{58.592{,}2225} = 1{,}0381$$

Cociente que valorado mediante la el algoritmo para calcular probabilidades de ***F*** en la HE;

$$F(1{,}0381;1;26) = \int_0^{Fc} f\left(\frac{S_1}{S_2};\upsilon_B;\upsilon_P\right)df = 0{,}3177$$

Da la misma probabilidad que en la prueba de "t". Además, puede comprobar que el estadístico $\mathbf{t_c}$ elevado al cuadrado es igual al estadístico $\boldsymbol{F_c}$: [$(\boldsymbol{t_c})^2$ 1,0189] = [$\boldsymbol{F_c}$ = 1,0381].

Muestra	Poza 1	Poza 2
1	905	1285
2	745	515
3	1151	1186
4	437	911
5	925	1131
6	831	1098
7	940	922
8	911	1043
9	1244	680
10	1039	592
11	682	1076
12	1214	889
13	622	1395
14	738	966
Suma	12384	13689
Unidades	14	14
Promedio General		931,18
Promedios	884,57	977,79
Varianzas	53.269,1868	63.915,2582
varianza P		58.592,2225
Grados Libertad p		26
Varianza b		60.822,3214
Estadístico F		1,0381
Probabilidad de F		0,3177

La HE ofrece una rutina para comparar dos varianzas de dos poblaciones. EL objetivo de la misma es diferente al presentado en esta diapositiva como prueba de F en la que se valoran directamente los promedios.

La Prueba de F que proporciona la HE valora que las varianzas de dos grupos no sean significativamente diferentes. No es recomendable utilizarla pues no da resultados iguales a las pruebas de t que ofrece la misma HE.

2.64 Resumen.

La *Distribución Normal* de las diferencias de las observaciones con respecto al promedio, proporciona al investigador herramientas poderosas para el análisis de resultados.

La Distribución Normal Estándar universalizó el uso de la teoría estadística para variables continuas al crear un sistema numérico independiente de las unidades en que se midan las variables.

Esta herramienta se hizo aun más poderosa con la aparición del *Teorema Central del Límite*, que abrió oportunidades, mediante el uso de los promedios, a infinidad de circunstancias en que las distribuciones de los datos no son normales.

Otras distribuciones relacionadas con ésta como son: la "t" de Studen, la F de Fisher y la χ^2 ampliaron la gama de posibilidades de análisis de resultados al poder combinar las probabilidades con Técnicas Estadísticas avanzadas.

REFERENCIAS SELECTAS:

7. Hillier Frederick S., y Lieberman Gerard j., *Introducción a la Investigación de Operaciones*. Capítulo 19. Segunda edición en español traducida de la cuarta edición en ingles. McGraw-Hill Interamericana de México, S. A. De C. V., 1990.
8. Kazmier Leonard. *Estadística* Aplicada a la Administración y a la Economía. Capítulo 21. Editorial McGraw-Hill, México 1993.
9. Miller Irwin, Freund John E., Johnson Richard A: *Probabilidad y Estadística para Ingenieros*. Capítulo 11. Traducido de la cuarta edición en ingles; Prentice-Hall Hispanoamericana, S. A. 1992.
10. Murray R. Spiegel: *Serie de compendios Schaum, Teoría y Problemas de Estadística.* Capítulos 7. Primera edición en español, traducido de la primera edición en ingles; Libros McGraw-Hill de México, S. A. De C. V., 1973.
11. Ostle Bernard: *Estadística Aplicada*. Capítulos 3 y 4. Primera edición en español traducida de la primera edición en ingles. Editorial Limusa, S. A., 1977.
12. Hoel Paul G. *Estadística Elemental*. Capítulo 5. Editorial Continental, S. A. México 1982.
13. Snedecor George W., y Cochran William G: *Statistical Methods*. Capítulos 2 y 3. Sexta edición; The Iowa State University, 1974.
14. Richard Larry E; LaCava Jerry J. Estadística en los negocios ¿Por qué y cúando?. Capítulo 6. Editorial McGraw-Hill Latinoamericana S. A. Bogotá Colombia, 1980.
15. Ríos Sixto. Iniciación estadística. Capítulo 6. Ediciones ICE, Madrid 1977.
16. Steel Robert G. D., Torrie James H: *Principles and Procedures of Statistics*. Capítulo 4. Primera edición; McGraw-Hill Book Company, Inc, 1960.

3 Estadística no Paramétrica.

En esta sección se utilizan los archivos:

E03_ENo_Paramétrica.P01.
E03_ENo_Paramétrica.W01.
E03_ENo_Paramétrica.X01

3.1 Menú.

3.2 Introducción.

Muchas veces, las distribuciones de datos en experimentos planificados, en estudios de exploración mediante muestreos y otros tipos de datos no presentan una distribución que pueda aproximarse mediante distribuciones de densidad como son La Normal, La Normal Estándar y sus distribuciones asociadas; La Binomial, La Poisson y sus distribuciones asociadas. En estos casos, *HACER INFERENCIAS MEDIANTE DISTRIBUCIONES TÍPICAS ES POCO CONFIABLE*.

La salida para el experimentador es utilizar la estadística de *DISTRIBUCIÓN LIBRE.*

La distribución libre es la que se crea ordenando a los datos por su magnitud ascendente. Esto es, la que asocia a cada dato con una posición en el conjunto total de datos. Visto de otra forma, la distribución libre esta asociada al número índice correlativo a la posición.

3.3 Las diferencias.

En los capítulos anteriores se habló de la media, la varianza y la desviación estándar para efectuar los reconocimientos, exploraciones y recomendaciones basándose en distribuciones de densidad como la Normal y la Binomial.

En la estadística no paramétrica el parámetro de posicionamiento hacia el centro de la distribución de mayor utilidad es la *Mediana* y como medida de dispersión el *Rango* o alguna variante de éste.

Una característica importante es que la distribución de datos debe corresponderse con el Orden Estadístico. Esto es, con el número que tiene una variable en un conjunto de datos ordenados ascendentemente.

3.4 La Mediana y el Rango.

La Mediana es una medida de posicionamiento central que parte la distribución ordenada de los datos en dos subconjuntos que tienen la misma cantidad de observaciones (Recordar que la media parte la distribución en dos partes con la misma probabilidad). Se ubica en la posición:

$$\widetilde{x} = x_{k+1};\ \text{si}\ \ n = 2k+1$$

Si el número de observaciones es impar. Y

$$\widetilde{x} = \frac{x_k + x_{k+1}}{2};\ \text{si}\ n = 2k$$

Si el número de observaciones es par.

El rango es la amplitud o recorrido de los datos. Se obtiene mediante: r = Máximo – Mínimo

3.5 Una distribución básica.

En la estadística no paramétrica también llamada de *DISTRIBUCIÓN LIBRE,* la distribución ordenada de los datos es fundamental. Si x_i es la i-ésima observación de la variable ***X***, la distribución de estas debe considerarse que el *Mínimo* = x_1, el Siguiente más pequeño como x_2, así sucesivamente hasta el valor más alto que se corresponde como *Máximo* = x_n.

Esto implica que los datos se consideren asociados a los estadísticos de orden.

La distribución de densidad o probabilidad específica para cada conjunto de datos se obtiene dividiendo cada número ordinal entre n:

$$p(x_i) = \frac{i}{n}$$

3.6 Ejemplo 3.1. De producción industrial.

Una empresa que quiere introducir cojinetes de bolas al país ofrece precios mucho más económicos. Para convencer a los posibles representantes nacionales ofreció conferencias sobre la empresa. Siendo los cojinetes de bolas (baleros o roles) artículos de alta precisión, los fabricantes hacían hincapié en los estrictísimos procesos del control de la calidad.

Entre estos se extrajeron los parámetros de control para generar un conjunto de datos relacionados con el control de la medida interior de un cojinete específico. La medida original era en pulgadas.

El primer paso en cualquier análisis de datos formal es asegurarse que la distribución de aproximación se ajusta a la distribución de los datos. Para esto, se usa generalmente la tabla de frecuencias y una prueba de Bondad de Ajuste.

3.7 El Intervalo de Clases.

En el control de la calidad y los procesos y en general en los estudios estadísticos, para tomar decisiones es insoslayable conocer la distribución de los datos.

Para elaborar la tabla de frecuencias debe determinarse la cantidad de clases en las que se van a asignar los datos por su valor. Para esto, se recomienda que el *Intervalo de Clase* esté entre ½ y ¼ veces la desviación estándar.

$$IC_{>} = \frac{S}{2} = \frac{0{,}0037}{2} = 0{,}0018;\ IC_{<} = \frac{0{,}0037}{4} = 0{,}0009$$

Dividiendo el rango por los IC, se obtiene un estimado del número de clases.

$NC <= \frac{r}{IC >} = \frac{0{,}0148}{0{,}0018} = 8;\ NC >= \frac{0{,}0148}{0{,}0009} = 16$ Usando un IC = 0,0012 se obtendrán aproximadamente 13 clases.

3.8 Los límites de clases.

Es conveniente que la primera y última clase estén vacías. Para el límite inferior de la primera clase use:

$LIC_1 = Min - IC - 0{,}0001 = 0{,}6064 - 0{,}0012 - 0{,}0001 = 0{,}6051$

Para obtener el punto medio agregue la Límite Inferior de la Clase la mitad del intervalo de clase:

$\bar{x}_1 = LIC_1 + \frac{IC}{2} = 0{,}6051 + \frac{0{,}0012}{2} = 0{,}6057$

El Límite superior de la clase se consigue sumando el IC al límite inferior de la clase.

$LSC_1 = LIC_1 + IC = 0{,}60510{,}0012 = 0{,}6063$

3.9 El Cuadro de Frecuencias.

El diámetro de los roles es una variable continua, por tanto el límite inferior de una clase tendrá el mismo valor que el límite superior de la precedente. Dado que el conteo lo efectuará la HE con base al límite superior de las clases no es necesario hacer distinción. Para el conteo de la primera clase se usa la siguiente instrucción:

$f_1 = \text{CONTAR.SI(\$B\$11:\$I\$40;"<= 0,6063")} = 0$

Para el conteo de la segunda clase y subsiguientes:

$f_2 = \text{CONTAR.SI(\$B\$11:\$I\$40;"<= 0,6075")} - \text{SUMA(\$D\$65:D65)} = 11$

Los conteos son correctos, sí la suma de frecuencias es igual al total de las observaciones (240).

3.10 El Cuadro de Frecuencias.

Agregue en la HE las frecuencias acumulativas ascendente sumado a la frecuencia acumulativa anterior la frecuencia observada.

Las frecuencias acumulativas descendentes, restando a 240 las frecuencias de las clases. Obtendrá un cuadro similar al que aquí se presenta.

Límites de Clases			Frecuencias		
Inferior	Medio	Superior	Observadas	Ascendente	Descendente
0,6051	0,6057	0,6063	0	0	240
0,6063	0,6069	0,6075	11	11	229
0,6075	0,6081	0,6087	12	23	217
0,6087	0,6093	0,6099	14	37	203
0,6099	0,6105	0,6111	23	60	180
0,6111	0,6117	0,6123	17	77	163
0,6123	0,6129	0,6135	30	107	133
0,6135	0,6141	0,6147	36	143	97
0,6147	0,6153	0,6159	28	171	69
0,6159	0,6165	0,6171	22	193	47
0,6171	0,6177	0,6183	14	207	33
0,6183	0,6189	0,6195	13	220	20
0,6195	0,6201	0,6207	13	233	7
0,6207	0,6213	0,6219	7	240	0
0,6219	0,6225	0,6231	0	240	0

3.11 El orden medio y la Mediana.

El orden medio se obtiene:

$$k_{(50\%)} = \frac{n+1}{2} = \frac{240+1}{2} = 120{,}5$$

O sea que la mediana se estimará promediando la observación 120 y 121.
Solicitando a la HE el valor de la observación 120 y 121:

$$k_{120} = \text{K.ESIMO.MENOR}(\$B\$11:\$I\$40;120) = 0{,}6141$$

$$k_{121} = \text{K.ESIMO.MENOR}(\$B\$11:\$I\$40;121) = 0{,}6141$$

La medina será:

$$\tilde{x} = \frac{0{,}6141 + 0{,}6041}{2} = 0{,}6041$$

Con datos agrupados en el cuadro de frecuencias:

$$\tilde{x} = LI_m + \left(\frac{\frac{n+1}{2} - S_{m-1}}{f_m} \right) IC = 0{,}6135 + \left(\frac{120{,}5 - 107}{36} \right) 0{,}0012 = 0{,}6140$$

En donde:

LI_m = Límite inferior de la clase en que caerá la Mediana;
S_{m-1} = Suma acumulativa hasta la clase anterior a donde se ubicará la mediana;
f_m = frecuencia observada de la clase en donde se ubicará la mediana;
IC = Intervalo de Clase.

Cuando la distribución de los datos es normal y la muestra mayor de 30 observaciones, el Error Típico para la mediana se estima en;

$$\sigma_{\tilde{x}} = \sigma \sqrt{\frac{\pi}{2n}}$$

En datos individuales, también puede tomar el promedio de los 4 o 5 valores medios para obtener un valor de la mediana más consistente.

3.12 Cuartiles: el primer cuartil.

Los cuartiles son las medidas que separan a la población en 4 subconjuntos con la misma cantidad de unidades. Una medida de dispersión muy usada es el rango intercuartílico. Esto es: la diferencia entre los valores que separan a los dos cuartos centrales de la población:

$$r_{(75-25)} = \tilde{x}_{75} - \tilde{x}_{25}$$

Naturalmente que habrá que conseguir los valores que separan el primero y el tercer cuartil:

$$\tilde{x}_{(25)} = LI_{C1} + \left(\frac{\frac{(n+1)P}{100} - S_{C1-1}}{f_{C1}} \right) IC = 0{,}6111 + \left(\frac{60{,}25 - 60}{17} \right) 0{,}0012 = 0{,}6111$$

3.13 Cuartiles: El tercer cuartil y el Rango Intercuatílico.

$$\tilde{x}_{(75)} = LI_{C3} + \left(\frac{\frac{(n+1)P}{100} - S_{C3-1}}{f_{C3}} \right) IC = 0{,}6159 + \left(\frac{180{,}75 - 171}{22} \right) 0{,}0012 = 0{,}6164$$

El rango intercuartílico:

$$r_{(75-25)} = \tilde{x}_{75} - \tilde{x}_{25} = 0{,}6164 - 0{,}6111 = 0{,}0053$$

Valor que significa: entre el 50% de la población centrada hay 0,0055 pulgadas de diferencia. En términos relativos:

$$v.r_{(75-25)} = \frac{(\tilde{x}_{75} - \tilde{x}_{25})\times 100}{\tilde{x}} = \frac{0{,}0053\times 100}{0{,}6140} = 0{,}87\%$$

Valor que indica una variación muy baja. O una población muy homogénea. Suponga que se quiere controlar la producción para que se mantenga en un rango de 90%.

3.14 Percentiles y rango percentílico.

El rango percentílico 90% indica al lector que el experimentador quiere poner de manifiesto o asegurar que su fábrica produzca dentro del el rango entre el percentil 5:

$$\tilde{x}_{(5)} = LI_{P(5)} + \left(\frac{\frac{(n+1)P}{100} - S_{p(5)-1}}{f_{p(5)}} \right) IC = 0{,}6075 + \left(\frac{12{,}05-11}{12} \right) 0{,}0012 = 0{,}6076\_$$

Y el percentil 95:

$$\tilde{x}_{(95)} = LI_{P(k)} + \left(\frac{\frac{(n+1)P}{100} - S_{p(k)-1}}{f_{p(k)}} \right) IC = 0{,}6207 + \left(\frac{238{,}45-233}{7} \right) 0{,}0012 = 0{,}6216$$

Para su análisis e interpretación, estos estadísticos pueden representarse en el grafico de las ojivas de las frecuencias relativas separando los rangos percentílicos.

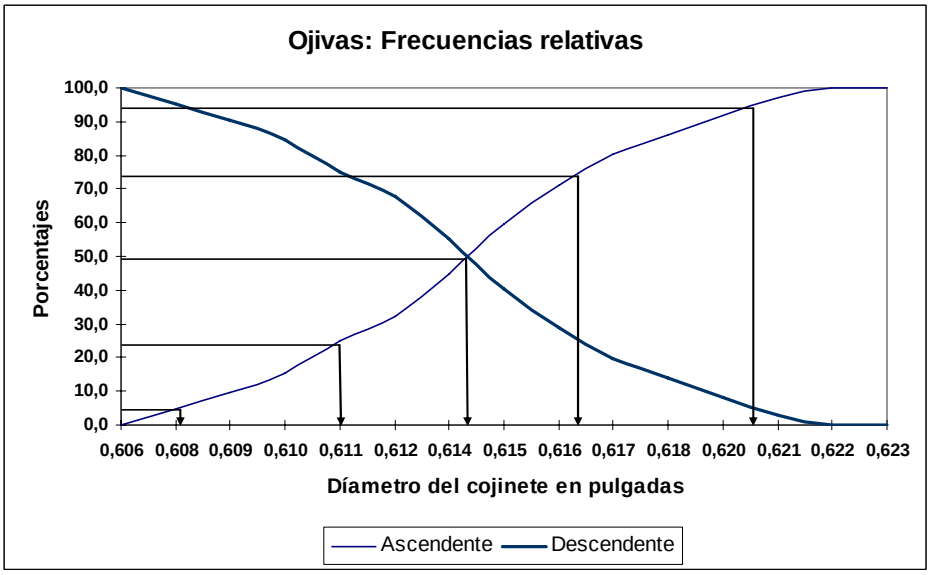
3.15 Las Ojivas.

En las ojivas se han señalado sobre la curva ascendente: La mediana, el primero y tercer cuarto y los percentiles 5 y 95.

Las curvas se parecen a una normal acumulativa.

En cuanto a rangos entre cuartiles y percentiles con relación a la mediana, parecen equilibrados.

Se puede utilizar la herramienta gráfica como instrumento de control de la producción.



3.16 Control de la calidad y proceso.

En la producción industrial se utilizan instrumentos gráficos para visualizar el proceso de la producción. Uno de estos se conoce como carta de control o gráfico de control de procesos o de

calidad. Éste consiste en representar mediante un gráfico el muestreo de la calidad a través del tiempo.

Las cartas que usan las estadísticas no paramétricas son sobre: la mediana y los rangos.

En el ejemplo los datos se han separado por secciones de 8 unidades, cada una de ellas corresponde a un muestreo sistemático con iniciación aleatoria tomadas a través del proceso de fabricación. Sobre estas muestras se llevan controles de la calidad, pero además sirven para el control del proceso de fabricación.

3.17 Estadísticos sobre medianas.

Para controlar las medianas deben obtenerse estadísticos sobre las medianas de cada muestra. Se calculan en una sección aparte de la HE. Los estadísticos se obtienen mediante instrucciones directas de la HE.

Mediana de medianas.

$\tilde{\tilde{x}}$ = MEDIANA(B183 : B212) = 0,6138.

Nota: Puede comprobar el resultado pidiendo el

=K.ESIMO.MENOR($B$183:$B$212;15,5) = 0,6138.

Esta instrucción es la que se utiliza para obtener los límites de la carta de control.

Para el percentil 5:

$\tilde{\tilde{x}}_{0,05}$ = K.ESIMO.MENOR($B$183 : $B$212;1,55) = 0,6098

Para el cuartil 1:

$\tilde{\tilde{x}}_{0,25}$ = K.ESIMO.MENOR($B$183 : $B$212;7,75) = 0,6132

Para el Cuartil 3:

$\tilde{\tilde{x}}_{0,75}$ = K.ESIMO.MENOR($B$183 : $B$212;23,25) = 0,6158

Para percentil 95:

$\tilde{\tilde{x}}_{0,95}$ = K.ESIMO.MENOR($B$183 : $B$212;29,45) = 0,6167

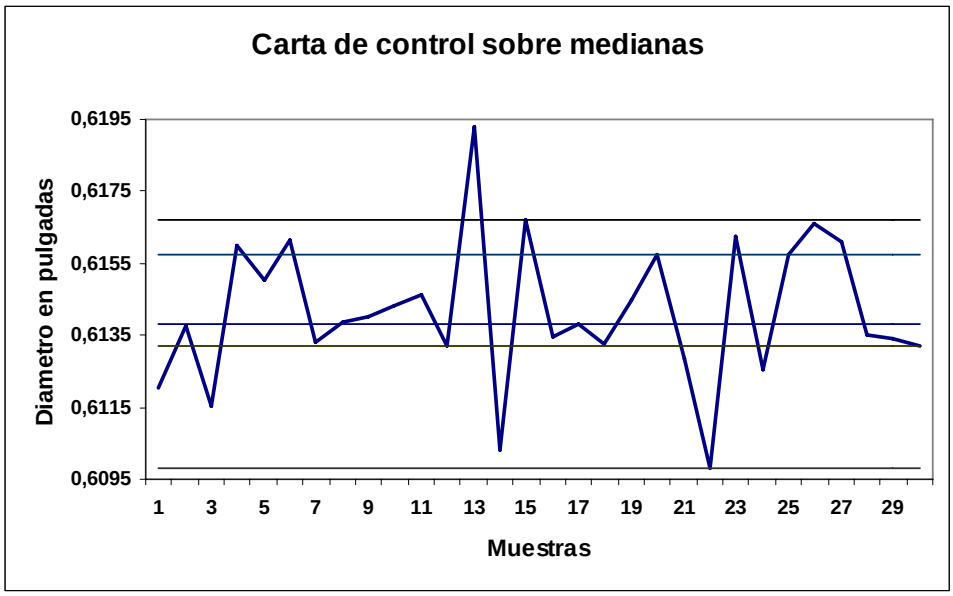
3.18 La Carta de Control.

La carta de control presenta una muestra fuera de control que excede el límite superior de control.

Las líneas de control no están equilibradas.

Lo más importante es que arriba de la mediana (línea centra en azul celeste) se encuentran dos tercios de las muestras.

En la HE puede observar como se acomodan los datos para elaborar el grafico de control del proceso o



de la calidad. Este consta de 6 ZONAS:

- ***a) La mediana de cada una de las 8 observaciones que se toman en cada muestra que se representa con una línea quebrada, dando idea del flujo del proceso;***
- ***b) El valor para el percentil 5;***
- ***c) El valor para el percentil 25;***
- ***d) El valor para la mediana de las medianas. Muchas veces esta valor está dado por el valor que se espera, en este ejemplo, el diámetro de control de los cojinetes de bolas;***
- ***e) El valor para el percentil 75;***
- ***f) Y el valor para el percentil 95.***

Aún cuando se esté utilizando la estadística no paramétrica, el sistema de control se fundamente en que éste se mantenga bajo control aleatorio, esto significa que los valores del parámetro muestral utilizado para el control oscile sin dirección definida. Se considera que el control estadístico se rompe cuando sucede alguno de los siguientes eventos:

- ***a) Un sólo valor excede los límites de control (percentil 5 y 95);***
- ***b) Dos o más valores se presenten en las zonas inmediatamente interiores (entre percentil 5 y 25, o entre percentil 75 y 95 o ZONA B) o más allá. La muestras 26, 27 y 28 se encuentran en la ZONA B superior indicando un evento fuera de control aleatorio;***
- ***c) Cuatro o más valores se ubiquen en la zona arriba del valor de parámetro de posición: De la muestras 15 a la 21 se ubican por arriba de la mediana general;***
- ***d) Ocho o más valores se ubiquen en la zona central o ZONAS C a ambos lados del parámetro de Posición, en este caso, la mediana general.***

3.19 Prueba de la Corrida.

Dados los resultados de la carta de control es conveniente efectuar una prueba llamada de *La Corrida* cuyo objetivo es determinar sí la muestra, en este caso, el proceso, se mantiene bajo control aleatorio.

Los pasos para la prueba son:

a) Regístrense las observaciones en el orden en que fueron recogidas;

b) Determínese la mediana de la muestra;

c) Indíquese con signo – las observaciones que estén por abajo la mediana y con signo + las que estén por arriba de la misma;

d).$n_1 = 15$ es el número de signos –, $n_2 = 15$ es número de signos +;

e) Cuéntese el número de corridas e identifíquense con $r = 15$;

f) La tabla estadística del signo proporciona los estadísticos. La posición es $10 < 15 < 22$. Esto indica que se acepte que la secuencia es aleatoria.

Una corrida se cuenta cuando un signo cambia, por tanto, el número de corridas indica el número de veces que se alternan los signos +, − ó blancos.

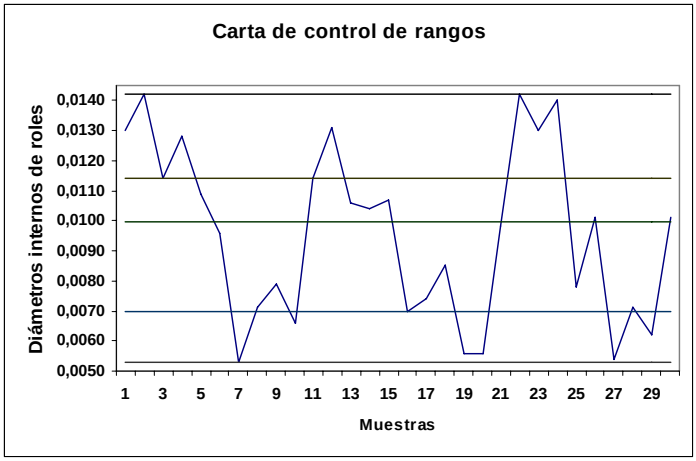
3.20 Carta de control sobre rangos:

La carta de control sobre medianas suele acompañarse por la carta de control de rangos.

Esta se elabora de manera similar a la anterior.

La prueba de la corrida $r = 9 < 10$ indica que no hay una secuencia aleatoria en la presentación de los rangos.

En la carta es notoria la secuencia que abarca aproximadamente siete u ocho



nuestras. Seguramente los ingenieros encargados del proceso buscarán el factor que propicia este comportamiento. Veamos que indica la prueba de la corrida.

En los fenómenos naturales como en los procesos de producción se espera que la aparición de una diferencia sea completamente aleatoria. Cuando hay una secuencia, como es el caso, se supone que el proceso se pone fuera de control aleatorio. Uno de los objetivos del control del proceso es precisamente encontrar o ubicar los factores que hacen que el sistema se ponga fuera de control aleatorio facilitando el mejoramiento de la calidad y aumentando la eficiencia de la producción.

3.21 El Intervalo de Confianza.

Cuando la muestra es pequeña, menos de 25, la distribución binomial proporciona toda la información para establecer intervalos de confianza: Por ejemplo para una muestra de tamaño 8:

La probabilidad que no salgan muestras fuera de rango: $P(0)=\binom{8}{0}0{,}5^8=0{,}0039$

La probabilidad que salga una muestra fuera de rango: $P(1)=\binom{8}{1}0{,}5^8=0{,}0313$

La probabilidad que salgan dos muestras fuera de rango: $P(2)=\binom{8}{2}0{,}5^8=0{,}1094$

El Acumulado es: 0,1445.

3.22 EL Intervalo de Confianza cola superior.

El número ordinal para la cola superior se obtiene mediante $n - r + 1 = 8 - 3 + 1 = 6$.

La probabilidad que salgan 6 muestras fuera de rango: $P(6)=\binom{8}{6}0{,}5^8=0{,}1094$

La probabilidad que salgan 7 muestras fuera de rango: $P(7)=\binom{8}{7}0{,}5^8=0{,}0313$

La probabilidad que salgan 8 muestras fuera de rango: $P(8)=\binom{8}{8}0{,}5^8=0{,}0039$

Pareciera que una unidad proporciona un intervalo confiable aceptable: $\Pr\{\tilde{x}_1 > \tilde{X} < \tilde{x}_8\}=92{,}97\%$

3.23 Intervalo de confianza con n grande.

Es evidente que la distribución binomial es muy apropiada para muestras < 25. Para muestras mayores se utiliza la aproximación de la normal a la distribución binomial (vea capítulo 5). En donde el intervalo de confianza estará dado por los ordinales:

$$\frac{n+1}{2}\pm\frac{z\sqrt{n}}{2}$$

Así, para 8, 30 y 240 muestras y un nivel de confianza de 0,05:

Tamaño de muestra		8	30	240
Probabilidad	0,025	-1,960	-1,960	-1,960
Número de orden medio		4,5	15,5	120,5
Número de orden inferior		**1,7**	**10,1**	**105,3**
Número de orden superior		**7,3**	**20,9**	**135,7**

De acuerdo al tamaño de muestra que se esté utilizando, se elegirán los valore límite apropiados.

Para el caso $n = 8$ en la carta de control Compárelos con los de la diapositiva 17.***Pruebas de Bondad de Ajuste: frecuencias y promedio.***

Antes de iniciar cualquier proceso de estimación es conveniente valorar la distribución que se va a utilizar para las predicciones, conclusiones y recomendaciones. La más conocida se valora utilizando la χ^2. Se procede como sigue:

1.- Obténgase el cuadro de frecuencias.

2.- Calcúlese el promedio:

$$n = \sum_{i=1}^{15} f_i = 0 + 11 + ... + 7 + 0 = 240$$

$$\bar{\bar{x}} = \frac{1}{n}\sum_{i=1}^{15} f_i \bar{x}_i = \frac{0(0{,}6057) + 11(0{,}6069) + ... + 7(0{,}6213) + 0(0{,}6225)}{240} = \frac{147{,}3336}{240} = 0{,}6139$$

3.25 Pruebas de Bondad de Ajuste: desviación estándar.

3.- La Desviación Estándar;

$$S^2 = \frac{\sum_{i=1}^{15} f_i (\bar{x}_i - \bar{\bar{x}})^2}{n-1} =$$

$$= \frac{0(0{,}6057 - 0{,}6139)^2 + 11(0{,}6069 - 0{,}6139)^2 + ... + 0(0{,}6225 - 0{,}6139)^2}{240 - 1} =$$

$$= \frac{0{,}0032}{239} = 1{,}3524\text{E-}05$$

$$S = \sqrt{1{,}3524\text{E-}05} = 0{,}0037$$

4.- Obténganse las probabilidades esperadas de cada clase usando la distribución Normal. Para la primera clase:

$$P(-\infty; LS_1) = 0 - \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{0{,}6063} e^{-\frac{(0{,}6063 - 0{,}6139)^2}{2(0{,}0037)^2}} dx =$$

$$= \text{DISTR.NORM}(0{,}6063; \bar{\bar{x}}; S; 1) = 0{,}01951$$

3.26 Pruebas de Bondad de Ajuste: probabilidad del intervalo.

5.- Para la clase 2 y hasta la penúltima:

$$P(LS_2; LI_2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{0{,}6075} e^{-\frac{(0{,}6075 - 0{,}6139)^2}{2(0{,}0037)^2}} dx - \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{0{,}6063} e^{-\frac{(0{,}6063 - 0{,}6139)^2}{2(0{,}0037)^2}} dx =$$

$$= \text{DISTR.NORM}(0{,}6075; 0{,}6139; 0037; 1) -$$

$$= -\text{DISTR.NORM}(0{,}6063; 0{,}6139; 0{,}0037; 1) = 0{,}02163$$

6.- Para la última clase debe calcular desde el límite inferior de la clase 15 hasta infinito:

$$P(LI_{15}; +\infty) = 1 - \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{0{,}6219} e^{-\frac{(0{,}6219 - 0{,}6139)^2}{2(0{,}0037)^2}} dx =$$

$$= 1 - \text{DISTR.NORM}(0{,}6219; 0{,}6139; 0{,}0037; 1) = 0{,}01470$$

Compruébese en la HE que la suma de las probabilidades de cada intervalo es 1.

3.27 Pruebas de Bondad: prueba de χ^2.

7.- Obténganse los valores esperados o frecuencias esperadas, multiplicando la probabilidad del intervalo por el número de observaciones;

$$fe_i = P(C)_i \times n;\ \ fe_1 = 0{,}01951 \times 240 = 4{,}7$$

8.- Efectúese la prueba de χ^2 sin corregir por continuidad. Se retiran la primera y la última clase que se agregaron con el objetivo de que los gráficos tengan mejor presencia:

$$\chi^2_{(15-1)} = \sum_{i=2}^{14} \frac{(fo_i - fe_i)^2}{fe_i} = \frac{(11-5{,}2)^2}{5{,}2} + \frac{(12-9{,}1)^2}{9{,}1} \ldots + \frac{(7-4{,}2)^2}{4{,}2} = 18{,}8134$$

9.- Valórese la $\chi^2 = 18{,}8134$:

$$F_{[18{,}8134;15\text{-}1]} = Y_0 \int_0^{18{,}8134} (18{,}8134)^{\frac{1}{2}(15-1)} e^{-\frac{1}{2}18{,}8134}\, d\chi = 0{,}0931$$

3.28 Resultados de la prueba con $\chi 2$.

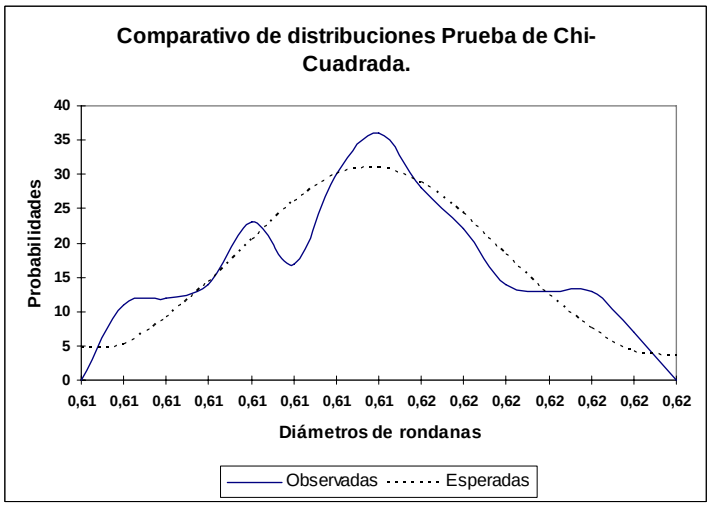
Límites de Clases			Frecuencias	Total	Devios	Probabilidad	Frecuencias	Chi-cuadrada
Inferior	Medio	Superior	Observadas	f * x	Cuadrático	Intervalo	Esperadas	Parcial
0,6051	0,6057	0,6063	0	0,0000	0,0000	0,01951	4,7	
0,6063	0,6069	0,6075	11	6,6759	0,0005	0,02163	5,2	6,5013
0,6075	0,6081	0,6087	12	7,2972	0,0004	0,03794	9,1	0,9201
0,6087	0,6093	0,6099	14	8,5302	0,0003	0,05989	14,4	0,0097
0,6099	0,6105	0,6111	23	14,0415	0,0003	0,08506	20,4	0,3275
0,6111	0,6117	0,6123	17	10,3989	0,0001	0,10871	26,1	3,1678
0,6123	0,6129	0,6135	30	18,3870	0,0000	0,12503	30,0	0,0000
0,6135	0,6141	0,6147	36	22,1076	0,0000	0,12939	31,1	0,7877
0,6147	0,6153	0,6159	28	17,2284	0,0001	0,12050	28,9	0,0292
0,6159	0,6165	0,6171	22	13,5630	0,0001	0,10097	24,2	0,2058
0,6171	0,6177	0,6183	14	8,6478	0,0002	0,07614	18,3	0,9992
0,6183	0,6189	0,6195	13	8,0457	0,0003	0,05166	12,4	0,0292
0,6195	0,6201	0,6207	13	8,0613	0,0005	0,03154	7,6	3,8950
0,6207	0,6213	0,6219	7	4,3491	0,0004	0,01733	4,2	1,9409
0,6219	0,6225	0,6231	0	0,0000	0,0000	0,01470	3,5	
					Suma	1,00000	Suma	18,8134
Tamaño de la muestra		240	Sumas de cuadrados		0,0032		Probabilidad	**0,0931**
Total de las clases		147,3336	Varianza		1,3524E-05			
Promedio		0,6139	Desviación Estándar		0,0037			

La probabilidad de la prueba 0,0931 ó 9,31% indica la similitud entre la distribución de los datos y la normal con la que se quiere aproximar. Dependiendo las condiciones de confiabilidad establecidas, por ejemplo para un 5%, las diferencias no son suficientes para rechazar la hipótesis nula:

Ho; *Distribución de Datos = Distribución Normal.* ***El gráfico comparativo.***

Las irregularidades de la distribución de los datos no son suficientes para que una distribución normal no pueda aproximar adecuadamente a la distribución de datos. Además, se nota equilibrada y sin curtosis.

No hay ningún impedimento para utilizar procesos de estadística no paramétrica cuando la distribución es normal, usualmente en busca de más información.



3.30 *Prueba de Kolmogorov-Smirnov.*

Una prueba de *Bondad de Ajuste* que se considera más precisa que la de χ^2 es la conocida como de *Kolmogorov-Smirnov*. Esta prueba compara las probabilidades acumulativas *esperadas* con las *observadas*.

La probabilidad observada se obtiene dividiendo las frecuencias observadas por el número de observaciones.

$$Po_i = \frac{f_i}{\sum_{i=2}^{14} f_i}; \quad Po_2 = \frac{11}{240} = 0{,}0458 \text{ Por tanto:}$$

$$D_2 = \left|Pe_2 - Po_2\right| = \left|0{,}02163 - 0{,}04583\right| = 0{,}02421$$

Se hacen las comparaciones sin considerar las clases 1 y 15 que se crearon con fines de graficación.

El proceso es el siguiente:

1. ***Calcule las distribuciones relativas acumulativas ascendentes observadas;***
2. ***Calcule las distribuciones relativas acumulativas ascendentes esperadas;***
3. ***En otra columna calcule las diferencias absolutas entre ambas distribuciones;***
4. ***La mayor diferencia absoluta será la base de comparación d_c = 0,02597.***

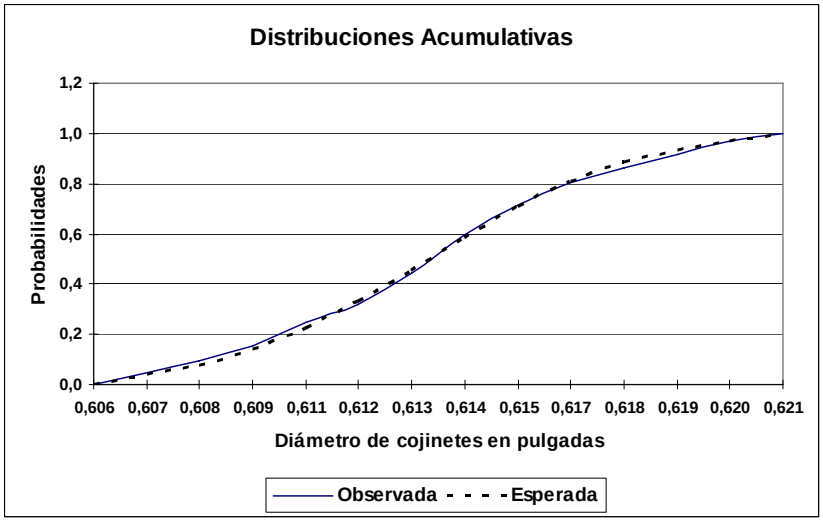
3.31 *Desarrollo de la prueba.*

El valor absoluto de la diferencia esperada D_c = 0,02597 es inferior al criterio D_t = 0,0879 (Tabla de 15 de la HE) por tanto, debe aceptarse la hipótesis nula. Y declarar que los datos se distribuyen *Normal*.

Límites de Clases			Frecuencia	Probabilidad		Diferencias
Inferior	Medio	Superior	Acumulativa	Observada	Esperada	Absolutas
0,6051	0,6057	0,6063				
0,6063	0,6069	0,6075	11	0,04583	0,04114	0,00469
0,6075	0,6081	0,6087	23	0,09583	0,07908	0,01675
0,6087	0,6093	0,6099	37	0,15417	0,13897	0,01520
0,6099	0,6105	0,6111	60	0,25000	0,22403	0,02597
0,6111	0,6117	0,6123	77	0,32083	0,33274	0,01191
0,6123	0,6129	0,6135	107	0,44583	0,45777	0,01194
0,6135	0,6141	0,6147	143	0,59583	0,58716	0,00867
0,6147	0,6153	0,6159	171	0,71250	0,70766	0,00484
0,6159	0,6165	0,6171	193	0,80417	0,80863	0,00447
0,6171	0,6177	0,6183	207	0,86250	0,88477	0,02227
0,6183	0,6189	0,6195	220	0,91667	0,93643	0,01976
0,6195	0,6201	0,6207	233	0,97083	0,96797	0,00286
0,6207	0,6213	0,6219	240	1,00000	1,00000	0,00000
0,6219	0,6225	0,6231				
			Valor calculado D =			**0,02597**
			Valor tabulado 5% = 1.36/Raiz(n)			**0,08779**
			Valor tabulado 1% = 1.63/Raiz(n)			**0,10567**

3.32 La representación gráfica.

En las distribuciones acumulativas las diferencias son menos notorias, además que según la prueba la distribución Normal o la Normal Estándar serían las herramientas Apropiadas para estudiar los resultados del conjunto de datos resultantes del muestreo de calidad y control de proceso de los cojinetes de bolas en la variable diámetro interior.



3.33 Las dos pruebas de Bondad de Ajuste.

Las pruebas de bondad de ajuste ofrecidas tienen características diferentes. La de χ^2 prueba diferencias de la distribución por clases considerando todas las diferencias. Posiblemente sea más exigente y muestre diferencias significantes antes que la de Kolmogorov-Smirnov.

La segunda prueba, se efectúa sobre la mayor diferencia absoluta entre las distribuciones acumulativas observada y esperada. Por su naturaleza, es menos sensible a las variaciones.

La recomendación de usar una u otra depende de la naturaleza de la experimentación y un análisis profundo de las consecuencias de hacer recomendaciones equivocadas.

Ver prueba de hipótesis del capítulo VII en este libro. En el ejemplo, pareciera ser más determinante la oscilación que muestra la carta de control de los rangos que se aprecia en la figura de la diapositiva 3.29, lo que orientaría a utilizar y la prueba de χ2.

3.34 Comparación de dos poblaciones.

Muchos experimentos consideran grados de mérito. Por ejemplo, la calificación de varios jueces en una exposición canina, el grado de calidad que asignan catadores a dos o tres tipos de cafés.

Calificaciones que se hacen en rangos de alto a bajo pueden valorarse mediante pruebas de distribución libre.

Se tratarán dos pruebas: la del *Signo* y la del *Rango con Signo.*

Para ilustrar se usará una prueba de degustación de dos salsas efectuada en la Escuela de Ingeniería Administrativa del Instituto tecnológico de Costa Rica.

3.35 Ejemplo y preparación.

Una empresa dedicada a preparar aderezos elaboró una salsa para competir con una que es muy buscada en el mercado.

Se prepararon pruebas en varios lugares del País, ésta consistía en ofrecerla a personas adultas sin consideración de sexo que salían de algún supermercado que se prestó para hacer la prueba.

Al azar se les asignaba el orden en que se les ponía la salsa en la mesa de degustación pidiéndoles que después de la prueba las calificaran, según su satisfacción, del 1 al 10 en una hoja numerada, con dos escalas señaladas como salsa 1 y salsa 2. Aparte se apuntaba a que salsa ***A*** o ***B*** correspondía según la asignación aleatoria. Los resultados de una prueba en la HE.

Prueba de χ^2 obtención del promedio.

Se consideró que la distribución que aproxima los datos es una binomial. Para valorar la *Bondad de Ajuste*, efectúe la prueba:

$$\bar{x} = \frac{\sum_{i=1}^{10} f_i x_i}{\sum_{i=1}^{10} f_i} =$$

$$= \frac{0(0)+1(0)+2(0)+3(1)+4(4)+5(7)+6(14)+7(16)+8(11)+9(4)+10(3)}{60} =$$

$$= \frac{404}{60} = 6{,}7$$

1.- Obténgase la distribución de frecuencias;

2.- Obténgase el promedio de calificaciones.

3.37 Prueba de χ^2: frecuencias esperadas.

3.-Obténgase la proporción de la respuesta *p. Recordar q* = 1 - *p*;

$$p = \frac{\bar{x}}{n} = \frac{6{,}7}{10} = 0{,}6733$$

4.-Obténgase la probabilidad esperada para cada evento, se ejemplifica con el 0.

$$P(0) = \binom{10}{0} 0{,}6733^0 0{,}3267^{10} = \text{DISTR.BINOM}(0;10;0{,}6733;0) = 0{,}00001$$

5.-Obténgase las frecuencias esperadas multiplicando la probabilidad esperada por el número de muestras. Por ejemplo, Para el evento 5;

$$fe_i = Pe_i \times \sum_{i=1}^{10} f_i;\ \ fe_5 = 0{,}12974 \times 60 = 7{,}8$$

6. Obténganse las χ^2 parciales (columna F);

$$\chi_i^2 = \frac{(fo_i - fe_i)^2}{fe_i}; \text{ej}: \chi_5^2 = \frac{(7-7{,}8)^2}{7{,}8} = 0{,}0790$$

3.38 Prueba de χ^2: valoración de la prueba.

7. Obténgase la χ^2 total;

$$\chi^2_{(11-1)} = \sum_{i=0}^{10} \frac{(fo_i - fe_i)^2}{fe_i} = 0{,}0008 + 0{,}0171 + ... + 2{,}9799 = 4{,}0775$$

8.- Valórese la χ^2 total;

$$F_{[4{,}0775;11\text{-}1]} = Y_0 \int_0^{4{,}0775} (4{,}0775)^{\frac{1}{2}(11-1)} e^{-\frac{1}{2}4{,}0775} d\chi = 0{,}9438$$

9.- La conclusión: La prueba indica que las distribuciones observada y esperada se asemejan en un 94,38%. Por tanto, no puede rechazarse la hipótesis nula y puede usarse la distribución binomial para analizar los resultados. La prueba completa en la siguiente diapositiva.

3.39 Desarrollo de la prueba de $\chi 2$ para bondad de ajuste..

La distribución Binomial Aproxima adecuadamente los Datos.

Calificación	Frecuencia Observada	Expansion Totales	Probabilidad Esperada	Frecuencia Esperada	Chi-Cuadrada parcial
0	0	0	0,00001	0,0	0,0008
1	0	0	0,00029	0,0	0,0171
2	0	0	0,00265	0,2	0,1587
3	1	3	0,01454	0,9	0,0186
4	4	16	0,05245	3,1	0,2311
5	7	35	0,12974	7,8	0,0790
6	14	84	0,22285	13,4	0,0296
7	16	112	0,26249	15,7	0,0040
8	11	88	0,20289	12,2	0,1131
9	4	36	0,09293	5,6	0,4455
10	3	30	0,01916	1,1	2,9799
		Suma	1,00000	Suma	4,0775
				Probabilidad	**0,9438**
Número de muestras N=		60	Ensayos o calificacviones n		10
Suma Total		404	Proporción = p		0,6733
Promedio		6,7	Proporción q		0,3267

3.40 Prueba de Kolmogorov-Smornov para la salsa B.

El estudiante deberá entender que a los tratamientos se les debe correr la misma prueba. En este caso, el cambio debe tomarse como parte de la enseñanza. Los pasos:

1.- Elabórese el cuadro de frecuencias.

2.- Obténgase el Promedio:

$$\bar{x} = \frac{\sum_{i=0}^{10} f_i x_i}{n} = \frac{323}{60} = 5{,}4$$

3.- Calcúlese la probabilidad esperada acumulada:

$$pe_i = \frac{f_i}{n}; \quad Pe_i = \frac{f_i}{n} + pe_{i-1}$$

4.- Calcule la probabilidad esperada:

$$Fe_i = \sum_{i=0}^{10} \binom{10}{i} 0{,}5^{10} = \text{DISTR.BINOM}(x;10;0{,}5383;1) = 0{,}0004$$

3.41 *Prueba de K-V: las diferencias absolutas.*

5.- Calcule las diferencias absolutas entre probabilidad esperada y observada:

$$D_i = |Fe_i - Pe_i|$$

$D_c = 0{,}0297 < D_t = 0{,}1756$ Acepte Ho. La prueba de Kolmogorov-Smirnov indicó que debe aceptarse la hipótesis nula Ho; La Distribución de los datos es Binomial.

Calificación	Frecuencia Observada	Expansion Totales	Probabilidad Observada	Probabilidad Esperada	Diferencia Abs(fe - fo)
0	0	0	0,0000	0,0004	0,0004
1	0	0	0,0000	0,0056	0,0056
2	2	4	0,0333	0,0325	0,0009
3	4	12	0,1000	0,1162	0,0162
4	13	52	0,3167	0,2869	0,0297
5	12	60	0,5167	0,5259	0,0092
6	13	78	0,7333	0,7581	0,0247
7	11	77	0,9167	0,9128	0,0039
8	5	40	1,0000	0,9804	0,0196
9	0	0	1,0000	0,9980	0,0020
10	0	0	1,0000	1,0000	0,0000
			Máxima D		**0,0297**
			Criterio 5%	1,36	**0,1756**

Número de muestras N=	60	Ensayos o calificacviones n	10
Suma Total	323	Proporción = p	0,5383
Promedio	5,4	Proporción q	0,4617

3.42 *Para decidir entre A y B se puede usar la prueba de z.*

Dado que ambas distribuciones son binomiales se pueden utilizar las ventajas de contar con una distribución de densidad para que simulen los datos, por ejemplo, la prueba de z. Donde

$$z_c = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} = \frac{0{,}6733 - 0{,}5383}{\sqrt{(0{,}6058)(0{,}3942)\left[\frac{1}{60} + \frac{1}{60}\right]}} = \frac{0{,}1350}{0{,}0892} = 1{,}5131$$

La probabilidad determinada es:

$$F(z = 1{,}5131) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{1{,}5131} e^{-\frac{1}{2}(1{,}5131)^2} dx = 0{,}0651$$

Insuficiente para declarar que hay diferencias en las preferencias entre la Salsa A y la Salsa B.

3.43 *La Prueba del Signo: generalidades.*

Si los catadores consideran a una salsa superior a otra, sin importar el puntaje. Consistentemente la nota a la salsa superior será superior.

Esta prueba es simple y puede valorarse mediante dos métodos:

El de Chi-cuadrada, Y;

El de La Prueba del Sigo.

Se efectuará primero la prueba del signo. Los pasos son los siguientes:

1.- Examine cada una de las parejas (x_i, y_i);

2.- Si $x_i > y_i$ Asigne el signo +; si $x_i < y_i$ asigne el signo -; si $x_i = y_i$ la pareja se descarta; =SI(B629=C629;"";SI(B629>C629;"+";"-")).

3.44 La prueba del signo: resultados y conclusión.

3.- El número *n* será el número de parejas no rechazadas: 37 signos + y 11 de signos -, por tanto $n = 48$;

4.- Denótese con la letra *r*, el número de veces que se presentó el signo menos frecuente. Este es $r = 11$;

5.- Regla de decisión: si el valor observado de r es igual o menor al valor tabulado para el nivel se significación elegido, *la hipótesis se rechaza.* En este caso $r_c = 11$ es menor que $r_{t(0,5)} = 16$, por tanto, la hipótesis deberá rechazarse.

CONCLUSIÓN:

La Salsa ***A*** fue calificada superior a la salsa ***B*** con significación del 5%.

3.45 La Prueba del Signo, material para una prueba de $\chi 2$.

El estudiante se habrá percatado que la prueba del signo proporciona material para elaborar una prueba de χ^2.

Considerando la hipótesis *Ho*; PA = PB = 0,5 se espera que la mitad de individuos elijan la Salsa A y la otra mitad la Salsa B, por tanto la frecuencia esperada será:

$$fe = \frac{n}{2} = \frac{48}{2} = 24$$

Las frecuencias observadas son $fo_1 = 37$ y $fo_2 = 11$. Por tanto la prueba será:

$$\chi^2_{2-1} = \frac{(|37-24|-0,5)^2}{24} + \frac{(|11-24|-0,5)^2}{24} = \frac{(|37-24|-0,1)^2}{48} = 13,0208$$

Que define una probabilidad de:

$$F_{[13,0208;\,2\text{-}1]} = Y_0 \int_0^{3,0208} (3,0208)^{\frac{1}{2}(2-1)} e^{-\frac{1}{2}3,0208} d\chi = 0,0003$$

De que las *Salsa **A** y **B*** se elijan por igual. O sea Ho; $P_A = P_B = 0,5$ se rechaza.

3.46 Prueba de Wincoxon o del rango con signo.

La prueba del signo, aun acompañada de la prueba de χ^2 no es la mejor alternativa para comparar dos poblaciones, máxime si la distribución de los datos es binomial, distribución que requiere muestras grandes.

La alternativa para distribuciones libres es la prueba del Rango con Signo, desarrollada por *Wilcoxon*, la prueba consiste en ordenar el conjunto de datos por su diferencia absoluta y asignar al rango pesado el signo de la diferencia relativa. Esta prueba considera además, la magnitud de la diferencia, entre mayor sea, mayor será el orden estadístico que entra en comparación.

Esta prueba puede considerarse como muy eficiente, sobre todo en poblaciones con distribución discreta o inespecífica.

En estadística, la prueba de Mann-Whitney-Wilcoxon, es una prueba no paramétrica para determinar si los puntos medios entre dos muestras de observaciones son iguales. La hipótesis nula es que las dos muestras están extraídas de una sola población, y por lo tanto se supone que los puntos medios sean iguales. Requiere las dos muestras ser independientes, y las observaciones ser ordenadas ascendentemente, es decir uno puede decir por lo menos, de cualquier dos observaciones, cuál es el mayor.

Es una de las pruebas se significación no paramétricas más conocidas. Fue propuesta, al parecer independientemente, por Mann y Whitney (1947) y Wilcoxon (1945), y por lo tanto a veces también se llama la prueba de Mann-Whitney-Wilcoxon (MWW) o la prueba de la rango con signo de Wilcoxon.

http://en.wikipedia.org/wiki/Mann-Whitney_U

3.47 Prueba de Wilcoxon: preparación en la HE..

Los pasos para desarrollar la prueba del signo son:

1.- Examine cada una de las parejas (x_i, y_i);

2.- Calcule la diferencia relativa $d_i = x_i - y_i$ sobre la columna 4 (Columna D de la HE);

3.- Calcule la diferencia absoluta $D_i = | x_i - y_i |$ =ABS(B709-C709) sobre la columna 5 (Columna E de la HE);

4.- Clasifique ascendentemente con base en la diferencia absoluta D_i;

5.- Agregue en la columna 6 (Columna F de la HE) el orden estadístico para los rangos absolutos, esto es 1,2,..,n;

6.- Sobre la columna 7 (Columna G de la HE) calcule los órdenes pesados. Esto es; el promedio para los rangos absolutos del mismo valor.

3.48 Prueba de Wilcoxon: orden ponderado.

Por ejemplo, con $D = 0$ hay $m = 12$ diferencias iguales, el orden ponderado es:

$$P_i = \frac{\sum_{i=nk}^{nk\neq(k-1)} n_i}{m} = \frac{1+2+3+4+5+6+7+8+9+10+11+12}{12} = 6{,}5$$

7.- En la columna 8 (H de la HE) ubicará los rangos ponderados para las diferencias relativas positivas;

8.- En las columna 9 (I de la HE) ubicará los rangos ponderados con el signo menos de las diferencias relativas;

9.- Ignore las diferencias absolutas o asigne 0 a cualquiera de las columnas;

10.- Al calce del cuadro, Sume las columnas 8 y 9 en la misma posición;

11.- El valor absoluto menor de ambas sumas se designará como indicador $T_c = |-331| = 331$.

3.49 Prueba de Wilcoxon: valoración..

12.- Compare el valor Tc con el criterio de la Tabla de la Prueba de Rangos con Signo de Wilcoxon. Como n > 25, el valor T_t se distribuye aproximadamente como *Normal Estándar* con media:

$$\mu = \frac{n(n+1)}{4} = \frac{60(60+1)}{4} = 915$$

Y varianza;

$$\sigma^2 = \frac{n(n+1)(2n+1)}{24} = \frac{60(60+1)(2\times 60+1)}{24} = 18.452{,}50$$

El estadístico z: $z_c = \frac{abs(T_t) - \mu_T}{\sigma_T} = \frac{|-331| - 915}{\sqrt{18.452{,}50}} = -4{,}2992$

3.50 Prueba de Wilcoxon: conclusión.

13. La probabilidad que determina el valor de calculada es:

$$P(z_c) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-4{,}2992} e^{-\frac{1}{2}(-4{,}2992)^2} dz = 0{,}000009$$

Evidentemente, se debe rechazar la hipótesis nula Ho; $P_A = P_B = 0{,}5$. O la probabilidad de que los catadores hayan elegido por igual a la Salsa ***A*** y a la Salsa ***B*** es prácticamente 0.

Si se observan las probabilidades de las tres pruebas que se han realizado para comparar dos poblaciones, esta es la que proporciona la probabilidad para la zona de rechazo más significante.

3.51 Prueba de Wilcoxon: recomendación.

Además se tiene la ventaja de considerar el signo de la comparación.

Esto implica, en experimentos en los que participa el humano que tiende a responder tanto por la estimulación de los sentidos como por la experiencia, pues, además de la diferencia de elección, se considera la posición.

Por ejemplo, una persona puede estar acostumbrada a calificar alto, si le pone un diez a la salsa ***A*** y un nueve a la ***B*** la diferencia será +1. Si otra persona muy exigente califica bajo y a la salsa ***A*** le da 1 y a la salsa ***B*** le da 0, la diferencia será de +1. Situaciones como estas no son consideradas en otras pruebas cuyo punto de apoyo es el promedio entrando a engrosar el error experimental.

Por esto, en ciertas condiciones la prueba del *Rango con Signo* es muy eficiente.

3.52 Resumen.

Como se ha visto, la *Estadística no Paramétrica* es una alternativa de análisis de datos y no una salida para situaciones "anormales" de las poblaciones.

En algunos casos, el uso de *Estadísticas no Paramétricas* es ventajoso al análisis de datos mediante las técnicas tradicionales análisis mediante *Estadísticas Paramétricas.*

Por estas razones, es preferible que se entienda el método como *Estadísticas de Distribución Libre* en las cuales la distribución de orden estadístico asociado a la magnitud de los datos proporcionan las bases para el desarrollo de esta importante y poco utilizada parte del análisis estadístico de las poblaciones.

REFERENCIAS SELECTAS:

1. Mood, Alexander M; Graybill, Franklin A; Boes, Duane c. Itroduction to the theory of Statistics. Capítulo 11.McGraw-Hill 1974.
2. Mood, Alexander M; Graybill, Franklin A; Boes, Duane c.Introducción a la Teoría de la Estadística. Capítulo 16. Aguilar, S. A. Madrid España. 1972.
3. Ostle, Bernard. Estadáitica Aplicada. Capítulo 15. Editorial LIMUSA, 1977.
4. Snedecor, George; Cochran, William. Statistical Methods. Capítulo 5. The Iowa State University Oress, Ames Iowa, EEUU, 1971.
5. Steel, Robert, G. D; Torrie, James H. Principles and Procedures of Statistics. Capítulo 21. McGraw-Hill 1960

4 La Distribución Binomial.

Los archivos para esta sección son:
E04_DBinomial_P01.ppn;
E04_DBinomial_W01.doc;
E04_DBinomial_X01.xls

Atribuido a Newton, el teorema fue en realidad descubierto por Abu Bekr ibn Muhammad ibn al-Husayn al-Karaji alrededor del año 1000.
http://es.wikipedia.org/wiki/Teorema_del_binomio

4.1 Menú.

4.2 Introducción.

Los especialistas en ciencias sociales, los analistas de marketing, los políticos requieren valorar proyectos que involucren cualidades y preferencias del público; el genetista requiere valorar características fenotípicas y genotípicas de poblaciones de moscas de la fruta; el ingeniero industrial necesita valorar los procesos de calidad; el médico las respuestas a tal o cual técnica quirúrgica. Y como estos podrían mencionarse ejemplos que involucren las más variadas áreas de la ciencia.

Mediante la estadística se conoce, por ejemplo: la preferencia para tal cuál candidato, la proporción de artículos defectuosos en una línea de producción de un producto, la preferencia del ama de casa por un producto de limpieza.

4.3 La Variable Cualitativa.

Los profesionales y ejemplos de la diapositiva anterior tienen en común que requieren del análisis de *Variables Cualitativas*. Variables que miden situaciones con salidas simples, o se posee la cualidad o no se posee sin que medie entre ambas ninguna escala.

Decisiones simples que se concretan a marcar un *Sí* o un *No* en un cuestionario, marcar con 1 si el sujeto experimentado posee la cualidad y un 0 si no la posee.

En general, interesa la probabilidad que una condición simple o la combinación de varias condiciones tienen en una situación determinada.

4.4 El Ensayo Simple.

El ensayo simple consiste en valorar las probabilidades de los individuos que poseen la cualidad *A* en la población.

En donde la cualidad *A* es la que interesa al proyecto. Obviamente, los elementos restantes de la población analizada se tornan complementarios al tamaño total de la población.

Así por ejemplo, a un floricultor puede interesarle reproducir gladiolos de color blanco y ningún otro, entonces, todos las gladiolos que posean el color blanco serán los elementos que poseen la cualidad *A*.

El floricultor sabe que dividiendo el número de gladiolos blancos multiplicándolas por 100 y dividiéndolas por todas los gladiolos obtiene el porcentaje *P* de gladiolos de color blanco.

4.5 La Probabilidad.

Pasar de porcentajes a probabilidades es cuestión casi de semántica. Intuitivamente sabemos lo que significa y como se calcula, no obstante, debe formalizarse la notación que se usará en la discusión.

Llamaremos *Evento* a la ocurrencia de la cualidad que interesa al proyecto, por tanto, la probabilidad del evento que interesa (gladiolos blancos) se define por: $P(E) = \dfrac{n_A}{n_A + (N - n_A)}$

En donde n_A representa a los gladiolos blancos y (N – n_A) a los de otro color. El denominador considera dos *Sucesos* u ocurrencias posibles.

4.6 Primera Regla de Probabilidad.

Regla 1. En un ensayo que tiene *k* posibles resultados igualmente verosímiles, en los cuáles uno y sólo uno puede aparecer en cada intento, la probabilidad de cualquier resultado es:

$\dfrac{1}{k}$ En donde *k* es el número de *Eventos Posibles* e igualmente creíbles.

En términos generales se dice que *k* es igual a *N*.

En adelante si no hay una indicación específica se usará N denotando al total de sucesos posibles o elementos de la población.

4.7 La Probabilidad Combinada.

Ahora supongamos que el floricultor ha cultivado gladiolos de 6 colores sin reparar en esta cualidad, por tanto no hay razones para suponer que las proporciones sean diferentes. Para su venta, los prepara en bolsas con media docena de bulbos. Es fácil deducir que la probabilidad que juegan en cada bolsa es igual para cada color y también se puede deducir que es:

$$P(E_1) = P(E_2) = P(E_3) = P(E_4) = P(E_5) = P(E_6) = \frac{1}{6}$$

Es posible deducir esta probabilidad porque las condiciones que entrañan en "juego" son *verosímiles*. O tienen la virtud de la *verosimilitud*, que resulta porque las situaciones involucradas en el ensayo o "juego" son creíbles y aceptables.

El estudiante debe entender que se trata una situación ideal, pues en la práctica, es muy probable que algún color de la flor del gladiolo esté asociado con una mayor capacidad de sobrevivir.

4.8 El Problema 4-1.

El floricultor quiere estar seguro de que las bolsas que vende contienen una muestra de los 6 colores de gladiolos. Para esto, siembra un poco más de 10 docenas de bulbos (considera los que no germinan) de gladiolos. Al momento de la cosecha toma los datos de los colores de 120 gladiolos por sus colores numerados del 0 = Blanco; 1 = Nacarado; 2 = Amarillo; 3 = Anaranjado; 4 = Rosado; 5 = Rojo.

Se espera que no haya diferencia entre las frecuencias de los colores. El estudiante entenderá que no habrá exactamente:

$$fe_i = \frac{120}{6} = 20$$

Gladiolos de cada color. La notación fe_i se usará para la frecuencia esperada.

4.9 La probabilidad de los eventos.

A la cantidad anterior se le llama frecuencia esperada. Aplicando la Regla 1, esto es, dividiendo esta frecuencia por las 120 unidades de la muestra se obtiene la probabilidad de cada color:

$$P(E_i) = \frac{20}{20+100} = \frac{2}{12} = \frac{1}{6}$$

Para estar acorde a los procedimientos estadísticos se establecen las hipótesis del ensayo:

$$Ho; P(E_i) = \frac{1}{6} \text{ contra } Ha; \text{algún o algunos } E_i \text{ diferentes.}$$

Que se valuará con una probabilidad de confianza de 95%.

Es el momento para que el estudiante abra su archivo EXCEL y cree un nuevo conjunto de datos que irá resolviendo paso a paso o diapositiva a diapositiva si lo prefiere poniendo especial atención a las diferencias que seguramente encontrará puesto que cada conjunto se genera de manera aleatoria.

El estudiante puede ignorar las etiquetas de las diapositivas para conseguir un trabajo continuo y propio.

Para estudiar las diferencias y concurrencias se recomienda crear un nuevo documento y guiarse por las diapositivas y el libro.

Tome su ritmo de aprendizaje, recuerde que usted es su propio profesor.

4.10 La Distribución de frecuencias observadas.

La prueba de Chi-Cuadrada.

$$\chi^2_{6-1} = \sum_{i=1}^{6} \frac{\left(\left|fo_i - fe_i\right| - 0{,}5\right)^2}{fe_i} =$$

$$= \frac{\left(\left|20-20\right| - 0{,}5\right)^2}{20} + \frac{\left(\left|24-20\right| - 0{,}5\right)^2}{20} + \ldots + \frac{\left(\left|16-20\right| - 0{,}5\right)^2}{20} = 6{,}4750$$

Determina una probabilidad en la función de densidad de:

$$F(6{,}4750;\,6-1)=$$

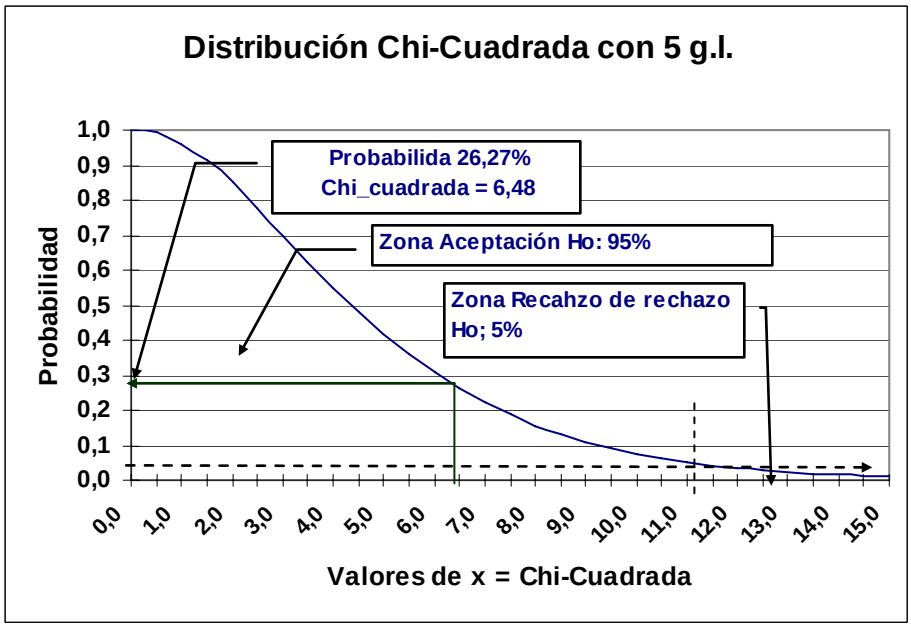
$$=Y_0\int_0^{6{,}4750}(6{,}4750)^{\frac{1}{2}(6-1)}e^{-\frac{1}{2}6{,}4750}\,d\chi=0{,}2667$$

Cantidad que estadísticamente se considera insuficiente para declarar diferencias de importancia entre las frecuencias.

Colores	Evento	FRECUENCIAS Observadas	FRECUENCIAS Esperadas	Chi-Cuadrada
Blanco	0	20	20	0,0125
Nacarado	1	24	20	0,6125
Amarillo	2	11	20	3,6125
Anaranjado	3	25	20	1,0125
Rosado	4	24	20	0,6125
Rojo	5	16	20	0,6125
		120	120	6,4750
			Probabilidad	**0,2627**

El estudiante debe suponer que no hay diferencias de ninguna clase que hagan suponer que un tipo de gladiolo posea mayor vitalidad que otros u otros.

4.11 Interpretación de la prueba.



La $\chi^2_{(n-1)}$ es una función de probabilidad en donde la variable calculada (χ^2_5) = 6,48 en el eje ***X*** define un área bajo la curva. Siguiendo la línea fucsia se llega a una probabilidad de 26,27% en el eje Y.

Para la prueba se definió una probabilidad de confianza de 95% señalada con verde brillante en el gráfico. Para que se declaren diferencias en las frecuencias, el valor de χ^2 calculada debió alcanzar un valor de algo más de 11 que define la zona de rechazo.

Por tanto debe aceptarse Ho.

4.12 Regla 2 de probabilidad.

Regla 2. <u>Regla de la aditividad</u>. Si un evento es satisfecho por uno de un grupo de posibilidades mutuamente excluyentes, la probabilidad del evento es la suma de las probabilidades de los resultados en el grupo.

En terminología matemática, la regla puede definirse como:

$P(E) = P(R_1 \text{ o } R_2 \text{ o } .... \text{ o } R_m) = P(R_1) + P(R_2) + ... + P(R_m)$

Mutuamente excluyentes significa que el que ocurra uno no interfiere con la ocurrencia del otro.

El floricultor quiere saber ¿cuál es la probabilidad de los colores blancos? Esto hace referencia a los gladiolos de color Blanco y Nacarado por tanto:

$$P(E)=P(B)+P(N)=\frac{20}{120}+\frac{20}{120}=\frac{1}{6}+\frac{1}{6}=\frac{2}{6}=\frac{1}{3}$$

4.13 La Independencia: planteamiento.

Cada uno los dos colores de tono blanco son producidos por bulbos de gladiolos genéticamente diferentes, aunque se parezcan, no se interfieren.

Sí la finca está dividida en sector I y sector II, en el primero se cultivan flores Blancas, Amarillas y Anaranjadas, en el otro, Nacaradas, Rosadas y Rojas. Para valorar la probabilidad de las flores de tono blanco (blancas y nacaradas) debe considerarse las dos fuentes de origen de los gladiolos. Primero se responde la siguiente pregunta

P: ¿Cuál es la probabilidad que un bulbo venga del sector I?:

R: Por la regla 1 cada color tiene una probabilidad de $\frac{1}{6}$ y en el Sector I se cultivan gladiolos de 3 colores. Por la regla 2:

$$P(I) = \frac{1}{6}Blancas + \frac{1}{6}Amarillas + \frac{1}{6}Anaranjadas = \frac{3}{6} = \frac{1}{2}$$

4.14 La independencia: conclusión..

P: ¿Cuál es la probabilidad de que un bulbo provenga del sector I y su matiz sea blanco?

R: Dos bulbos cumplen con tener un matiz blanco: el mismo blanco y el nacarado: por la Regla 1 es $\frac{1}{6}Blanco + \frac{1}{6}Nacarado = \frac{2}{6} = \frac{1}{3}$.

Pero se está condicionando a que provenga del Sector I cuya probabilidad es de ½. Por tanto, la probabilidad pedida es:

$$P(E) = P(A:B) = P\left[\frac{1}{6}Blanco + \frac{1}{6}Nacarado\right] \cap P\left[\frac{1}{6}Blanco + \frac{1}{6}Amarillos + \frac{1}{6}Anaranjados\right] =$$

$$= \frac{1}{3} \times \frac{1}{2} = \frac{1}{6}$$

La división en P(A /B) indica presentación de un suceso condicionado a la presentación del otro.

Pues una única variedad cumple con las dos condiciones y es la que produce gladiolos Blancos. En Términos de conjuntos:

$$P(A:B) = P(A \cup B) - P(A \cap C) = P(Blancas) + P(Nacaradas) - P(Blancas)$$

4.15 Ampliación de la regla II.

La ampliación de la regla 2 para eventos que no se excluyen dice que:

La probabilidad de eventos que no son mutuamente excluyentes es igual a suma de las probabilidades de cada evento, menos la probabilidad de los eventos comunes. En términos de conjuntos: $P(E) = P(A \cup B) - P(A \cap B)$

La regla se enuncia pero no es indispensable para deducir la fórmula del Binomio.

El floricultor tiene dudas sobre la calidad agrícola del sector I y decide hacer un ensayo planificado. Toma 10 bolsas de bulbos al azar y planta 1 bulbo en 6 repeticiones; la hipótesis se plantea en términos de las frecuencias:

Ho; $n_{(\text{Blancas})} = n_{(\text{Amarillas})} = n_{(\text{Anaranjadas})} = 3$

En cada repetición de 6 bolsas.

4.16 El valor esperado y el ensayo.

Se mencionó que las frecuencias observadas difícilmente se presentan con valores idénticos a los esperados. El juego de efectos aleatorios impide que esto ocurra. En este ensayo, las bolsas se llenan con 6 bulbos tomados al azar, por tanto, el ensayo deberá corroborar los valores esperados.

El resultado indica que la sección I produce tanto como la sección II. Y además, las bolsitas de gladiolos que se ofrecen al mercado llevan, generalmente, gladiolos de 6 colores. Esto se puede verificar usando las frecuencias por tipo de bulbo.

	R E P E T I C I O N E S						Nº Gadiolos	Frecuencia	Chi-Cuadrad
Bolsas	r1	r2	r3	r4	r5	r6	B, Am ,An	Esperada	Parcial
1	0	2	3	0	0	5	4	3	0,0833
2	4	1	2	2	3	4	3	3	0,0833
3	4	0	4	1	4	2	3	3	0,0833
4	4	0	3	0	0	1	4	3	0,0833
5	0	2	4	4	0	4	3	3	0,0833
6	2	1	2	1	4	4	4	3	0,0833
7	0	1	4	2	0	4	4	3	0,0833
8	1	4	3	3	2	0	3	3	0,0833
9	4	3	5	4	1	4	1	3	0,7500
10	2	5	0	5	2	3	3	3	0,0833
						Sumas	32	30	1,5000
								Probabilidad	**0,9971**

4.17 Prueba de frecuencias.

La hipótesis que ahora se valora es:

Ho; ni = 10

La prueba indica que la probabilidad de que las frecuencias observadas difieran de las esperadas es 11,95%. En esencia, se le puede asegurar al comprador de gladiolos que es muy probable que obtenga gladiolos de 6 colores en cada bolsa de 10 bulbos.

Color del Gladiolo	Núero de la Variable x	FRECUENCIAS		Chi-Cudrada
		Observada	Esperada	Parcial
Blanco	0	13	10	0,6250
Nacarado	1	8	10	0,2250
Amarillo	2	11	10	0,0250
Anaranjado	3	7	10	0,6250
Rosado	4	17	10	4,2250
Rojo	5	4	10	3,0250
Sumas		60	60	8,7500
			Probabilidad	**0,1195**

4.18 La probabilidad del evento.

Dado que el ensayo se realiza de una manera probabilística, la probabilidad para el evento que se supone con menos calidad agrícola, por la regla dos de probabilidad es:

$$P(E_1)+P(E_2)+\ldots+P(E_{10})=\frac{4+3+3+4+3+4+4+3+1+3}{6\times 10}=\frac{32}{60}=0{,}5333$$

Las frecuencias parecen no diferir y la probabilidad de los bulbos que provienen del sitio I es muy cercana al 50% esperado. La apreciación del floricultor no parece confirmarse con las pruebas obtenidas, pues en todos los casos apuntan a que las frecuencias del sitio I no difieren con las frecuencias del sitio II. Dicho de otra manera, se deben al azar.

4.19 La probabilidad condicional.

Lo primero que debe entenderse es que se han hecho dos pruebas sobre números esperados. Queda por estudiar la de los porcentajes esperados. Para llegar a la respuesta deben estudiarse uno o dos conceptos adicionales de probabilidad.

Para la venta a mayoristas, los gladiolos se empacan en cajas de 5 bolsas. De tanto en tanto, se abre una caja, se abre una bolsa para revisarle algunas condicionantes como es peso, tamaño y ausencia visual de enfermedades en los bulbos. Al llegar a almacén que más consume, hace la misma prueba. Suponiendo que se eligió la misma caja: ¿Cuál es la probabilidad de que se elija la misma bolsa?

En este ejemplo, se está condicionando la probabilidad al hecho de que la bolsa a la que se le hicieron las pruebas de salida se reempaca y manda al mercado dando oportunidad a que la misma bolsa sea elegida en la prueba que hace el almacén.

4.20 Muestreo con reemplazo.

El permitir que la misma unidad pueda ser utilizada nuevamente cambia las probabilidades a un MUESTREO CON REEMPLAZO.

En el primer muestreo, una bolsa tomada al azar de 5 posibles tiene una probabilidad, por la regla 1 de:

$$P(E)=\frac{1}{k}=\frac{1}{5}$$

En el segundo muestreo la probabilidad es la misma: un quinto. Pero, se está condicionando a que en el muestreo de entrada al almacén se elija, por supuesto al azar, la misma bolsa de gladiolos que se eligió en el muestreo de control de la calidad de salida de sitio de producción.

Esto es más fácil de visualizar creando el MARCO MUESTRAL.

4.21 Marco de muestreo k^2.

En el primer muestreo la probabilidad de elegir cualquier bolsa es $\frac{1}{5}$. Supóngase que se eligió la bolsa identificada con la letra *c*.

		Muestreo de Llegada				
		1	2	3	4	5
Muestreo de Salida		a	b	c	d	e
1	a	a,a	a,b	a,c	a,d	a,e
2	b	b,a	b,b	b,c	b,d	b,e
3	c	c,a	c,b	**c,c**	c,d	c,e
4	d	d,a	d,e	d,c	d,d	d,e
5	e	e,a	e,b	e,c	e,d	e,e

Al Condicionar el ensayo a que la misma bolsa sea elegida en el segundo muestreo, una y solo una combinación es posible; está iluminada en verde (en negrita). El número de sucesos posibles es de $k \times k = 5 \times 5 = 25$. Por tanto:

$$P(c,c)=\frac{1}{5}\times\frac{1}{5}=\frac{1}{25}$$

4.22 Más ejemplos de probabilidad condicionada.

Hay condicionamiento de lugar y de posición. Pues es importante para las probabilidades que se crean en este tipo de ensayos. Por ejemplo:

¿Cuál es la probabilidad que no se escoja la bolsa *c*?

Inspeccionando el marco de muestreo y aplicando la regla II se puede ver:

$$P(x\neq c)=P(a,a)+P(a,b)+...+P(b,e)+P(e,a)+...+P(e,e)=$$

$$=\frac{1+1+....+1+1+1+...+1}{25}=\frac{16}{25}$$

Visto de otro modo: la probabilidad de que no se elija *c* en la salida es $\frac{4}{5}$ y que no se elija en la llegada es $\frac{4}{5}$. Por Tanto;

$$P(x\neq c)=\frac{4}{5}\times\frac{4}{5}=\frac{16}{25}$$

4.23 Regla III de probabilidad.

Estos resultados nos permiten ampliar la regla 2 a la regla del producto o regla 3 de probabilidad.

Regla III. Regla de la multiplicación. *En una serie de ensayos independientes, sobre las mismas unidades, la probabilidad del evento es igual al producto de las probabilidades de los eventos individuales.*

En términos matemáticos,

$P(E_1 \text{ y del } E_2 \ldots \text{ y del } E_m) = P(E_1)P(E_2)\ldots P(E_m)$

En la práctica la suposición que los ensayos son independientes, se basa en que los resultados son igualmente *verosímiles*, más que en la justificación del conocimiento de las circunstancias que rodean al ensayo.

Esta recopilación a cerca de probabilidades proporciona, solamente, un bagaje de conocimiento de probabilidades mínimo pero necesario para trabajar y entender la distribución *Binomial*.

4.24 La Distribución Binomial.

El floricultor quiere confirmar que el Sitio I de la finca es de mejor calidad agrícola que el Sitio II. De ser así, sembraría los gladiolos de más demanda en el Sitio I. Encarga a una ingeniera agrónoma la prueba.

Esta compró 35 bolsas de gladiolos del floricultor en varios comercios para hacer la prueba en unidades experimentales de dos macetas. Los resultados se muestran en la HE.

Recordando que los gladiolos con flores: *Blancas, Amarillas y Anaranjadas* se siembran en el Sitio I (S_1), serán marcadas como el suceso de interés con un 1 y 0 si florece de otro color (S_2).

De la floración la agrónoma encontró los siguientes resultados.***Distribución de frecuencias.***

Parece claro que la proporción de bulbos de gladiolos que vienen del sitio I es la misma de los que vienen del sitio II.

Por tanto, se puede pensar que las frecuencias esperadas serán para cada clase, por la regla:

$$fe_i = \frac{105}{4} = 26{,}25$$

Ó 25%. La probabilidad de que en una maceta aparezca una flor de un bulbo cultivado en el sitio I es 0,5 o 50%. Pero qué además también aparezca en la maceta 2, por la regla III es $0{,}50^2 = 0{,}25$.

Evento x	Combinación en Macetas	Frecuencias Observadas	Gladiolos Cualidad
0	S_2 - S_2	28	0
1	S_2 - S_1	24	24
1	S_1 - S_2	25	25
2	S_2 - S_2	28	56

Numero ensayos por muestras	210
Suma de Gladiolos con la cualidad	105
Tamaño del ensayo	2
Suma de gladiolos del sitio I	105
Promedio de Gladiolos	1,0000
Proporción de Gladiolos Sitio I	0,5000
Proporción de Gladiolos Sitio II	0,5000

4.26 La estructura del Binomio.

Ya aparece evidente la estructura del binomio, la probabilidad de que en la maceta 1 se siembre un bulbo preveniente del sitio II es 0,5 y la Probabilidad de que en la maceta 2 también se siembre un bulbo del sitio II es 0,5. Por tanto, la probabilidad para los eventos, del ensayo son, por la regla 3:

$$P(S\_II \text{ y } S\_II) = q \times q = 0{,}5 \times 0{,}5 = 0{,}25 = q^2$$

$$P(S\_I \text{ o } S\_II) = (p \times q) + (q \times p) = 0{,}5 \times 0{,}5 + 0{,}5 \times 0{,}5 = 2(p \times q) = 0{,}5 = 2pq$$

$$P(S\_I \text{ y } S\_I) = p \times p = 0{,}5 \times 0{,}5 = 0{,}25 = p^2$$

Está demostrado que $p^2 + 2pq + q^2 = (p + q)^2 = 1$

4.27 La conclusión para ensayos con 3 repeticiones.

Supóngase que en forma paralela se realizó la misma prueba utilizando unidades experimentales de 3 macetas. Los bulbos de los sitios se arreglan de la siguiente manera:

Que aparezca 0 o que aparezcan 3 se combina de una sola manera. Que aparezcan dos de tres se arreglan de 3 formas.

Gladiolos S1 Evento x	MACETAS 1	2	3	Códigos
0	SITIO 2	SITIO 2	SITIO 2	000=S2,S2,S2
1	SITIO 2	SITIO 2	SITIO 1	001=S2,S2,S1
1	SITIO 2	SITIO 1	SITIO 2	010=S2,S1,S2
1	SITIO 1	SITIO 2	SITIO 2	100=S1,S2,S2
2	SITIO 1	SITIO 1	SITIO 2	110=S1,S1,S2
2	SITIO 1	SITIO 2	SITIO 1	101=S1,S2,S1
2	SITIO 2	SITIO 1	SITIO 1	011=S2,S1,S1
3	SITIO 1	SITIO 1	SITIO 1	111=S1,S1,S1

4.28 Aplicando las reglas de probabilidad.

En este segundo ensayo las probabilidades para el sitio I fue de p = 0,55, consecuentemente q = 0,45. Aplicando las reglas de probabilidad se desarrolla el siguiente cuadro.

Gladiolos S1 Evento x	MACETAS 1	2	3	Regla 3 de Probabilidad	Regla 2 de Probabilidad
0	0,45	0,45	0,45	0,09113	0,09113
1	0,45	0,45	0,55	0,11138	
1	0,45	0,55	0,45	0,11138	
1	0,55	0,45	0,45	0,11138	0,33413
2	0,55	0,55	0,45	0,13613	
2	0,55	0,45	0,55	0,13613	
2	0,45	0,55	0,55	0,13613	0,40838
3	0,55	0,55	0,55	0,16638	0,16638
			Sumas	1,00000	1,00000

La aparición de flores Blancas, Amarillas o Anaranjada, de la maceta 2 está condicionada al evento de la maceta 1 y la 3 a las anteriores. Por tanto, la probabilidad final del evento es la multiplicación de las probabilidades de cada suceso o maceta (regla 3). Cuando hay combinación de colores se aplica la regla 2 de la suma.

4.29 La fórmula del Binomio.

La fórmula del binomio ya aparece claramente:

$$F(x) = p^3 + 3p^2q + 3pq^2 + q^3 = 1$$

Que puede inducirse a n repeticiones:

$$F(x) = \sum_{i=0}^{n} \binom{n}{x} p^x q^{n-x}; \text{i} = 0,1,...,\text{n}$$

En el cuarto término se aplica la regla 3:

$$P(x=0) = P(s_2)P(s_2)P(s_2) = 0{,}45 \times 0{,}45 \times 0{,}45 = 0{,}09113$$

Para el evento que aparezca una flor del sitio 1 se aplica la regla 2 y 3:

$$P(x=1) = P(s_2)P(s_2)P(s_1) + P(s_2)P(s_1)P(s_2) + P(s_1)P(s_2)P(s_2) =$$
$$= 3(0{,}45 \times 0{,}45 \times 0{,}55) = 3(0{,}11138) = 0{,}33413$$

4.30 Reglas involucradas.

Para el evento que aparezcan 2 flores del sitio 1 en 2 de 3 macetas:

$$P(x=2) = P(s_1)P(s_1)P(s_2) + P(s_1)P(s_2)P(s_1) + P(s_2)P(s_1)P(s_1) = 3(0{,}55 \times 0{,}55 \times 0{,}45)$$
$$= 3(0{,}13613) = 0{,}40838$$

Para el evento en que únicamente aparecen flores del sitio I:

$$P(x=3) = P(s_1)P(s_1)P(s_1) = 0{,}55 \times 0{,}55 \times 0{,}55 = 0{,}16638$$

Aplicando la fórmula condensada: el factorial de 0 es 1 y cualquier número elevado a la potencia 0 es igual a la unidad $y^0 = 1$:

$$F(x=0) = \frac{3!}{(0!)(3-0)!} 0{,}55^0 0{,}45^{3-0} = 1(1)0{,}09113 = 0{,}09113$$

$$F(x=1) = \frac{3!}{(1!)(3-1)!} 0{,}55^1 0{,}45^{3-1} = 3(0{,}55)(0{,}20250) = 0{,}33413$$

La fórmula del binomio ya se percibe con claridad para expandirla a *n* repeticiones.

4.31 La Formula del Binomio.

$$F(x) = \sum_{i=0}^{n} \binom{n}{x} p^x q^{n-x}; \text{i} = 0,1,...,\text{n}$$

Distribución de densidad o de probabilidad acumulativa Binomial. Claramente depende de la manera en que operan los sucesos en cada evento para obtener la probabilidad del mismo y pueda expresarse como las combinaciones de x en n repeticiones.

$$n\text{C}x = \binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Y las probabilidades binomiales de las frecuencias relativas p y q en la expresión.

$$p^x q^{n-x}$$

4.32 Valuando una hipótesis.

La ingeniera agrónoma tiene que decidir y recomendar con base a los estudios de los resultados de los ensayos. Antes, debe valorar las hipótesis:

H*o*; $P_{(S1)} = 0{,}5$; Contra H*a*; $P_{(S1)} \neq 0{,}5$

Decide estimar las probabilidades binomiales para $n = 3$; $p = q = 0{,}5$, extrapolar las frecuencias esperadas multiplicando la probabilidad de cada evento por $\boldsymbol{N} = 100$ y comparar las frecuencias mediante la prueba de χ^2.

4.33 Desarrollo de la prueba.

Obteniendo el estadístico χ^2:

$$\chi^{1}_{(4-1)} = \sum_{i=0}^{3} \frac{\left(\left|fo_i - fe_i\right| - 0{,}5\right)^2}{fe_i} =$$

$$= \frac{\left(\left|14 - 12{,}5\right| - 0{,}5\right)^2}{12{,}5} + \frac{\left(\left|26 - 37{,}5\right| - 0{,}5\right)^2}{37{,}5} + \frac{\left(\left|41 - 37{,}5\right| - 0{,}5\right)^2}{37{,}5} + \frac{\left(\left|19 - 12{,}5\right| - 0{,}5\right)^2}{12{,}5} = 6{,}4267$$

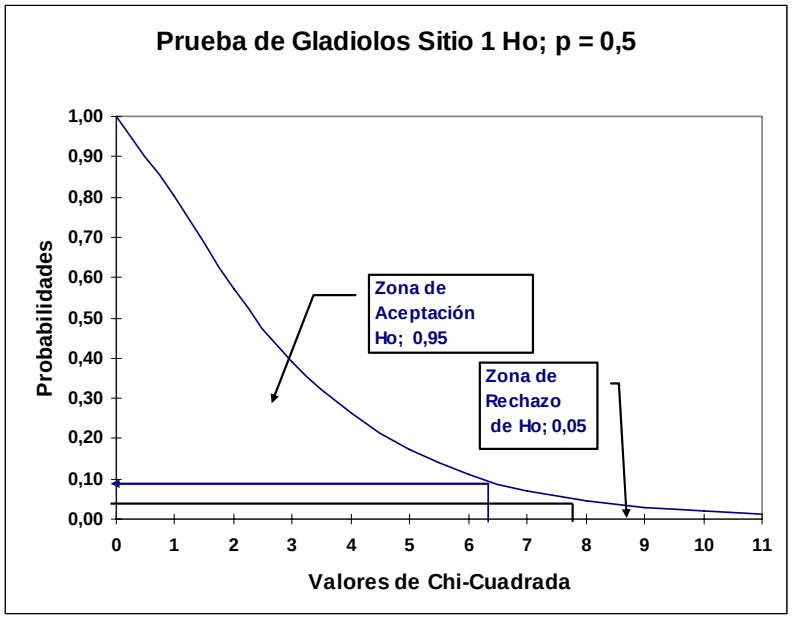
Con probabilidad:

$$F_{[6,4267;\,4\text{-}1]} =$$

$$= Y_0 \int_0^{6,4267} (6{,}4267)^{\frac{1}{2}(4-1)} e^{-\frac{1}{2}6,4267}\, d\chi$$

$$= 0{,}0926$$

En el gráfico puede observarse lo cerca que está de entrar a la zona de rechazo de la hipótesis nula. No Obstante debe aceptarse y reportar que no hay diferencias en la proporción que producen el Sitio I y el Sitio II; y recomendar más pruebas.



4.34 Muestreo en población con distribución binomial.

La estructura anterior puede aplicarse a poblaciones que se distribuyen binomialmente. Para explicar el proceso se utilizará el siguiente ejemplo.

Una empresa caficultora quiere diversificarse, envasar su propio café y exportarlo a la Comunidad Europea. Antes de ponerlo a la venta debe ajustarse a las normas de pesos y medidas. Para el café se exige un peso no inferior al 0,5% reportado en el envase. La empresa envasa en bolsas de 500 gramos. Por tanto, el peso no puede ser inferior a 497,5 gramos en el 2% de las muestras.

El Ingeniero Industrial encargado de la producción decidió elaborar cartas de control de pesos.

4.35 Método de muestreo.

Para controlar la calidad y el proceso se ha propuesto un método de muestreo sistemático con iniciación aleatoria de 6 unidades después de 100 envasadas.

La máquina envasadora lleva un conteo de las bolsas que prepara. Para iniciar el muestro el Ingeniero entrega a los operarios en número que tienen que sumar al que guarda la máquina desde que paró para iniciar las tomas de 6 paquetes que son pesados inmediatamente anotando el peso de cada envase y señalando el peso promedio en la carta de control.

Después del primer muestreo, se toman muestras cada que la máquina envasa 100 bolsas más, ella misma avisa mediante un timbre.

4.36 La Hipótesis.

En la HE se presenta el registro de 124 muestras de 6 bolsas de café con el peso de cada una. Se registra el peso, si este es inferior al de la norma (497,5 gr.) se anota 1 en el registro de fallas, que es el archivo que se muestra en la HE.

La hipótesis que debe valorarse es la siguiente: ¿la distribución de fallas puede aproximarse mediante una Binomial con P = 0,005?

Ho; $P = 0{,}005$

Esta valoración se efectuará mediante la comparación de las frecuencias observadas con las frecuencias esperadas usando la aprueba de χ^2. Lo primero que debe hacerse es sumar las fallas de cada muestra y obtener las frecuencias esperadas.

4.37 Cálculos de la Binomial:

Para $x = 0$ envases fuera de norma de 6 muestras:

$$P(x=0;0{,}005) = \frac{6!}{0!(6-0)!}0{,}005^{0}0{,}995^{(6-0)} = 0{,}97037$$

Para $x = 1$ envases fuera de norma de 6 muestras:

$$P(x=1;0{,}005) = \frac{6!}{1!(6-1)!}0{,}005^{1}0{,}995^{(6-1)} = 0{,}02926$$

Para $x = 2$ envases fuera de norma de 6 muestras:

$$P(x=2;0{,}005) = \frac{6!}{2!(6-2)!}0{,}005^{2}0{,}995^{(6-2)} = 0{,}00037$$

Los siguientes cálculos de probabilidades se omiten pues son muy cercanas a cero y no influyen en el resultado. Estas probabilidades se multiplican por el número de muestras consideradas para obtener las frecuencias esperadas y aplicar la χ^2 modificada.

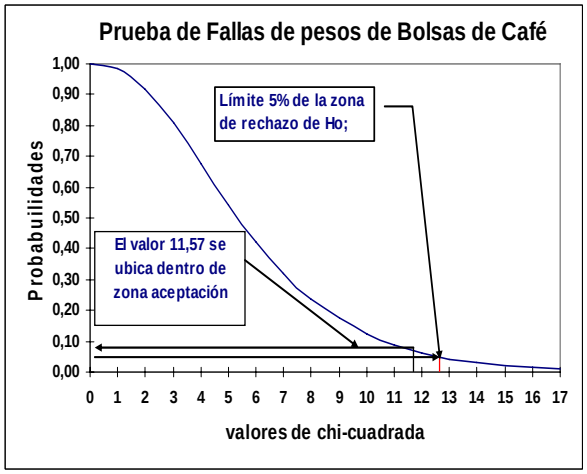
$$\chi^2_{(7-1)} = \sum_{i=0}^{6}\frac{(fo_i - fe_i)}{fe^i} = \frac{(114-120{,}3)^2}{120{,}3} + \frac{(10-3{,}6)^2}{3{,}6} + \frac{(0-0{,}00037)^2}{0{,}00037} + ...\chi^2_6 = 11{,}5704$$

La probabilidad que determina este estadístico es:

$$F_{[11{,}5704;\,4\text{-}1]} = Y_0\int_0^{11{,}5704}(11{,}5704)^{\frac{1}{2}(6-1)}e^{-\frac{1}{2}11{,}5704}d\chi = 0{,}0723$$

4.38 Desarrollo de la prueba.

Eventos x	Frecuencias Observads	Número de Fallas	Probabilidad Binomial	Frecuencias Esperadas	Chi-cuadrada parcial
0	114	0	0,97037	120,3	0,3326
1	10	10	0,02926	3,6	11,1919
2	0	0	0,00037	0,0	0,0456
3	0	0	0,00000	0,0	0,0003
4	0	0	0,00000	0,0	0,0000
5	0	0	0,00000	0,0	0,0000
6	0	0	0,00000	0,0	0,0000
Sumas	124	10	1,00000	124,0	11,5704
				Probabilidad	**0,0723**



El valor de $\chi^2 =$ 11,57 se ubica dentro de la zona de aceptación de la Hipótesis nula: los datos se distribuyen Binomial P = 0,005 debe aceptarse la hipótesis nula.

La proporción de fallas es de: $\hat{p}_f = \frac{\sum_{i=1}^{744} x_i}{744} = \frac{10}{124 \times 6} = 0{,}0134$

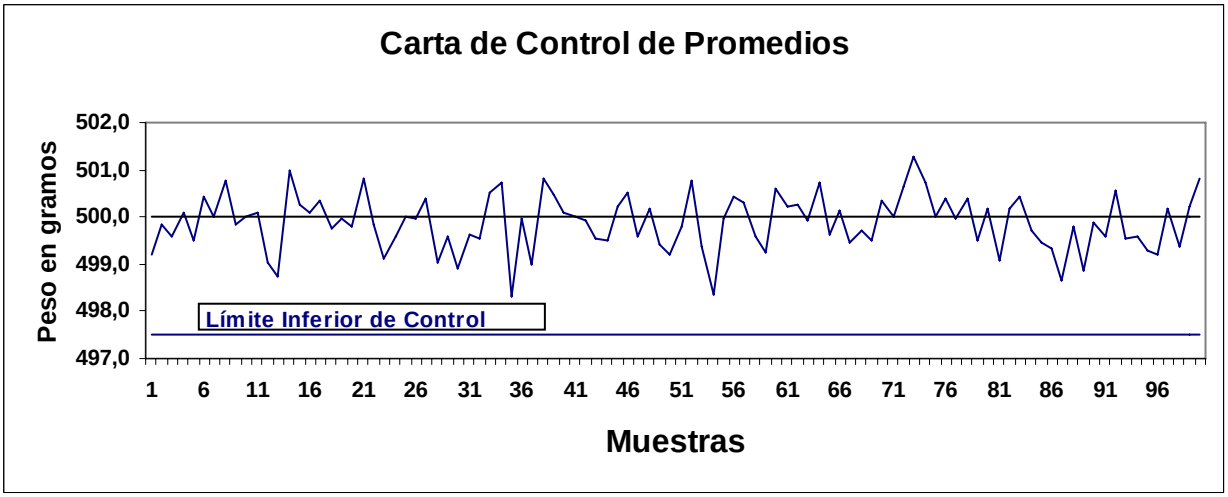
Ó 1,34%. Inferior al 2% que fija la norma.

4.39 La otra parte de la prueba.

Aun cuando la variable peso no corresponde a una distribución de cualidades, los proyectos o exigencias del mercado suelen incluir ambos tipos. Recuerde que el promedios de muestras fuera de norma no puede exceder al 2% (el porcentaje estimado de fallas fue de 1,34%.).

El análisis se ofrecerá mediante la carta de control de los últimos 100 muestreos. El sistema parece estar bajo control estadístico y no hay ninguna muestra que llegue al Límite Inferior de Control.

La recomendación es: Exportar el café.



4.40 Media y Desviación Estándar de la Distribución Binomial.

Es conveniente deducir los parámetros de la las distribuciones, cualesquiera que sean estas, para aprovechar las ventajas que proporciona la aplicación de la teoría estadística desarrolladas sobre el análisis paramétrico, esto es, sobre la media y la varianza.

La distribución binomial tiene la característica de presentar media y desviación estándar, de proporciones, porcentajes y de números de individuos que poseen la cualidad bajo estudio.

Esta característica de la distribución puede confundir un poco al estudiante. Se tratará de aclarar el uso de uno u otro parámetro durante el desarrollo del ejemplo.

No se debe pasar por alto que se opera con distribuciones discretas.

4.41 Problema 4-2. Ejemplo sobre ambiente.

Un proyecto cuyo objetivo fue delinear un plan de acción para desarrollar el respeto al ambiente en niños de primaria, panificó en primer lugar, obtener una semblanza de la idea que tenían los escolares de quinto año sobre conceptos relacionados al cuidado y respeto al medio ambiente. Se realizó una encuesta en escuelas primarias estatales.

La elección se realizó bajo una técnica de conglomerados por distritos y después eligiendo al azar a uno de los grupos se quinto año de la escuela.

Los niños respondían marcando en casillas identificadas con "Sí" cuando reconocían en la pregunta la trascendencia positiva hacia el ambiente y "No" cuando creían lo contrario.

Los datos de las 5 preguntas se muestran en la HE.

4.42 La Hipótesis del Proyecto.

El conocimiento del niño sobre temas ambientales se reflejaría en la distribución de frecuencias, el menor conocimiento ocurriría si las frecuencias apuntaran a 5 respuestas negativas y el mayor cuando apuntara hacia 5 positivas. Es evidente que la distribución de frecuencias se podía aproximar mediante la distribución binomial.

Los directores del proyecto estarían satisfechos si el promedio fuera superior a 2,5 respuestas positivas o el porcentaje fuera superior al 55% de respuestas afirmativas.

Las 457 encuestas de vaciaron en la HE para su informatización, el 1 bajo un encabezado significa un "Sí" en la encuesta (únicamente se refieren los resultados de esta pregunta).

4.43 Estadísticas dato por dato y de frecuencias.

Cuando se hace un estudio que considera 5 respuestas posibles a un tema y estas respuestas hacen referencia a poseer una cualidad mediante un 1 y no poseerla con un cero, se crean dos conjuntos de estadísticas descriptivas:

Las que se obtienen de las respuestas individuales de cada encuestado a cada pregunta específica y las que se obtienen de las sumas de cada encuestado, esto es, de las distribuciones de frecuencias. Estas se irán tratando a medida que se utilicen. Por el momento la proporción p = Promedio = Suma Total / Conteo.

Estadísticas Directas	Datos	Frecuencias
Promedio	0,5714	2,8568
Desviación Típica	0,0102	0,0544
Mediana	1	3
Moda	1	3
Desviación Estándar	0,4950	1,1856
Máximo	1	5
Mínimo	0	0
Rango	1	5
Coeficiente Asimetría	-0,2886	-0,0870
Curtosis	-1,9183	-0,5506
Suma Total	1357	1357
Cuenta	2375	475

El cuadro se obtiene operando instrucciones para cada estadístico. El de datos corresponde al conjunto de las 5 preguntas, el de frecuencias a la suma de las respuestas de cada encuestado. Esta manera de obtener los estadísticos servirá de referencia a los que se obtienen mediante datos agrupados y muestran de manera más apropiada la distribución de datos. Se espera que las frecuencias se distribuyan binomial.

4.44 Estadísticas de la distribución de frecuencias.

La otra alternativa para obtener estadísticas descriptivas es utilizando La Distribución de Frecuencias que en estos problemas tienen clases cerradas correspondientes a los eventos de *x* que han sucedido.

Así, x = 0 indicará el número de estudiantes que no señalaron ninguna respuesta positiva; x = 1, significa el número de estudiantes que señalaron una repuesta positiva y así sucesivamente hasta x = 6, para aquellos eventos en que se marcaron seis respuestas positivas.

Evento x	Frecuencias Observadas	Probabilidad Binomial	$P(x)\times x_i$	$P(x)(x_i-\bar{x})^2$	$P(x_i)\left(\frac{x_i-\bar{x}}{s}\right)^3$	$P(x_i)\left(\frac{x_i-\bar{x}}{s}\right)^4$
0	8	0,01447	0,0000	0,11808	-0,2490	0,6427
1	56	0,09643	0,0964	0,33249	-0,4556	0,7645
2	116	0,25709	0,5142	0,18875	-0,1194	0,0924
3	151	0,34270	1,0281	0,00702	0,0007	0,0001
4	104	0,22841	0,9136	0,29849	0,2518	0,2601
5	40	0,06089	0,3045	0,27970	0,4424	0,8568
		1,00000			-0,1290	2,6166

Estadísticas Descriptivas	
Suma de Frecuencias	475
Promedio	2,8568
Mediana	2,8841
Moda	2,9268
Varianza	1,2245
Desviación Estándar	1,1066
Desviación Típica	0,4949
Asimetría	-0,0003
Curtosis	-0,0008

4.45 Cálculos e igualdades.

En primer lugar debe obtenerse la probabilidad binomial para cada evento. Para esto se requiere conocer *n* o número de ensayos, *p = proporción de aciertos*. *n* = 5 y la proporción promediando los datos. Eso es:

$$p=\frac{\sum_{i=1}^{2375}x_i}{2.375}=\frac{1.357}{2.375}=0,5714$$

El promedio de la sumas de todos los que respondieron Sí en las preguntas entre el total de todos los ceros más los unos es igual a la proporción que respondió "Sí".

Después debe calcularse las probabilidades binomiales para la proporción obtenida, esto con el objeto de eliminar variaciones aleatorias y considerar que la distribución de los datos se puede aproximar mediante una binomial.

Es estudiante puede comprobar que el resultado es idéntico al obtenido utilizando las frecuencias agrupadas:

$$np=\frac{\sum_{i=0}^{5}f_ix_i}{5}=\frac{0\times8+1\times56+2\times116+3\times151+4\times104+5\times40}{5}=\frac{1.357}{475}=2,8568$$

Para obtener p basta dividirlo por n.

$$p=\frac{2,8568}{5}=0,5714$$

Asumiendo que la distribución es binomial, puede obtenerse mediante la fórmula que utiliza las probabilidades:

$$np=\sum_{x=0}^{5}fo_i\binom{5}{x}0,5714^{x}0,4286^{5-x}$$

4.46 La función de probabilidad.

El uso de la función de la HE permite calcular las probabilidades del evento o el acumulado. Para el caso interesan las probabilidades por evento:

$$P(x=0)=\frac{5!}{0!5!}0{,}5714^{0}0{,}4286^{5}=\text{DISTR.BINOM}(0;5;0{,}5714;0)=0{,}01447$$

$$P(x=1)=\frac{5!}{1!4!}0{,}5714^{1}0{,}4286^{4}=\text{DISTR.BINOM}(1;5;0{,}5714;0)=0{,}09643$$

$$P(x=2)=\frac{5!}{2!3!}0{,}5714^{2}0{,}4286^{3}\,\text{DISTR.BINOM}(2;5;0{,}5714;0)=0{,}25709$$

$$P(x=3)=\frac{5!}{3!2!}0{,}5714^{3}0{,}4286^{2}=\text{DISTR.BINOM}(3;5;0{,}5714;0)=0{,}34270$$

$$P(x=4)=\frac{5!}{4!1!}0{,}5714^{4}0{,}4286^{1}=\text{DISTR.BINOM}(4;5;0{,}5714;0)=0{,}22841$$

$$P(x=5)=\frac{5!}{5!0!}0{,}5714^{5}0{,}4286^{0}=\text{DISTR.BINOM}(5;5;0{,}5714;0)=0{,}06089$$

Sumando los eventos se abra comprobado que los cálculos de cada uno de los eventos es correcto si:

$$\sum_{xi=0}^{5}\binom{n}{x}p^{x}q^{n-x}=1$$

4.47 Estadística descriptiva: número promedio.

El número de muestras se obtiene sumando las frecuencias:

$$m=\sum_{i=0}^{5}f_i=8+56+116+151+104+40=475$$

El promedio se obtiene usando la modalidad de multiplicar la probabilidad de cada evento por el número del evento:

$$np=\sum_{x=0}^{5}fo_i\binom{5}{x}0{,}5714^{x}0{,}4286^{5\text{-}x}=8\times0{,}01447+56\times0{,}09643+\ldots+40\times0{,}06089=$$

$$=2{,}8568$$

O el promedio de individuos por muestra:

$$\bar{\bar{x}}=\frac{\sum_{j=1}^{475}\sum_{i=1}^{5}x_{ij}}{475}=\frac{1.375}{475}=2{,}8568$$

Las igualdades de la distribución binomial se empiezan a aparecer. Así, el promedio de proporciones se puede calcular mediante la fórmula:

$$p=\frac{\bar{x}_n}{n}=\frac{np}{n}=\frac{2{,}8568}{5}=0{,}5714$$

4.48 Estadísticas descriptivas: la proporción es una media.

Esto significa que en la Distribución Binomial tendremos dos promedios: $\bar{x}_p=p$; para proporciones y porcentajes.

$\bar{x}_n = np$; para número de individuos o eventos.

La mediana no es un estadístico que se acostumbre calcular para distribuciones cualitativas, no obstante, en ciertas circunstancias puede ser de utilidad, pero recuerde que es una mediada para variables continuas.

La Mediana. Notará que el límite inferior de la clase mediana se fija en x =2,5, considerando las frecuencias como un número continuo:

$$\tilde{x} = LI_{\tilde{x}} + \left(\frac{\frac{n+1}{2} - S_{(m-1)}}{f_m} \right) IC = 2{,}5 + \left(\frac{238-180}{151} \right) 1 = 2{,}8841$$

4.49 Estadísticas descriptivas: la Varianza.

De la moda se puede decir lo mismo que de la mediana.

$$\vec{x} = LI\vec{x} + \left[\frac{f_m - f_{m-1}}{(f_m - f_{m-1}) + (f_m - f_{m+1})} \right] IC = 2{,}5 + \left[\frac{151-116}{(151-116)+(151-104)} \right] 1 = 2{,}9268$$

Trabajando con porcentajes o probabilidades es más fácil entender el coeficiente de asimetría y su efecto sobre la forma de la distribución:

- ***La proporción centrada es de 0,5, o sea, que 50% de los datos están debajo de 50% y el resto sobre 50%; en el ejemplo, la proporción es de 0,5714 indicando que hay una tendencia a acumular valores altos, por arriba del promedio;***
- ***El promedio de números para que fuera centrado debería ser 2,5, pero es de 2,8568;***
- ***La mediana se corre a la derecha de la media y antes de la moda, el valor de esta es: 2,8841;***
- ***La moda se corre a la derecha de la mediana con un valor de 2,9268;***
- ***El coeficiente de asimetría es negativo –0,0003, indicando que la cola izquierda es más alta;***
- ***Un coeficiente de asimetría negativo indicará una cola izquierda larga y una acumulación de valores altos.***

La varianza se ha calculado de manera similar como se calculó el promedio, multiplicando la desviación cuadrática por la probabilidad del evento:

$$s_n^2 = \sum_{i=0}^{5} p(x)_i \left(\bar{x}_i - \bar{\bar{x}}\right)^2 = 0{,}01447(0-2{,}8568)^2 + 0{,}09643(1-2{,}8568)^2 + 0{,}25709(2-2{,}8568)^2 +$$
$$+ 0{,}34270(3-2{,}8568)^2 + 0{,}22841(4-2{,}8568)^2 + 0{,}06089(5-2{,}8568)^2 =$$
$$= 1{,}2245$$

Se puede comprobar que:

$$s_n^2 = npq = 5 \times 0{,}5714 \times 0{,}4286 = 1{,}2245$$

Al igual que con la media se tienen dos varianzas calculadas a partir de las proporciones p:

$$s_p^2 = pq = 0{,}5714 \times 0{,}4286 = 0{,}2449$$

4.50 Estadísticas descriptivas: Desviaciones Estándar.

Las desviaciones estándar correspondientes:

$$s_n = \sqrt{s_n^2} = \sqrt{1{,}2245} = 1{,}1066$$

$$s_p = \sqrt{s_p^2} = \sqrt{0{,}2449} = 0{,}4949$$

Con las desviaciones típicas (distribuciones de estimadores) debe tenerse cuidado pues la *desviación estándar de proporciones* es la *desviación típica de números de eventos*. Además, la mayoría de los libros la refieren como desviación estándar o error estándar.

$$s_{\bar{n}} = \frac{s_n}{\sqrt{n}} = \frac{1{,}1066}{\sqrt{5}} = 0{,}4949$$

$$s_{\bar{p}} = \frac{s_p}{\sqrt{n}} = \frac{0{,}4949}{\sqrt{5}} = 0{,}2213$$

4.51 Estadísticas descriptivas: Asimetría o Sesgo..

El Coeficiente de Asimetría:

$$ca = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{15} f_i \left(\frac{\bar{x}_i - \bar{\bar{x}}}{s} \right)^3 = \frac{475}{(474)(473)} (-0{,}1290) = \text{-}0{,}0003$$

El tercer momento a_3 , que es la suma de la fórmula anterior se obtiene mediante:

$$a_3 = \frac{q-p}{\sqrt{npq}} = \frac{0{,}4286 - 0{,}5714}{\sqrt{5 \times 0{,}5714 \times 0{,}4286}} = -0{,}1290$$

El Coeficiente de Curtosis;

$$cc = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{15} f_i \left(\frac{\bar{x}_i - \bar{\bar{x}}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)} =$$

$$= \frac{475(476)}{(474)(473)(472)} (2{,}6166) - \frac{3(474)^2}{(473)(472)} = \text{-}0{,}0008$$

4.52 Estadística descriptiva: Curtósis o alargamiento.

El cuarto momento que es la suma de la fórmula anterior se puede calcular mediante:

$$a_4 = 3 + \frac{1 - 6pq}{npq} = 3 + \frac{1 - 6 \times 0{,}5714 \times 0{,}4286}{5 \times 0{,}5714 \times 0{,}4286} = 2{,}6166$$

A estas alturas, el estudiante habrá percibido que las estadísticas descriptivas de la distribución binomial pueden calcularse directamente si se conoce p o q. Aun cuando pueden obtenerse usando las tablas de frecuencias que permiten una panorámica del problema mediante el histograma y elaborar pruebas de para corroborar la similitud entre las distribuciones observadas en las esperadas en la prueba de *BONDAD DE AJUSTE*.

Otras alternativas para valorar hipótesis sobre proporciones, porcentajes y eventos se verán en los dos siguientes capítulos.

4.53 Probando la hipótesis Ho; P < 0,55.

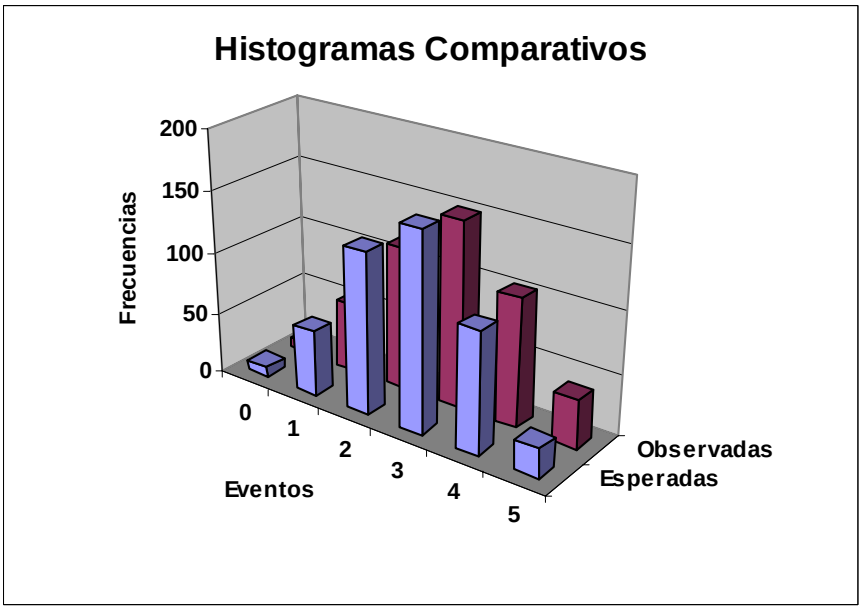
Por el momento la prueba que se está en capacidad de utilizar es comparar las frecuencias observadas contra las frecuencias esperadas que se obtendrán con la proporción p =

Evento x	Probabilidad Binomial	FRECUENCIAS Esperadas	FRECUENCIAS Observadas	Chi-cuadrada Parciales
0	0,01845	8,8	8	0,00802
1	0,11277	53,6	56	0,06994
2	0,27565	130,9	116	1,59144
3	0,33691	160,0	151	0,45487
4	0,20589	97,8	104	0,33253
5	0,05033	23,9	40	10,17203
Sumas	1,00000	475,0	475	12,62884
			Probabilidad	**0,02712**

*0,55*La probabilidad de la prueba:

$$F_{[12,62884;\,6\text{-}1]} = Y_0 \int_0^{12,62884} (12,62884)^{\frac{1}{2}(6-1)} e^{-\frac{1}{2}12,62884}\, d\chi = 0,02712$$

4.54 Resultado e interpretación de la prueba.



Ésta prueba tiene mucha similitud con la prueba de bondad de ajuste, la diferencia está en que para esta última se utilizan los estimados de los datos (p = 0,5714). En esta prueba de proporciones se utiliza la distribución teórica que se obtendría con la proporción que se quiere valorar P = 0,55. La probabilidad de la prueba de 0,02712 ó 2,71% indica que la hipótesis de que P = 0,55 debe rechazarse. Puesto que p = 0,5714 debe concluirse que los niños tienen más conocimiento sobre protección del medio ambiente del esperado.

En el gráfico es notorio que la distribución observada se corre a valores que dan mejor calificación.

4.55 La Binomial en el Control de la Calidad y el Proceso.

La Distribución Binomial es muy útil en el control estadístico de la calidad y de los procesos, especialmente cuando las fallas no son frecuentes o cuando es necesario usar pocas muestras debido a la destrucción del material.

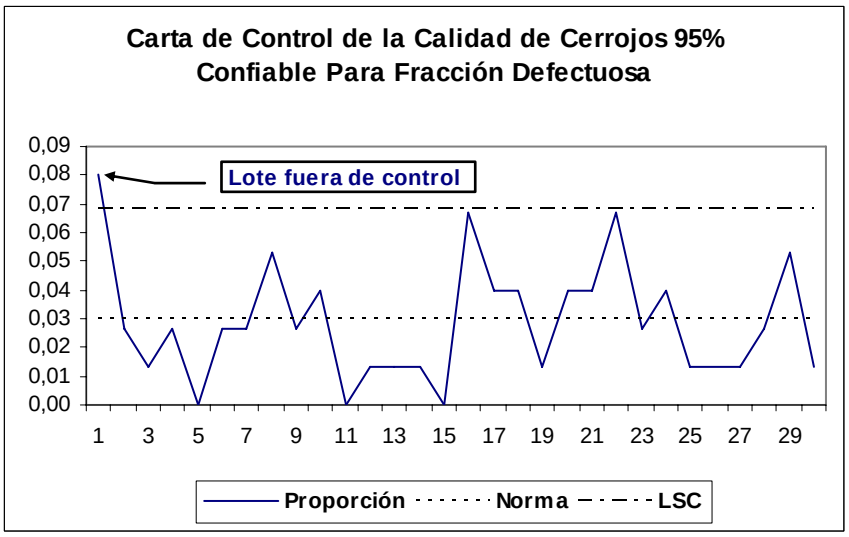
Una empresa que fabrica cerrojos quiere mantener una calidad de menos de 3% defectuosos, para mantener esta calidad, de control se toma una muestra de 75 cerrojos cada 4 horas de operación. El control de proceso y calidad se valoran mediante el gráfico de control de la fracción defectuosa. Los datos cerrojo a cerrojo y promedio se presentan en la HE.

4.56 Gráfico de control de la calidad sobre proporciones.

El gráfico muestra lo esencial en el control de la calidad: El límite aceptado en la proporción de fallas de 3%, y el límite superior de control.Un intervalo confiable que se obtiene al considerar un valor z = 1,96 que determina un límite 95% confiable. El uso de este en una distribución discreta como es la binomial será objeto de estudio del siguiente capítulo. Es simple observar que sólo una muestra cae fuera del límite de control.

El Límite superior de control es un intervalo confiable acotado en la cola superior. Esto es. Para proporciones se obtiene resolviendo:

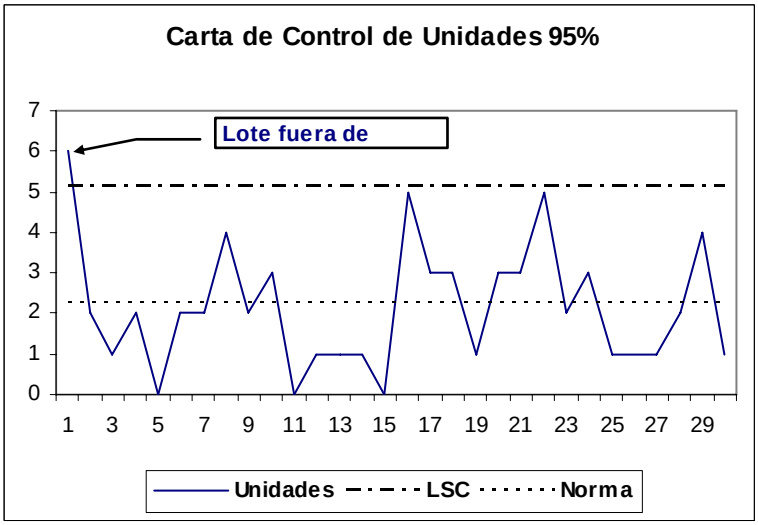
$$LSC_{95\%} = \left\{ Norma + z\sqrt{\frac{pq}{n}} \right\} = 0,03 + 1,96\sqrt{\frac{0,03 \times 0,97}{75}} = 0,069 \approx 0,07$$



4.57 *El gráfico de Control de la Calidad sobre número de cerrojos defectuosos.*

De la misma manera se puede obtener un gráfico de control de la calidad o del proceso utilizando los números de cerrojos defectuosos en lotes de 75 unidades exploradas cada 4 horas.

Lo dicho de la diapositiva anterior vale para esta, con el comentario adicional que para que se pueda hacer inferencia sobre números, *La Distribución Binomial* debe poderse aproximar con una *Distribución Normal Estándar*.



La proporción aceptable de fallas es 0.03, por tanto, el número de cerrojos con defectos aceptable en 75 inspeccionados será:

$$np = 75 \times .03 = 2{,}25$$

Unidades. Y el intervalo confiable para 95% es:

$$LSC_{95\%} = \left\{Norma + z\sqrt{npq}\right\} = 2{,}25 + 1{,}96\sqrt{75 \times 0{,}03 \times 0.97} = 5{,}15$$

4.58 *Conclusión y Resumen.*

En este capítulo se ha tratado lo referente a las bases de probabilidad mínimas para comprender los alcances de la *Distribución Binomial* y las estadísticas descriptivas de la distribución.

En los dos capítulos siguientes se tratará de la inferencia estadística usando *La Distribución Binomial* haciendo uso de la aproximación que se puede efectuar mediante *La distribución Normal Estándar*.

También se abordará la aproximación que se hace a la *Distribución Binomial* por medio de *La Distribución Poisson.*

REFERENCIAS SELECTAS:

1. Hillier Frederick S., y Lieberman Gerard j., Introducción a la Investigación de Operaciones. Capítulo 19. Segunda edición en español traducida de la cuarta edición en ingles. McGraw-Hill Interamericana de México, S. A. De C. V., 1990.
2. Miller Irwin, Freund John E., Johnson Richard A: Probabilidad y Estadística para Ingenieros. Capítulo 11. Traducido de la cuarta edición en ingles; Prentice-Hall Hispanoamericana, S. A. 1992.
3. Murray R. Spiegel: Serie de compendios Schaum, Teoría y Problemas de Estadística. Capítulos 7. Primera edición en español, traducido de la primera edición en ingles; Libros McGraw-Hill de México, S. A. De C. V., 1973.
4. Ostle Bernard: Estadística Aplicada. Capítulos 3 y 7. Primera edición en español traducida de la primera edición en ingles. Editorial Limusa, S. A., 1977.
5. Snedecor George W., y Cochran William G: Statistical Methods. Capítulo 8 y 9. Sexta edición; The Iowa State University.
6. Steel Robert G. D., Torrie James H: Principles and Procedures of Statistics. Capítulo 4. Primera edición; McGraw-Hill Book Company, Inc, 1960.

5 La χ^2: Aproximado la Binomial con la Normal.

Los archivos para esta sección son:
E05_ChCuadrada_P01.ppn;
E05_ChCuadrada_W01.doc;
E05_ChCuadrada_X01.xls

5.1 Menú.

5.2 Introducción.

Desde el primer capítulo se ha utilizado una distribución estadística para efectuar pruebas de variables discretas que hacen mención a cantidades de individuos, tal como son las frecuencias observadas de las distribuciones de datos cuando se agrupan, para percibir su imagen o buscar mecanismos para describir e inferir adecuadamente la población que se estudia.

Esta distribución, además de utilizarse como mecanismo para evaluar la aproximación de las frecuencias observadas a las desarrolladas con una distribución probabilística denominadas frecuencias esperadas, se utiliza como una especie de puente que une la estadística de variables continuas con la estadística que trata de distribuciones resultantes de cualidades.

5.3 La Distribución de $\chi 2$.

La distribución estadística identificada con la letra griega *Chi o ji* está definida por la estructura matemática:

$$\chi^2 = \frac{ns^2}{\sigma^2} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{\sigma^2}$$

En donde el numerador se conoce como *Suma de Cuadrados* de una muestra grande expresada, y perfectamente podría sustituirse por: $(n-1)s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2$

En el numerador sin perder ninguna de sus propiedades. La principal, la de ser una distribución de densidad.

5.4 Intervalo de confianza.

Por ser una distribución de densidad y su forma, permite estimar intervalos de confianza para la desviación estándar para niveles de significación α sea 0,05 0 0,01 que son los usuales o cualesquiera que sirvan al proyecto. Considerando el intervalo confiable para una χ^2 cualquiera:

$$\Pr\left[\chi^2_{(n-1;\alpha)} < \frac{ns^2}{\sigma^2} < \chi^2_{(n-1;\alpha)}\right] = \alpha$$ Ecuación que se transforma fácilmente en:

$$\Pr\left[\frac{s\sqrt{n}}{\chi^2_{(n-1;\alpha)}} < \sigma < \frac{s\sqrt{n}}{\chi^2_{(n-1;\alpha)}}\right] = \alpha$$

El intervalo de confianza con probabilidad $1 - \alpha$ para la desviación estándar.

5.5 La aproximación a La Normal Estándar.

Quizá la característica se deba a que sí se toma el lado positivo de la normal estándar, la ecuación:

$$\sqrt{\chi^2} - \sqrt{\upsilon - 1}$$

Se distribuye muy aproximadamente como una *Normal Estándar* acumulativa cuando el número de *Chi-Cuadradas* parciales es mayor a 30. Y cuando los grados de libertad son $\upsilon = 2$:

$$z = \sqrt{\chi^2}$$

ATENCIÓN: La probabilidad desde el valor de χ^2 hasta infinito es el doble de la probabilidad de z hasta infinito. Esto es, χ^2 refleja la probabilidad de las dos colas en una sola. Por esto:

$$P(\chi^2) = 2P(z)$$

5.6 La Binomial.

En otras palabras, la prueba de χ^2 siempre hace referencia a una prueba *Normal Estándar* de dos colas.

$$P(\chi^2) = P\left(z_{-\alpha/2}\right) + P\left(z_{+\alpha/2}\right)$$

La del extremo inferior y la del extremo superior. Antes de utilizar esta importante igualdad, se debe aclarar que se va a trabajar con las variables cualitativas que caracterizan a la Binomial aunque se haya iniciado con una distribución para variables continuas como es la χ^2. Recordaremos que la función de probabilidad de la binomial esta definida por:

$$F(x) = \sum_{x=0}^{n}\binom{n}{x}p^x q^{n-x}; \text{i} = 0,1,...\text{n}$$

5.7 Aproximando la Binomial con la Normal Estándar.

Aún cuando *La Distribución Binomial* es una herramienta de deducción e inducción poderosa, no posee la flexibilidad, los alcances y sobre todo la simpleza de uso de *La Distribución Normal Estándar*.

La teoría estadística ha desarrollado todo un bagaje cultural y tecnológico con estas dos distribuciones pues cubren una gran cantidad de aspectos relacionados con la aplicación práctica de la teoría estadística en el análisis de poblaciones y en el diseño de experimentos planificados.

Para iniciar la deducción del hecho de aproximar una distribución discreta mediante una continua, supóngase que un evento ocurre con una proporción $p = 0{,}5$ que se calcularán para $n = 10$ sucesos.

5.8 Probabilidades por evento.

La probabilidad de que $x = 0$ individuos de $n = 10$ seleccionados al azar posean la característica de interés se calcula mediante:

$$P(x=0)=\binom{10}{0}p^0q^{10-0}=\frac{10!}{0!(10-0)!}0{,}5^0 0{,}5^{10-0}=1(0{,}00098)=0{,}00098$$

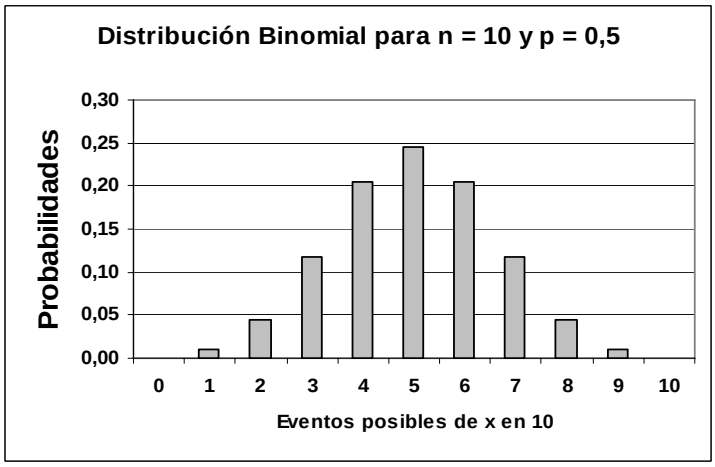
La probabilidad de que $x = 1$ individuos de $n = 10$ seleccionados al azar posean la característica de interés:

$$P(x=1)=\binom{10}{1}p^1q^{10-1}=\frac{10!}{1!(10-1)!}0{,}5^1 0{,}5^{10-1}=10(0{,}00098)=0{,}00977$$ De la misma manera pueden calcularse los eventos restantes hasta obtener el cuadro que se muestra en la siguiente diapositiva.

5.9 El cuadro de probabilidades e Histograma.

Como era de esperarse la suma de las probabilidades individuales de los eventos x es igual a la unidad. El Gráfico representado por barras muestra una forma muy parecida a la normal. Operando los eventos como si se tratase de una variable continua se calcula la tabla de frecuencias para obtener las estadísticas descriptivas.

Tamaño muestra		10	
Prooprción interés p =		0,5	
Proporción complementaria		0,5	
Evento x	Número Combinaiones	Probabilidad Binomial	Propabilidad Evento
0	1	0,00098	0,00098
1	10	0,00098	0,00977
2	45	0,00098	0,04395
3	120	0,00098	0,11719
4	210	0,00098	0,20508
5	252	0,00098	0,24609
6	210	0,00098	0,20508
7	120	0,00098	0,11719
8	45	0,00098	0,04395
9	10	0,00098	0,00977
10	1	0,00098	0,00098
		Sumas	1,00000



5.10 Distribución de Frecuencias y Estadísticos.

Para obtener el promedio y la varianza se utilizó la alternativa de multiplicar la probabilidad de cada evento por el número de aciertos que indica. Como se sabe, el promedio se puede obtener mediante:

$= np = 10 \times 0{,}5 = 5$

Y la varianza:

$\sigma^2 = npq = 10 \times 0{,}25 = 2{,}5$

Con una:

$s = 1{,}5811.$

Evento x	Frecuencias o probabilidad	$p_i x_i$	$p_i(x_i - \bar{x})^2$
0	0,00098	0,00000	0,02441
1	0,00977	0,00977	0,15625
2	0,04395	0,08789	0,39551
3	0,11719	0,35156	0,46875
4	0,20508	0,82031	0,20508
5	0,24609	1,23047	0,00000
6	0,20508	1,23047	0,20508
7	0,11719	0,82031	0,46875
8	0,04395	0,35156	0,39551
9	0,00977	0,08789	0,15625
10	0,00098	0,00977	0,02441
Muestras			10
Promedio			5
Varianza			2,5
Desviación Estándar			1,5811

5.11 Límites estandarizados.

Para poder comparar *La Distribución Binomial* con *La Normal Estándar* es necesario suponer que los eventos ocurren de una manera continua. Para esto debe considerarse que el límite superior de una clase se une al inferior de la siguiente sin solución de continuidad. Esto es, de la mitad entre un evento y otro.−0,5, 0,5, 1,5, 2,5, 3,5 y así sucesivamente hasta el intervalo 10,5.

Después se obtiene las variables estandarizadas de ambos límites, la probabilidad acumulada que determinan en *La Distribución Normal Estándar* de manera que la diferencia de la probabilidad del límite superior menos la probabilidad del límite inferior determinan la probabilidad del intervalo. Los cálculos en la siguiente diapositiva.

Recordar que la variable estándar se calcula mediante:

$$z_i = \frac{x_i - \bar{x}}{s}; \text{ para } x = 2; z_{LI1} = \frac{0{,}5 - 5}{1{,}5811} = -2{,}8460$$

$$z_{LS1} = \frac{1{,}5 - 5}{1{,}5811} = -2{,}2136$$

La probabilidad para el intervalo de 0,5 a 1,5 se obtiene utilizando la función de densidad de la Normal Estándar que proporciona, restando a la probabilidad acumulada que define la variable z mayor, o del límite superior la del límite inferior, esto es:

$$F(x = 1{,}5) - F(x = 0{,}5 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{1{,}5} e^{-\frac{1}{2}\left(\frac{1{,}5-5}{1{,}5811}\right)^2} dx - = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0{,}5} e^{-\frac{1}{2}\left(\frac{0{,}5-5}{1{,}5811}\right)^2} dx =$$

$$= 0{,}0134 - 0{,}0022 = 0{,}0112$$

De esta manera se calculan las probabilidades para todos los intervalos excepto el primero y el último. El primero debe considerar la probabilidad que arrastra desde menos infinito al punto x = −0,5, por tanto:

$$F(-0{,}5) - F(-\infty) = 0 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-0{,}5} e^{-\frac{1}{2}\left(\frac{-0{,}5-5}{1{,}5811}\right)^2} dx = 0{,}0022$$

El último debe considerar la probabilidad para x = 10,5 hasta más infinito:

$$F(-\infty) - F(10{,}5) = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+10{,}5} e^{-\frac{1}{2}\left(\frac{10{,}5-5}{1{,}5811}\right)^2} dx = 0{,}0022$$

Ambos coinciden puesto que la distribución está perfectamente equilibrada con p = q = 0,5.

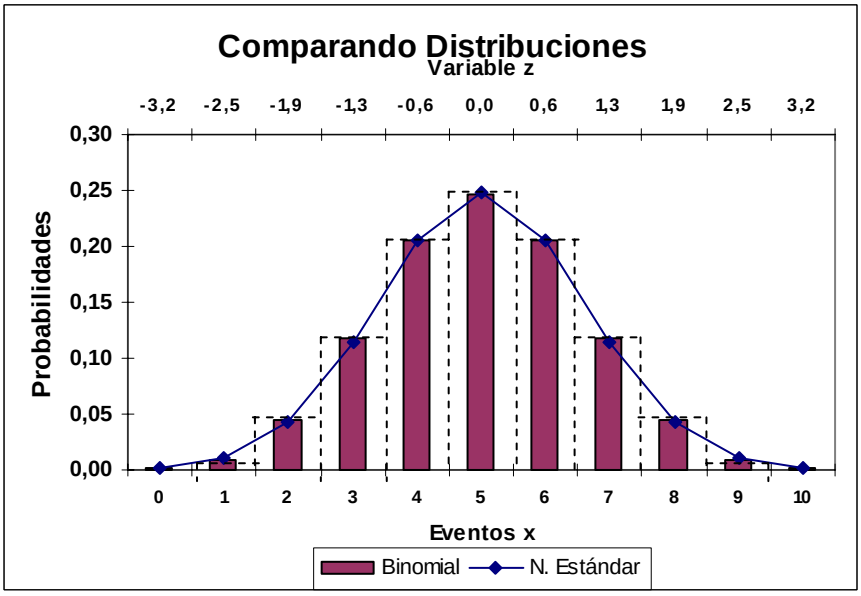
5.12 Probabilidades esperadas.

Se notan pequeñas diferencias entre las probabilidades puntuales de *La Distribución Binomial* y las probabilidades de los intervalos de *La Distribución Normal Estándar*. Veamos como compensarlas.

LÍMITES		VALORES z		Probabilidades Acumuladas		Probabilidad	
Inferior	Superior	Inferior	Superior	Inferior	Superior	Binomial	N. Estándar
-0,5	0,5	-3,4785	-2,8460	0,0000	0,0022	0,0010	0,0022
0,5	1,5	-2,8460	-2,2136	0,0022	0,0134	0,0098	0,0112
1,5	2,5	-2,2136	-1,5811	0,0134	0,0569	0,0439	0,0435
2,5	3,5	-1,5811	-0,9487	0,0569	0,1714	0,1172	0,1145
3,5	4,5	-0,9487	-0,3162	0,1714	0,3759	0,2051	0,2045
4,5	5,5	-0,3162	0,3162	0,3759	0,6241	0,2461	0,2482
5,5	6,5	0,3162	0,9487	0,6241	0,8286	0,2051	0,2045
6,5	7,5	0,9487	1,5811	0,8286	0,9431	0,1172	0,1145
7,5	8,5	1,5811	2,2136	0,9431	0,9866	0,0439	0,0435
8,5	9,5	2,2136	2,8460	0,9866	0,9978	0,0098	0,0112
9,5	10,5	2,8460	3,4785	0,9978	1,0000	0,0010	0,0022
					Sumas	1,0000	1,0000

5.13 Diferencias de probabilidades.

En la figura de la derecha se pueden apreciar las diferencias entre las distribuciones: *La Binomial* está representada por las columnas sólidas mientras *La Normal Estándar* con los rectángulos punteados que representan el área bajo la curva del intervalo.



Por ejemplo la probabilidad que 8 o más individuos posean la cualidad, mediante la binomial es: P(8,9,10) = 0,0439 + 0,0098 + 0,0010 = 0,0547. Para aproximar con

la Normal Estándar considérese la probabilidad acumulativa hasta el límite 7,5 que es de 0,9431, por tanto la probabilidad solicitada es 1 – 0,9431 = 0,0569 una aproximación muy aceptable.

5.14 Deduciendo las diferencias: superiores.

Pero usualmente se tomará la probabilidad directa al número, esto es 8:

$$z_8 = \frac{8-5}{1{,}5811} = 1{,}8974$$

Que determina una probabilidad acumulativa de:

$$P(z_8) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{1{,}8974} e^{-\frac{1}{2}\left(\frac{8-5}{1{,}5811}\right)} dz = 0{,}9711$$

Por tanto la probabilidad estimada es de 1 – 0,9711 = 0,0289 una pobre aproximación. Veamos otro ejemplo considerando la probabilidad de que se encuentren hasta 4 individuos con la cualidad. Para la probabilidad binomial P(x = 0;1 ;2 ;3 ;4) = 0,0010 + 0,0098 + 0,0437 +0,2051 = 0,3770 o 37,70%. Con la *Normal Estándar* tomaremos el límite 4,5.

5.15 Deduciendo las diferencias: inferiores.

La probabilidad para el límite 4,5 es de 0,3779 o 37,79% también una aproximación muy aceptable. Tomado el valor neto de 4:

$$z_4 = \frac{4-5}{1{,}5811} = -0{,}6325$$

Que determina una probabilidad acumulativa de:

$$P(x=4) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-0{,}6325} e^{-\frac{1}{2}\left(\frac{4-5}{1{,}5811}\right)} dz = 0{,}2635$$

Nuevamente se tiene una pobre aproximación. Tome en cuenta que nunca se van a tener fracciones de números en la distribución binomial, por tanto, el ajuste que se deba hacer tendrá que efectuarse sobre la variable estandarizada.

5.16 El ajuste por continuidad.

En general, deberá restarse media unidad a cada evento de un elemento de la distribución binomial para que su variable estandarizada pueda aproximarse adecuadamente con la distribución Normal Estándar, esto es: $z_c = \frac{\left(|x_i - \bar{x}| - 0.5\right)}{s} = -\frac{\left(|4-5| - 0{,}5\right)}{1{,}5811} = -0{,}3162$

La ecuación anterior se generaliza como: $z_c = \frac{|x - np| - \frac{1}{2}}{\sqrt{npq}}$ En términos de números, promedio y varianza de la Distribución Binomial.

5.17 Como ejemplo un problema de muestreo.

En una encuesta a estudiantes del Instituto Tecnológico se apreció que las mujeres repiten menos cursos que los hombres. Un grupo de estudiantes, para su trabajo de final de curso decidió, entre otras cosas, comprobar esta hipótesis haciendo una encuesta a 328 estudiantes a los que se les

preguntaba si habían reprobado algún curso, si era así, se marcaba con 1 la encuesta y la hoja electrónica. Los resultados se muestran en el cuadro.

Interesa la clase de estudiantes que perdieron cursos.

5.18 La Hipótesis.

La oficina de registro reporta que el 63% de los estudiantes perdieron algún curso. Entonces, si los hombres y las mujeres repiten cursos con la misma proporción, la hipótesis se plantea como:

Ho; $P_H = P_M \leq 0{,}63$; contra *Ha*; $P_H = P_M > 0{,}63$

Puesto que la proporción de hombres y mujeres no es la misma, las pruebas deben efectuarse por separado para hombres y para mujeres. Si los hombres reprueban igual que las mujeres, las pruebas deben ser iguales, esto es:

Ho; $P_H = P_M \leq 0{,}63$, contra *Ha*; $P_H = P_M > 0{,}63$

5.19 La prueba en los varones.

El valor de *z* calculada para los varones es: $z_c = \dfrac{|157 - 0{,}63 \times 221| - 0{,}5}{\sqrt{221 \times 0{,}63 \times 0{,}37}} = 2{,}4062$

$$P(z_c) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{2{,}4062} e^{-\frac{1}{2}\left(\frac{157 - 0{,}63(221)}{\sqrt{221(0{,}63)(0{,}37)}}\right)} dz = 0{,}9919$$

Para obtener la probabilidad de la zona de rechazo este valor se resta de 1: $P(\alpha) = 1 - 0{,}9919 = 0{,}0081$.

Esto significa, con mucha certeza, que la probabilidad de que la proporción con que no pasan cursos los hombres es mayor a 63%. Por tanto debe rechazarse la hipótesis nula referente a los varones.

5.20 La prueba en las mujeres.

Procediendo de manera similar en las mujeres:

$$z_M = \frac{|55 - 0{,}63 \times 107| - 0{,}5}{\sqrt{107 \times 0{,}63 \times 0{,}37}} = -2{,}3848$$

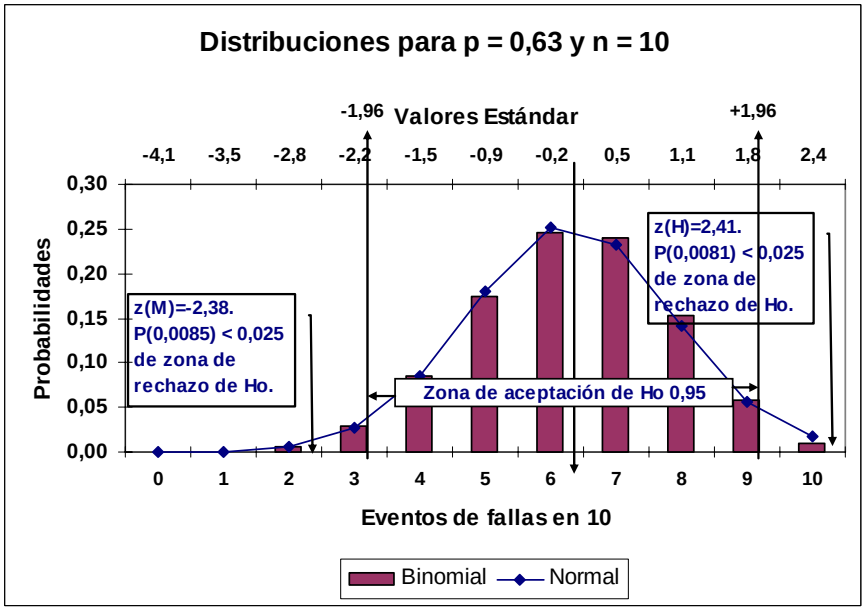
Es importante conservar el signo pues la diferencia entre el valor observado de 55 cursos perdidos observados contra $0{,}63 \times 107 = 64{,}41$ es negativa y la probabilidad acumulativa que determina es de:

$$P(z_c) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-2{,}3848} e^{-\frac{1}{2}\left(\frac{55 - 0{,}63(107)}{\sqrt{107(0{,}63)(0{,}37)}}\right)} dz = 0{,}0085$$

En la cola del lado izquierdo. También indica una razón poderosa para asegurar que la proporción de mujeres que pierden curso es inferior a 63%. Se rechaza la hipótesis nula para las mujeres.

5.21 Una prueba de dos colas.

En ambos casos se ha rechazado la hipótesis nula, sin embargo las implicaciones son diferentes. Haciendo un diagrama de Las Distribuciones Binomial y Normal para 10 individuos. La proporción con que reprueban cursos las mujeres es significativamente inferior a 63% pues el valor *z* es inferior a –1,96, valor estandarizado que delimita la zona de rechazo en la cola inferior. Por otro lado, la proporción de hombres es significativamente superior a 63% pues el valor de *z* cae por arriba del 1,96 que delimita la zona de rechazo de la hipótesis nula en 2,5%. Puesto que puede haber casos que caigan en cualquiera de los extremos se llaman pruebas de dos colas.



5.22 La prueba de χ2 en hombres.

La prueba de Chi-Cuadrada opera sobre todas las clases involucradas en este caso los que pierden cursos y lo que no lo pierden. (Las diferencias se deben al redondeo ver HE)

$$\chi^2_{(2-1)} = \frac{\left(|x_1 - np| - 0{,}5\right)^2}{np} + \frac{\left(|x_2 - nq| - 0{,}5\right)^2}{nq} =$$

$$= \frac{\left(|157 - 0{,}63\times 221| - 0{,}5\right)^2}{0{,}63\times 221} + \frac{\left(|61 - 0{,}37\times 221|\right)}{0{,}37\times 221} = 2{,}1422 + 3{,}6475 = 5{,}7896$$

La probabilidad para el estadístico:

$$F(5{,}7896;\, 2\text{-}1) = Y_0 \int_0^{5{,}7896} (5{,}7896)^{\frac{1}{2}(2-1)} e^{-\frac{1}{2}5{,}7896}\, d\chi = 0{,}0161$$

Como se mencionó en la diapositiva 6

$$z = \sqrt{\chi^2_{(2-1)}} = \sqrt{5{,}7896} = 2{,}4062$$

La probabilidad de z:

$$P(-z_c) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-2{,}4062} e^{-\frac{1}{2}\left(\frac{64 - 0{,}63(221)}{\sqrt{221(0{,}63)(0{,}37)}}\right)} dz = 0{,}00806$$

$$P(+z_c) = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+2{,}4062} e^{-\frac{1}{2}\left(\frac{157 - 0{,}63(221)}{\sqrt{221(0{,}63)(0{,}37)}}\right)} dz = 1 - 0{,}99194 = 0{,}00806$$

Y las probabilidades como en la diapositiva 7.

$$P\left[\chi^2_{(2-1)} = 0{,}0161\right] =$$

$$= P\left[(z = -2{,}4062) = 0{,}00806\right] + P\left[(z = 2{,}4062) = 0{,}00806\right] = 0{,}0161$$

5.23 Prueba de $\chi 2$ en las mujeres.

Se recuerda que las diferencias en probabilidades puede deberse a los redondeos. Sustituyendo valores en la ecuación de χ^2:

$$\chi^2_{(2-1)} = \frac{\left(|x_1 - np| - 0{,}5\right)^2}{np} + \frac{\left(|x_2 - nq| - 0{,}5\right)^2}{nq} =$$

$$= \frac{\left(|55 - 0{,}63 \times 107| - 0{,}5\right)^2}{0{,}63 \times 107} + \frac{\left(|52 - 0{,}37 \times 107|\right)}{0{,}37 \times 107} = 2{,}1043 + 3{,}5829 = 5{,}6872$$

La probabilidad para este valor es de:

$$F(5{,}6872;\, 2\text{ - }1) = Y_0 \int_0^{5{,}6872} (5{,}6872)^{\frac{1}{2}(2-1)} e^{-\frac{1}{2}5{,}6872}\, d\chi = 0{,}0171$$

La raíz cuadrada de χ^2 es el valor de ***z*** para ambas colas y las probabilidades:

$$P(-z_c) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-2{,}3848} e^{-\frac{1}{2}\left(\frac{52 - 0{,}63(107)}{\sqrt{107(0{,}63)(0{,}37)}}\right)} dz = 0{,}0085$$

$$P(+z_c) = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+2{,}3848} e^{-\frac{1}{2}\left(\frac{55 - 0{,}63(107)}{\sqrt{107(0{,}63)(0{,}37)}}\right)} dz = 1 - 0{,}99146 = 0{,}0085$$

La probabilidad de $\chi 2$ es la suma de las probabilidades de las dos colas de *z*.

$$P\left[\chi^2_{(2-1)} = 0{,}0171\right] = P\left[(z = -2{,}3848) = 0{,}0085\right] + P\left[(z = 2{,}3848) = 0{,}0085\right] = 0{,}0171$$

5.24 Recomendación.

Para aplicar la prueba de χ^2 debe tenerse la precaución de valorar si la prueba de hipótesis involucra valores en ambas colas de una *Distribución Normal*, caso en el que la probabilidad de χ^2 considera las probabilidades de la suma de cada una de las colas de la distribución normal. Si la prueba es de una sola cola, entonces, la probabilidad que determina la prueba de χ^2 debe aplicarse a una prueba de ***z*** de una sola cola.

Otro punto a tomar en cuenta es el uso de la corrección de *Yates* o *corrección por continuidad*. Usualmente el nivel de confianza es más pequeño cuando no se corrige, esto implica disminuir la probabilidad de cometer error I pero aumenta la probabilidad de cometer error II, la recomendación es que lo use siempre que los valores esperados no sean muy pequeños, caso en que se disturba la prueba. Es conveniente que valore sus riesgos de cometer errores estadísticos (vea prueba de hipótesis).

5.25 Intervalo de Confianza para varones.

Suponga que en El Tecnológico se inscribieron 5.575 varones y 2.791 mujeres en el semestre. ¿Cuántos alumnos y alumnas que pierden cursos se esperan con un nivel de confianza de 95%?

El intervalo de confianza para individuos está definido por:

$$\Pr\left\{(x = np) - z\sqrt{npq} \geq X \leq (x = np) + z\sqrt{npq}\right\} = 95\%$$

Para hombres con una proporción de cursos perdidos de: 0,71

$$\Pr\left\{4.103 - 1{,}96\sqrt{5.775 \times 0{,}71 \times 0{,}29} \geq X \leq 4.103 + 1{,}96\sqrt{5.775 \times 0{,}71 \times 0{,}29}\right\} = 95\%$$

$$\Pr\left\{4.035 \geq X \leq 4.170\right\} = 95\%$$

$$\Pr\left\{4.035 \geq X \leq 4.170\right\} = 95\%$$

Se espera que entre 4.035 y 4.170 alumnos pierdan algún curso.

5.26 Intervalo de Confianza de proporciones para mujeres.

Ahora se efectúa la estimación usando el intervalo de confianza para proporciones (y porcentajes):

$$\Pr\left\{ p - z\sqrt{\frac{pq}{n}} \geq P \leq p + z\sqrt{\frac{pq}{n}} \right\} = 1 - \alpha$$

La proporción de alumnas que pierden cursos es de:

$$\Pr\left\{ 0{,}514 - 1{,}96\sqrt{\frac{0{,}514 \times 0{,}486}{2.791}} \geq P \leq 0{,}514 + 1{,}96\sqrt{\frac{0{,}514 \times 0{,}486}{2.791}} \right\} = 0{,}95$$

$$\Pr\{0{,}4955 \geq P \leq 0{,}5326\} = 0{,}95$$

Multiplicando Estos porcentajes por el número de alumnas matriculadas se obtiene la estimación solicitada:

$$\Pr\{0{,}4955 \times 2.791 \geq N_M \leq 0{,}5326 \times 2.791\} = 0{,}95$$

$$\Pr\{1.383 \geq N_M \leq 1.486\} = 0{,}95$$

5.27 La aproximación a La Binomial.

Cuando de utiliza *La Distribución Binomial* en la estimación de probabilidades se hace referencia a casos muy concretos, en este ejemplo para una muestra de 10 alumnos:

¿Cuál es la probabilidad de que menos del 20% de alumnos no pierdan cursos?

Primero $n = 0{,}2 \times 10 = 2$, esto significa que no lo pierda ninguno, que lo pierda 1 y que los pierdan 2. Por la primera regla de probabilidades sería:

$$P[E = \{P(x=0) + P(x=1) + P(x=2)\}] =$$

$$= 0{,}000048 + 0{,}000819 + 0{,}006273 = 0{,}007140$$

Aproximadamente $0{,}0071 \times (5.775 + 2.179) = 61$ individuos. Se pude preguntar, la probabilidad de 8 (80%) o más pierdan cursos.

$$P[E = \{P(x=8) + P(x=9) + P(x=10)\}] =$$

$$= 0{,}152876 + 0{,}057845 + 0{,}009849 = 0{,}220571$$

Aproximadamente $0{,}220571 \times (5.775 + 2.179) = 1.889$ alumnos. Pero encontrar por ejemplo el 25% de la población resulta complicado con este método.

5.28 Experimentos con muestras pareadas.

En una universidad, para llevar el curso de diseño de experimentos el alumno debe llevar o haber cumplido un curso avanzado de cálculo. Un grupo de profesores opina que el requisito no es necesario y decidieron efectuar un experimento con los alumnos que habían cumplido el primer curso de cálculo. A 112 pares de alumnos de sexo masculino se les autorizó el curso de estadística, a la mitad de ellos además se les obligó a llevar el segundo curso de cálculo sin enterarlos de la finalidad del cambio en la política curricular. La hipótesis que se va a valorar dice:

La cantidad de alumnos que ganan el curso de estadística es igual para aquellos alumnos que llevan el curso de cálculo avanzado como aquellos que no lo llevan.

5.29 Los resultados de la experiencia.

En cada par de muestras pueden darse los siguientes sucesos: el que recibe calculo pierda estadística (0) y el que no recibe cálculo también pierda (0); el que recibe cálculo pierda (0) el que no lo recibe gane (1); el que recibe cálculo gane (1) el que no lo recibe pierda (0); que ambos ganen estadística (1). El resumen de los eventos se muestra en el cuadro aledaño.

Con Cálculo	Sin Cálculo	Eventos
Pierde	Pierde	10
Pierde	Gana	16
Gana	Pierde	35
Gana	Gana	51
Suma		112

Los sucesos en que los dos alumnos pierden o ganan no proporcionan información útil para valorar la hipótesis.

Por tanto interesa los sucesos en que los grupos cruzados.

5.30 Las clases útiles para la prueba.

En el cuadro las clases que interesan a la hipótesis.

Con Cálculo	Sin Cálculo	Eventos
Pierde	Gana	16
Gana	Pierde	35
Suma		51

El promedio de individuos se obtiene dividiendo el total por 2. Esta cantidad es la esperada, Dicho de otra forma, Si las proporciones de los grupos que reciben calculo y los que no lo reciben fueran parecidos. Efectuando la prueba de χ^2.

$$\chi^2_{(2-1)} = \frac{\left(|16-25{,}5|-0{,}5\right)^2}{25{,}5} + \frac{\left(|35-25{,}5|-0{,}5\right)}{25{,}5} = 6{,}3529$$

Resta obtener la probabilidad que determina el estadístico.

5.31 La conclusión mediante la χ^2.

$$F(6{,}3529;\nu\text{-}1) = Y_0\int_0^{6{,}3529}(6{,}3529)^{\frac{1}{2}(2-1)}e^{-\frac{1}{2}6{,}3529}d\chi = 0{,}0117$$

Una probabilidad de 1,17% de que la cantidad de alumnos que ganan el curso de estadística recibiendo el curso de cálculo y no recibiéndolo sean iguales. En otras palabras, debe rechazarse la hipótesis con un nivel de significación de 1,17%.

Se va a mostrar la aproximación de *La Normal Estándar* mediante el criterio de las variables estandarizadas. Primero debe entenderse que la proporción de individuos en uno y otro grupo debe ser de 0,5 o 50%. Usando esta consideración, las pruebas de ***z*** se ofrecen en la siguiente diapositiva:

5.32 La conclusión usando el criterio de z.

En la comparación se tienen dos valores de ***z***, uno corresponde a la cola inferior y otro a la cola superior.

$$-z = -\frac{|16-0{,}5\times 51|-0{,}5}{\sqrt{51\times 0{,}5\times 0{,}5}} = \text{-}2{,}5205;\ P(-z_c) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\text{-}2{,}5205} e^{-\frac{1}{2}\left(\frac{16-0{,}5(51)}{\sqrt{51(0{,}5)(0{,}5)}}\right)}dz = 0{,}0059$$

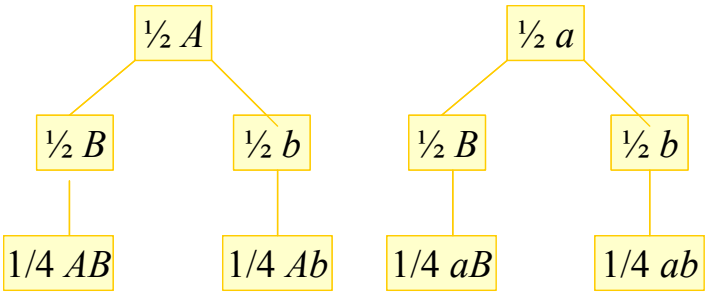
$$z = \frac{|35-0{,}5\times 51|-0{,}5}{\sqrt{51\times 0{,}5\times 0{,}5}} = 2{,}5205 \qquad P(z_c) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+2{,}5205} e^{-\frac{1}{2}\left(\frac{35-0{,}5(51)}{\sqrt{51(0{,}5)(0{,}5)}}\right)}dz = 0{,}0059$$

Un valor de **z** positivo y otro negativo. El primero determina una probabilidad de menos infinito a -*z* de 0,0059. El segundo determina una probabilidad de +*z* hasta más infinito de 0,0059. La suma de ambas es igual a la probabilidad de χ^2 y se puede comprobar que $z^2 = (\pm 2{,}5205)^2 = 6{,}3529 = \chi^2_1$.

5.33 *Comparación de clases.*

El ejemplo típico para este tipo de comparaciones lo constituyen los experimentos de la genética mendeliana clásica. Por ejemplo *La Ley de la Transmisión Independiente*. Mendel investigó la descendencia de híbridos respecto a varios caracteres alternativos. Polinizó una planta femenina de semillas lisas y amarillas (*AA* y *BB* ambos caracteres dominantes) con polen de una planta con guisantes rugosos y verdes (*aa* y *bb*). La primera generación híbrida formó semillas lisas y amarillas. Estas semillas los cruzó entre sí y la segunda generación híbrida produjo semillas de 4 clases en las cantidades siguientes: 315 lisas y amarillas, 101 rugosas y amarillas, 109 lisas y verdes y 32 rugosas y verdes. Según Mendel, esta proporción se debía a la transmisión independiente de las unidades hereditarias que rigen estos caracteres. Es decir, en la formación de los gametos del híbrido ½ contienen *A* y ½ *a*; similarmente, ½ contienen *B* y ½ *b*. Cada pareja se transmite de modo independiente de la otra de manera que el híbrido produce 4 clases de gametos aproximadamente en igual número.

5.34 *Las frecuencias esperadas por Mendel.*



Las cuatro clases de gametos se unen al azar y forman nueve tipos genéticos en la segunda generación. Con dominancia se forman solo cuatro tipos distintos. Estos se presentan en las proporciones del cuadro.

La prueba para este tipo de experimentos consiste en comparar las frecuencias esperadas con las frecuencias observadas.

Y la hipótesis: *Ho*; $fe_i \sim (9{:}\ 3\ {:}3\ {:}1)$.

Tipos Genéticos	Frecuencia	Fenotipo	Relación
1/16 AABB + 2/16 AABb + 2/16 AaBB + 4/16 AaBb	0,5625	Lisa-Amarilla	9
1/16 Aabb + 2/16 Aabb	0,1875	Lisa-Verde	3
1/16 aaBB + 2/16 aaBb	0,1875	Rugosa-Amarilla	3
1/16 aabb	0,0625	Rugosa-Verde	1

5.35 *La prueba de χ^2 para comparar clases.*

La χ^2 se calcula para cada una de las clases:

$$\chi^2_{(4-1)} = \frac{\left(|315 - 312{,}8| - 0{,}5\right)^2}{312{,}8} + \frac{\left(|108 - 104{,}3| - 0{,}5\right)^2}{104{,}3} + \frac{\left(|101 - 104{,}3| - 0{,}5\right)^2}{104{,}3} +$$

$$+ \frac{\left(|32 - 34{,}8| - 0{,}5\right)^2}{34{,}8} = 0{,}3293$$

$$F(0{,}3293; 4-1) = Y_0 \int_0^{0{,}3293} (,3293)^{\frac{1}{2}(4-1)} e^{-\frac{1}{2}0{,}3293}\, d\chi = 0{,}9544$$

La probabilidad indica que la similitud entre las frecuencias observadas y esperadas es de 95,44%. Por tanto, se acepta la hipótesis de que las frecuencias del cruzamiento siguen la relación 9: 3 :3 : 1.

5.36 Tablas de contingencia 2 ×2.

Este tipo de problemas ocurre frecuentemente en proyectos de investigación. En muchos experimentos controlados se pretende comparar dos procedimientos o tratamientos basándose en resultados obtenidos de muestras independientes sin que el investigador haya efectuado apareamientos de unidades. La comparación de proporciones de diferentes grupos también es muy común en estudios que no son tan estrictos como los experimentales, por ejemplo, en manufacturas se quiere saber la proporción de artículos defectuosos provenientes de dos suplidoras del artículo; o en ingeniería de seguridad automotriz, comparando las proporciones de daños corporales sufridos en accidentes automovilísticos por los pasajeros que utilizan cinturón de seguridad y los que no lo usaron.

5.37 Ejemplo del oculista.

Una clínica muy famosa de Barcelona España quiere recomendar una técnica, más costosa, para operar cataratas que en apariencia reduce la opacidad del ojo que algunos pacientes presentan después de la operación. Encarga a un profesional que está haciendo su postgrado que investigue, las ventajas y desventajas de esta técnica. Con la información que posee la clínica ha encontrado 1.325 expedientes útiles.

Técnicas	Presentó Opacidad		Suma Técnica
	No	Si	
Nueva Técnica	395	42	437
Otras Técnicas	752	136	888
Suma Problema	1.147	178	1.325

Bajo la hipótesis nula cabría esperar que la proporción de individuos que presentan el problema fuera igual. Adviértase que *Ho* es equivalente a afirmar que la no presentación de opacidad es *independiente* de la técnica, es decir, las clasificaciones son independientes.

5.38 Las proporciones esperadas.

El primer paso para la prueba es encontrar las frecuencias esperadas de cada una de las cuatro clases. Para esto considérese que, sí los resultados no dependen de la alguna condición, las proporciones marginales contienen toda la información. Se habla de proporciones marginales a las que se obtienen de dividir los totales para cada clase por el total general. Esto es:

La proporción marginal para la hilera 1 es:

$$p_{1\cdot} = \frac{n_{1\cdot}}{n_{\cdot\cdot}} = \frac{437}{1.325} = 0{,}3298$$

La proporción marginal para la hilera 2 es:

$$p_{2\cdot} = \frac{n_{2\cdot}}{n_{\cdot\cdot}} = \frac{888}{1.325} = 0{,}6702$$

Técnicas	Presentó Opacidad		Suma Técnica
	No	Si	
Nueva Técnica			0,3298
Otras Técnicas			0,6702
Suma Problema	0,8657	0,1343	1,0000

La proporción marginal para la columna 1 es: $p_{\cdot 1} = \frac{n_{\cdot 1}}{n_{\cdot\cdot}} = \frac{1.147}{1.325} = 0{,}8657$

La proporción marginal para la columna 2 es: $p_{\cdot 2} = \frac{n_{\cdot 2}}{n_{\cdot\cdot}} = \frac{178}{1.325} = 0{,}1343$

Multiplicando las proporciones marginales correspondientes se obtienen las proporciones de cada uno de los cuatro eventos;

Las proporciones esperadas para la clase: (No presento opacidad) y (Nueva Técnica) es: $p_{11} = p_{1\cdot} \times p_{\cdot 1} = 0{,}3298 \times 0{,}8657 = 0{,}2855$. ***La proporción esperada para la clase (Sí presentó opacidad) y (Mueva Técnica) es:*** $p_{12} = p_{1\cdot} \times p_{\cdot 2} = 0{,}3298 \times 0{,}1343 = 0{,}0443$. ***La proporción esperada para la clase (No presentó opacidad) y (Otras Técnicas) es:*** $p_{2\cdot} \times p_{\cdot 1} = 0{,}6702 \times 0{,}8657 = 0{,}5802$. ***Y la proporción esperada para la clase (Si presentó opacidad) y (Otras Técnicas) es:*** $p_{2\cdot} \times p_{\cdot 2} = 0{,}6702 \times 0{,}1343 = 0{,}0900$.

Técnicas	Presentó Opacidad		Suma
	No	Si	Técnica
Nueva Técnica	0,2855	0,0443	0,3298
Otras Técnicas	0,5802	0,0900	0,6702
Suma Problema	0,8657	0,1343	1,0000

5.39 *La ecuación para obtener las frecuencias esperadas.*

Multiplicando estas frecuencias por el gran total se obtienen los valores esperados.

$$fe_i = (p_{ij})n..$$

Para calcular las frecuencias esperadas en más simple utilizar los totales marginales divididos por el gran total:

Los eventos esperados de cada clase se obtienen multiplicando los eventos marginales dividiéndolos por el total de eventos. De esta manera, el número de individuos esperados en la clase: (No presento opacidad) y (Nueva Técnica) que se identificará como n_{11} ***es:***

Técnicas	Presentó Opacidad		Suma
	No	Si	Técnica
Nueva Técnica	378,3	58,7	437
Otras Técnicas	768,7	119,3	888
Suma Problema	1.147	178	1.325

$$n_{11} = \frac{n_{1\cdot} \times n_{\cdot 1}}{n..} = \frac{437 \times 1.147}{1.325} = 378{,}3 \text{ ;}$$

Este número proviene del producto:

$$p_{11}n = p_{1\cdot} \times p_{\cdot 1} \times n = \left(\frac{437}{1.325} \times \frac{1.147}{1.325}\right)1.325 =$$

$$= (0{,}3298 \times 0{,}8657)1.325 = 0{,}2855 \times 1.325$$

De la manera corta para la clase (Sí presentó opacidad) y (Mueva Técnica) = n_{12}***:***

$$n_{12} = \frac{n_{1\cdot} \times n_{\cdot 2}}{n..} = \frac{437 \times 178}{1.325} = 58{,}7$$

Para la clase (No presentó opacidad) y (Otras Técnicas) = n_{21}***:***

$$n_{21} = \frac{n_{2\cdot} \times n_{1\cdot}}{n..} = \frac{888 \times 1.147}{1.325} = 768{,}7$$

Para la clase (Si presentó opacidad) y (Otras Técnicas) = n_{22}***:***

$$n_{22} = \frac{n_{2\cdot} \times n_{\cdot 2}}{n..} = \frac{888 \times 178}{1.325} = 119{,}3$$

Las cantidades $n_{1\cdot}$ es el total de la hilera 1; $n_{2\cdot}$ es el total de la hilera 2; $n_{\cdot 1}$ es el total de la columna 1; $n_{\cdot 2}$ es el total de la columna 2 y $n..$ es el total general.

5.40 *La χ^2 en tablas de contingencia.*

La ecuación de la χ^2 ya es notoria:

$$\chi^2_{(h-1)(c-1)} = \sum_{i=1}^{c}\sum_{j=1}^{h}\frac{\left(\left|fo_{ij} - n..p_{ij}\right| - 0{,}5\right)^2}{n..p_{ij}} = \sum_{i=1}^{c}\sum_{j=1}^{h}\frac{\left(\left|fo_{ij} - fe_{ij}\right| - 0{,}5\right)^2}{fe_{ij}}$$

La nueva técnica quirúrgica requiere de equipo costoso y un método de cirugía que alarga el tiempo de operación, por esto, es necesario establecer un criterio de al menos 99% de confianza para la prueba. Por otro lado, debe estarse muy seguro para recomendar la técnica como de uso generalizado a los médicos.

En este caso, hay suficientes expedientes (información) para asegurar una probabilidad de cometer error II baja y se fija de una vez la probabilidad del cometer error I en 1%.

5.41 La prueba de χ^2.

La prueba se efectúa según lo acostumbrado. Se le recuerda que la corrección de *Yates* es opcional pero indispensable en este caso.

$$\chi^2_{(2-1)(2-1)} = \frac{\left(\left|395 - 378{,}3\right| - 0{,}5\right)^2}{378{,}3} + \frac{\left(\left|42 - 58{,}7\right| - 0{,}5\right)}{58{,}7} + \frac{\left(\left|752 - 768{,}7\right| - 0{,}5\right)^2}{768{,}7} +$$
$$+ \frac{\left(\left|136 - 119{,}3\right| - 0{,}5\right)^2}{119{,}3} = 7{,}7116$$

La probabilidad de que las diferencias entre las proporciones de opalescencia con la técnica nueva ($p_A = 0{,}0443$) y las otras ($p_B = 0{,}0900$) se deba a efectos fortuitos es de:

$$F\left(7{,}7116;(2-1)(2-1)\right) = Y_0\int_0^{7{,}7116}(7{,}7116)^{\frac{1}{2}((2-1)(2-1))}e^{-\frac{1}{2}7{,}7116}d\chi = 0{,}0055$$

Menos del 1%. En otras palabras, se rechaza la hipótesis:

Ho; $p_{NT} = p_{OT}$

La reducción en la opalescencia en lo ojos después de la operación de cataratas *depende* de la nueva técnica quirúrgica. Por esto, a esta prueba también se le llama *Prueba de Independencia de Eventos.*

5.42 Tablas de contingencia h × c.

La prueba de χ^2 para tablas de contingencia 2 × 2 se extiende fácilmente a grupos de *h* hileras y *c* columnas. En ecuación:

$$\chi^2_{(h-1)(c-1)} = \sum_{i=1}^{c}\sum_{j=1}^{h}\frac{\left(\left|fo_{ij} - n..p_{ij}\right| - 0{,}5\right)^2}{n..p_{ij}} = \sum_{i=1}^{c}\sum_{j=1}^{h}\frac{\left(\left|fo_{ij} - fe_{ij}\right| - 0{,}5\right)^2}{fe_{ij}}; 1 = 1,2,...,c;\ j = 1,2,...,h$$

Es la suma de cada una de las χ^2 (de cada una de las celdas de la tabla) valuada en la distribución de χ^2 con $(h - 1)(c - 1)$ grados de libertad mediante la función de distribución acumulativa de la HE o de las tablas estadísticas de ésta función.

El corrector por continuidad no es obligado, sin embargo, conveniente, más, si el número de grados de libertad es bajo o la variable es estrictamente discreta.

5.43 Ejemplo de Tabla de Contingencia h × c.

Un ejemplo clásico para la prueba de χ^2 en tablas de contingencia $h \times c$ se refiere a una encuesta para determinar si la edad de los conductores a partir de los 18 años, edad en que en Costa Rica se les otorga licencia de conducir, tiene efecto sobre el número de accidentes de automóviles.

Como trabajo de tema a los estudiantes del curso de Estadística Básica se les encargó probar la hipótesis:

El número de accidentes es independiente de la edad del conductor con un nivel de confianza de 0,05.

Los datos provienen de 4.527 registros del historial de conductores con licencias de tipo B1 (vehículos de carga ligera) del Ministerio de Transportes, Departamento de Tránsito.

5.44 Los datos y los totales marginales.

Los datos de accidentes por edad de los conductores se muestran en el siguiente cuadro.

Para estimar los valores esperados np_{ij} no es necesario calcular las proporciones marginales pues, tomando los mismos totales por hilera y columna, y aplicando la siguiente fórmula se obtienen las frecuencias esperadas o individuos esperados. Se ejemplifica para la celda de la hilera 2 y columna 3.

Número de Accidentes	EDAD DE LOS CONDUCTORES						Sumas Nº Accidentes
	18-20	21-30	31-40	41-50	51-60	61-70	
0	**295**	**698**	846	826	750	639	4.054
1	23	75	60	51	66	46	321
2	7	31	27	22	15	17	119
más de 2	1	7	10	6	3	**6**	33
S. Edades	326	811	943	905	834	708	4.527

$$fe_{23} = n..p_{23} = \frac{n_{2.} \times n_{.3}}{n..} = \frac{321 \times 943}{4.527} = 66{,}9$$

5.45 Las Frecuencias esperadas.

De la misma manera se calculan las frecuencias esperadas de todas las celdas. Los resultados en el cuadro.

El tercer paso de esta sencilla prueba se consigue calculando las Chi-Cuadradas para cada celda y sumándolas por hileras y columnas. Lo usual es sumarlas hacia alguno de los márgenes y volverlas a sumar:

Frecuencias Esperadas

Número de Accidentes	EDAD DE LOS CONDUCTORES						Sumas Nº Accidentes
	18-20	21-30	31-40	41-50	51-60	61-70	
0	**291,9**	**726,3**	844,5	810,4	746,9	634,0	4.054,0
1	23,1	57,5	**66,9**	64,2	59,1	50,2	321,0
2	8,6	21,3	24,8	23,8	21,9	18,6	119,0
más de 2	2,4	5,9	6,9	6,6	6,1	**5,2**	33,0
S. Edades	326,0	811,0	943,0	905,0	834,0	708,0	4.527,0

$$\chi^2_{(4-1)(6-1)} = \left[\frac{\left(|295 - 291{,}9| - 0{,}5\right)^2}{291{,}9} + \ldots + \frac{\left(|639 - 634{,}0| - 0{,}5\right)^2}{634{,}0} = 1{,}4058 \right] + \ldots +$$

$$+ \left[\frac{\left(|1 - 2{,}4| - 0{,}5\right)}{2{,}4} + \ldots + \frac{\left(|6 - 5{,}2|\right)}{5{,}2} = 2{,}5030 \right] = 19{,}2270$$

5.46 La valuación del estadístico de χ^2.

Finalmente se suman las Chi-Cuadradas del margen para obtener un estadístico general con el valor de 19,2270.

El resultado de los cálculos en el cuadro.

$$F_{[19,2270;(4-1)(6-1)]} = Y_0 \int_0^{19,2270} (19,2270)^{\frac{1}{2}(4-1)(6-1)} e^{-\frac{1}{2}19,2270} d\chi =$$

$$= \text{DISTR.CHI}(19,2270; 3\times 5) = 0,2036$$

Chi-cuadradas

Número de Accidentes	EDAD DE LOS CONDUCTORES						Sumas Nº Accidentes
	18-20	21-30	31-40	41-50	51-60	61-70	
0	**0,0225**	**1,0613**	0,0013	0,2798	0,0093	0,0316	1,4058
1	0,0064	5,0218	0,6061	2,5022	0,6846	0,2731	9,0942
2	0,1335	3,9543	0,1182	0,0699	1,8819	0,0663	6,2240
más de 2	0,3232	0,0585	1,0031	0,0014	1,0945	**0,0223**	2,5030
	0,4855	10,0959	1,7286	2,8533	3,6703	0,3933	19,2270
						Probabilidad	**0,2036**

5.47 Resultado de la prueba.

La prueba indica que no hay evidencias estadísticas para rechazar que:

El número de accidentes es independiente de la edad del conductor .

Recuerde que para considerar la probabilidad significativa debería ser menor o igual a 0,05. La indicada por la función de la HE es de 0,2036 o 20,36% valor que orienta a considerar que las diferencias se deben al azar. En otras palabras, no hay evidencias para asegurar que el número de accidentes de tránsito depende de la edad del conductor.

Es Importante hacer ver al estudiante que la prueba no dice nada sobre las diferencias entre edades. Considerando las dos clases: entre número de accidentes las diferencias parecen evidentes pues la mayor proporción corresponde a conductores que no han tenido accidentes; no así entre edades.

5.48 Pruebas de porcentajes.

	EDAD DE LOS CONDUCTORES						
Nº Accidentes	18-20	21-30	31-40	41-50	51-60	61-70	Suma Frec.
0	0	0	0	0	0	0	0
1	23	75	60	51	66	46	321
2	14	62	54	44	30	34	238
3	3	21	30	18	9	18	99
S. Accidentes	40	158	144	113	105	98	658
S. Registros	326	811	943	905	834	708	4.527
Proporción	0,1227	0,1948	0,1527	0,1249	0,1259	0,1384	0,1454

Hay varias alternativas para analizar el comportamiento de las clases o totales marginales o aun dentro de un grupo. Considérese la proporción de accidentes por edad del conductor, esta es la única prueba factible pues la inclusión de la clase 18 a 20 con 3 años de observaciones modifica el espacio muestral de las otras clases que es de 10 años y además, la última es abierta. En este caso, el análisis sobre las proporciones para cada edad es el indicado.

La pregunta que surge es:

¿La proporción de accidentes por edad es igual?

5.49 Recordando la aproximación con la Normal.

Recuérdese que la cantidad:

$$z_c = \frac{|x - np| - \frac{1}{2}}{\sqrt{npq}}$$

Es una variable con distribución Normal Estándar. Dividiendo por n:

$$z_c = \frac{\frac{1}{n}\left(|x - np| - \frac{1}{2}\right)}{\frac{1}{\sqrt{n}}\frac{1}{\sqrt{n}}\sqrt{npq}} = \frac{|p - P| - \frac{1}{2n}}{\sqrt{\frac{pq}{n}}}$$

5.50 La prueba estadística de DMS (Diferencia Mínima Significativa).

Si se contrastan los valores estandarizados *z* de dos proporciones se obtendría una nueva variable estandarizada:

$$z_d = z_1 - z_2 = \frac{|p_1 - p_2|}{\sqrt{\overline{pq}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Que puede escribirse como:

$$z_d\sqrt{\overline{pq}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = |p_1 - p_2|$$

Si se elige *z* para un nivel de significación específico α el estadístico se convierte en una prueba llamada *Diferencia Mínima Significativa*:

$$z_\alpha\sqrt{\overline{pq}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = DMS$$

5.51 La regla de decisión en el contraste.

La regla de decisión es:

Si el valor absoluto de la diferencia entre proporciones es mayor o igual que *DMS* debe rechazarse la hipótesis nula:

Ho; $P_1 = P_2$

La forma alternativa a la prueba es ubicar el punto que define el valor $\mathbf{z}_d$ en La Distribución Normal Estándar:

$$P(z_1 - z_2) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z_1 - z_2} e^{-\frac{1}{2}\left(\frac{p_1 - p_2}{\sqrt{\overline{pq}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}\right)} dz$$

Y decidir según el valor que indique la probabilidad obtenida, si se ubica dentro del intervalo de confianza para la prueba se aceptará la hipótesis, de otro modo se rechazará.

Para el ejemplo se considerará un nivel de significación de 0,05.

5.52 Pruebas de contrastes alternados.

La *Diferencia Mínima Significativa* para la prueba varía debido a que varía el número de observaciones por contraste, por ejemplo contrastando la proporción del grupo 18-20 con la proporción del grupo 21-30 se obtiene:

$$DMS_{(18-20\text{ vs }21\text{-}30)} = 1{,}96\sqrt{0{,}1242\left(\frac{1}{326}+\frac{1}{811}\right)} = 0{,}04530$$

Puesto que el valor absoluto de la diferencia entre proporciones: |0,1227 - 0,1948| = 0,0721 es mayor que 0,0453 deber rechazarse la hipótesis de que las proporciones son iguales. Usando la prueba alternativa

$$z_d = z_1 - z_2 = \frac{0{,}1227 - 0{,}1948}{\sqrt{0{,}1242\left(\frac{1}{326}+\frac{1}{811}\right)}} = -3{,}1204$$

$$P(z_1 - z_2) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{-3{,}1204} e^{-\frac{1}{2}\left(\frac{0{,}1227-0{,}1948}{0{,}0231}\right)} dz = 0{,}0009$$

Este valor debe ser inferior a 0,025 para declarar diferencias significativas y menor a 0,01 para que sea altamente significativa.

La fórmula de las desviaciones de las diferencias entre dos proporciones considera una varianza común a la población estudiada. Puesto que denotada como

$$S = \overline{p}\,\overline{q}$$

En donde las proporciones se obtiene del total de accidente, así:

$$S_G^2 = \frac{n_A}{n..}\left(1-\frac{n_A}{n..}\right) = \frac{658}{4.527}\left(1-\frac{658}{4.527}\right) = 0{,}1454\times 0{,}8546 = 0{,}1242$$

5.53 Todos los contrastes.

Los conductores entre 21 y 30 años sufren más accidentes que los conductores de otras edades. Los resultados no son contradictorios a la prueba anterior pues en este caso no se habla de número de accidentes que si toma en cuenta la prueba en tabla de contingencia. Esto es, no hay interacción entre edad y número de accidentes. En todas las clasificaciones de accidentes la diferencia se manifiesta.

	Proporciones			Observaciones					
Contraste	1	2	Diferencia	n1	n2	DMS	zd	Probabilidad	Resultado
18-20 vs 21-30	0,1227	0,1948	-0,0721	326	811	0,04530	-3,1204	0,0009	**
18-20 vs 31-40	0,1227	0,1527	-0,0300	326	943	0,04438	-1,3250	0,0926	ns
18-20 vs 41-50	0,1227	0,1249	-0,0022	326	905	0,04462	-0,0950	0,4622	ns
18-20 vs 51-60	0,1227	0,1259	-0,0032	326	834	0,04512	-0,1390	0,4447	ns
18-20 vs 61-70	0,1227	0,1384	-0,0157	326	708	0,04624	-0,6663	0,2526	ns
21-30 vs 31-40	0,1948	0,1527	0,0421	811	943	0,03308	2,4952	0,0063	**
21-30 vs 41-50	0,1948	0,1249	0,0700	811	905	0,03340	4,1051	0,0000	**
21-30 vs 51-60	0,1948	0,1259	0,0689	811	834	0,03407	3,9652	0,0000	**
21-30 vs 61-70	0,1948	0,1384	0,0564	811	708	0,03553	3,1114	0,0009	**
31-40 vs 41-50	0,1527	0,1249	0,0278	943	905	0,03215	1,6976	0,0448	ns
31-40 vs 51-60	0,1527	0,1259	0,0268	943	834	0,03284	1,6000	0,0548	ns
31-40 vs 61-70	0,1527	0,1384	0,0143	943	708	0,03435	0,8151	0,2075	ns
41-50 vs 51-60	0,1249	0,1259	-0,0010	905	834	0,03316	-0,0613	0,4756	ns
41-50 vs 61-70	0,1249	0,1384	-0,0136	905	708	0,03466	-0,7666	0,2217	ns
51-60 vs 61-70	0,1259	0,1384	-0,0125	834	708	0,03530	-0,6951	0,2435	ns

5.54 Resumen.

Se inició con el estudio la distribución χ^2 (Chi-cuadrada) que permite relacionar eventos o número de individuos, una variable eminentemente discreta con *La Distribución Normal Estándar* para variables continuas.

Después se entró a observar como hay paralelismo entre *La Distribución Binomial* que emula las distribuciones de cualidades con *La Distribución Normal Estándar* que emula características cuantitativas, encontrando que la aproximación de la normal es convincente.

Si hay pocas observaciones es conveniente hacer una corrección por continuidad de 0,5 unidades o medio intervalo de clase.

5.55 Conclusión.

La distribución de variables cualitativas y más específicamente en su expresión racional como proporciones o porcentajes puede ser modulada, considerando unas pocas restricciones fáciles de cumplir mediante las herramientas que proporciona *La Distribución Normal Estándar.*

No debe olvidarse, que para ciertas condiciones de experimentación, *La Distribución Binomial* es la mejor opción. Sobre todo, cuando el tema experimentado obliga a manejar tamaños de muestra muy reducidos, sea por el costo de la experimentación o porque implica la destrucción de material experimental.

En el desarrollo del tema se abrieron muchas alternativas para el experimentador.

REFERENCIAS SELECTAS.

1. Hillier Frederick S., y Lieberman Gerard j., *Introducción a la Investigación de Operaciones*. Capítulo 19. Segunda edición en español traducida de la cuarta edición en ingles. McGraw-Hill Interamericana de México, S. A. De C. V., 1990.
2. Miller Irwin, Freund John E., Johnson Richard A: *Probabilidad y Estadística para Ingenieros*. Capítulo 11. Traducido de la cuarta edición en ingles; Prentice-Hall Hispanoamericana, S. A. 1992.
3. Murray R. Spiegel: *Serie de compendios Schaum, Teoría y Problemas de Estadística.* Capítulos 7. Primera edición en español, traducido de la primera edición en ingles; Libros McGraw-Hill de México, S. A. De C. V., 1973.
4. Ostle Bernard: *Estadística Aplicada*. Capítulos 3 y 7. Primera edición en español traducida de la primera edición en ingles. Editorial Limusa, S. A., 1977.
5. Snedecor George W., y Cochran William G: *Statistical Methods*. Capítulo 8 y 9. Sexta edición; The Iowa State University.
6. Steel Robert G. D., Torrie James H: *Principles and Procedures of Statistics*. Capítulo 4. Primera edición; McGraw-Hill Book Company, Inc, 1960.

6 La Distribución Poisson.

6.1 Los archivos para esta sección son:

E06_DPoisson_P01.pps;
E06_DPoisson_W01.doc;
E06_DPoisson_X01.xls;

6.2 Menú.

6.3 Introducción.

Se ha observado que la distribución binomial se parece mucho a una normal cuando n crece para algún valor fijo de p. Además, para que se tenga una buena aproximación a la normal, la magnitud de n depende del valor de p, si éste está alrededor ½ se requiere una pequeña muestra (n), cuando p es mayor o menor a ½, la cantidad de elementos estudiados debe aumentarse, una regla general muy conservadora indica:

La aproximación mediante *La Distribución Normal Estándar* será adecuada si la media $\mu = np$ es mayor a 15 individuos.

En muchas aplicaciones se estudian eventos que ocurren con poca frecuencia de manera que *n* tiene que ser muy grande para encontrar unos pocos elementos con el atributo que se estudia, son poblaciones en que el promedio *np* es menor de 15 independientemente de la cantidad de observaciones, esto es, la proporción *p* es muy baja.

6.4 El descubridor.

Para estos casos, una aproximación diferente fue propuesta por *S. D. Poisson* (París, 1837). Este matemático trabajó en el límite de la distribución binomial, esto es:

Cuando al mismo tiempo *n* tiende a infinito y *p* tiende a cero, de manera que el promedio $\mu = np$ se hace constante.

La expresión binomial para tales sucesos tiende a la forma simplificada,

$$P(x) = \frac{\mu^x}{x!} e^{-\mu}, \; x = 0, 1, 2, \ldots.$$

En donde $e = 2{,}71828...$, es la base de los logaritmos neperianos. La fórmula se usa para calcular las probabilidades de la distribución denominada con el nombre del autor, Poisson cuyos términos iniciales son:

$$P(0) = e^{-\mu}: \; P(1) = \mu e^{-\mu}: \; P(2) = \frac{\mu^2}{2} e^{-\mu}: \; P(3) = \frac{\mu^3}{(2)(3)} e^{-\mu}$$

Siméon Denis Poisson. Nació el 21 de de junio de 1781 en Pithiviers, Francia y murió el 25 de abril de 1840 en Sceaux (cerca de París), Francia. Los trabajos más importantes eran series de e integrales definidos y sus avances en las series de Fourier. Este trabajo dio los fundamentos para trabajos más modernos en esta área por Dirichlet y Riemann.

http://www-history.mcs.st-and.ac.uk/~history/Mathematicians/Poisson.html

6.5 Proceso Poisson.

En general, un proceso aleatorio es un procedimiento físico controlado completamente o en partes por algún tipo de mecanismo aleatorio. Puede ser una sucesión de lanzamientos de dados, mediciones de la calidad de productos que salen de una línea de producción, las vibraciones de las alas del Concorde, el ruido de una señal de radio, el movimiento de partículas en una solución o muchos otros fenómenos.

Lo que caracteriza a tales procesos, es su dependencia del tiempo. El hecho de que ciertos eventos suceden o no, a intervalos regulares o en un intervalo continuo de tiempo.

Situaciones como: la ocurrencia de imperfecciones en un rollo de casimir, la medición de la radiación de un contador Geiger, el número de artículos defectuosos de una línea de producción, las entradas de correos electrónicos en un computador servidor.

El modelo matemático que aproxima la distribución de estos datos es la *Distribución Poisson.*

6.6 La Probabilidad de x éxitos en tiempo T.

Divídase el intervalo en *n* partes iguales de longitud Δt, de forma tal que:

$$T = n\Delta t$$

Además debe suponerse que:

- La probabilidad de un éxito durante un intervalo de tiempo muy pequeño esta dada por $\mu\Delta t$ donde μ es el promedio de éxitos por unidad de tiempo;

- La probabilidad de más de un éxito durante cada uno de los pequeños intervalos de tiempo es despreciable;
- La probabilidad de un éxito durante uno de tales intervalos no depende de lo que sucedió antes o de lo que sucederá después;

Suposiciones que satisfacen las condiciones fundamentales de la *Distribución Binomial*.

6.7 Problema 1. De llegadas.

Una empresa que ofrece el servicio de conexiones de computadoras particulares a la Red Mundial De Computadoras (INTERNET) tiene muchos problemas con los Virus Cibernéticos.

El esfuerzo es de tal magnitud que lo sobreasa y decide contratar los servicios de una empresa dedicada a la "Cacería de nuevos Virus Cibernéticos"

El primer paso de esta empresa fue determinar el número de nuevas agresiones por día.

Después de 10 días de monitoreo decidió iniciar el estudio.

Los programadores y analistas de esta empresa dedicada a la "Cacería de Virus", como sistema, suelen hacerse preguntas que involucran probabilidades, para enfrentar el problema al tenor de *¿cuáles son las probabilidades*?

a) De que entren cuatro nuevos virus por día.

b) Que entren 10 virus en dos días consecutivos.

6.8 Obteniendo el promedio.

En la HE se presentan los resultados de los 10 primeros días de monitoreo en la entrada de virus. El primer paso es encontrar el promedio mediante un cuadro de frecuencias. El evento mínimo es de 2 y el máximo es de 10 nuevos virus en 24 horas. La Distribución de Frecuencias se presenta a continuación:

Evento x	Frecuencia Observada	Frecuencia por evento
2	1	2
3	0	0
4	2	8
5	1	5
6	1	6
7	3	21
8	0	0
9	0	0
10	2	20
Suma de Días		10
Suma Total de Virus		62
Promedio		6,2

En donde el promedio se obtiene mediante,

$$\bar{x} = \frac{\sum_{i=2}^{10} f_i x_i}{\sum_{i=2}^{10} f_i} = \frac{1\times 2 + 0\times 3 + ... + 2\times 10}{1 + 0 + ... + 2} = \frac{62}{10} = 6{,}2$$

6.9 Las respuestas.

En la primera pregunta el evento que se quiere explorar se refiere a la entrada de exactamente 4 virus en un día cualquiera. La respuesta se consigue aplicando la ecuación Poisson para $x = 4$ y la media 6,2. La probabilidad es de 12,49%.

$$P(x=4) = \frac{6{,}2^4 e^{-6{,}4}}{4!} = 0{,}1249$$

En la segunda el número de observaciones se duplica. Por una de las propiedades de la distribución Poisson, la proporción se mantiene dando una media de 12,4 sucesos en 48 horas. La probabilidad de que entren 10 virus se estima en 9,75%.

$$P(x=10) = \frac{12{,}4^{10} e^{-12{,}4}}{10!} = 0{,}0975$$

Se puede comprobar que la proporción de entradas de virus por día se obtiene promediando las $24 \times 10 = 240$ observaciones obteniendo 12,4 virus en 48 horas.

$$p = \frac{\sum_{i=1}^{240} x_i}{240} = 0{,}2583$$

$$\bar{x}_{48} = np = 48 \times 0{,}2583 = 12{,}4$$

6.10 Problema 2. Fallas.

Una empresa dedicada a la fabricación de envases de hojas de lata ha comprado una máquina que suelda las latas con estaño usando un proceso electrolítico.

La empresa vendedora asegura 0,02 imperfecciones por hora y la capacidad de fabricar 1.000 envases de galón en la misma cantidad de tiempo.

El instructivo de la máquina muestra las respuestas a las siguientes preguntas. Las probabilidades de encontrar imperfecciones:

- Una imperfección en 3 horas;
- Al menos dos imperfecciones en 5 horas;
- Cuando más una imperfección en 15 horas.

El departamento de control de la empresa toma 10 muestras tres veces al día. Los resultados de 100 muestras se presentan en la HE. Las respuestas a las preguntas con estos datos son:

6.11 Promedio de fallas en la soldadura.

En los procesos industriales mecanizados la cantidad de producto fabricado es muy regular y podría usarse como variable *x* la cantidad de envases en lugar del tiempo y decir 3.000 envases, 5.000 envases o la producción de 15.000 envases, aunque sea preferible usar el tiempo como unidad.

La tabla de frecuencias, promedios y proporciones es:

Evento x	Frecuencia Observada	Frecuencia por Evento
0	78	0
1	21	21
2	1	2
3	0	0
Suma de Días		100
Suma Total Inperfecciones		23
Promedio		0,23
Tamaño de muestra		10
Proporción		0,023

$$\bar{x} = \frac{\sum_{i=0}^{3} f_i x_i}{\sum_{i=0}^{3} f_i} = \frac{78(0) + 21(1) + 1(2) + 0(3)}{78 + 21 + 1 + 0 = 100} = \frac{23}{100} = 0{,}23$$

$$p = \frac{\bar{x}}{n} = \frac{0{,}23}{10} = 0{,}023$$

Estadísticos obtenidos del ejercicio de la empresa.

6.12 Las respuestas del problema 2.

a) Si la proporción de fallas es de 0,023 por hora, en 3 horas se espera un promedio de:

$$\lambda_3 = n \times p = 3 \times 0{,}023 = 0{,}069$$

Se quiere saber la probabilidad de obtener una falla en ese tiempo. Para esto se aplica la probabilidad Poisson:

$$P(x=1)=\frac{\lambda^1 e^{-\lambda}}{1!}=\frac{0{,}069^1 e^{-0{,}069}}{1}=$$

$$=\text{POISSON}(1;0{,}069;0)=0{,}0644$$

b) Al menos dos imperfecciones en 5 horas. Primero se obtiene el promedio para 5 horas:

$$\lambda_5 = n \times p = 5 \times 0{,}023 = 0{,}115$$

Al solicitar la probabilidad de 2 o más soldaduras malas es más fácil obtenerla por probabilidades complementarias:

$$P(x \geq 2)=1-\left(\frac{0{,}115^0 e^{-0{,}115}}{0!}+\frac{0{,}115^1 e^{-0{,}115}}{1!}\right)=$$

$$=1-(0{,}8914+0{,}1025)=$$

$$=1-\text{POISSON}(1;0{,}115;1)=0{,}0061$$

c) Para esta debe obtenerse el promedio de fallas en 15 horas:

Y probabilidad de:

$$\lambda_{15} = n \times p = 15 \times 0{,}023 = 0{,}345$$

$$P(x<2)=\left(\frac{0{,}345^0 e^{-0{,}345}}{0!}+\frac{0{,}345^1 e^{-0{,}345}}{1!}\right)=$$

$$=(0{,}8914+0{,}1025)=$$

$$=\text{POISSON}(1;0{,}345;1)=0{,}9526$$

6.13 *Análisis de Colas y Líneas de Espera.*

Se forma una línea de espera siempre que la *Demanda Actual* del servicio supera la *Capacidad Actual* de proporcionarla.

Muchas veces es difícil predecir cuándo llegarán las unidades en busca de un servicio y cuánto tiempo tendrán qué esperar para darles ese servicio.

El proporcionar demasiado servicio provoca costos excesivos. Por otro lado, carecer de la capacidad de servicio suficiente causa colas excesivamente largas que también alargan los costos por servicios insatisfechos, quedando además, costos sociales y empleados ociosos cuando la cola desparece.

La teoría de las colas, en sí, no resuelve el problema, pero contribuye con información vital que se requiere para tomar decisiones concernientes a la predicción de algunas características sobre la línea de espera, tal como el tiempo promedio de espera.

6.14 *Proceso básico de las colas.*

- Los *Clientes* que requieren un servicio a través del tiempo son una *Fuente de Entrada* que pasa a ocupar un lugar en *El Sistema de Colas* cuando el cliente se une a una de estas.
- En determinado momento se selecciona un cliente de la *Cola* mediante una regla conocida como *Disciplina de Cola* o *Disciplina de Servicio.*
- Pasado un tiempo, mediante un mecanismo específico se le proporciona el servicio solicitado. El cliente después de recibirlo *Sale del Sistema de Cola.* Esquemáticamente arriba;
- En el sistema de colas, el tiempo que tarda en satisfacerse el servicio es fundamental, como lo es en el *Proceso Poisson.*

6.15 Problema 3. Atención de emergencias.

La sala de emergencias de una Clínica de Seguridad Social proporciona cuidados médicos a presuntos pacientes que requieren servicio. La clínica es pequeña y el departamento es atendido por un médico. Por esto, el servicio se ofrece siempre que el doctor esté disponible.

Ha últimas fechas han aumentado las emergencias, sobre todo al iniciar la mañana y al anochecer y los accidentados deben esperar, en ocasiones, mucho tiempo para ser atendidos.

El doctor de emergencias ha hecho una propuesta a la dirección del hospital para tratar dos casos simultáneamente. La dirección, ha encargado al departamento de operaciones que estudie el problema y ofrezca soluciones.

El Ingeniero Administrador inició el estudio reuniendo datos históricos para planificar la operación del año siguiente. Reconoció que el problema de emergencias se podría aproximar con modelos de *Teoría de Colas.*

6.16 La Distribución de Frecuencias.

El análisis de los datos históricos de un año indicó que la entrada de casos a emergencias puede considerarse aleatoria.

Emergencias	FRECUENCIAS OBSERVADAS			EXPANSIÓN DE TOTALES = f * x		
x	Mañana	Tarde	Resto Día	Mañana	Tarde	Resto Día
0	5	0	48	0	0	0
1	24	5	97	24	5	97
2	44	34	97	88	68	194
3	86	41	70	258	123	210
4	56	66	36	224	264	144
5	61	50	9	305	250	45
6	48	46	8	288	276	48
7	25	48	0	175	336	0
8	10	40	0	80	320	0
9	3	18	0	27	162	0
10	3	17	0	30	170	0
Sumas	365	365	365	1499	1974	738
Promedios				4,107	5,408	2,022

De acuerdo a los médicos responsables el ingeniero dividió el día en tres clases: de 6 a 10 de la mañana, de las 16 a las 21 horas y el resto del día. La tabulación de emergencias se hace en el siguiente cuadro así como la estimación de promedios por hora: 4,107; 5,408 y 2,022 emergencias por hora, para la mañana, tarde y resto del día respectivamente.

6.17 Los promedios de emergencias por hora.

En la distribución Poisson la media es un parámetro muy importante, pues proporciona la base para la estimación. Cuando además se cuenta con datos para obtener el cuadro de frecuencias u ocurrencia de eventos es posible hacer pruebas para verificar la calidad de aproximación que se obtendrá con la distribución de probabilidad Poisson. Los cálculos de promedios obtenidos a partir de las tablas de frecuencias son:

Por la mañana: $\bar{x}_m = \dfrac{\sum_{i=0}^{10} f_i x_i}{\sum_{i=0}^{10} f_i} = \dfrac{5(0)+24(1)+...+3(10)}{5+24+...+3} = \dfrac{1.499}{365} = 4{,}107$

Por la Tarde: $\bar{x}_t = \dfrac{\sum_{i=0}^{10} f_i x_i}{\sum_{i=0}^{10} f_i} = \dfrac{0(0)+5(1)+...+17(10)}{0+5+...+17} = \dfrac{1.974}{365} = 5{,}408$

Resto del día: $\bar{x}_r = \dfrac{\sum_{i=0}^{10} f_i x_i}{\sum_{i=0}^{10} f_i} = \dfrac{48(0)+97(1)+...+0(10)}{48+97+...+0} = \dfrac{738}{365} = 2{,}022$

6.18 Prueba de Bondad de Ajuste para la mañana.

En la prueba de *Bondad de Ajuste* se comparan mediante la prueba de χ^2 las frecuencias observadas contra las esperadas obtenidas usando la distribución Poisson.

En donde la frecuencia esperada se obtiene multiplicando la probabilidad Poisson por el número de observaciones:

Emergencias x	Frecuencias Observada	Frecuencias Esperada	Prueba Ajuste	Prueba Chi-Cuadr.
0	5	6,0	-1,0	0,1690
1	24	24,7	-0,7	0,0184
2	44	50,7	-6,7	0,8765
3	86	69,4	16,6	3,9941
4	56	71,2	-15,2	3,2484
5	61	58,5	2,5	0,1078
6	48	40,0	8,0	1,5850
7	25	23,5	1,5	0,0974
8	10	12,1	-2,1	0,3511
9	3	5,5	-2,5	1,1378
10	3	3,5	-0,5	0,0767
Sumas	365	365		11,6622
			Probabilidad	**0,3083**

Para obtener las probabilidades esperadas en una distribución Poisson se utilizo la función integrada de la HE en la modalidad de datos individuales multiplicada por el total de días explorado de 365.. Se ejemplifica con x = 0 y x = 1:

$$fe_0 = \left(\frac{4{,}107^0 e^{-4{,}107}}{0!}\right)365 = \text{POISSON}(0;4{,}107;0) * 365 = 6{,}0$$

$$fe_1 = \left(\frac{4{,}107^1 e^{-4{,}107}}{1!}\right)365 = \text{POISSON}(1;4{,}107;0) * 365 = 24{,}7$$

Una vez que se tienen las frecuencias esperadas se procede a obtener el estadístico de Chi-Cuadrada:

$$\chi^2_{(11-1)} = \sum_{i=0}^{10} \frac{(fo_i - fe_i)}{fe_i} = \frac{(5-6{,}0)^2}{6{,}0} + \frac{(24-24{,}7)^2}{24{,}7} + ... + \frac{(3-3{,}5)^2}{3{,}5} = 11{,}6622$$

$$F_{[11,6622;11\text{-}1]} = Y_0 \int_0^{11,6622} (11,6622)^{\frac{1}{2}(11-1)} e^{-\frac{1}{2}11,6622} d\chi =$$

$$= \text{DISTRIBUCION.CHI}(11,6622;11-1) = 0,3083$$

La Probabilidad de la prueba indica que debe aceptarse la hipótesis nula:

Ho; La distribución Poisson aproxima adecuadamente los datos.

6.19 Prueba de Bondad de Ajuste para la tarde.

Procediendo de la misma forma para la distribución de emergencias para la tarde (se menciona que cuando había 10 o más emergencias de contaba como 10. Esto obligó a un ajuste obteniendo la clase 10 por diferencia en este y el caso anterior).

La prueba de Chi-cuadrada se estimó en:

$$\chi^2_{(11-1)} = \sum_{i=0}^{10} \frac{(fo_i - fe_i)}{fe_i}$$

$$= \frac{(0-1,6)^2}{1,6} + \frac{(5-8,8)^2}{8,8} + \ldots + \frac{(17-17,9)^2}{17,9} = 17,4681$$

Con una probabilidad de aceptación de:

$$F_{[17,4681;11\text{-}1]} =$$

$$= Y_0 \int_0^{17,4681} (17,4681)^{\frac{1}{2}(11-1)} e^{-\frac{1}{2}17,4681} d\chi =$$

$$= \text{DISTRIBUCION.CHI}(17,4681;11-1)$$

$$= 0,0646$$

Emergencias	Frecuencias		Prueba	
x	Observada	Esperada	Ajuste	Chi-Cuadr.
0	0	1,6	-1,6	1,6351
1	5	8,8	-3,8	1,6699
2	34	23,9	10,1	4,2562
3	41	43,1	-2,1	0,1030
4	66	58,3	7,7	1,0219
5	50	63,0	-13,0	2,6977
6	46	56,8	-10,8	2,0615
7	48	43,9	4,1	0,3826
8	40	29,7	10,3	3,5893
9	18	17,8	0,2	0,0015
10	17	17,9	-0,9	0,0494
Sumas	365,0	365,0		17,4681
			Probabilidad	0,0646

La Probabilidad indica que debe aceptarse la hipótesis nula, aunque con un nivel muy próximo a la significación:

Emergencias	Frecuencias Obsevadas		Frecuencias Esperadas		Frec. Relativas Acumula.		Diferencia
x	Observación	Acumuladas	Observación	Acumulada	Observada	Esperada	Absoluta
0	0	0	1,6	1,6	0,0000	0,0045	0,0045
1	5	5	8,8	10,5	0,0137	0,0287	0,0150
2	34	39	23,9	34,4	0,1068	0,0942	0,0126
3	41	80	43,1	77,5	0,2192	0,2123	0,0069
4	66	146	58,3	135,8	0,4000	0,3720	0,0280
5	50	196	63,0	198,8	0,5370	0,5447	0,0077
6	46	242	56,8	255,6	0,6630	0,7004	0,0374
7	48	290	43,9	299,5	0,7945	0,8207	0,0262
8	40	330	29,7	329,2	0,9041	0,9020	0,0021
9	18	348	17,8	347,1	0,9534	0,9508	0,0026
10	17	365	17,9	365,0	1,0000	1,0000	0,0000
				Máxima desviación absoluta			0,0374
				Valor Crítico(0,05; 365) = 1,63/Raiz(n)			0,0853

Ho; La distribución Poisson aproxima adecuadamente los datos.

La prueba resultó muy ajustada, se podría confirmar la Bondad de Ajuste mediante la prueba de Kolmogorov-Smirnov descrita en Estadísticas no Paramétricas (capítulo III).

La prueba de utiliza las distribuciones de probabilidad acumulativas. En el siguiente cuadro se muestran los datos necesarios para la prueba.

La regla de decisión indica que debe aceptarse la hipótesis nula puesto que la desviación máxima observada d = 0,0374 es menor que el criterio de comparación $D_{(0,05;\ 365)} = \frac{1,36}{\sqrt{365}}$ ***debido a que la cantidad de observaciones es mayor a 35 según se indica en la tabla de valores críticos correspondientes. Ver capítulo III.***

6.20 Prueba de Bondad de Ajuste para resto de día.

Procediendo de la misma forma en la distribución de emergencias para el resto del día.
La prueba de Chi-cuadrada se estimó en:

$$\chi^2_{(11-1)} = \sum_{i=0}^{10} \frac{(fo_i - fe_i)}{fe_i}$$

$$= \frac{(48-48,3)^2}{48,3} + \frac{(97-97,7)^2}{97,7} + ... + \frac{(0-0,1)^2}{0,1} = 6,2314$$

Emergencias x	Frecuencias		Prueba	
	Observada	Esperada	Ajuste	Chi-Cuadr.
0	48	48,3	-0,3	0,0022
1	97	97,7	-0,7	0,0052
2	97	98,8	-1,8	0,0322
3	70	66,6	3,4	0,1760
4	36	33,7	2,3	0,1636
5	9	13,6	-4,6	1,5609
6	8	4,6	3,4	2,5415
7	0	1,3	-1,3	1,3246
8	0	0,3	-0,3	0,3348
9	0	0,1	-0,1	0,0752
10	0	0,0	0,0	0,0152
Sumas	365	365,0		**6,2314**
			Probabilidad	**0,7955**

Con una probabilidad de aceptación de:

$$F_{[6,2314;11\text{-}1]} = Y_0 \int_0^{6,2314} (6,2314)^{\frac{1}{2}(11-1)} e^{-\frac{1}{2}6,2314} d\chi =$$

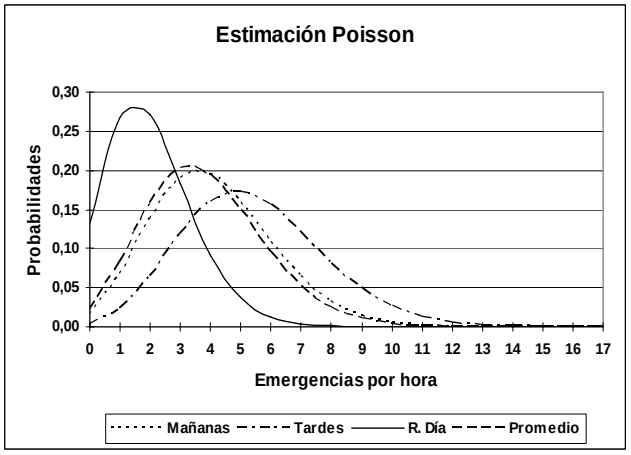
$$= \text{DISTRIBUCION.CHI}(6,2314;11-1) = 0,7955$$

E indica que debe aceptarse la hipótesis nula:

Ho; La distribución Poisson aproxima adecuadamente los datos.

6.21 Las distribuciones de probabilidad.

Emergencia x	PROMEDIOS Y ESTIMACIONES POISSON PARA			
	4,107 Mañanas	**5,408** Tardes	**2,022** R. Día	**3,846** Promedio
0	0,0165	0,0045	0,1324	0,0214
1	0,0676	0,0242	0,2677	0,0822
2	0,1388	0,0655	0,2706	0,1580
3	0,1900	0,1181	0,1824	0,2026
4	0,1951	0,1597	0,0922	0,1948
5	0,1602	0,1727	0,0373	0,1498
6	0,1097	0,1557	0,0126	0,0960
7	0,0644	0,1203	0,0036	0,0527
8	0,0330	0,0813	0,0009	0,0254
9	0,0151	0,0489	0,0002	0,0108
10	0,0062	0,0264	0,0000	0,0042
11	0,0023	0,0130	0,0000	0,0015
12	0,0008	0,0059	0,0000	0,0005
13	0,0002	0,0024	0,0000	0,0001
14	0,0001	0,0009	0,0000	0,0000
15	0,0000	0,0003	0,0000	0,0000
16	0,0000	0,0001	0,0000	0,0000
17	0,0000	0,0000	0,0000	0,0000



Se ha probado que la distribución Poisson aproxima apropiadamente la distribución de frecuencias de las emergencias en el día. Las estimaciones incluyendo el promedio se muestran en el cuadro con el que se elabora el gráfico.

Debe notar que a medida que el promedio es mayor, la probabilidad se desplaza a la derecha hasta parecerse a una distribución *Normal* sesgada a la derecha.

6.22 Algunas respuestas al caso.

Si el médico tarda en promedio 24 minutos por paciente, esto es 2,5 pacientes por hora:

¿Cuál es la probabilidad de que a cualquier hora del día lleguen más de 2,5 pacientes?

El problema estriba en que la distribución Poisson se refiere a sucesos discretos y usualmente, tratando de promedios se hace referencia a datos continuos. Para el promedio del día de 3,846 pacientes por hora la probabilidad que se busca considerando las distribuciones acumuladas está entre:

$P(x \leq 2) = 0{,}2616$ y $P(x \leq 3) = 0{,}4642$

Lo más simple es efectuar una extrapolación lineal multiplicando por la fracción de evento. Así la probabilidad para 2,5 pacientes por hora se estimaría en:

$P(2{,}5) = P(2) + [P(3) - P(2)][2{,}5 - 2] =$

$= 0{,}2229 + (0{,}4642 - 0{,}2229)0{,}5 = 0{,}3629$

Por tanto

$P(x > 2{,}5) = 1 - 0{,}3629 = 0{,}6371$

La probabilidad de encontrar cola en emergencias en el día es 63,71%

Recuerde la cantidad de pacientes que atiende el médico por hora.

Promedios x	4,107 Mañanas	5,408 Tardes	2,022 R. Día	3,846 Promedio
0	0,0165	0,0165	0,0165	0,0165
1	0,0676	0,0676	0,0676	0,0676
2	0,1388	0,1388	0,1388	0,1388
3	0,1900	0,1900	0,1900	0,1900
4	0,1951	0,1951	0,1951	0,1951
5	0,1602	0,1602	0,1602	0,1602
6	0,1097	0,1097	0,1097	0,1097
7	0,0644	0,0644	0,0644	0,0644
8	0,0330	0,0330	0,0330	0,0330
9	0,0151	0,0151	0,0151	0,0151
10	0,0062	0,0062	0,0062	0,0062
11	0,0023	0,0023	0,0023	0,0023
12	0,0008	0,0008	0,0008	0,0008
13	0,0002	0,0002	0,0002	0,0002
14	0,0001	0,0001	0,0001	0,0001
Estadístico z	-0,7929	-1,2505	0,3362	-0,6862
Igual o menos 2,5	0,2139	0,1055	0,6316	0,2463
Mas de 2,5	0,7861	0,8945	0,3684	0,7537

La extrapolación puede efectuarse utilizando la aproximación mediante la distribución Normal Estándar considerando la particularidad de que en la distribución Poisson la media y la varianza tienden a tener el mismo valor. Así para el promedio de todo el día la probabilidad que lleguen exactamente 2,5 pacientes en una hora sería:

$$z_{\bar{x}} = \frac{2{,}5 - 3{,}846}{\sqrt{3{,}846}} = -0{,}6862$$

Con una probabilidad de:

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{2{,}5} e^{-\frac{1}{2}\left(\frac{2{,}5-3{,}846}{\sqrt{3{,}846}}\right)^2} dx = 0{,}2463$$

O 24,63%. Y la probabilidad que lleguen más de 2,5 es el complemento 1 − 0,2463 = 0,7535 ó 75,35%.

6.23 Continuando con las respuestas.

Para la mañana la probabilidad de que lleguen más de 2,5 pacientes por hora es:

$P(x \le 2{,}5) = 0{,}2229 + [0{,}4149 - 0{,}2229](0{,}05) = 0{,}3179$

Y

$P(x > 2{,}5) = 1 - 0{,}3179 = 0{,}6821$

Muy próxima al promedio general. Para la Tarde:

$P(x \le 2{,}5) = 0{,}0942 + [0{,}2123 - 0{,}0942](0{,}05) = 0{,}1533$

Y

$P(x > 2{,}5) = 1 - 0{,}1543 = 0{,}8467$

Para el resto del día:

$P(x \le 2{,}5) = 0{,}6707 + [0{,}8531 - 0{,}6707](0{,}05) = 0{,}7619$

Y

$P(x > 2{,}5) = 1 - 0{,}7619 = 0{,}2381.$

De estas condiciones, pareciera que la única parte crítica ocurre por la tarde. Sin embargo, estudiando el problema se encontró que, como en todos los hospitales las emergencias se manejan de manera muy particular.

6.24 Las soluciones son particulares.

NOTA D-24

Promedios x	5,0000 Premura	4,107 Mañanas	5,408 Tardes	2,022 R. Día	3,846 Promedio
0	0,0067	0,0165	0,0045	0,1324	0,0214
1	0,0337	0,0676	0,0242	0,2677	0,0822
2	0,0842	0,1388	0,0655	0,2706	0,1580
3	0,1404	0,1900	0,1181	0,1824	0,2026
4	0,1755	0,1951	0,1597	0,0922	0,1948
5	0,1755	0,1602	0,1727	0,0373	0,1498
6	0,1462	0,1097	0,1557	0,0126	0,0960
7	0,1044	0,0644	0,1203	0,0036	0,0527
8	0,0653	0,0330	0,0813	0,0009	0,0254
9	0,0363	0,0151	0,0489	0,0002	0,0108
10	0,0181	0,0062	0,0264	0,0000	0,0042
11	0,0082	0,0023	0,0130	0,0000	0,0015
12	0,0034	0,0008	0,0059	0,0000	0,0005
13	0,0013	0,0002	0,0024	0,0000	0,0001
14	0,0005	0,0001	0,0009	0,0000	0,0000
Estadístico z		-0,1786	0,0816	-0,5956	-0,2309
Probabilida Presión < 5		0,4291	0,5325	0,2757	0,4087
Probabilidad Espera > 5		0,5709	0,4675	0,7243	0,5913

En los hospitales hay muchas maneras de cubrir estas situaciones. Nuevos datos indican que el tiempo promedio que necesita un doctor de 24 minutos para atender a un paciente se convierte en 12 cuando tiene 6 pacientes (5 esperando turno).

En estos momentos, la enfermera de recepción clasifica a los pacientes en tres grupos, los que son verdaderas emergencias, los que presentan una emergencia media y los que pueden esperar por tiempo largo. Pasa a los pacientes del grupo 1, después a los del dos y así hasta finalizar la cola. Con esta nueva información se calculan las probabilidad de que lleguen mas pacientes de los 60/12 = 5 que pueden atenderse por hora.

Las probabilidades se obtiene fácilmente de las acumulativas o las individuales siendo de:

23,18% Para la mañana;
45,53% Para la tarde;
1,74% Para el resto del día;
Y de 19,12% en promedio.

Información suficiente para tomar decisiones en lo tocante a las emergencias del hospital.

Utilizando la extrapolación mediante la Normal Estándar (en el cuadro de arriba).

6.25 Planes de Control de la Calidad.

En muchos planes de *Control de la Calidad* o de *Control de Proceso* se utiliza la distribución *Poisson*. En estos casos, los eventos que no alcanzan la norma de calidad suelen medirse como fallas por unidad de tiempo.

Dada la naturaleza de la producción manufacturada, que se supone un proceso continuo desde que se inicia la elaboración de la primera unidad hasta que se empaca la última de la línea. Es, a todas luces, incosteable revisar todos los artículos, se establecen planes de muestreo de *Control de la Calidad* o *De Control del Proceso.*

Por su naturaleza, una falla de calidad es un evento raro, por tanto, entra dentro de la competencia de la distribución *Poisson*.

Cuando la media y la varianza de una distribución de datos tienen un valor muy próximo o cuando el evento es raro, el proceso de fallas puede aproximarse mediante la distribución *Poisson*.

Al ser, los procesos de fabricación continuos, las fallas por unidad de tiempo suelen reportarse por unidades producidas en el entendido que hay una relación directa entre la cantidad que se produce y el tiempo en que se fabrica.

6.26 El problema 4. Envases de hoja de lata.

El departamento de control de la calidad de la empresa ENVASES HL, S, A. Tiene por costumbre tomar tres muestras de 10 envases por jornada de 8 horas de trabajo en un plan de muestreo secuencial con iniciación aleatoria (ver sección de técnicas de muestreo). Al final del día hace un recuento de las fallas que acumula en un reporte que todos los días actualiza con los últimos 33 reportes en una hoja de control de la calidad que incluye, además, un control del proceso recopilando el conjunto de factores que concurren cuando el proceso de *pone Fuera de Control Estadístico.*

Al inicial el día, el supervisor de la producción tiene en su mano las dos cartas de control de las máquinas de soldado para que efectúe los ajustes pertinentes y tome las acciones que considere necesarias.

El reporte de fallas en la HE.

6.27 Las Distribuciones Binomial y Poisson.

En el control de fallas están involucradas dos distribuciones que aunque son de la misma familia, su utilidad es diferente:

- *La Binomial*: Que se utiliza para establecer planes de control sobre las proporciones o porcentajes;
- *La Poisson*: Que se utiliza para establecer planes de control sobre el número de fallas.

Usualmente se utiliza sólo una de ellas dependiendo de los objetivos del control.

De manera general las cartas de control son alternativas pictóricas para mostrar el comportamiento estadístico de los errores mediante intervalos confiables con probabilidad específica en una sucesión de muestras para determinar el momento en que el sistema de producción deja su variación aleatoria debido a la presencia de un factor que provoca la falla. Siendo la ubicación y control de estos factores el objetivo del sistema.

6.28 La Binomial para Proporciones.

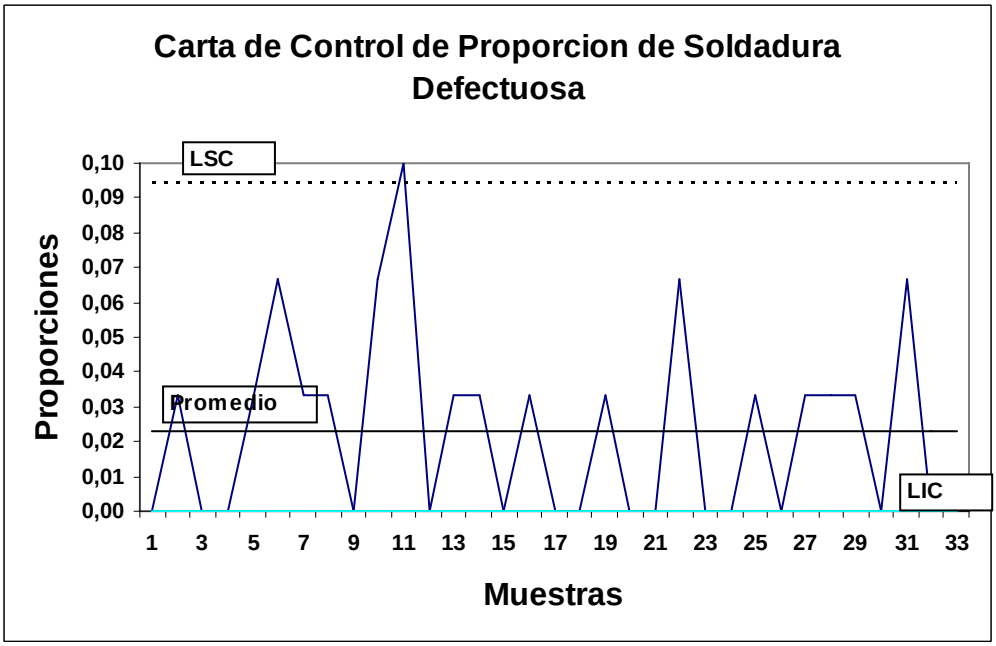
Los límites para la *Carta de Control 99% Confiable* de proporciones se definen como:

$$LSC = p + z\sqrt{\frac{p(1-p)}{n}} = 0{,}023 + 2{,}576\sqrt{\frac{0{,}023(1-0{,}023)}{30}} = 0{,}094$$

$$LIC = p - z\sqrt{\frac{p(1-p)}{n}} = 0{,}023 - 2{,}576\sqrt{\frac{0{,}023(1-0{,}023)}{30}} = -0{,}048$$

Por definición, no puede haber proporciones negativas por tanto, *El Límite Inferior de Control* se define en 0.

El gráfico muestra a la observación 11 fuera de control estadístico y más allá de la norma al sobrepasar el *Límite Superior de Control*.



6.29 La Poisson para el número de fallas.

Los límites para la *Carta de Control 99% Confiable* para observaciones se define como:

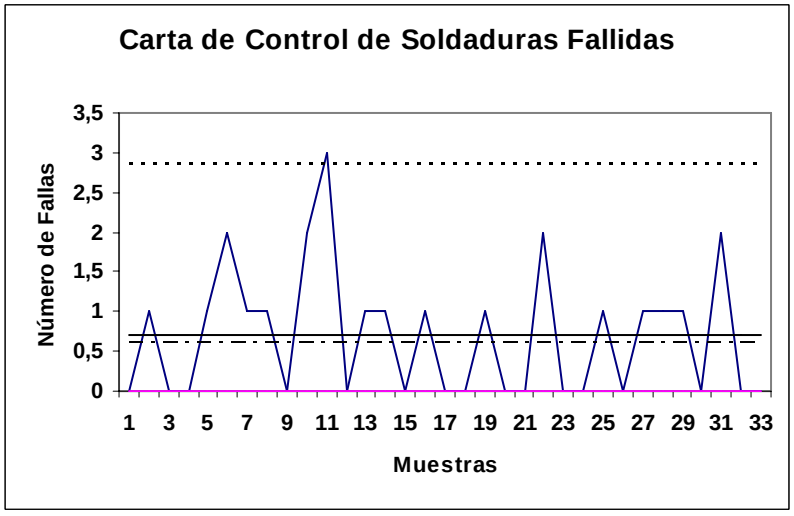
$$LSC = \lambda + z\sqrt{\lambda} = 0{,}7 + 2{,}576\sqrt{0{,}70} = 2{,}85$$

$$LIC = \lambda - z\sqrt{\lambda} = 0{,}7 - 2{,}576\sqrt{0{,}70} = 0{,}30$$

Se agrega el promedio de fallas para 30 observaciones de 0,02 × 30 = 0,6 que aparece como la línea de color café oscuro (la más cercana al eje *x*).

Es evidente que el control de fallas es más estricto que el de proporciones pues 5 muestras presentan situaciones fuera de control estadístico.

El *Límite Inferior de Control* únicamente tiene fines informativos para este caso, pues es obvio que debe ser 0.



6.30 Propiedades de la Distribución Poisson.

A estas alturas el estudiante está en capacidad de deducir las dos propiedades de las variables con distribución Poisson:

- Primera: *El valor de la varianza de la distribución es muy próxima al valor de la media* μ;
- Segunda: *Si una serie de variables independientes X1, X2, X3,,, cada una con distribución Poisson con medias* μ_1, μ_2, μ_3*,,, su suma tendrá una distribución Poisson con una media igual a la suma de las medias (*$\mu_1 + \mu_2 + \mu_3$*,,,).*

Debe verse la distribución *Poisson* como una variante de la Binomial sólo qué ésta tiene dos alternativas para el promedio de medias y varianzas:

Para números:

$$\mu = np; \quad \sigma^2 = np(1-p)$$

Para proporciones que se puede ampliar a porcentajes:

$$\mu = p; \quad \sigma^2 = p(1-p)$$

Es fácil deducir la primera propiedad, recuérdese que $p + q = 1$ y $q = 1 - p$, por tanto, es de esperarse que la varianza tienda a aproximarse al promedio cuando *p* es muy pequeño y *q* tiende a 1.

6.31 La segunda propiedad.

La segunda propiedad se deduce de la anterior. Ya que por sus características de medir eventos raros, la pequeña magnitud de la *proporción* ***p*** tiende a ser constante y para un conjunto de variables ***X***, independientemente del taño de ***n***, el promedio de ellas será:

$$\mu_G = n_1 p + n_2 p + \ldots + n_m p = p\sum_{i=1}^{m} n_i = n_G p$$

Por estas dos propiedades se ha visto que la distribución de Poisson es útil para aproximarse al fenómeno cuando el promedio μ es pequeño, aún cuando el valor de *n* no está bien definido y aún, sí los valores ***n*** y ***p*** se presume que varían de una muestra a otra.

El ejemplo que ha servido desde hace mucho tiempo para ejemplificar el uso de la distribución de *Poisson* se debe a *Ladislaus von Bortkewitch* (1.898) que se refiere al número de hombres en la caballería Prusiana que mueren en un año por coces de caballos. Si se tienen $N = 200$ observaciones una por cada 10 batallones en 20 años, Cualquier día un hombre está expuesto a la pequeña probabilidad de que sea coceado, pero no está claro que valor tiene *n* y también *p* que puede ser variable para cada individuo.

Bortkewitch reporta 122 muertes por tanto la proporción de 200 batallones año analizados.

$$\lambda = \frac{122}{200} = 0{,}61$$

El cuadro de frecuencias muestra los datos del ejemplo en el que se ajusta el valor esperado para la clase de 3 o más muertes. Y el promedio se toma directo pues no cierra en 3, mediante:

$$\lambda = \frac{109\times 0 + 65\times 1 + 22\times 2 + 4\times 3}{200} = \frac{121}{200} = 0{,}605$$

Muertes x	Observados	Esperados	Contrib. a Chi-Cuadrada
0	109	108,7	0,001
1	65	66,3	0,025
2	22	20,2	0,157
3 o más	4	4,8	0,137
Sumas	200	200,0	0,320
		Probabilidad	0,956

Que provoca diferencias que no son de importancia.

La valuación de la χ^2 indica que la probabilidad de aproximación de la distribución esperada es de 0,956 ó 95,6%. Una excelente aproximación.

6.32 Problema 5. Tradicional del proceso Poisson.

Una empresa de la construcción reporta los siguientes accidentes laborales de 155 ingenieros entre 30 y 35 años durante su primer año de trabajo:

Accidentes de T	0	1	2	3	4 o más
Nº de personas	80	61	13	1	0

La dirección de la empresa proyecta pagar un plan de seguros médicos a los ingenieros jóvenes, pues su trabajo en el campo los expone más a los accidentes.

La aseguradora solicitó la información mencionada para entregar un plan de seguros apropiado.

La gerencia desea saber si la frecuencia de accidentes está dentro de límites aceptables.

El análisis se inicia encontrando el promedio de accidentes de los ingenieros por un año, para esto se utiliza la conocida tabla de frecuencias y promedios para datos agrupados.

Como en el caso anterior, es posible conocer el número de accidentes si tener que estimarlos mediante el cuadro de frecuencias.

Nº Accidentes x	Frecuencia Observada	Frecuencia por evento
0	80	0
1	61	61
2	13	26
3	1	3
4	0	0
Estadísticos		
Nº Observaciones		155
Nº de Accidentes		90
Promedio de accidentes		0,5806

6.33 La Distribución de Frecuencias.

El promedio de accidentes se obtiene dividiendo el número de ingenieros accidentados entre el total de ingenieros en el análisis.

Con este promedio se obtiene la distribución de frecuencias esperadas para la efectuar una prueba de *Bondad de Ajuste* comparando las frecuencias observadas con las esperadas mediante una prueba de χ^2. Para el caso

será suficiente una probabilidad de confianza de 95%.

$$\lambda = \frac{\sum_{i=0}^{4} f_i x_i}{\sum_{i=0}^{4} f_i} = \frac{80(0)+61(1)+13(2)+1(3)+0(4)}{80+61+13+1+0} = \frac{90}{155} = 0{,}5806$$

6.34 Prueba de Bondad de Ajuste.

Frecuencias esperadas:

$$fe_i = n\left(\frac{\lambda^x e^{-\lambda}}{x!}\right); \quad fe_2$$

$$= 155\left(\frac{0{,}5806^2 e^{-0{,}5806}}{2!}\right) = 14{,}6$$

Nº Accidentes	Probabilidad	Frecuencias		Bondad
x	Poisson	Observada	Esperada	Ajuste
0	0,5595	80	86,7	0,5220
1	0,3249	61	50,4	2,2488
2	0,0943	13	14,6	0,1795
3	0,0183	1	2,8	1,1831
4	0,0027	0	0,4	0,4108
5	0,0003	0	0,0	0,0477
6	0,0000	0	0,0	0,0046
7	0,0000	0	0,0	0,0004
Sumas	1,0000	155	155,0	**4,5969**
			Probabilidad	**0,7090**

La prueba de χ^2:

$$\chi^2_{(8-1)} = \frac{(80-86{,}7)^2}{86{,}7} + \frac{(61-50{,}4)^2}{50{,}4} + \ldots + \frac{(0-0{,}00..)^2}{0{,}00..} = 4{,}5969$$

La probabilidad que determina el estadístico:

$$F_{[4{,}5909;\,8\text{-}1]} = Y_0 \int_0^{4{,}5909} (4{,}5909)^{\frac{1}{2}(8-1)} e^{-\frac{1}{2}4{,}5909} d\chi = 0{,}7090$$

El valor de 70,90% está dentro de la zona 95% de aceptación de la hipótesis nula. Por tanto, esta debe aceptarse

6.35 Comprobando la primera propiedad: $\mu = \sigma 2$.

El Promedio:

$$\mu = \sum_{i=0}^{7} P_i x_i = 0{,}559537 \times 0 + 0{,}324893 \times 1 +$$

$$+ \ldots + 0{,}000002 \times 7 = 0{,}5806$$

En donde $\boldsymbol{P}_i$ debe entenderse como la probabilidad del evento.

La varianza:

Nº Accidentes	Probabilidad	Estimadores	
x	Poisson	Promedio	Varianza
0	0,559537	0,0000	0,1886
1	0,324893	0,3249	0,0571
2	0,094324	0,1886	0,1900
3	0,018256	0,0548	0,1069
4	0,002650	0,0106	0,0310
5	0,000308	0,0015	0,0060
6	0,000030	0,0002	0,0009
7	0,000002	0,0000	0,0001
Estimadores		0,5806	0,5806

$$\sigma^2 = \sum_{i=0}^{7} P_i(x_i - \lambda)^2 = 0{,}559537(0 - 0{,}5806)^2 + 0{,}324893(1 - 0{,}5806)^2 +$$

$$+ \ldots + 0{,}000002(7 - 0{,}5806)^2 = 0{,}5806$$

Mediante esta prueba práctica se demuestra la primera propiedad de la Distribución Poisson: el promedio y la varianza tienen el mismo valor λ.

6.36 *Uso conjunto de la Poisson y la Normal.*

Una empresa procesadora de leche de vaca ha decidido cambiar sus máquinas envasadoras en cartón sanitario inducidos por los vendedores de maquinas que aseguran una mayor precisión de llenado necesaria para cubrir las normas internacionales que piden una variación promedio inferior a 5% y cuando más un 2% de envases por debajo de la norma. Se estableció un programa de control de la calidad en el que se muestrean 5 unidades por cuatro veces, en una jornada de trabajo de 8 horas iniciando la toma de la primera muestra de forma aleatoria eligiendo un número entre el minuto 0 y el 120 para la primera muestra, las siguientes se toman en secuencia de 2 horas. Los resultados de 100 muestreos, o 20 días se presentan en la HE. Los estadísticos obtenidos directamente:

Promedio: $\mu = \dfrac{\sum_{i=1}^{N} x_i}{N} = \dfrac{9}{500} = 0{,}018$

Varianza: $\sigma^2 = \dfrac{\sum_{i=1}^{N} (x_i - \mu)^2}{N} = p(1-p) = 0{,}018$

6.37 *El promedio es suficiente.*

En la distribución Poisson el promedio es lo único que se requiere conocer para efectuar estimaciones, puesto que se ha comprobado que $\mu = \sigma^2$.

Para muestras de tamaño $n = 5$ se espera un promedio de $5 \times 0{,}018 = 0{,}09$ unidades fuera de norma. Comprobaremos este promedio utilizando la distribución de frecuencias para las muestras de tamaño $n = 5$.

El promedio:

$$\lambda = \frac{92 \times 0 + 7 \times 1 + 1 \times 2}{92 + 7 + 1} = \frac{9}{100} = 0{,}09$$

Evento x	Frecuencia Observada	Totales x Evento
0	92	0
1	7	7
2	1	2
Sumas	100	9

Nº Muestras	100
Promedio fuera norma	0,0900

6.38 *Prueba de Bondad de Ajuste 95%*

Aún corriendo el riesgo de ser reiterativo, es necesario valuar, sí las frecuencias esperadas se pueden aproximar mediante la distribución Poisson para $n = 5$.

Evento x	Probabilidad Poisson	Frecuencias Observada	Frecuencias Esperada	Prueba B. Ajuste
0	0,9139	92	91,4	0,0040
1	0,0823	7	8,2	0,1826
2	0,0037	1	0,4	1,0718
3	0,0001	0	0,0	0,0111
4	0,0000	0	0,0	0,0002
5	0,0000	0	0,0	0,0000
Sumas	1,0000	100,0000	100,0000	1,2697
			Probabilidad	**0,9380**

$$P(x=2)=\frac{\lambda^x e^{-\lambda}}{x!}=$$

$$=\frac{0{,}09^2 e^{-0{,}09}}{2!}=0{,}0037$$

La frecuencia esperada para x = 2:

$$fe_2 = P(x=2)\times m = 0{,}0037\times 100 = 0{,}4$$

El Estadístico de Chi-Cuadrada:

$$\chi^2_{(6-1)}=\frac{(92-91{,}4)^2}{91{,}4}+\frac{(7-8{,}2)^2}{8{,}2}+\ldots+\frac{(0-0{,}00..)^2}{0{,}00..}=1{,}2697$$

La probabilidad que determina el estadístico.

$$F_{[1{,}2697;\,6\text{-}1]}=Y_0\int_0^{1{,}2697}(1{,}2697)^{\frac{1}{2}(6-1)}e^{-\frac{1}{2}1{,}2697}d\chi=0{,}9380$$

Valor que evidentemente se encuentra en la zona 95% de aceptación de la Hipótesis Nula. Por tanto, se puede utilizar la Distribución Poisson para estudiar la distribución de los datos.

La probabilidad de 93,80% indica que debe aceptarse: que las frecuencias observadas del experimento se pueden aproximar con la Distribución Poisson sin duda alguna.

6.39 Prueba de z para defectuosos.

La empresa asegura una probabilidad de que cuando más 2% de envases están fuera de la norma. Para el caso:

$$P(x\geq 2)=1-\big(P(x=0)+P(x=1)+P(x=2)\big)=1-\big(0{,}9139+0{,}0823+0{,}0037\big)=0{,}0038$$

ó, 38%. Muy por debajo de la norma.

Para valorar la hipótesis: Ho; p ≤ 0,02 o $x \leq np$

Se puede utilizar la distribución de z puesto que $N = 500$, una muestra definitivamente grande.

$$z=\frac{|x-np|-0{,}5}{\sqrt{npq}}=-\frac{|9-500\times 0{,}02|-0{,}5}{\sqrt{500\times 0{,}02\times 0{,}98}}=-0{,}1597$$

Estadístico que determina una probabilidad de:

$$F(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0{,}1597}e^{-\frac{1}{2}\left(\frac{9-500\times 0{,}02}{\sqrt{500\times 0{,}02\times 0{,}98}}\right)^2}dx=0{,}4366$$

Ó ,66% de que el número de envases fuera de la norma sea diferente a 10 o $p \neq 0{,}02$. Por tanto, se acepta que la maquinaria trabaja dentro de lo ofrecido.

6.40 Prueba de z sobre la proporción.

También se puede probar la hipótesis usando directamente las proporciones:

$$z=\frac{|p-P|-\frac{1}{2n}}{\sqrt{\frac{pq}{n}}}=-\frac{|0{,}018-0{,}02|-\frac{1}{2\times 500}}{\sqrt{\frac{0{,}02\times 0{,}98}{500}}}=-0{,}1597$$

La empresa distribuye en cajas de 12 envases:

¿Cuál es la probabilidad de que no se encuentren envases por debajo de 0,02%?

El promedio para $n = 12$ es $12 \times 0{,}018 = 0{,}216$.

$$\lambda_{12} = np = 12\times 0{,}018 = 0{,}216$$

La probabilidad que no se encuentre ninguno fuera de la norma es:

$$P(x=0)=\frac{0{,}216^{0}e^{-0{,}216}}{0!}=0{,}8057$$

¿Cuál es la probabilidad de encontrar en una caja de 12 litros 3 envases por debajo de la norma?

Esta probabilidad se encuentra restando de 1 la probabilidad acumulada hasta 3 envases fuera de norma:

$$F(x\le 2)=1-\sum_{x=0}^{2}\frac{\lambda^{x}e^{-\lambda}}{x!}=1-\text{POISSON}(2;0{,}216;1)=1-0{,}9986=0{,}0014$$

6.41 Normas en una variable continua.

En términos de la variable continua, la estimación por intervalos permite responder a la interrogante que pudiera ser fuente de controversias en caso de una inspección:

¿Cuál es la probabilidad asegurada para una caja de 12 litros?

Esto es:

$$\Pr\left\{\bar{x}-z\frac{s}{\sqrt{n}}\ge\bar{X}\le\bar{x}+z\frac{s}{\sqrt{n}}\right\}=0{,}98$$

$$\Pr\left\{1.000-2{,}326\frac{1{,}488}{\sqrt{12}}\ge\bar{X}\le 1{,}000+2{,}326\frac{1{,}488}{\sqrt{12}}\right\}=0{,}98$$

$$\Pr\left\{999\ge\bar{X}\le 1{,}001\right\}=98\%$$

Las cajas de productos de esta empresa ofrecidos en envases de cartón sanitario presentarán promedios entre los límites indicados con probabilidad de 98%.

Un mecanismo similar se sigue para establecer límites de control estadístico sobre los promedios diarios de 20 unidades.

6.42 Límites en cartas de control.

a) La norma internacional establece como límite de promedios con una variación inferior al 5%.

b) Algunas Cartas de Control utilizan 6 zonas para valorar si el proceso se mantiene en condiciones de variación estadísticas.

El valor *z* que define el valor de control para el 5% inferior es –1,645 y para 95% 1,645. Con estos se establecen los límites extremos de control:

El Límite Superior de Control.

$$LSC=\bar{x}+1{,}645\frac{1{,}488}{\sqrt{5}}=1001{,}1$$

El Límite Inferior de Control:

$$LIC=\bar{x}-1{,}645\frac{1{,}488}{\sqrt{5}}=999{,}0$$

Límite Tercer Cuatilo:

$$LMS=\bar{x}+0{,}674\frac{1{,}488}{5}=1000{,}6$$

Límite Primer Cuartilo:

$$LMI=\bar{x}-0{,}674\frac{1{,}488}{\sqrt{5}}=999{,}4$$

6.43 Objetivos de la Carta de Control.

Se ha mencionado que el objetivo de la Carta de Control es valorar si el sistema se mantiene bajo control estadístico. Una regla empírica indica que aparece un factor que altera el proceso estadístico cuando:

1. Un punto aparece más allá de la zona A;
2. Dos puntos consecutivos aparecen en la zona B;
3. Cuatro puntos consecutivos ocurren en la zona C.

Cuando esto ocurre, el personal involucrado en la producción debe darse a la tarea de localizar la fuente o factor que ha alterado proceso estadístico de la producción.

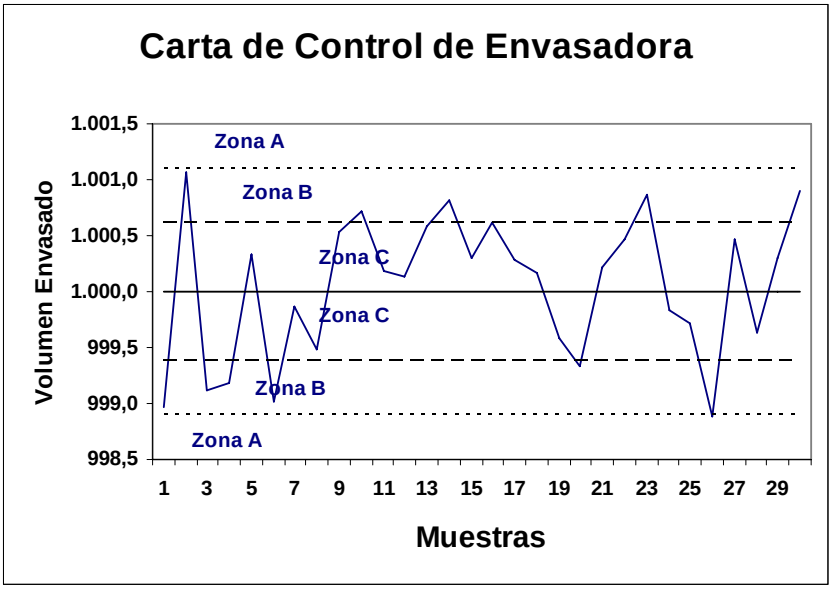
El ubicar estos factores y la regularidad con que se presentan, ofrecen a los encargados de la producción o del control de la calidad, hacer de la producción una función acuciosamente estudiada.

6.44 Valoración de la Carta de Control

La muestra 2: Alcanzan la zona A por la parte superior. Se valora como falla; acción, buscar la causa).

La muestra 26: Alcanza la zona A inferior y hay una observación considerada falla; acción, buscar la causa.

Las muestras 3 y 4 se presentan en la zona B inferior. No hay falla; acción, atención.



6.45 Control conjunto.

Muchos casos de control de la producción y de la calidad de un producto maquinado o manufacturado requiere de controles conjuntos sobre elementos que no cumplen ciertas normas de producción o calidad.

Usualmente, el control ejercido sobre la variable cualitativa que califica a una unidad como “No Aceptable” y “Aceptable” es mucho más estricta que la variable cuantitativa que mide o determina la magnitud de la falla.

La base estadística que da soporte a los procesos de control de la calidad y proceso es la presentación aleatoria de las diferencias en las unidades fabricadas.

Los Sistemas de control de procesos mediante variables cualitativas y cuantitativas son complementarios.

6.46 Resumen

La Distribución Poisson se adapta muy apropiadamente a variables cualitativas en donde interesan eventos poco frecuentes.

El proceso Poisson se caracteriza por su dependencia del tiempo. Un hecho ocurre o no en intervalos continuos de tiempo; La probabilidad de más de un evento cierto en un espacio de tiempo es despreciable; La ocurrencia de un evento no depende de lo que pasó antes ni de lo que pasará después.

Las cualidades de la distribución Poisson le permiten adaptarse a múltiples circunstancias, incluso cuando el espacio muestral no está muy bien definido.

Se adapta muy apropiadamente a los procesos productivos en los que interesan fallas o aciertos raros. Pero sobre todo, por la constancia de las cantidades producidas por unidad de tiempo.

La distribución Poisson es complemento ideal para el muestreo secuenciado que se hace a intervalos regulares de tiempo o de unidades producidas.

REFERENCIAS SELECTAS:

1. Hillier Frederick S., y Lieberman Gerard j., Introducción a la Investigación de Operaciones. Capítulo 19. Segunda edición en español traducida de la cuarta edición en ingles. McGraw-Hill Interamericana de México, S. A. De C. V., 1990.
2. Miller Irwin, Freund John E., Johnson Richard A: Probabilidad y Estadística para Ingenieros. Capítulo 11. Traducido de la cuarta edición en ingles; Prentice-Hall Hispanoamericana, S. A. 1992.
3. Murray R. Spiegel: Serie de compendios Schaum, Teoría y Problemas de Estadística. Capítulo 7. Primera edición en español, traducido de la primera edición en ingles; Libros McGraw-Hill de México, S. A. De C. V., 1973.
4. Ostle Bernard: Estadística Aplicada. Capítulos 4, 5 y 6. Primera edición en español traducida de la primera edición en ingles. Editorial Limusa, S. A., 1977.
5. Snedecor George W., y Cochran William G: Statistical Methods. Capítulos 2, 3 y 8. Sexta edición; The Iowa State University, 1974.
6. Steel Robert G. D., Torrie James H: Principles and Procedures of Statistics. Capítulos 19 y 20. Primera edición; McGraw-Hill Book Company, Inc, 1960.

7 Prueba de Hipótesis.

En esta sección se utilizan los archivos:
E07_Prueba_Hipótesis_P01.pps;
E07_Prueba_Hipótesis_W01.doc;
E07_Prueba_Hipótesis_X01.xls.

7.1 Menú.

Introducción:
La hipótesis Estadística.
Ejemplo 1: Del Oftalmólogo: proporciones y porcentajes.
Análisis de las Consecuencias del planteamiento.
Errores y probabilidades de los errores.
La Herramienta para Controlar los Errores.
Pruebas de Hipótesis en Variables Continuas: Muestreo.
Prueba de Hipótesis en Experimentos Planificados.
Prueba de Hipótesis en Control de Procesos y Calidad.
La Carta de Control Estadístico.
Abrir Hoja Electrónica: Charla con los Ejemplos.

7.2 Definición de Hipótesis

Suposición previa con respecto a una situación desconocida cuya verdad está, por consiguiente, sujeta a investigación por un método adecuado, ya sea por deducción lógica de consecuencias que se puedan comprobar, por investigación experimental directa o por descubrimiento de hechos no conocidos hasta ahora y sugeridos por la hipótesis.

En la hipótesis tiene significado indagar sobre la verdad objetiva de esta, y el interés principal al hacer una hipótesis es: descubrir una que sea verdadera.

A menudo sólo es posible estimar la verdad de una hipótesis sobre una base de probabilidad.

7.3 Experimento.

Es la prueba de una hipótesis en condiciones controladas.

Es uno de los instrumentos característicos del científico.

Le permite hacer preguntas a la naturaleza y recibir contestación a las mismas.

Estas contestaciones le plantean nuevos problemas, cuyas soluciones requieren experimentos más complicados en busca de diferencias menores, mediante técnicas perfeccionadas, planes detallados y mejor análisis de resultados.

7.4 Los resultados experimentales.

Los resultados experimentales suelen expresarse numéricamente.

Se emplean métodos estadísticos para resumir los datos en estimas de magnitud de tales efectos, promedios, varianzas, y relaciones entre variables medidas.

Junto con estas pruebas se emplean las pruebas de significación, que permiten al investigador profundizar más allá de las muestras y hacer inferencias acerca de los parámetros de las poblaciones de las cuales se ha extraído la muestra.

La estadística ayuda al investigador a comprender el grado de imperfección de sus datos experimentales.

7.5 La Hipótesis Estadística.

Una hipótesis es una proposición concreta que se prueba para ayudar a esclarecer el objetivo del experimento.

La proposición debe concretarse de manera que tenga dos únicas salidas:

La que se valorará denominada hipótesis nula identificada como *Ho*:

Y la hipótesis alternativa identificada como *Ha*.

La prueba indicará uno de dos únicos resultados:

Se acepta la hipótesis Nula *Ho*;

Se rechaza (o no se acepta) la hipótesis nula.

7.6 La hipótesis nula y la naturaleza.

Al plantear una hipótesis debe considerarse que se está haciendo una pregunta a la naturaleza, quién va a 'responder'.

Al valorar la hipótesis mediante los resultados experimentales el investigador aceptará o rechazará la hipótesis. Conclusiones que pueden estar o no acordes a lo qué sucede en la naturaleza. En esta prueba pueden surgir cuatro situaciones:

1.- Aceptar una hipótesis nula acorde al estado natural del material experimentado;

2.- Aceptar una hipótesis nula diferente al estado natural;

3.- Rechazar una hipótesis nula acorde al estado natural;

4.- Rechazar una hipótesis nula diferente al estado natural.

Cuyos efectos deben ser concienzudamente estudiados antes de iniciar la experiencia.

7.7 Dos Situaciones.

En el contraste de la Hipótesis Nula con la Hipótesis Alternativa entran el juego dos situaciones primarias de acuerdo a cómo el usuario use las recomendaciones:

Los puntos 1 y 3 llevan a recomendaciones correctas. Los usuarios de tales recomendaciones habrán obtenido lo que se les anuncia;

Los puntos 2 y 4 llevan a recomendaciones que cuando los usuarios las pongan en práctica se encontrarán con resultados equivocados.

¿Cuánto daño ocasiona al investigador o al usuario estas recomendaciones equivocadas?

El esquema tradicional de estas cuatro situaciones se muestra en la siguiente diapositiva.

7.8 Esquema de una Prueba de Hipótesis.

Definición:

Error I; Aceptar una hipótesis nula que es falsa; la estadística los controla con probabilidad α.

Error II; Rechazar una hipótesis nula que es verdadera; la estadística lo controla con probabilidad β.

Hipótesis: Ho; A = B	**Respuesta de la Naturaleza**	
Conclusión de la prueba.	**A = B**	**A ≠ B**
Aceptar Ho: *a* = *b* .	**No hay Consecuencias**	**Se Comete Error I** **Probabilidad α**
Rechazar Ho: *a* = *b*	**Se Comete Error II** **Probabilidad β**	**No Hay Consecuencias.**

7.9 Ejemplo 1. Del Oftalmólogo.

Un doctor especialista en oftalmología estudia el efecto de dos técnicas de cirugía para remover cataratas con respecto a la opacidad de la cápsula del cristalino que se llega a presentar después de la cirugía. Estas técnicas se denominarán ***A*** y ***B*** con proporciones o porcentajes $\boldsymbol{P_1}$ y $\boldsymbol{P_2}$ de casos de ojos sanos y $\boldsymbol{Q_1}$ y $\boldsymbol{Q_2}$ de casos de ojos con opalescencia que son o serían reales. Estos parámetros serán estimados mediante las proporciones o porcentajes $\boldsymbol{p_1}$ y $\boldsymbol{p_2}$ resultantes de la exploración de casos mediante un muestreo —interesan los casos sin opacidad o positivos.

7.10 El Objetivo del Cirujano.

Consideraciones. El doctor piensa que cuando se operan cataratas con la técnica ***B*** la proporción de ojos que presentan opalescencia después de la cirugía es menor que cuando se operan con la Técnica ***A***. Las técnicas la ha señalado de acuerdo a su aparición en la cirugía de ojos, esto es, apareció primero la técnica ***A*** y posteriormente la técnica ***B***.

Costo: La técnica ***B*** es más costosa para el paciente pues requiere equipo especial.

El Objetivo del cirujano es probar sin lugar a dudas que los pacientes que se operan con la Técnica ***B*** tienen menos probabilidades de presentar opalescencia que cuando se operan con la Técnica ***A***.

7.11 Las Hipótesis del ejemplo del oftalmólogo.

Para simplificarse los cálculos el cirujano estableció la hipótesis en término del problema como qué:

La Técnica ***A*** = Técnica ***B***.

El siguiente paso consistió en determinar qué sus datos eran de tipo Cualitativo. Pues los pacientes operados presentaban o no, la opalescencia.

En seguida debió elegir el parámetro que le indicaría, sin dudas, las diferencias entre las técnicas. Puedo optar por estudiar el número de individuos, la proporción o el porcentaje que presentaban opalescencia. Optó por usar las proporciones. De manera resumida usando la forma simplificada:

Ho; $P_A = P_B$; contra *Ha*; $P_A < P_B$.

7.12 Análisis de las Consecuencias.

Hipótesis: Ho; $P_A = P_B$	**Respuesta de la Naturaleza**	
Conclusión del cirujano.	**Técnica A = Técnica B**	**Técnica A < Técnica B**
Aceptar Ho: $p_A = p_B$ .. Recomendar que las técnicas son iguales.	**No hay Consecuencias**	**Los pacientes no se beneficiarán con la nueva técnica y el cirujano será criticado cuando otros investigadores reviertan la conclusión, pues seguramente se seguirá investigando.**
Rechazar Ho: $p_A = p_B$ Recomienda la técnica *B* basándose en las pruebas.	**Los pacientes estarán pagando más por una técnica que en realidad no los beneficia (usualmente no se dan cuenta). El médico será criticado por hacer una recomendación errada, siempre que haya nuevos experimentos, de otra forma, la técnica *B* nunca se usará por el hecho de ser más costosa.**	**No hay Consecuencias.**

7.13 Probabilidad para los Errores Estadísticos..

Después de prever las consecuencias de formular recomendaciones que no van a estar de acuerdo a lo que ocurrirá en la práctica el investigador, en este caso el doctor oftalmólogo, debe acotar la probabilidad con la que está dispuesto a aceptar errores.

Debe tomar en cuenta que entre más estricto sea con la probabilidad de cometer errores más costoso será el experimento.

El error de aceptar una hipótesis nula que es falsa o error I se controla con probabilidad α, conocida también como probabilidad se *significación*, e indicará la proporción o porcentaje de fallas que el investigador está dispuesto a aceptar.

El error de rechazar una hipótesis nula que es cierta o error II se controla con probabilidad β, también conocido como *precisión*, e indica la proporción o porcentaje de aproximación al parámetro que se estima.

Ambas probabilidades se aprecian perfectamente en el Intervalo de Confianza.

El intervalo de confianza está definido por:

$$\Pr\left\{ p - z\sqrt{\frac{pq}{n}} \le P \ge p + z\sqrt{\frac{pq}{n}} \right\} = \alpha$$

$$\Pr\{p - \text{precisión} = d \le P \ge p + \text{precisión} = d\} = \text{confianza}$$

7.14 El Error Típico de la Proporción.

El intervalo de confianza es una herramienta de la estadística inferencial que permite estimar un parámetro dentro de ciertos límites y bajo una probabilidad de acertar. Se puede usar si la distribución de los estimadores es normal o se está trabajando con promedios —ver teorema central del límite— y con una cantidad suficiente de datos.

En este problema los parámetros son proporciones cuyos estimadores se distribuyen normal alrededor del parámetro, sí la cantidad de datos es grande. El estimador del promedio es *p*. Y el error típico de las proporciones se define por:

$$S_p = \sqrt{\frac{pq}{n}}$$

El error típico es la desviación estándar de estimadores, esto es, de una cantidad determinada de proporciones p_i obtenidos de varios muestreos en la misma población. Note que depende de n, que es el tamaño o las unidades que se han inspeccionado en cada muestra.

7.15 El Intervalo de Confianza.

El intervalo de confianza para la proporción *P* esta definido por:

$$\Pr\left\{\hat{p} - z\sqrt{\frac{pq}{n}} \le P \ge \hat{p} + z\sqrt{\frac{pq}{n}}\right\} = \alpha; \text{ o } \Pr\left\{\hat{p} - z\sqrt{\frac{pq}{n}} \ge P \le \hat{p} + z\sqrt{\frac{pq}{n}}\right\} = 1 - \alpha$$

Y

$$d = \left|\pm z\sqrt{\frac{pq}{n}}\right|$$

Es la precisión.

En la primera ecuación: se asegura que el parámetro *P* estará fuera de los límites de ±*d* con una probabilidad α.

En la segunda ecuación: se asegura que el parámetro *P* estará dentro de los límites ± *d* con una probabilidad 1 − α o *probabilidad del Intervalo de Confianza.*

7.16 La Probabilidad α.

En el intervalo de confianza es evidente que los elementos dentro del paréntesis están sujetos a la probabilidad α.

Esta la define el investigador, usualmente es un valor inferior a 0,05 o 5%. Entre menor sea este valor, la probabilidad de cometer el error I se reduce.

Y esta probabilidad define el valor de la variable estandarizada ***z*** para una distribución normal estándar. Puesto que los estimadores ***p*** pueden ser menores, iguales o mayores que *P* el intervalo de confianza es una prueba de *dos colas*, en donde se estima una probabilidad $\frac{\alpha}{2}$ para cada cola de la distribución Normal Estándar.

Por ejemplo para un nivel de significación del 5%, el valor $\mathbf{z}_{(0,025)} = -1{,}96$.

7.17 La Probabilidad β.

La probabilidad β define el grado de aproximación del estimador al parámetro. En el intervalo de confianza este valor depende de *n* y de la magnitud de *p* en la siguiente ecuación:

$$d = \left|\pm z\sqrt{\frac{pq}{n}}\right|$$

La única aproximación del parámetro es ***p.*** Para tener una medida porcentual de la aproximación del intervalo de confianza, bastaría obtener el estimado:

$$B = \frac{d}{\hat{p}}$$

Como estimación de la probabilidad *β.*

7.18 Intervalo Confiable 99% para Técnica A.

El cirujano decidió utilizar un nivel de confianza de 1%, éste determina un valor $\boldsymbol{z}_{(0,01)} = -2,5758$. El intervalo de confianza para la técnica ***A*** es:

$$\Pr\left\{0,8484 - 2,5758\sqrt{\frac{0,8484 \times 0,1516}{2.526}} \geq P_A \leq 0,8484 + 2,5758\sqrt{\frac{0,8484 \times 0,1516}{2.526}}\right\} = 0,99$$

$$\Pr\{0,8300 \geq P_A \leq 0,8668\} = 0,99$$

La diferencia entre parámetro y estimador se estima en:

$$d = |\hat{p}_A - P_A| = \left|z\sqrt{\frac{pq}{n}}\right| = \left|2,5758\sqrt{\frac{0,8484 \times 0,1516}{2.526}}\right| = 0,0184$$

La probabilidad de aproximación de ***B*** como aproximación de β;

$$B = \frac{d}{\hat{p}_A} = \frac{0,0184}{0,8484} = 0,0217$$

7.19 Intervalo Confiable 99% para Técnica B.

El Intervalo de confianza:

$$\Pr\left\{0,8990 - 2,5758\sqrt{\frac{0,8990 \times 0,1010}{1.861}} \geq P_B \leq 0,8990 - 2,5758\sqrt{\frac{0,8990 \times 0,1010}{1.861}}\right\} = 0,99$$

$$\Pr\{0,8810 \geq P_B \leq 0,9170\} = 0,99$$

La diferencia entre parámetro y estimador se estima en:

$$d = |\hat{p}_A - P_A| = \left|z\sqrt{\frac{pq}{n}}\right| = \left|2,5758\sqrt{\frac{0,8990 \times 0,1010}{1.861}}\right| = 0,0180$$

La probabilidad de aproximación de ***B*** como estimación de β;

$$B = \frac{d}{\hat{p}_A} = \frac{0,0180}{0,8990} = 0,0200$$

7.20 Comparación mediante Intervalos de Confianza 99%

El estudiante podrá percatarse que el límite superior de la Técnica ***A*** con un porcentaje estimado en 86,68% está por abajo del al límite inferior para la técnica ***B*** de 88,10%.

Esto significa que los límites de confianza 99% no se traslapan, por tanto podría rechazarse la hipótesis *Ho*; $P_A = P_B$ a favor de la hipótesis alternativa *Ha*; $P_A < P_B$.

Sin embargo, es preferible usar la prueba directa involucrando las dos proporciones en una hipótesis conjunta como:

Ho; $(P_A - P..) - (P_B - P..) = P_A - P_B = 0$

Contra

Ha; $P_A \neq P_B$.

En donde la diferencia estandarizada se distribuye como ***z***.

7.21 La Diferencia Mínima Significativa.

Para comparar dos proporciones o promedios se aprovecha el hecho que la diferencia de las variables estandarizadas como las proporciones de distribuye Normal Estándar. Esto es:

$$z_d = \frac{P_A - P_B}{S_d}$$

En Donde

$$S_d = \sqrt{\overline{pq}\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}$$

O bien, usando el criterio de *Diferencia Mínima Significativa*:

$$DMS = z_\alpha S_d$$

Donde $\mathbf{z}_\alpha = 2{,}5758$ es el valor de **z** para el nivel de confianza $\alpha = 0{,}01$ que se está utilizando en este estudio. El criterio de decisión dice: Sí el valor absoluto de la diferencia $p_A - p_B$ es mayor o igual a *DMS* debe rechazar la hipótesis nula.

7.22 La prueba y el resultado.

Conociendo las fórmulas y el criterio se procede con la prueba. La desviación típica de las diferencias es:

$$S_d = \sqrt{(0{,}8698 \times 0{,}1302)\left(\frac{1}{2.526} + \frac{1}{1.861}\right)} = 0{,}0103$$

El valor estandarizado de la diferencia es:

$$z_d = \frac{p_A - p_B}{S_d} = \frac{0{,}8484 - 0{,}8900}{0{,}0103} = \frac{-0{,}0506}{0{,}0103} = -4{,}9229$$

La probabilidad se da en la cola izquierda puesto que la proporción de ojos sin opalescencia de la técnica ***B*** es mayor que con la técnica ***A***.

$$F(z = -0{,}0506) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-0{,}0506} e^{-\frac{1}{2}\left(\frac{-0{,}0506}{0{,}0103}\right)^2} dx = 4{,}27E-07$$

La probabilidad de que sean iguales es 0,000000427. Y la DMS.

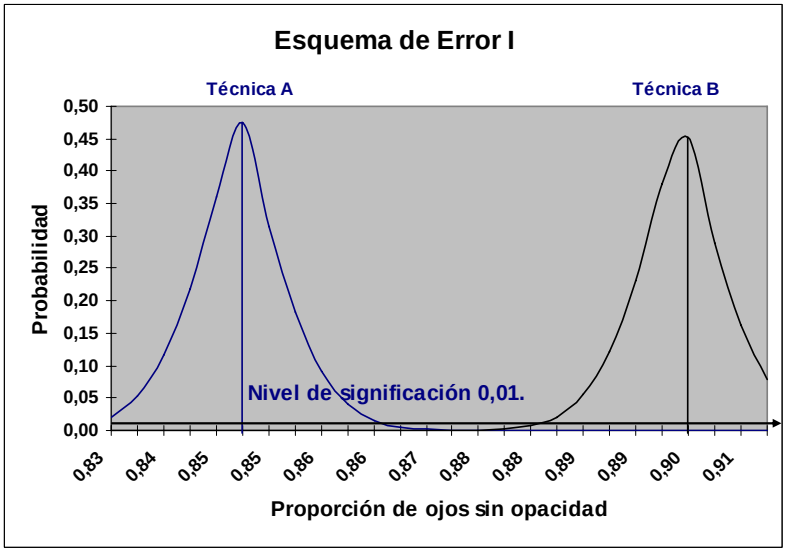
$$DMS = \left|-2{,}5758 \times 0{,}0103\right| = 0{,}0265$$

Puesto que 0,0506 > 0,0265 debe rechazarse *Ho*; $P_A = P_B$ con diferencias altamente significativas ($\alpha = 0{,}01$).

7.23 El error I y Probabilidad α.

Como se mencionó la probabilidad para protegerse del error I la define el investigador. Esta opera sobre la distribución de los estimadores en este caso las proporciones

La línea roja (se tira desde el 1% del eje Y paralela al eje *X*) indica el nivel de significación para la prueba de proporciones. Si las colas se tocan por debajo de esta línea la hipótesis debe rechazarse y la distribución más



a la derecha será superior a la otra. El nivel de la prueba asegura que una de cada 100 resultará diferente.

7.24 La Parte Fija del Error II.

Explicar el error II es más complicado y se requiere un punto de referencia. Una prueba es válida si las poblaciones que se comparan provienen de una misma población que han estado sujetas a efectos de factores que las alteran. Por esto, se usa la media general como punto de referencia.

La magnitud α que se permita al error I tendrá consecuencias inversamente proporcionales sobre el error II. La *DMS* es el criterio en la prueba que contrasta dos proporciones considerando la parte fija del error II.

$$DMS = z_{\alpha} S_d = 0{,}0265$$

Estos ejemplos no dejarán lugar a duda: para $\alpha = 0{,}05$, $\boldsymbol{z} = 1{,}96$; para $\alpha = 0{,}01$, $\boldsymbol{z} = 2{,}58$.

7.25 La Parte Variable del Error II.

La componente variable del error II es el diferencial que indica la aproximación del estimador al parámetro: $p - P$. Considerando que $P = 0{,}8698$ la proporción promedio y que el estimador es $p_A = 0{,}8484$

$$\beta_A = \frac{|p_A - P|}{P_A} = \frac{|0{,}8484 - 0{,}8698|}{0{,}8484} = 0{,}0253 \text{ ó } 2{,}53\%$$

Alrededor de P_A. Al ser inferior a 2,65% de la *DMS*, deberá usarse P. Para el estimador P_B:

$$\beta_B = \frac{|p_B - P|}{P_B} = \frac{|0{,}8990 - 0{,}8698|}{0{,}8990} = 0{,}0324 \text{ ó } 3{,}24\%$$

Alrededor de P_B. Al ser mayor a 2,65% de la *DMS*, deberá decirse que el factor ***B*** ha producido cambios tales en la población que ha creado una población bien diferenciada, o nueva.

7.26 Esquematizando el Error II.

El error II ocurre al rechazar una hipótesis que es verdadera (*Ho*; $P_A = P_B$.)

En la población ***A*** el límite inferior para la prueba es de $p_A - \text{DMS} = 0{,}8484 - 0{,}0265 = 0{,}8219$. Por tanto, las proporciones que no son consideradas vienen desde menos infinito a este punto, con una probabilidad de 0,4333 o 43,33% de cometer error II.

Puesto que el área que queda fuera del intervalo confiable obtenido con la técnica A queda a la izquierda de la media general, la probabilidad de que no se consideren valores de la población cuando se comete error II, o se rechaza una hipótesis nula verdadera es de:

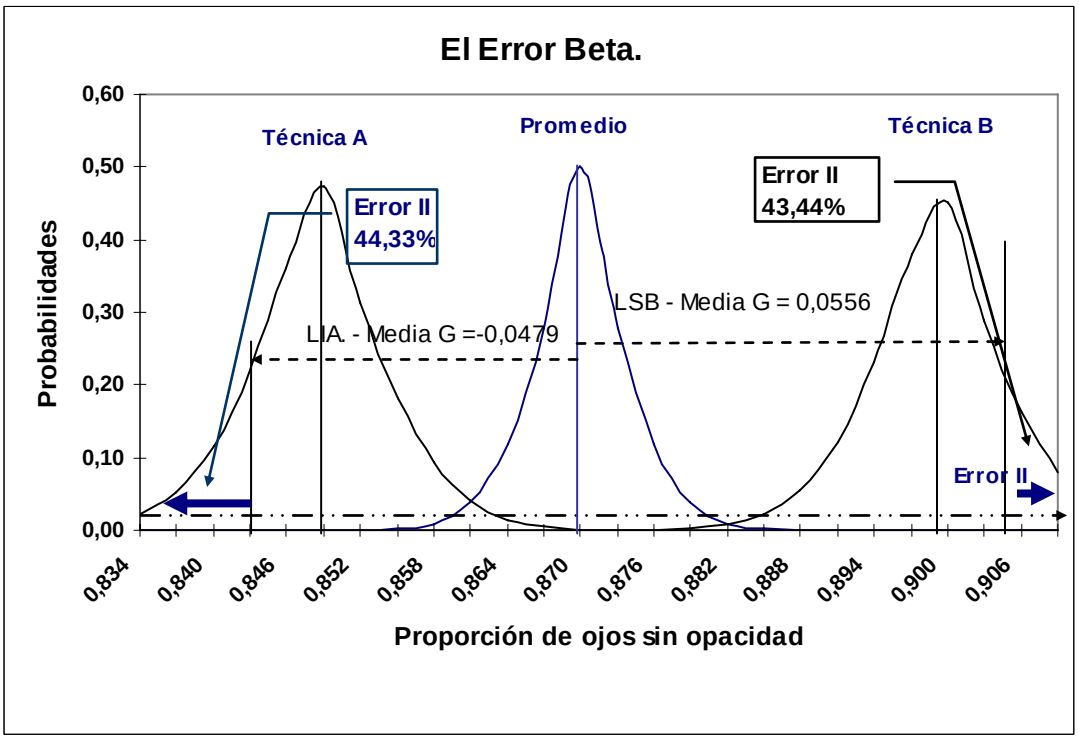
$$F(z = -0{,}1425) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-0{,}1425} e^{-\frac{1}{2}\left(\frac{0{,}8219 - 0{,}8698}{0{,}3365}\right)^2} dx = 0{,}4433$$

En la población ***B*** el límite superior para la prueba es $p_B + \text{DMS} = 0{,}8990 + 0{,}0265 = 0{,}9255$. Por tanto, las proporciones que no son consideradas van desde este punto hasta más infinito, con una probabilidad de 0,4344 o 43,44%.

En este caso la media de la población B se ubica a la derecha de la media general que se está considerando como parámetro de la población. En el contraste de la Población A con la Población B, las proporciones que no se consideran se ubican por arriba del punto que se obtiene al sumar a la proporción de la población B más la DMS hasta más infinito.

$$F(z=0{,}5656)=1-\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+0{,}5656} e^{-\frac{1}{2}\left(\frac{0{,}9255-0{,}8698}{0{,}3365}\right)^2} dx = 1-0{,}5656 = 0{,}4344$$

En general, en una comparación, a medida que la media que se compara se aleja de la media general la probabilidad de cometer error II aumenta.



7.27 *La Herramienta para controlar los Errores.*

Nuevamente, a partir del intervalo de confianza y de poblaciones con distribución normal se ha deducido la herramienta para controlar el error II llegando a:

Deducción de la fórmula del tamaño de muestra a partir de un intervalo de confianza operado para un lado de la desigualdad.

$$\left\{(p-P)=z\sqrt{\frac{pq}{n}}\right\}=\alpha$$

Operando en el interior del paréntesis de corchetes se despejará n:

$$\left\{(p-P)=z\frac{\sqrt{pq}}{\sqrt{n}}\right\}=\alpha$$

Multiplicando ambos miembros por: $\dfrac{\sqrt{n}}{(p-P)}$

$$\left\{\sqrt{n}=\frac{z\sqrt{pq}}{(p-P)}\right\}=\alpha$$

Elevando al cuadrado ambos miembros de la igualdad:

$$\left\{ n = \frac{z^2 pq}{(p-P)^2} \right\} = \alpha$$

Finalmente, teniendo presente que α el nivel de significación es únicamente la referencia para la variable z de la distribución Normal Estándar, se obtiene la ecuación para calcular el tamaño de muestra para poblaciones grandes.

$$n = \frac{z^2 pq}{d^2}$$

En donde **z** controla el error I, o error de significación y d_2 es el diferencial de aproximación que controla el error II.

El cirujano quiere estar seguro que con un nivel de confianza de 99%, las proporciones que representan los ojos sin opacidad no difieran más de 0,025 o 2,5%. La respuesta a:

¿Qué cantidad de ojos intervenidos para retirar la catarata de los ojos se deben considerar?

$$n = \frac{-2{,}5758(0{,}5 \times 0{,}5)}{0{,}025^2} = 2.654$$

7.28 La Precisión y el tamaño de muestra n.

El uso de la máxima variación posible para proporciones que ocurre cuando p y q son iguales o $p = 0{,}5$. El cirujano se percató que con las 4.387 sus resultados eran correctos. Poco después le entro la duda preguntándose ¿sí debía considerar cada grupo por separado? no cumpliría para ninguno de los dos grupos. Consultó a un estadígrafo quién le dijo: Obtenga la precisión con los datos que obtuvo. Así para el grupo operado con la técnica ***A***.

$$d_A = \left| 2{,}5758 \sqrt{\frac{0{,}8484 \times 0{,}1516}{2.526}} \right| = 0{,}0184 \text{ ó } 1{,}84\%$$

Para el grupo ***B***:

$$d_B = \left| 2{,}5758 \sqrt{\frac{0{,}8990 \times 0{,}1010}{1.861}} \right| = 0{,}0180 \text{ ó } 1{,}80\%$$

Cantidades que se vuelven a estudiar pero en otro contexto. Ambas resultaron inferiores al 2,5% designado como límite para la prueba.

Nota: La variación máxima se utiliza como alternativa a un muestreo piloto. Ver Técnicas de Muestreo.

7.29 Prueba de Hipótesis en Variables Continuas.

Usualmente, el tema de Prueba de Hipótesis se hace basándose en variables continuas con distribución normal usualmente distribución de promedios soportándose en el *Teorema Central del Límite*, generalmente en problemas de muestreo haciendo uso, como en el problema anterior, del Intervalo de Confianza.

En las siguientes diapositivas se abordarán dos circunstancias muy relacionadas con la Prueba de Hipótesis:

Los errores de decisión en las variables continuas que poco diferirán de lo hasta aquí tratado;

Y los errores de decisión en un Muestreo Completo al Azar.

7.30 El Problema de Muestreo.

La empresa Empacadora del Valle Verde, S. A. Ubicada en Cartago Costa Rica, tienen que programar sus compras de guisante (chícharo o arveja) para cumplir con los compromisos adquiridos del producto enlatado. Encarga al ingeniero agrónomo responsable del programa de asistencia a los agricultores indique a departamento de compras la cantidad de arvejas que debe adquirir fuera de la zona.

Las condicionantes del problema son las siguientes:

- Los compromisos de la empresa están presupuestados en 50.000 toneladas más o menos 5.000 Toneladas;
- Al programa se han asociado 570 agricultores con 25 hectáreas sembradas de arveja por agricultor.
- Es el primer año que opera el programa y no se sabe el rendimiento que se pueda tener.

7.31 Muestra de estimación.

El agrónomo encargado del proyecto decide efectuar una muestra aleatoria de estimación de 20 agricultores. Los elige al azar de la lista, toma su vehículo y los visita para obtener una estimación del rendimiento. En conjunto y revisando el sembradío que se encuentra en floración obtienen una estimación del rendimiento con la que el Ingeniero efectúa su proyección.

Muestra	Agricultor	qq / Ha.	Kg / Ha
1	534	128	2.560
2	494	112	2.240
3	414	138	2.760
4	248	105	2.100
5	27	113	2.260
6	299	142	2.840
7	465	121	2.420
8	321	120	2.400
9	207	157	3.140
10	501	124	2.480
11	356	149	2.980
12	420	104	2.080
13	114	149	2.980
14	170	134	2.680
15	78	150	3.000
16	551	127	2.540
17	457	140	2.800
18	285	101	2.020
19	288	115	2.300
20	353	103	2.060

7.32 La Estimación del rendimiento

El agrónomo consideró un nivel de confianza del 0,05. Por tanto, el intervalo confiable 95% para el promedio de Kg. / ha es:

$$\Pr\left\{2.532 - 1{,}96 \times 78{,}6919 \geq \overline{X} \leq 2.532 + 1{,}96 \times 78{,}6919\right\} = 95\%$$

$$\Pr\left\{2.378 \geq \overline{X} \leq 2.686\right\} = 95\%$$

Para estimar el rendimiento total y efectuar los presupuestos se obtiene el número total de Ha:

$$Total\_Ha = 25Ha \times 570A = 14.250\text{Ha}$$

Multiplicando por los promedios se obtiene el rendimiento total esperado:

$$\Pr\left\{33.883 \geq \sum_{i=1}^{14.250} x_i \leq 38.279\right\} = 95\%$$

La empresa deberá comprar producto para procesar 50.000 TM.

7.33 La precisión de la estima.

La precisión de la estima para el intervalo confiable 95% es de:

$$d = \pm z\frac{s}{\sqrt{n}} = 1{,}96 \times 78{,}6919 = 154 \text{ Kg/Ha}$$

$$B\% = \frac{d \times 100}{\overline{x}} = \frac{154 \times 100}{2.532} = 6{,}09\%$$

Que le parece alta al agrónomo (esperaba 5%). Dado que son aproximaciones de cosecha y no hay una medida real, decide usar el límite inferior para recomendar la compra de 50.000 – 33.883 = 16.117 TM de guisantes.

Considera que sería mejor comerciar el excedente, que quedarse corto en la estimación. Como ejercicio calcula el tamaño de muestra para β = 0,05.

$$n = \frac{z^2 s^2}{d^2} = \frac{1.96^2 \times 123.848{,}421}{(0{,}05 \times 2.532)^2} = 29{,}7 \equiv 30$$

Con pesos reales debería tomar el faltante para 30 muestras.

En la HE se muestran los resultados de los 570 agricultores en toneladas métricas por agricultor.

7.34 Los parámetros.

El estudiante habrá comprendido que los parámetros se obtienen considerando la cosecha de cada agricultor. Estos se resumen en el cuadro de estadísticas descriptivas.

Al agrónomo interesa el total cosechado de 36.468,369 Toneladas. La estimación se encuentra entre los límites confiables 33.883 y 38.279 toneladas. Sumando a la cosecha obtenida las 16.117 toneladas compradas la planta tendrá que procesar 52.585, Un exceso de 2.585 toneladas fácilmente comerciable.

TM	
Media	63,980
Error típico	0,384
Mediana	63,723
Moda	58,510
Desviación estándar	9,1650
Varianza de la muestra	83,9964
Curtosis	-0,0069
Coeficiente de asimetría	0,0262
Rango	55,630
Mínimo	35,814
Máximo	91,444
Suma	36.468,369
Cuenta	570

7.35 El Error de Estimación.

El promedio estimado mediante la muestra se calculo en 36.081 para estimar 36.468 toneladas. La diferencia con el peso final fue de 387 toneladas.

Los cálculos se obtuvieron de la estimación del rendimiento por hectárea en donde el promedio se calculó en 2.532 y una desviación típica de 78,6919 kilos de guisante por hectárea. El promedio final o Parámetro fue de 2.559 kilos por hectárea y la diferencia de 0,343 kilos es el error de estimación.

En este caso se puede verificar el error de estimación puesto que fue posible obtener los valores reales, pero en la mayoría de los casos esto no es posible y se debe confiar *en proyección o inferencia estadística*.

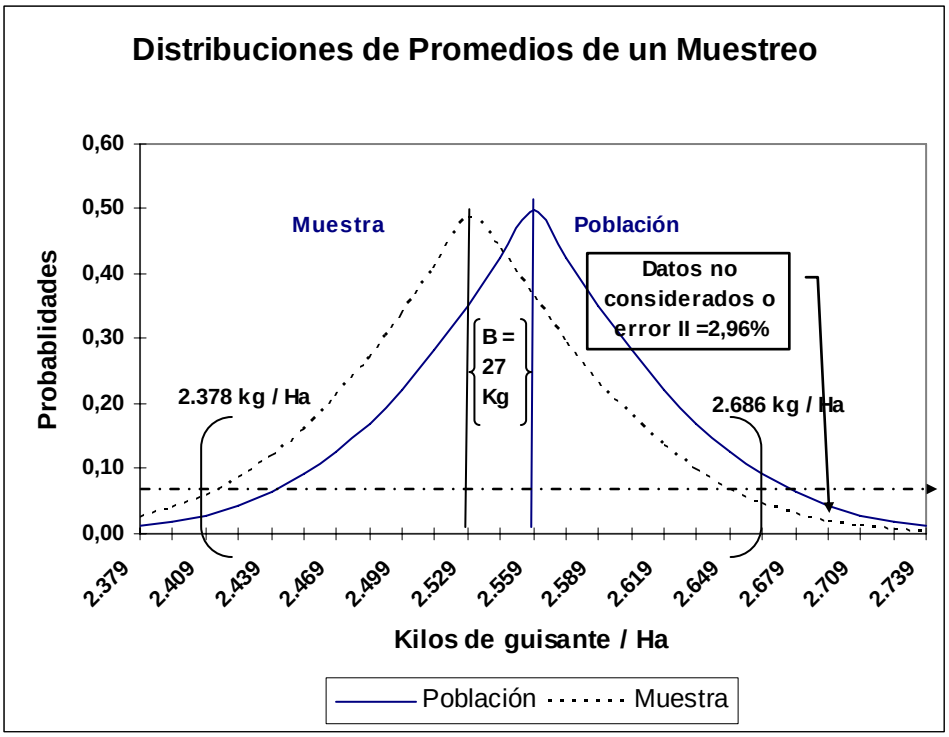
7.36 Comparando distribuciones

En este caso es notorio que la distribución desarrollada con la información de la muestra está por debajo de la distribución real de los datos.

En estas condiciones el error II se mostrará sobre los promedios a la derecha.

La diferencia de –27 kilos o 1,06% de error de estimación. En condiciones de inferencia, en donde no es posible obtener el valor paramétrico, esta cantidad es el error II también llamado error del consumidor, es además mucho más difícil de controlar.

Con la muestra habrían quedado fuera de la estimación los promedios extremos de la cola derecha puesto que el promedio de la población fue mayor, esto es, promedio superiores a 2,686 Kg. / Ha.



7.37 Prueba de hipótesis en Experimentos Planificados.

En las experiencias planificadas las inferencias que pueden hacerse de los resultados dependen de la manera como se ha realizado el experimento. Por esto, la prospectiva de una experiencia debe incluir una descripción detallada de los propósitos del mismo.

Los Objetivos del Experimento. Debe definirse claramente el propósito del experimento en términos de las cuestiones que se desea resolver, las hipótesis que se desea probar, los detalles que se desean estudiar o los efectos que se desean estimar. El plan debe considerar en qué extensión se aplicarán las inferencias deducidas de los datos.

Esto implica un análisis exhaustivo de las hipótesis para estimar las consecuencias de las recomendaciones emanadas de los resultados, tanto para el investigador como para el usuario.

7.38 El Problema de Experimentación.

Una empresa que fabrica recipientes de hoja de lata quiere homogenizar el equipo de troquelado y rolado que ha venido usando. Tiene la oferta de 5 empresas que fabrican troqueladoras y roladoras.

En la actualidad tiene 4 troqueladoras de diferentes marcas que operan cuatro cuadrillas. La gerencia de la empresa COENLA, S.A., consigue que las empresas oferentes faciliten la operación de equipos similares a los que ofrecen, para qué las cuatro cuadrillas pueden operar los equipos.

Puesto que las cuadrillas de operarios son un factor operativo necesario para que el trabajo se lleve a cabo, se planifica una experimentación con un diseño en Bloques Incompletos. Una de las variables analizadas es el consumo de las toneladas de hoja de lata en la fabricación de envases de más demanda.

7.39 El nivel de seguridad.

El objetivo de la experimentación será: *determinar la máquina que ofrezca mayor rentabilidad para la empresa*.

En Experimentación, usualmente se elige entre niveles de seguridad de 0,05 ó 5% y 0,01 ó 1%.

Debe recordarse que el nivel de confianza y la precisión son inversos. Si se quiere mucha confiabilidad la precisión baja, fenómeno que puede contrarrestarse aumentando el número de observaciones. En este experimento no es posible puesto que el número de repeticiones por tratamiento es de 4, una por cada cuadrilla, situación que orienta a usar confiabilidad de 5%.

Considerando los estudios de costo de equipo, mantenimiento y operación, durabilidad, capacidad de operación etc., la posibilidad de que la empresa haga una elección equivocada es baja, esto también encamina a usar un nivel de confianza de 5%.

7.40 Los datos de campo y El ANDEVA.

Los datos de campo o resultados en el procesado de toneladas métricas de hoja de lata por mes se muestran en el siguiente cuadro:

Maquinas	C U A D R I L L A S				Suma Máquinas
	1	2	3	4	
A	892	722	789	806	3.209
B	1.045	858	772	842	3.517
C	746	635	707	638	2.726
D	817	605	973	813	3.208
E	1.221	822	806	922	3.771
S. Cuadrillas	4.721	3.642	4.047	4.021	16.431

En ANDEVA muestra diferencias significativas entre máquinas y cuadrillas.

ANÁLISIS DE VARIANZA

Origen de las variaciones	*Suma de cuadrados*	*Grados de libertad*	*Promedio de los cuadrados*	*Estadístico de F*	*Probabilidad de F*	*Valor crítico para F*
Maquinas	153.549,70	4	38.387,4250	3,7435	**0,0336**	3,2592
Cuadrillas	120.842,95	3	40.280,9833	3,9282	**0,0364**	3,4903
Error	123.052,30	12	10.254,3583			
Total	397.444,95	19	20918,15526			

El modelo de bloques incompletos se procesa como si fuera un diseño en bloques usual. La diferencia consiste en que cada cuadrilla deberá operar a todas las máquinas y únicamente hay 4 cuadrillas. En un modelo completo, deberían haber 20 cuadrillas. EL modelo del diseño es:

$$Y_{IJ} = \mu.. + \rho_i + \tau_j + \varepsilon_{ij}$$

En donde ρ es el efecto de la cuadrilla o bloque; τ el efecto de las maquinarias y ε la variación remanente o error.

7.41 La Hipótesis Sobre Los Promedios.

Un objetivo de la experimentación de la producción de las máquinas troqueladoras es conocer ¿Qué máquinas(s) producen más? Usando como indicador la cantidad de hoja de lata sin desperdicios que se ha procesado en un mes de operación, para este fin se plantean las hipótesis:

$$Ho; \mu_i = \mu_j$$

$$Ha; \mu_i > \mu_j$$

Para toda i > j, en promedios ordenados descendentemente. La prueba se ha planteado con un nivel de confianza de 0,05.

Para efectuar los contrastes se usará la prueba de *DMS* que como se ha visto es un intervalo de confianza modificado.

$$DMS = t\left(\sqrt{Se\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}\right); o\ P\left[t_{(\alpha;gle)}\right] = P\frac{\left|\mu_i-\mu_j\right|}{\sqrt{Se\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}}$$

Utilizando el criterio de *DMS*.

$$DMS = 2{,}179\left[101{,}2638\sqrt{\frac{2}{4}}\right] = 156{,}01$$

EL criterio de definición para la prueba de *DMS* dice:

Si el valor absoluto de la diferencia de los promedios en el contraste es igual o mayor al valor de *DMS*, deberá rechazar la hipótesis de que los promedios son iguales al nivel de significación de la prueba, en este caso 0,05 0 5%.

Usando la probabilidad, primero se calcula el estadístico "t" que definirá el área de probabilidad. Para el contraste de la máquina E contra la máquina C:

$$t = \frac{922-682}{71{,}6043} = 3{,}6485$$

Y la probabilidad:

$$F(3{,}6485;12;2) = Y_0\int_0^{3{,}6485}\left(1+\frac{3{,}6485^2}{12}\right)^{\frac{12+1}{2}} dt = 0{,}0033$$

En este caso la lectura es directa, el nivel de significación alcanzado por el contraste es de 0,33%. Las diferencias entre promedios de producción pueden considerarse reales.

7.42 El Contraste de los Promedios.

La prueba para contratar los promedios se resume en el cuadro siguiente. En donde el único resultado claro es que la máquina identificada con la letra ***C*** es la que menos cantidad de hoja de lata procesa. Las diferencias entre las otras máquinas no son definitorias.

Contraste	Promedios i	Promedios j	Diferencia Absoluta	Estadístico t	Probabilidad Ho;	Resultado Contraste
E vs B	943	879	64	0,8868	0,3926	NS
E vs A	943	802	141	1,9622	0,0734	NS
E vs D	943	802	141	1,9657	0,0729	NS
E vs C	**943**	**682**	**261**	**3,6485**	**0,0033**	**
B vs A	879	802	77	1,0754	0,3034	NS
B vs D	879	802	77	1,0788	0,3019	NS
B vs C	**879**	**682**	**198**	**2,7617**	**0,0172**	*
A vs D	802	802	0	0,0035	0,9973	NS
A vs C	802	682	121	1,6864	0,1175	NS
D vs C	802	682	121	1,6829	0,1182	NS

7.43 La precisión general.

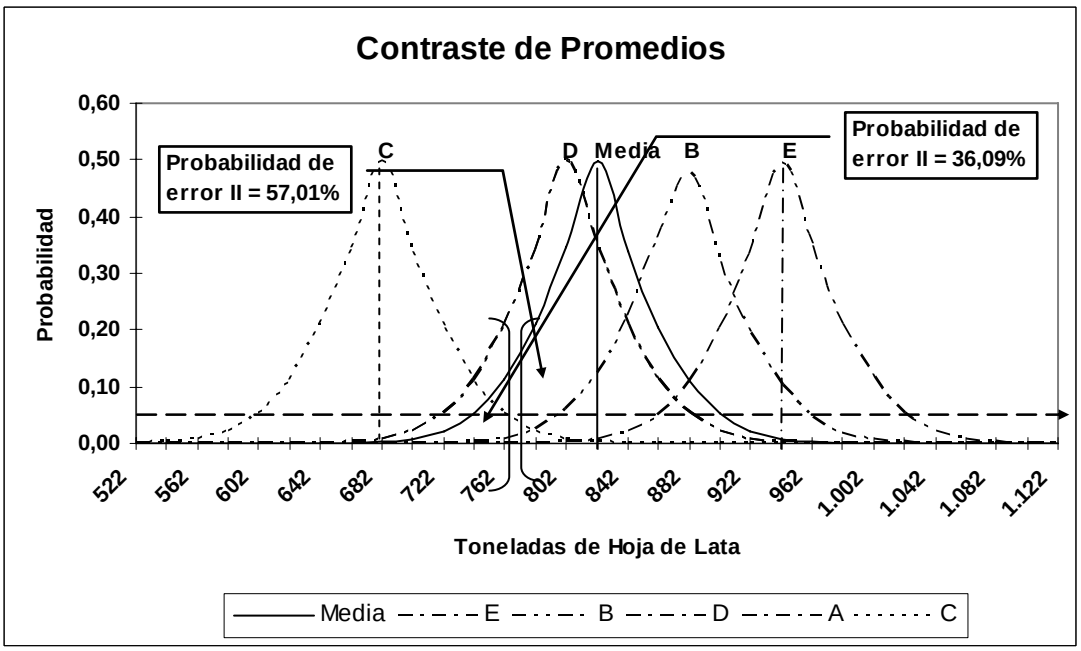
En el ANDEVA la varianza se conoce como cuadrado medio del error. La precisión general, suponiendo un promedio de 822 TM se estima en:

$$d = t\frac{S_E}{\sqrt{n_o}} = 2{,}179\frac{101{,}2638}{\sqrt{4}} = 110{,}32$$

La precisión del diseño se mantiene entre 711,23 ≥ 822 ≤ 931,87. Notará que los promedios de las Máquinas ***C*** en la cola inferior y ***E*** en la cola superior desbordan los límites. Y el promedio de la máquina ***D*** se aproxima mucho al promedio general.

En los diseños planificados se espera que algún o algunos de los niveles del factor se separe lo suficiente de la media general para crear otra población. La consecuencia es que entre más separados estén los promedios, la probabilidad de cometer error II aumenta pero el nivel de significación disminuye. Se crítica que a la prueba de *DMS* no compensa esta pérdida de precisión de la estima.

7.44 Comparando las Distribuciones de Promedios.



El gráfico muestra la posición relativa de la distribución de los promedios incluyendo el promedio general.

La línea roja que parte del eje Y en 0,05 indica el nivel de significación, cuando las curvas se cortan en o por debajo de la línea habrá diferencias significativas entre promedios.

Al cometer error II se ha rechazado una hipótesis nula que es verdadera. Para la máquina C: $\bar{y}_c + DMS = 638 + 156{,}01 = 794{,}01$ por tanto, $\beta = 1 - P(z = -0{,}1765) = 1 - 0{,}4299 = 0{,}5701$, el corchete que se abre hacia la derecha.

Para la máquina E, $\bar{y} - DMS = 922 - 156{,}01 = 765{,}99$ por tanto, $\beta = P(z = -0{,}3561) = 0{,}3609$, el corchete que se abre hacia la izquierda.

Para la Máquina C, la probabilidad de no contemplar valores del promedio real se estima en:

$$F(z = -0{,}1765) = 1 - \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{-0{,}1765} e^{-\frac{1}{2}\left(\frac{794{,}01-822}{101{,}26}\right)^2} dx = 1 - 0{,}4299 = 0{,}5701$$

Pues la distribución la producción de la máquina C se estimó, equivocadamente, por debajo de la distribución 'real', por tanto, los valores de la distribución 'real' que no se consideran en la estimación de C son aquellos superiores a 794,01 toneladas de hoja de lata.

Para la máquina E, la probabilidad de no contemplar valores del promedio 'real' se estima en

$$F(z=-0{,}3561)=1-\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{-0{,}3561} e^{-\frac{1}{2}\left(\frac{765{,}99-822}{101{,}26}\right)^2} dx = 0{,}3609$$

Puesto que la distribución de la máquina E se estimó por arriba de la distribución 'real', los valores que quedan sin estimar son aquello más bajos hasta el 765,99 en la distribución 'real'.

7.45 Prueba de Hipótesis en el Control de Procesos.

Cuando se habla de procesos industriales los errores estadísticos suelen nombrarse de manera diferente. Así, el Error I se conoce como el error para el fabricante y el error II el error para el consumidor.

En la mercadotecnia globalizada, estos errores suelen estar determinados mediante estándares de uso universal. Por ejemplo, para exportar a o producir en Europa las empresas deben sujetarse a los estándares que fija la Comunidad Económica Europea.

Estos estándares obligan al fabricante a utilizar en sus procesos de control de la calidad un nivel de confiabilidad α como probabilidad permisible al fabricante y un nivel β como probabilidad de que el producto llegue al consumidor en con una calidad determinada.

7.46 Herramienta para Control.

Una herramienta que ha demostrado ser útil en el control de los procesos y la calidad es la conocida CARTA DE CONTROL. Un gráfico que se interpreta fácilmente fundamentado en: *un muestreo acucioso, un intervalo de confianza permanente y el proceso aleatorio que debe mantener la fabricación*.

El gráfico se divide en seis zonas que permiten apreciar si el sistema se mantiene bajo control aleatorio. Esto significa que los promedios de los muestreos son independientes entre sí. Cuando aparece algún factor que altere ese devenir errático de los promedios, el sistema sale del control aleatorio para entra a una fase en que se puede determinar él o los factores que hacen que el sistema deje de ser totalmente aleatorio.

7.47 El Ejemplo de Control de Calidad.

Una empresa que se dedica a procesar lácteos y otros productos, quiere exportar a Centro América y otros países. Para esto ha establecido un control de calidad basado en las exigencias de los estándares de Panamá, los más restrictivos de la zona. Tienen dudas en cuanto a cumplir los estándares para productos envasados en cartón emplasticado como son los lácteos fluidos y jugos de frutas.

Por norma, el departamento de control de la calidad hace un muestreo de cinco envases cada determinado tiempo para cada una de las envasadoras (ver muestreo sistemático con iniciación aleatoria) en donde una de las variables que se analiza es el volumen del líquido.

Se toma un promedio de las cinco muestras y se elabora una carta de control de las últimas 30 muestras que inmediatamente es entregada al supervisor de producción para que la analice.

7.48 Los estándares de envasado.

Los estándares para productos envasados en cartón recubiertos con plástico son de:

5% de probabilidad de muestras fuera de norma:

Hasta un 0,25% por abajo del volumen declarado en el envase.

Esto significa que el nivel de confianza con el que la empresa debe entregar productos es de 95%. Para envases de un litro, esto significa para el productor que 19 de cada 20 envases debe contener más de 997,5 mililitros.

A la empresa no le conviene que los envases lleven más producto para tener un "colchón". El proceder de esta manera lea ha causado pérdidas en años precedentes.

Por tanto, las envasadoras deben llenar entre 997,5 y 1.002,5 mililitros en el 95% de los envases de 1 litro que salen a la venta.

7.49 Elaborando la carta de control

1. Determinar la carta de control;
2. Obtener la muestra, valorar y registrar en la HE por unidad;
3. Obtener estadísticos descriptivos de las últimas 30 muestras = 30 × 5 = 150 observaciones;
4. Crear los valores percentiles probabilísticos $z(2,5)$, $z(25)$, $z(50)$, $z(75)$, $z(95)$;
5. Obtener los valores que demarcan las zonas: Los valores P(2,5) y P(95) definen El Límite Inferior de Control y El Límite Superior de Control respectivamente.

$$\bar{x}_o = \vec{X} + z(p)\frac{s}{\sqrt{n}}$$

6. Acomodar los datos de manera apropiada para graficar;
7. Graficar y detallar el gráfico.

7.50 Los Valores Límites de Zonas.

Límite Inferior de Control o percentil 2,5. Además limita la zona B inferior al promedio y la Zona A inferior al mismo.

$$LSC = \overline{X} - z_{(2,5)}\frac{S}{\sqrt{n}} = 1.000 - 1{,}96\frac{2{,}727}{\sqrt{5}} = 997{,}6$$

Percentil 25, limita la zona C inferior al promedio:

$$\widetilde{x}_{25} = \overline{X} - z_{(25)}\frac{S}{\sqrt{n}} = 1.000 - 0{,}674\frac{2{,}727}{\sqrt{5}} = 999{,}2$$

Percentil 50:

$$\widetilde{x}_{50} = \overline{X} - z_{(50)}\frac{S}{\sqrt{n}} = 1.000 - 0{,}000\frac{2{,}727}{\sqrt{5}} = 1.000{,}0$$

Percentil 75, limita la zona C superior al promedio:

$$\widetilde{x}_{75} = \overline{X} + z_{(75)}\frac{S}{\sqrt{n}} = 1.000 + 0{,}674\frac{2{,}727}{\sqrt{5}} = 1.000{,}8$$

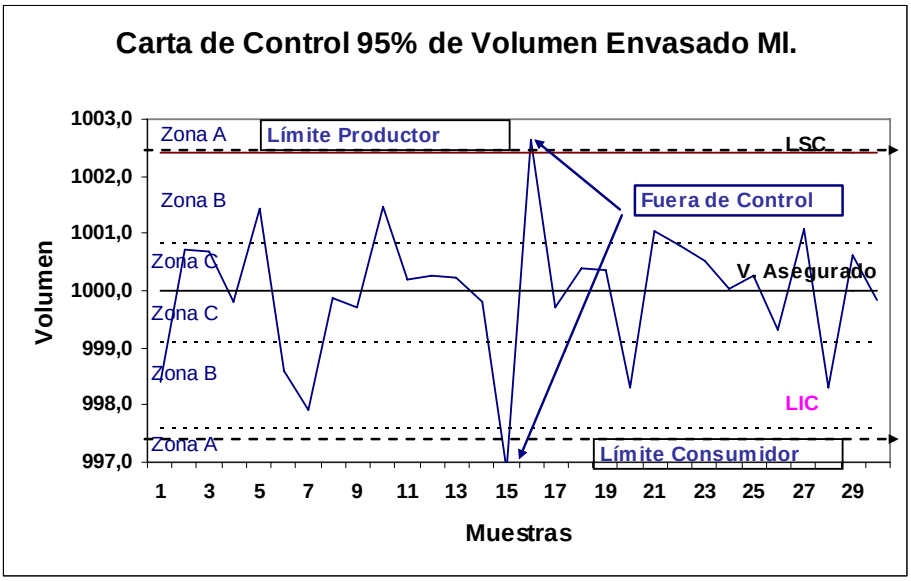
Límite Superior de Control o percentil 97,5. Además limita la zona B superior al promedio y la Zona A superior al mismo.

$$LSC = \overline{X} + z_{(97,5)}\frac{S}{\sqrt{n}} = 1.000 + 1{,}96\frac{2{,}727}{\sqrt{5}} = 1.002{,}4$$

7.51 El Grafico de Control Estadístico.

Algunas leyendas no son necesarias en una carta de control, por ejemplo el nombre de las zonas o el valor prometido en el envase.

Dos puntos salen de la zona A, uno inferior y otro superior, pero se aceptan 1 de cada 20.



No obstante, el sistema se puso fuera de control estadístico y deben investigase las causas. El sistema está fuera de control sí: ocho puntos sucesivos aparecen entre las dos zonas C; cuatro puntos sucesivos se presentan en la zona B superior o en la zona B inferior. Un punto cae en cualquiera de las zonas A.

Cuando los promedios se rebasan los límites de control las consecuencias son diferentes, si rebasa el *Límite Superior de Control* se perjudica el productor. Si los promedios son inferiores al *Límite inferior de Control* se perjudica al consumidor.

7.52 Estadísticas descriptivas.

Es indispensable que en cualquier análisis obtenga por lo menos las estadísticas descriptivas.

Siempre que esté utilizando los intervalos de confianza es muy conveniente que la distribución de los datos sea normal. Aunque para hacer inferencias sobre promedios no es indispensable.

Por ejemplo, media mediana y moda ofrecen valores muy aproximados con un corrimiento a la derecha provocando una asimetría negativa de –0,113 o cola derecha alargada aunque el valor absoluto de este menor a 0,58 valor crítico. El coeficiente de curtosis negativo de –0,129 está entre los límites –0,55 a 0,65. Ambos indican que la distribución de los datos puede considerarse normal. Entonces, puede hacerse inferencia sobre datos particulares. En el gráfico puede contar más datos arriba de la media.

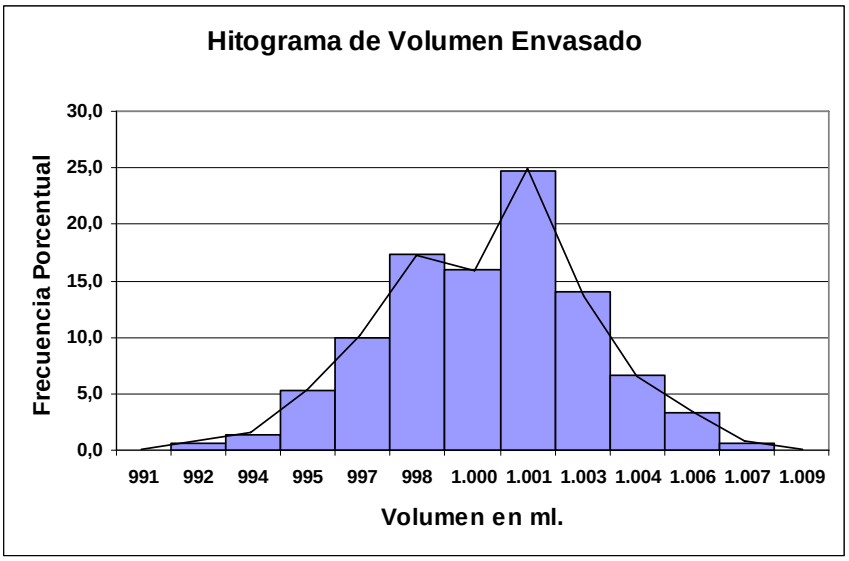
Estadísticas Descrptivas:	
Datos	150
Datos por grupo	5
Promedio	1.000,0
Desviación Típica	1,2
Mediana	1.000,3
Moda	1.001,2
Desviación estándar	2,727
Varianza	7,439
Coeficiente de curtósis	-0,113
Coeficiente de asimetría	-0,129
Máximo	1.006,4
Mínimo	992,8
Rango	1,01370
Suma total	149.995,9

7.53 Ayudas gráficas: Histogramas.

Pueden utilizarse las ayudas gráficas para percibir la forma general de la distribución de los datos.

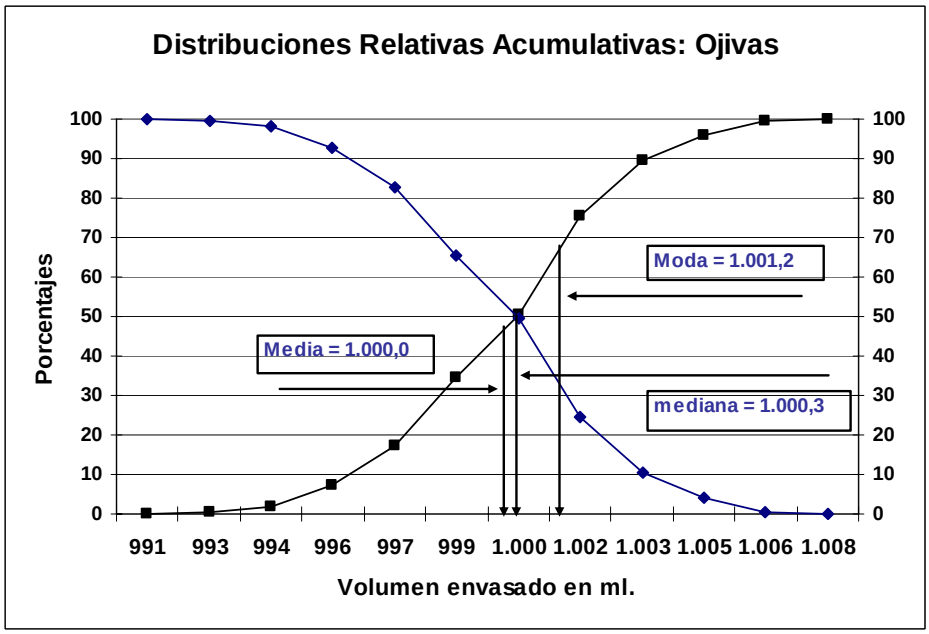
Las más conocidas son el histograma y polígono de frecuencias.

La figura confirma lo indicado en la diapositiva anterior. La posición de media, mediana y moda hacen, pensar por ejemplo, qué puede haber una deriva a la derecha provocada por ajustar la máquina para que llene un poco más los envases. Aun así la distribución de datos debe considerarse normal.



7.54 Ayudas Gráficas: Las Ojivas.

Las ojivas o distribuciones de frecuencias acumulativas permites observar la deriva modal con más precisión. Es aparente que muchos datos a la izquierda de la mediana o Percentil 50% deben compensar menos datos más pesados a la derecha de la moda.



Como se apuntó, la deriva modal indica que hay algún factor que está jalando la cúspide de la distribución a la derecha. Podría ser que cuando los operarios perciben que hay menos líquido calibran hacia arriba sin considerar que el proceso puede ser aleatorio.

7.55 Comentarios finales.

Se han explorado cuatro casos en el que la tiene que analizarse resultados de experiencias mediante procedimientos estadísticos en los que se establecen hipótesis que deben valorarse para hacer recomendaciones.

Estas recomendaciones tienen implicaciones sobre el o los investigadores y los usuarios que se han valorado como ejercicio regular en el planteamiento de cualquier investigación.

Se recomienda al estudiante tener presente el análisis de las consecuencias que tendrán las recomendaciones que deba hacer en su práctica profesional.

REFERENCIAS SELECTAS:

1. Miller Irwin, Freund John E., Johnson Richard A: Probabilidad y Estadística para Ingenieros. Capítulo 11. Traducido de la cuarta edición en ingles; Prentice-Hall Hispanoamericana, S. A. 1992.
2. Murray R. Spiegel: Serie de compendios Schaum, Teoría y Problemas de Estadística. Capítulos 7. Primera edición en español, traducido de la primera edición en ingles; Libros McGraw-Hill de México, S. A. De C. V., 1973.
3. Ostle Bernard: Estadística Aplicada. Capítulos 3 y 4. Primera edición en español traducida de la primera edición en ingles. Editorial Limusa, S. A., 1977.
4. Snedecor George W., y Cochran William G: Statistical Methods. Capítulos 2 y 3. Sexta edición; The Iowa State University, 1974.
5. Steel Robert G. D., Torrie James H: Principles and Procedures of Statistics. Capítulo 4. Primera edición; McGraw-Hill Book Company, Inc, 1960.

8 Muestreo sin restricciones en la aleatorización.

Los archivos para esta sección son:
E08_Muestreo_Irrestricto_P01.pps;
E08_Muestreo_Irrestricto_W01.doc;
E08_Muestreo_Irrestricto_X01.xls;
E08_Muestreo_Irrestricto_X02.xls;
E08_Muestreo_Irrestricto_X03.xls

8.1 Presentación.

O Muestreo Simple al Azar;

Es una técnica estadística para explorar poblaciones que permite obtener información con niveles de confianza conocidos y precisión determinada.

8.2 Los usuarios.

No sólo los investigadores de las ciencias físicas, biológicas y sociales sino los ingenieros, directores de empresas, funcionarios de estado, analizadores de mercado requieren de información confiable.

Puede ser una simple enumeración o datos muy costosos.

8.3 El problema.

Del Físico al leer un contador Geiger;
Del Ingeniero al someter a prueba un material de construcción;
Del Agrónomo al medir el rendimiento de un cereal híbrido;
Del Químico al determinar la concentración de un ácido;
Del Estadígrafo al estudiar la opinión pública sobre un candidato a solicitud de un partido político;
Del Especialista en marketing al estudiar la aceptación de un producto;
Y otros por el estilo, suelen resolverse mediante *Técnicas de Muestreo*.

8.4 La planificación.

Al planificar una muestra, debe especificarse claramente la forma en que los datos que se han de registrar satisfarán los propósitos del reconocimiento.

La población de la que deben obtenerse las muestras debe estar explícitamente definida.

El método debe ser eficiente y conducir a un análisis imparcial

Los elementos que deben incluirse en la población dependen de los propósitos del reconocimiento.

8.5 El muestreo probabilístico.

Elimina aspectos subjetivos en la elección de la muestra;

Se conocen todas las posibles muestras distintas;

La elección de la muestra se hace de modo aleatorio de acuerdo con una probabilidad preestablecida.

Y, el método de análisis está predeterminado sin ambigüedad.

Solamente con tales muestras es posible obtener Inferencias con respecto a la población que sean confiables y de precisión mesurable.

8.6 Muestreo Aleatorio Simple.

Es el método de elegir una *muestra* de *n* elementos de una *población* de *N* elementos tal que cada una estas muestras tenga igual probabilidad de ser elegida.

Se elige al *azar* el primer elemento, después, también al *azar* un segundo elemento y así sucesivamente hasta obtener los *n* elementos.

Como que un *elemento* no puede aparecer más de una vez en la *muestra*, esta es una forma de *muestreo sin remplazar* las unidades extraídas.La razón muestral o fracción muestral es $\frac{n}{N}$.

8.7 La finalidad.

Frecuentemente, la finalidad de elegir una muestra es estimar el valor medio de una característica en la población. Si y_i es el valor de la característica en la i-ésima unidad, su valor medio y varianza son: $\mu = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{N}(y_1 + y_2 + ... + y_N)$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu)^2 = \frac{1}{N}\left[(y_1 - \mu)^2 + (y_2 - \mu)^2 + ... + (y_N - \mu)^2\right]$$

Si x_i es el valor de la i-ésima unidad muestreada, su valor medio y varianza de la n unidades exploradas son: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + ... + x_N)$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_n - \bar{x})^2\right]$$

Las cantidades $\bar{x}$ y s^2 valor medio muestral y varianza muestral se llaman estadísticos muestrales, son el primero y segundo momentos muestrales y son, también, los estimadores de los parámetros de la población μ y σ^2.

Las más de las veces, desconocidos en la población y aproximados mediante los estimadores.

8.8 Los Estimadores.

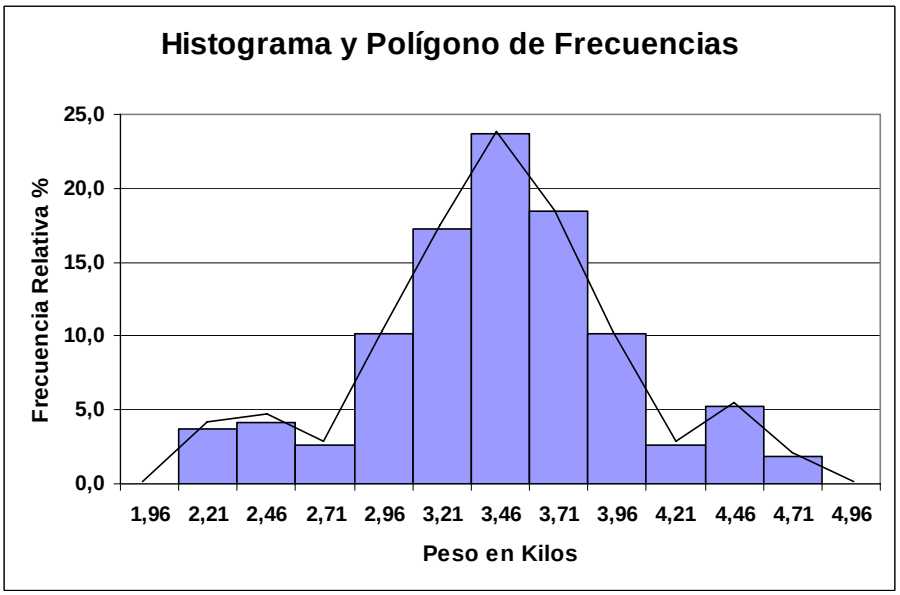
Los estimadores son variables aleatorias. Si se obtienen repetidas muestras de magnitud *n* de la población, y se calcula el promedio $\bar{x}$ de cada una, se obtendrá una población de promedios con su distribución propia que diferirá de la distribución de las observaciones *X*. La población de promedios $\bar{x}$ tendrá el mismo promedio μ que el de la población y una varianza de promedios igual a:

$$\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}\frac{N-n}{N-1}$$

La segunda fracción se aproxima a 1 cuando ***N*** es grande. O sea, que la varianza de promedios $\bar{x}$ es aproximadamente la fracción $\frac{1}{n}$ de la varianza de la población original. A medida que *n,* la magnitud de la muestra aumenta, la distribución de los promedios $\bar{x}_i$ se concentran alrededor del valor de la media poblacional μ, aumentando la precisión de la estimación del valor $\bar{x}$ como estimación del valor medio de la población.

8.9 Distribución de Frecuencias.

Las observaciones de una muestra, además de proporcionar estimaciones de los parámetros de la población, se usan también para obtener estimaciones de la función de frecuencias de la población. Esta estimación se consigue dividiendo el recorrido o rango de las observaciones muestrales en varios intervalos de largo *IC* (Intervalo de clase) y contando el número de observaciones que ocurren en cada intervalo.



Estos números se dividen por *n* para obtener las frecuencias relativas cuya suma es 1 o 100%. Que se grafican para obtener un figuras llamadas *Histograma* y *Polígono de Frecuencias.*

8.10 Una herramienta poderosa.

El Teorema central del límite: Una población definida por sus parámetros, media μ y varianza finita σ^2. Y siendo $\bar{x}$ la media de una muestra aleatoria de tamaño *n*, de esa población, la distribución de frecuencias de la variable estandarizada:

$$\delta_i = \frac{\bar{x}_i - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\sqrt{n}\left(\bar{x}_i - \mu\right)}{\sigma}$$

Se aproxima a la *Distribución de Frecuencias de la Normal Estándar* (media 0 y varianza 1) a medida que *n* crece.

Al menos la distribución de los promedios se puede aproximar mediante una *Distribución de Probabilidad* perfectamente conocida que es la *Normal Estándar* en donde, cualquier intervalo bajo la curva determina una probabilidad. Esta distribución, también aproxima convenientemente a la distribución de proporciones.

8.11 Un ejemplo que considera diferentes tipos de variables.

Dos jóvenes pasantes de Sociología pretenden efectuar un estudio sobre pesos y medidas de niñas y niños recién nacidos y sus familias en un hospital de Seguridad Social de la cabecera de una provincia de Costa Rica durante un año calendario.

Aun cuando el estudio involucró más variables, para el ejemplo interesan, el sexo, el peso y la estatura de los recién nacidos unitocos (un producto) tomados directamente de los registros del hospital y, Nivel Económico, Hábitos Higiénicos y Hábitos nutricionales obtenidos mediante entrevista, de la muestra de las familias seleccionadas para el estudio.

El peso en kilogramos ejemplifica la inferencia de variables continuas, la estatura de variables discretas y el sexo de los infantes y las variables familiares, de atributos.

8.12 Uso de los archivos del ejemplo.

Por la magnitud de las bases de datos que se operan es necesario distribuir la carga en tres Libros Electrónicos llamados:

E08_Muestreo_Irresticto_X03.xls archivo que contiene el generador de datos. Este archivo crea el conjunto de datos que simulan la encuesta. Estos se copian al siguiente libro.

X08_Muestreo_Irresticto_X02.xls que se utiliza para generar la muestra. *a*) copiar el archivo generado y pegar en la hoja que llamará ***Base_D*** con *Edición / Pegado especial / Valores; c*) Obtener la muestra aleatoria que deberá colocar en la hoja llamada Muestra, y; *d*) copiar la ***Muestra*** a el libro E08_Muestreo_Irresticto_X01.xls a la hoja con el mismo nombre.

E08_Muestreo_Irresticto_X01.xls en donde se pasa una copia de la muestra que se utilizará para el análisis y elaboración de los cuadros de resultados.

8.13 Delimitando los alcances de la investigación.

Objetivo primario:

Explorar a las familias de niños recién nacidos en un Hospital de la Seguridad Social en una cabecera de provincia de Costa Rica en las variables: Peso al nacimiento, Talla, Sexo, Nivel de Ingresos, Hábitos Higiénicos, Hábitos Nutricionales.

Objetivos secundarios:

Caracterizar la población de niños recién nacidos en un Hospital de la Seguridad Social en una cabecera de provincia de Costa Rica.

Población:

Los niños unitocos recién nacidos desde el 1 de enero hasta el 31 de diciembre de un año específico.

Probabilidad del muestreo:

Confiabilidad de 99% con un nivel de precisión de 2,5% sobre el promedio de peso.

8.14 Fuente de la información.

La información se obtendrá de los registros sistematizados del hospital.

El departamento de computación separó de la base de datos universal, aquella que cumpliera los requerimientos del estudio eliminando datos confidenciales. Un total de 4.591 registros que trasladó a un CD (disco compacto)

8.15 Aclaración.

El estudio consta de dos partes:

- La caracterización de la población de niños recién nacidos que se podría efectuar con precisión mediante un censo en lo tocante a peso, talla y sexo.
- La exploración mediante el muestreo de variables socioeconómicas como: Nivel Económico, Hábitos Higiénicos y Hábitos nutricionales de las familias que salgan elegidas para ser entrevistadas.

8.16 Método estadístico.

La exploración se efectuará mediante la *Técnica Estadística* de *Muestreo Simple al Azar.*

Se obtendrá una muestra preliminar de 30 unidades para determinar el tamaño de muestra que cumpla con un 99% de confianza y una precisión del 2,5% sobre el promedio de la variable de peso al nacimiento.

Se utilizará la prueba de χ^2 para la determinación de diferencias entre clases con nivele de confianza de 95%.

8.17 La Muestra Piloto.

Para obtener la muestra piloto se programa la hoja electrónica para que proporcione una secuencia de números aleatorios entre 1 y 4.591, inclusive.

El cuadro muestra los registros seleccionados, la referencia del número en el listado y el peso del niño. Las unidades seleccionadas se marcan mediante color verde claro en el listado de la población, esto con el fin de no seleccionarlas de nueva cuenta cuando se complete la muestra definitiva.

N° Muestra	Item	Peso Kilogramos	N° Muestra	Item	Peso Kilogramos	N° Muestra	Item	Peso Kilogramos
1	3231	3,199	11	1390	3,999	21	4584	4,313
2	855	4,056	12	3951	3,35	22	1428	2,399
3	2318	3,346	13	3710	3,586	23	996	4,243
4	1608	3,15	14	930	3,321	24	212	3,301
5	4572	3,098	15	3539	4,519	25	2375	2,105
6	253	3,347	16	1828	2,348	26	1485	3,464
7	336	3,838	17	544	3,36	27	853	3,341
8	237	3,156	18	3626	3,394	28	2612	3,242
9	2422	3,245	19	3381	3,779	29	3959	3,216
10	2043	3,152	20	1207	3,751	30	1808	3,127

Cuando no se tiene información sobre el promedio y la varianza de la población que se va a muestrear, el proceso consiste en seleccionar al azar un número de unidades próximas al valor que se requiere, en el ejemplo 30. El número a la izquierda se refiere al número de la muestra, el central al número de registro o item y el tercero a la variable de interés que es el peso de los bebés al nacer registrado en su hoja cínica sistematizada.

N°	1 - 10	11-20	21-30	31-40
1	3.231	1.390	4.584	2.734
2	855	3.951	1.428	1.085
3	2.318	3.710	996	3.183
4	1.608	930	212	3.579
5	4.572	3.539	2.375	2.607
6	253	1.828	1.485	2.467
7	336	544	853	2.881
8	237	3.626	2.612	3.857
9	2.422	3.381	3.959	4.450
10	2.043	1.207	1.808	2.141

Los números aleatorios generados mediante la instrucción:

=ENTERO(ALEATORIO()*4590)+1

entrada a la HE para crear números aleatorios entre 1 y 4591, el resultado se muestra en el cuadro adjunto.

Procedimiento para seleccionar la muestra:

1. ***Crear un cuadro como el inicial de este punto. En la primera columna el número de muestra, en la segunda el número que identifica la muestra en la base de daros, peso del bebé que es la variable que se usará de referencia para valorar la precisión del muestreo;***

2. ***Tomar el primer número aleatorio, en el ejemplo 3.231;***
3. ***Ubicarlo en la base de datos;***
4. ***Copiar el dato y marcar con color la muestra. Pegarlo en el cuadro de muestra;***
5. ***Marcar el número aleatorio elegido y seleccionar el siguiente.***

Recuerde que la muestra es sin reemplazo, esto significa que una muestra no puede ser seleccionada más de una vez.

8.18 El tamaño de la muestra.

En este caso, la finalidad del *Muestreo Piloto* fue determinar el tamaño de la muestra con una precisión de 2,5% sobre el peso promedio y confiabilidad de 99% utilizando la ecuación:

$$n = \frac{z^2 \times s^2}{d^2} = \frac{2{,}5758^2 \times 0{,}2886}{(3{,}3915 \times 0{,}025)^2} = 266$$

Muestras a recolectar tanto en las variables directas cono en las variables familiares. Nuevamente se generan unos 300 números aleatorios para completar la muestra a 266. En este caso, las unidades seleccionadas se marcan en azul celeste.

Varianza de la muestra:

$$s^2 = \frac{\sum_{i=1}^{20}(x_i - \bar{x})^2}{20-1} = \text{VAR(Rango)} = 0{,}289$$

Estadístico z para un nivel de confianza de 99%:

$$z_{(99)} = \text{DISTR.NORM.ESTAND.INV}(0{,}995) = 2{,}576$$

Precisión de 2,5% sobre el promedio;

Promedio:

$$\bar{x} = \frac{1}{20}\sum_{i=1}^{20} x_i = \text{PROMEDIO(Rango)} = 3{,}392$$

Precisión:

$$d = \bar{x} \times 0{,}025 = 0{,}085$$

Para poblaciones pequeñas se considera el efecto de corregir por finitud, o sea, asume que la población es pequeña y se utiliza:

$$n' = \frac{z^2 \times s^2}{d^2 + \frac{z^2 \times s^2}{N}} = \frac{n}{\left(1 + \frac{n}{N}\right)} = \frac{266}{\left(1 + \frac{266}{4.591}\right)} = 251$$

No parece que 15 muestras menos tengan mucha influencia sobre el costo de esta investigación y sí lo podría tener sobre la precisión.

El muestreo piloto sirve, también para probar los cuestionarios, el sistema de almacenamiento de datos, logística de la encuesta, preparación de encuestares, sistema de computación como ejemplo.

8.19 Completando la muestra.

El proceso de completar la muestra es idéntico al de obtener la *Muestra Piloto*. El estudiante deberá imaginar que cada registro que selección correspondería a un cuestionario. No podrá de ninguna manera alterar el orden en que aparecen los números aleatorios, sí por ejemplo ordenara ascendentemente, lo números aleatorios mayores no se elegirían. Es conveniente listarlos para facilitarse la toma de la muestra.

Al localizar la unidad que indica el número aleatorio y siempre que no haya sido seleccionada se marca con algún color de fondo, copiando toda la información de la hilera a la *Hoja Muestra* inmediatamente debajo de la muestra anterior.

El proceso se detiene cuando se han conseguido las 266muestras incluyendo las 30 de la *Muestra Piloto.*

En cuanto tenga la muestra abre el archivo E08_Muestreo_Irresticto_X01.xls. ***Copie y peque la muestra en la HE con el mismo nombre.***

Estadístico	0=Ma/1=fe	Centímetros	Kilogramos	N.E. 1	N.E. 2	N. E. 3	N.E. 4	HH. 1	HH. 2	HH. 3	NN. 1	NN.2	NN.3
Media	0,511	53,060	3,448	0,188	0,305	0,312	0,195	0,113	0,718	0,169	0,331	0,470	0,199
Error típico	0,031	0,095	0,033	0,024	0,028	0,028	0,024	0,019	0,028	0,023	0,029	0,031	0,025
Mediana	1,000	53,000	3,441	0,000	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,000
Moda	1,000	52,000	3,344	0,000	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,000
Desviación estándar	0,501	1,548	0,539	0,391	0,461	0,464	0,397	0,317	0,451	0,376	0,471	0,500	0,400
Varianza de la muestra	0,251	2,396	0,291	0,153	0,213	0,215	0,158	0,100	0,203	0,141	0,222	0,250	0,160
Curtosis	-2,013	-0,846	0,388	0,585	-1,280	-1,344	0,388	4,093	-1,058	1,159	-1,488	-2,000	0,296
Coeficiente de asimetría	-0,045	0,077	-0,086	1,606	0,854	0,816	1,544	2,462	-0,975	1,775	0,723	0,121	1,514
Rango	1	6	2,777	1	1	1	1	1	1	1	1	1	1
Mínimo	0	50	2,085	0	0	0	0	0	0	0	0	0	0
Máximo	1	56	4,862	1	1	1	1	1	1	1	1	1	1
Suma	136	14114	917,206	50	81	83	52	30	191	45	88	125	53
Cuenta	266	266	266	266	266	266	266	266	266	266	266	266	266

8.20 Estadística descriptiva.

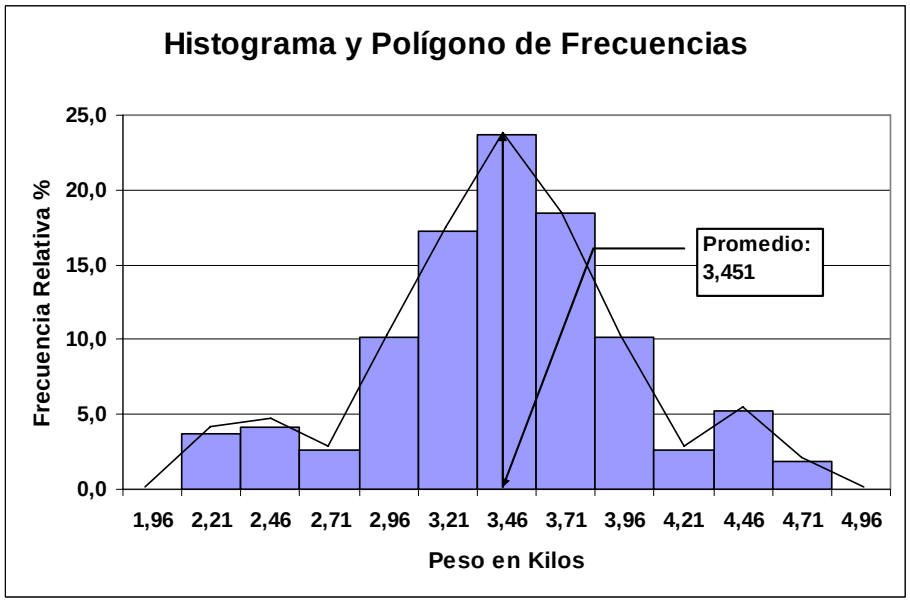
En las variables cualitativas, el promedio que deberá interpretarse como proporción es exacto, el resto de los estadísticos como la varianza, serán aproximados. Por ejemplo, para el sexo:

$$p = \bar{x} = \frac{\sum_{I01}^{266} x_i}{266} = \frac{1+0+0+...+0+1}{266} = 0{,}511$$

Para Nivel Económico 1 p_{NE1} =0,188 o 18,8%.

8.21 Métodos indirectos de comparación.

El Histograma de Frecuencias es una forma práctica de observar directamente, si las frecuencias de las muestras se distribuyen Normal (con forma de campana), o campana de Gaus.



Recuerde que para elaborar un Histograma es necesario conocer la distribución de frecuencias. Esto es, una división del recorrido o rango en una cantidad determinada de clases de igual tamaño. Una regla empírica dice que el Intervalo de Clase sea un valor de ½ a ¼ de la desviación estándar, esto es:

El coeficiente de curtosis 0,3883 indica una curva ligeramente achatada y el coeficiente de asimetría de –0,0858 una cola izquierda ligeramente más larga, siempre con respecto a una *Distribución Normal* sin que las diferencias sean significantes o importantes.

Al dividir el Recorrido de los Datos o Rango por los Intervalos de clase mínimo y Máximo se estima el número de clases mínimo y Máximo respectivamente:

$$IC_{+} = \frac{s}{2} = \frac{0{,}5390}{2} = 0{,}269\ \text{Kg.}; \quad IC_{-} = \frac{0{,}5390}{4} = 0{,}135\ \text{Kg}$$

Otra regla empírica indica que debe considerarse un número de clases entre 7 cuando hay pocos datos y 15 cuando hay muchos dejando a criterio del investigador cuánto es mucho o poco. En el ejemplo se determina un intervalo de clase de 0,250 Kg esperando unas 11 clases. Agregando la primera y la última en ceros, necesarias para representar adecuadamente la distribución de frecuencias se obtendrán 13.

$$NC_{-} = \frac{2{,}777}{0{,}269} = 10\ \text{Clases.} \quad NC_{+} = \frac{2{,}277}{0{,}135} = 21\ \text{Clases.}$$

También es conveniente que el límite inferior de la primera clase sea igual a mínimo menos el intervalo de clase:

$$LI_1 = Mínimo - IC = 2{,}085 - 0{,}250 = 1{,}835$$

El límite medio o punto medio de la primera clase se obtiene sumando al límite Inferior de la primera clase la mitad del intervalo de clase u obteniendo el punto medio del límite inferior más el mínimo:

$$\bar{x}_1 = LI_1 + \frac{IC}{2} = \frac{LI_1 + Mínimo}{2} = 1{,}835 + \frac{0{,}250}{2} = \frac{1{,}835 + 2{,}085}{2} = 1{,}960$$

El límite Superior de la primera clase es el mínimo;

$$LS_1 = Mínimo = 2{,}085$$

Para crear los límites de las clases se suma a cada uno de los límites anteriores un intervalo de clase, así para el siguiente y más, hasta que la última penúltima clase contenga al valor máximo, agregando la última.

El estudiante debe recordar que se está trabajando con una variable continua y que no hay discontinuidad entre los límites máximos de una clase y el mínimo de la subsecuente. Y como el conteo se le dejará a la HE, no es necesario separarlas en el cuadro de frecuencias. La función para la primera clase es:

$$f_1 = \text{CONTAR.SI(Muestra!\$E\$3:\$E\$268;"< 2,085")} = 0$$

Y para las subsecuentes

$$f_2 = \text{CONTAR.SI(Muestra!\$E\$3:\$E\$268;"< 2,335")} - \text{SUMA(\$D\$38:D38)} = 10$$

Variando manualmente el criterio que es el límite superior, dentro de la función lógica

En las investigaciones por muestro se prefiere presentar resultados universales, por esto es más conveniente mostrar un Histograma y Polígono de Frecuencias relativo utilizando las frecuencias relativas o porcentuales:

$$fr_2 = \frac{f_2 \times 100}{\sum_{i=1}^{c} f_i} = \frac{10 * 100}{266} = 3{,}8\%$$

Cuadro de Frecuencias.

Límtes de Claseses			Frecuencias	
Inferior	Medio	Superior	Observadas	Relativas
1,835	1,960	2,085	0	0,0
2,085	2,210	2,335	10	3,8
2,335	2,460	2,585	11	4,1
2,585	2,710	2,835	7	2,6
2,835	2,960	3,085	27	10,2
3,085	3,210	3,335	46	17,3
3,335	3,460	3,585	63	23,7
3,585	3,710	3,835	49	18,4
3,835	3,960	4,085	27	10,2
4,085	4,210	4,335	7	2,6
4,335	4,460	4,585	14	5,3
4,585	4,710	4,835	5	1,9
4,835	4,960	5,085	0	0,0
		Suma	266	100,0

También debe recordarse que las barras del Histograma se unen y el polígono de frecuencias se detalla porque los datos con de una variable continua. Los de una variable discreta o cualitativa debe graficarse con barras separadas y sin el Polígono de Frecuencias.

8.22 *Prueba de normalidad.*

Limites		Frecuencias Observadas			Frecuencias Esperadas		Desviacion		Aporte a
Inferior	Superior	Por Clase	Acumulativa	Rel. Acumul	de la Clase	Acumulativa	Absuta	F. Esperada	Chi-cuadrada
1,835	2,085	0	0	0,0000	0,0053	0,0053	0,0053	1,4	1,4139
2,085	2,335	10	10	0,0376	0,0131	0,0184	0,0191	3,5	12,1215
2,335	2,585	11	21	0,0789	0,0342	0,0527	0,0263	9,1	0,3921
2,585	2,835	7	28	0,1053	0,0720	0,1247	0,0195	19,2	7,7217
2,835	3,085	27	55	0,2068	0,1223	0,2470	0,0402	32,5	0,9380
3,085	3,335	46	101	0,3797	0,1674	0,4144	0,0347	44,5	0,0489
3,335	3,585	63	164	0,6165	0,1849	0,5993	0,0173	49,2	3,8884
3,585	3,835	49	213	0,8008	0,1647	0,7640	0,0368	43,8	0,6146
3,835	4,085	27	240	0,9023	0,1184	0,8823	0,0199	31,5	0,6401
4,085	4,335	7	247	0,9286	0,0686	0,9510	0,0224	18,3	6,9423
4,335	4,585	14	261	0,9812	0,0321	0,9831	0,0019	8,5	3,4906
4,585	4,835	5	266	1,0000	0,0169	1,0000	0,0000	4,5	0,0557
		266			Máxima desviación absoluta		0,0402	266,0	38,2679
					Criterio		**0,0834**		0,0001

Las sociólogas del estudio quieren probar que los niños de menor peso al nacer provienen de familias de pocas ventajas sociales. Para esto, necesitan conocer el peso que separe al 25% de los pesos más ligeros.

Se puede proceder de dos formas:

- Determinando el valor que separa al primer cuarto o 25% de los datos o primer cuartil;
- Utilizando la distribución Normal Estándar y separar el 25% mediante probabilidades.

Para este segundo caso, es insoslayable que la distribución de los pesos al nacimiento sea normal puesto que se hará inferencia sobre datos particulares. Por tanto, es necesario hacer una prueba que asegure que la aproximación mediante la *Distribución Normal Estándar* es posible.

8.23 *Método Directo.*

El método directo es una prueba de normalidad que compara las frecuencias esperadas con la observadas mediante una prueba que puede ser la de Chi-cuadrada o la de Kolmogorov-Smirnov conocidas como *Pruebas de Bondad de Ajuste*. Se usarán ambas.

En la prueba de K-S, puesto que la Máxima Diferencia Absoluta $d = 0{,}0402$ es menor al criterio $D_{(0{,}05;\ 266)} = 0{,}0834$ debe aceptarse que la distribución observada puede aproximarse por una Normal Estándar. La prueba de χ^2 indica diferencias que la distribución de las observaciones no debe ajustarse con una Normal.

Esta disyuntiva puede presentarse con cualquier conjunto de datos. La prueba de χ2 es más sensible a las variaciones de las clases. La prueba de K-S se está considerando más confiable por estar menos sujeta a variaciones imprevistas en una o dos clases. En este caso se seguirá este segundo criterio. Considerando pues, que la distribución de datos es normal.

Se pueden utilizar los estimados obtenidos directamente en las estadísticas descriptivas. Sin embargo, es preferible utilizar estimadores de datos agrupados puesto que el interés es la distribución de frecuencias. EL promedio o primer momento muestral se obtiene mediante:

$$\bar{\bar{x}} = \frac{\sum_{i=1}^{c} f_i \bar{x}_i}{\sum_{i=1}^{c} f_i} = \frac{0\times 1{,}960 + 10\times 2.210 + \ldots + 5\times 4.710 + 0\times 4.960}{0+10+\ldots+5+0} = \frac{917{,}860}{266} = 3{,}451$$

La Varianza o segundo momento muestral:

$$s^2 = \frac{\sum_{i=1}^{c} f_i(\bar{x}_i - \bar{\bar{x}})^2}{n-1} = \frac{0(1{,}960-3{,}451)^2 + 10(2{,}210-3{,}451)^2 + \ldots + 0(4{,}960-3{,}451)^2}{266-1} =$$

$$= \frac{75{,}7265}{265} = 0{,}2858$$

La desviación estándar es la raíz cuadrada de la varianza:

$$s = \sqrt{s^2} = \sqrt{0{,}2859} = 0{,}5346$$

El coeficiente de asimetría:

$$ca = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{c} f_i \left(\frac{\bar{x}_i - \bar{\bar{x}}}{s} \right)^3 =$$

$$= \frac{266}{265 \times 264} \left[0\left(\frac{3{,}451 - 3{,}451}{0{,}5346} \right)^3 + 10\left(\frac{2{,}210 - 03{,}451}{0{,}5346} \right)^3 + \ldots + 0\left(\frac{4{,}960 - 3{,}451}{0{,}5346} \right)^3 \right] =$$

$$= -0{,}1141$$

Valor que está dentro del intervalo criterio –0,24 a 0,24 para un nivel de confiabilidad del 5%. Este depende del tercer momento muestral. El coeficiente de curtosis o alargamiento:

$$cc = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{c} f_i \left(\frac{\bar{x}_i - \bar{\bar{x}}}{s} \right)^4 - 3\left(\frac{(n-1)^2}{(n-2)(n-3)} \right) =$$

$$= \frac{266(267)}{(265)(264)(263)} 852{,}1342 - 3\left(\frac{(265)^2}{(264)(263)} \right) = 0{,}2550$$

Valor que indica un ligero un alargamiento de la forma de la distribución de datos que se mantiene dentro del rango significativo –0,38 a 0,44. Considerando estos valores la distribución de los datos puede considerarse normal. No obstante se procede a efectuará la prueba de Bondad de Ajuste de Kolmogorov-Smirnov que compara la distribución de probabilidad acumulativa observada contre la esperada.La probabilidad acumulativa esperada se obtiene dividiendo la frecuencia acumulativa entre el total de observaciones. Así por ejemplo para la tercera clase;

$$G_3(x) = \frac{21}{266} = 0{,}079$$

Estadístico de datos agrupados

Punto Medio	Frecuencia Observada	Frecuencia * p. Medio	Desviaciones Cuadráticas	Desviaciones Cúbicas	Desviaciones Cuárticas
1,960	0	0,000	0,0000	0,000	0,0000
2,210	10	22,100	15,3909	-124,995	290,0849
2,460	11	27,060	10,7942	-69,998	129,7133
2,710	7	18,970	3,8394	-18,614	25,7889
2,960	27	79,920	6,4986	-20,871	19,1547
3,210	46	147,660	2,6629	-4,194	1,8878
3,460	63	217,980	0,0056	0,000	0,0000
3,710	49	181,790	3,2971	5,599	2,7168
3,960	27	106,920	7,0061	23,363	22,2633
4,210	7	29,470	4,0368	20,068	28,5084
4,460	14	62,440	14,2644	94,257	177,9814
4,710	5	23,550	7,9304	65,382	154,0347
4,960	0	0,000	0,0000	0,000	0,0000
					852,1342

Tamaño muestra	266	Sumas de Cuadrados	75,7265
Suma Total	917,860	Varianza	0,2858
Promedio	3,451	D. Estándar	0,5346
		C. Asimetría o sesgo	-0,1141
		C. Curtosis	0,2550

La distribución esperada para cada intervalo de clase se obtiene utilizando restando a la probabilidad Normal Acumulada del Límite Superior de la clase la probabilidad Normal acumulada hasta el Límite Inferior de la, por ejemplo para la clase 3 se obtiene mediante:

$$P(C_3)=\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{2,585} e^{-\frac{1}{2}\left(\frac{2,585-3,451}{0,5346}\right)^2} dx-\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{2,335} e^{-\frac{1}{2}\left(\frac{2,335-3,451}{0,5346}\right)^2} dx=$$

$$=\text{DISTR.NORM(B762;\$C\$752;\$F\$752;1)}$$

$$-\text{DISTR.NORM(A762;\$C\$752;\$F\$752;1)}=0,0527-0,0184=0,0342$$

Recuerde que el límite superior de una clase es igual al límite inferior de la clase siguiente como se puede comprobar en la HE. Así se obtienen las probabilidades para cada intervalo. Con excepción del primero en el que se toma la probabilidad desde infinito hasta el límite superior de la clase 1; y el último que calcula restando de 1 la probabilidad acumulativa hasta el límite inferior de la última clase. Estas probabilidades de cada clase se acumulan para obtener la probabilidad acumulada esperada.

Finalmente, se calcula el valor absoluto de la diferencia de probabilidad acumulada observada menos la probabilidad acumulada esperada, ejemplificando con la clase 3;

$$D_3=\left|G_3-F_3\right|=\left|0,0789-0,0527\right|=0,0263$$

Como criterio de comparación se toma la mayor diferencia absoluta, que corresponde a la clase 5 denominándola:

$$D=Max\left|G_i-F_i\right|=\left|0,2068-0,2470\right|=0,0402$$

Valor que se compara con el criterio de las tablas para la Prueba de Kolmogorov-Smirnov para un nivel confiable al 5%:

$$D_t=\frac{1,36}{\sqrt{n=266}}=0,0834$$

La regla dice que si el valor calculado es mayor o igual a valor tabulado la hipótesis que dice:

Ho; La distribución de aproximación o esperada es igual a la observada;

Debe rechazarse. En este caso se debe aceptar. Esta prueba de Bondad de Ajuste se está utilizando más que la de Chi-cuadrada por ser más estable a cambios bruscos en alguna clase.

La prueba de Chi-Cuadrada:

$$\chi^2_{c-1}=\sum_{i=1}^{c}\frac{\left(fo_i-fe_i\right)}{fe_i}=\frac{(0-1,4)^2}{1,4}+\frac{(10-3,5)^2}{3,5}+\ldots+\frac{(5-4,5)^2}{4,5}=$$

$$=1,4139+12,1215+\ldots+0,0557=38,2679$$

Que se valora mediante la función de densidad proporcionada por la HE y ejemplificada mediante:

$$F_{[38,2679;\,12\text{-}1]}=Y_0\int_0^{38,2679}(38,2679)^{\frac{1}{2}(12-1)}e^{-\frac{1}{2}38,2679}d\chi=0,0001$$

Probabilidad que indica que la distribución de los datos no puede aproximarse mediante una distribución normal estándar.

La diferencia entre ambas pruebas se hace patente en este ejemplo, la prueba de Kolmogorov-Smirnof indica que la distribución de frecuencias acumulativas si puede aproximarse mediante una normal; la prueba de χ^2 indica que la distribución de frecuencias absolutas no puede aproximarse mediante una normal.

En estas disyuntivas, la experiencia del investigador o del profesional en estadística es trascendente. En general significa, en el primer caso, usar la estadística paramétrica; en el segundo, la estadística de distribuciones libres. A medida que se adelante se indicará cuál se está ocupando.

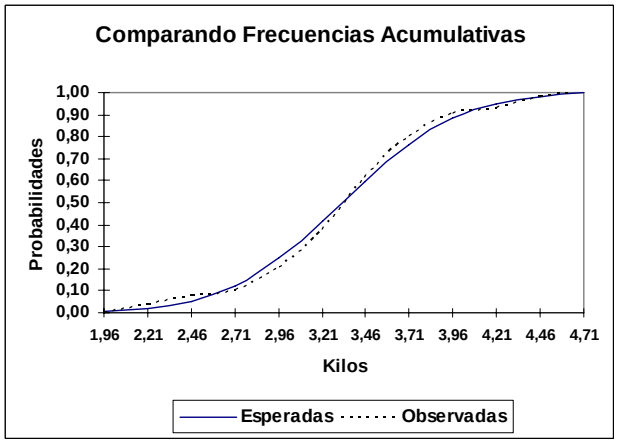
8.24 Consideraciones adicionales.

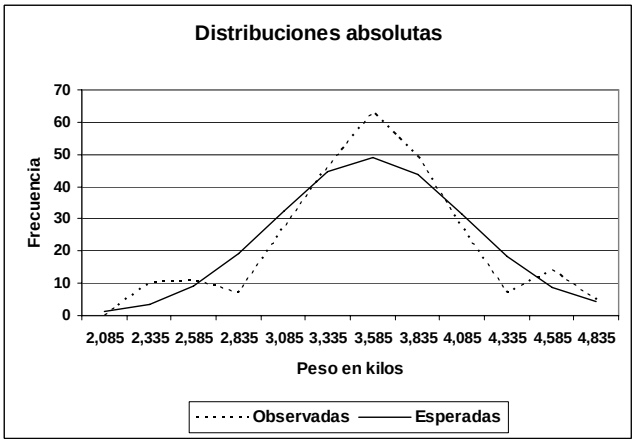
La figura que compara las dos distribuciones ofrece una visión condensada de la aproximación de ambas distribuciones acumulativas.

Como es de esperar, la distribución teórica es la sigmoides típica de la Normal acumulativa, la observada, muestra irregularidades cuya magnitud es irrelevante por tanto, la población de los pesos de bebés puede aproximarse mediante la Distribución Normal Estándar.

El gráfico de las distribuciones absolutas muestra las diferencias más marcadas. Como se mencionó, el la prueba indica no utilizar la normal para aproximar la distribución de frecuencias observadas para efectuar extrapolación.

Una alternativa de análisis es usar técnicas de estadística no paramétrica.





8.25 *El 25% de la población con menos peso.*

Puesto que la distribución es Normal, puede utilizarse el criterio de probabilidad usando:

$$X_{25} = \bar{\bar{x}} + z_{(0,25)}s = 3,451 - 0,674 \times 0,5346 = 3,090$$

O utilizar los estadísticos de orden calculando el primer cuartil mediante (no Paramétrica o NP):

$$\tilde{x}_{25} = LI_{0,25} + \left[\frac{\frac{n+1}{4} - S_{0,25}}{f_{0,25}}\right] IC = 2,835 + \left[\frac{(266+1)0,25 - 55}{46}\right] 0,250 = 2,899$$

O solicitarlo a la HE mediante la función:

$$x_{25} = \text{REDONDEAR(CUARTIL(Rango;1);3)} = 3,160$$

Nuevamente la experiencia del investigador es de mucha importancia.

8.26 *Los intervalos confiables para los cuartiles.*

El Teorema Central del Límite ofrece ciertas ventajas para trabajar con promedios, pero para trabajar con cuartiles, es indispensable que los datos se distribuyan normal. Para el primer cuarto de la población se estima mediante:

$$\Pr\left[x_{25\%} - z\left(\frac{1,3626 \times s}{\sqrt{n}}\right) \geq X_{25\%} \leq x_{25\%} + z\left(\frac{1,3626 \times s}{\sqrt{n}}\right)\right] = 1 - \alpha$$

$$\Pr\left[3,090 - 1,96\left(\frac{1,3626 \times 0,5346}{\sqrt{266}}\right) \geq X \leq 3,090 + 1,96\left(\frac{1,3626 \times 0,5346}{\sqrt{266}}\right)\right] = 0,95$$

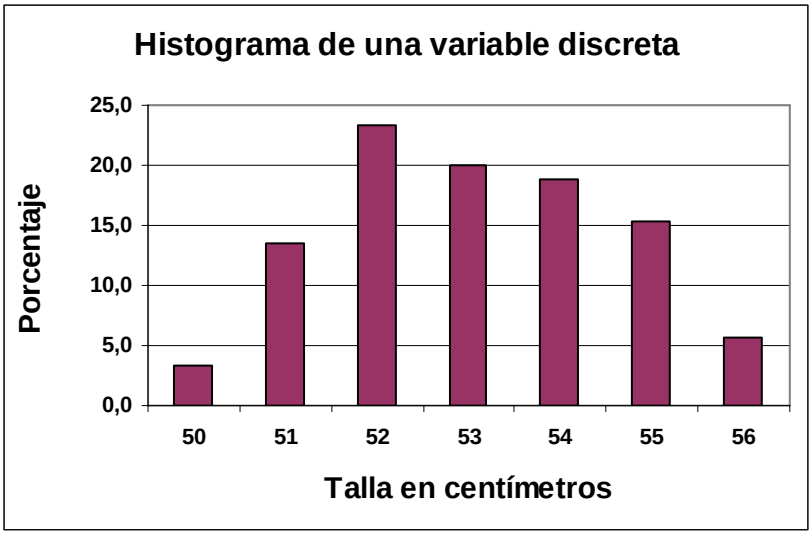
$$\Pr\left[3,002 \geq X_{25\%} \leq 3,178\right] = 95\%$$

En estas estimaciones por intervalo debe tenerse cuidado en el tamaño de la muestra que quiere estimarse. En este caso, se toma el total de la muestras.

La HE hace un conteo directo de los datos ordenados. Notará que el intervalo confiable deja por fuera el valor estimado mediante datos agrupados. Es posible que reduciendo los intervalos de clase en la distribución de frecuencias podría cambiar los resultados.

8.27 Las Variables Discretas.

Las *Variables Discretas*, cuando la muestra es de más de 30 observaciones se pueden operar como variables continuas, máxime si la variable es una discreta funcional, esto es, se tomaron números enteros pero en realidad es una medida continua, con todas las implicaciones de estas. Se muestra el Histograma separando las barras, indicando con esto la calidad discreta de la variable.



La preparación del cuadro de frecuencias difiere con el anterior en la facilidad de la rutina que cuenta celdas al cumplir la condición de búsqueda:

f_i = CONTAR.SI($D$392 : $D$657;"= 53")

Para cualquier valor.

8.28 Las Variables Cualitativas.

En los estudios mediante muestreo estadístico, la exploración de variables *Cualitativas* es muy importante, incluso de mayor interés que las cuantitativas, sobre todo en encuestas de opinión y estudios de mercado.

En general, las variables Cualitativas se engloban en el marco de la *Distribución Binomial* que también puede ser aproximada convenientemente por la *Normal Estándar.*

Muchas veces la acción de separar la información en pocos grupos permite una panorámica explícita del comportamiento incuso de *Variables Continuas o Métricas*.

En adelante se presentan Cuadros de Resultados que involucran Variables Cualitativas.

En los estudios mediante muestreo como en cualquier proyecto de investigación, deben establecerse hipótesis relevantes al estudio buscando la manera de responderlas como resultado del análisis de los datos. Los métodos de muestreo usan mucho de cuadros de resultados, que deben planificarse cuidadosamente para que los datos relevados, sea por medio de reportes sistematizados en cuanto a sexo, talla y peso de los bebés, o mediante cuestionarios que deben ser obtenidos por encuesta directa, mixta o indirecta en el autoencuesta.

En adelante, es estudiante deberá estar conciente que cada cuadro de resultados fue concebido, detallado y acompañado de las hipótesis que responden a los motivos del estudio.

8.29 Peso y talla por sexo.

Para elaborar el cuadro de resultados mostrado en esta diapositiva, los datos tabulados en la HE de la encuesta y relevantes al caso se procede de la siguiente manera:

1. ***Sé copiar a una zona separada de la HE los datos de sexo, peso y talla manteniendo la relación;***

2. ***Sé clasificar los datos teniendo cuidado de sombrear todas las variables para que en la clasificación no se separes, por sexo;***
3. ***Sé llevan a cabo las operaciones indicadas.***
4. ***Se comprueban los resultados.***

Estadístico	Niños	Niñas	Total
Números	130	136	266

Estadístico	Niños	Niñas	Total	Comprobar
Porcentajes	48,9	51,1	100	
Talla:				
Promedio	52,95	53,17	53,06	53,060
D. Estándar	1,377	1,693	1,548	
Peso:				
Promedio	3,460	3,437	3,448	3,448
D. Estándar	0,5290	0,5501	0,5390	

La comprobación de los promedios se obtiene por métodos indirectos:

$$\bar{x}_T = \frac{n_1 \times \bar{x}_1 + n_2 \times \bar{x}_2}{n_1 + n_2} = \frac{130 \times 52{,}95 + 136 \times 53{,}06}{130 + 136} = 53{,}060$$

$$\bar{x}_P = \frac{n_1 \times \bar{x}_1 + n_2 \times \bar{x}_2}{n_1 + n_2} = \frac{130 \times 3{,}460 + 136 \times 3{,}437}{130 + 136} = 3{,}448$$

Los promedios son correctos.

El agrupamiento produce proporciones, tallas y pesos diferentes. Es posible que se esté interesado en comparar los grupos. Las hipótesis involucradas en el cuadro anterior son:

Proporciones: *Ho*; $P_1 = P_2 = 0{,}5$; *Ha*; $P_1 \neq P_2 \neq 0{,}5$

Tallas: *Ho*; $\mu_1 = \mu_2 = 53{,}06$; *Ha*; $\mu_1 \neq \mu_2 \neq 53{,}06$

Pesos: *Ho*; $\mu_1 = \mu_2 = 3{,}448$; *Ha*; $\mu_1 \neq \mu_2 \neq 3{,}448$

8.30 Contrastando hipótesis sobre porcentajes.

Las hipótesis sobre proporciones, porcentajes o números se pueden contrastar utilizando la aproximación a la *Distribución Binomial* mediante la *Distribución Normal Estándar* a través de la *Distribución de* χ^2. Aún cuando frecuentemente la proporción de niñas es ligeramente mayor a la de los niños, estadísticamente no difieren. Por tanto, no hay razón para dudar que la proporción de niñas o niños sea 0,5. Dicho de otra manera, la proporción esperada es de 0,5. Por tanto, el cociente:

$$z_c = -\frac{|x - xp| - 0{,}5}{\sqrt{npq}} = \frac{|130 - 266 \times 0{,}5| - 0{,}5}{\sqrt{266 \times 0{,}5 \times 0{,}5}} = -0{,}3066$$

$$F(-0{,}3066 =) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-0{,}3066} e^{-\frac{1}{2}(-0{,}3066)^2} dx = 0{,}3796$$

Probabilidad que no es suficiente para declarar que la proporción de niños difiera de 0,5. Por tanto se acepta la hipótesis para esta clase.

El estudiante puede comprobar que para el número de niñas dará el mismo resultado. Así como utilizando la prueba de χ^2.

$$\chi^2_{2-1} = \frac{\left(|130 - 133| - 0{,}5\right)^2}{133} + \frac{\left(|136 - 133| - 0{,}5\right)^2}{133} = 0{,}0940$$

$$F_{[0,0940;\,2\text{-}1]} = Y_0 \int_0^{0,0940} (0,0940)^{\frac{1}{2}(2-1)} e^{-\frac{1}{2}0,0940} d\chi = 0,7592$$

El estudiante sabrá que elevando al cuadrado el valor de z obtendrá el de la χ^2:

$$z^2 = \chi_1^2 \therefore -0,3066^2 = 0,0940$$

Y también que la prueba de χ^2 indica la suma de probabilidades en una prueba de z de dos colas, esto es:

$$2P(z) = P(\chi_1^2) \therefore 2 \times 0,3796 = 0,7592$$

La valoración de la hipótesis de talla de los bebés se usará la prueba de t que facilita la HE. Sin embargo recuerde que:

$$t_c = \frac{|\bar{x}_1 - \bar{x}_2|}{S_p} = \frac{52,95 - 53,17}{\sqrt{\frac{(130-1)1,377^2 + (136-1)1,693^2}{130+136-2}}\sqrt{\frac{1}{130}+\frac{1}{136}}} = \frac{|-0,22|}{0,1897} = 1,1751$$

La probabilidad del estadístico:

$$F(1,1751;264) = Y_0 \int_{\infty}^{1,1751} \left(1 + \frac{1,1751^2}{266\text{-}1}\right)^{\frac{266+1}{2}} dt = 0,2410$$

Probabilidad que indica una similitud de promedios de 3-24,10%
Para los pesos de los bebés:

$$t_c = \frac{|\bar{x}_1 - \bar{x}_2|}{S_p} = \frac{|3,460 - 3,437|}{\sqrt{\frac{(130-1)0,2798 + (136-1)0,3026}{130+136-2}}\sqrt{\frac{1}{130}+\frac{1}{136}}} = \frac{0,022}{0,0655} = 0,3379$$

La probabilidad del estadístico:

$$F(0,3416;264) = Y_0 \int_{\infty}^{0,3416} \left(1 + \frac{0,3416^2}{266\text{-}1}\right)^{\frac{266+1}{2}} dt = 0,7357$$

La probabilidad de que las diferencias entre los promedios se deban a cuestiones aleatorias dado que es de una magnitud de 73,57%, valor que entra dentro de la zona de aceptación de la Hipótesis Nula.

Para responder a esta pregunta debe acomodar los datos de manera que se asocie el nivel económico cualificado del 1 a 4 con las variables cuantitativas. Para esto se procede de la siguiente manera.

8.31 El nivel económico y las variables cuantitativas.

La intención de este cuadro es presentar resultados concernientes a varias hipótesis importantes para la investigación:

- ¿El igual la proporción para cada grupo? Ho; $P_1 = P_2 = P_3 = P_4 = P_w$.
- ¿El nivel económico no tiene relación con la talla de los Bebés? Ho; $\mu_{T1} = \mu_{T2} = \mu_{T3}$

Nivel	Porcentaje de			Promedios	
Económico	Niños	Niñas	Ambos	Talla	Peso
1	9,0	9,8	18,8	52,64	2,695
2	15,0	15,4	30,5	53,01	3,267
3	15,8	15,4	31,2	53,34	3,622
4	9,0	10,5	19,5	53,10	4,177
Sumas	48,9	51,1	100,0	53,06	3,448

Nivel	Número de			Promedios	
Económico	Niños	Niñas	Ambos	Talla	Peso
1	24	26	50	52,64	2,695
2	40	41	81	53,01	3,267
3	42	41	83	53,34	3,622
4	24	28	52	53,10	4,177
Sumas	130	136	266	53,06	3,448

$= \mu_{T4}$.

- ¿El nivel económico no tiene relación con la talla de los Bebés? Ho; $\mu_{P1} = \mu_{P2} = \mu_{P3} = \mu_{P4}$.

El Detalle de las pruebas estadísticas para contrastar las hipótesis se muestran en la HE.

Copie a un sector diferente de la HE desde en número de muestra hasta los niveles económicos. Clasifique por nivel económico N1 en primer lugar, N2 en segundo y N3 en tercero.

8.32 Prueba para proporciones.

La proporción esperada para cada uno de los grupos es de P = 0,25. La prueba de la χ^2 proporcionará el criterio mediante:

$$\chi^2_{4-1} = \sum_{i=1}^{4} \frac{\left(\left|fo_i - fe_i\right| - 0{,}5\right)^2}{fe_i} = \frac{\left(\left|50 - 66{,}5\right| - 0{,}5\right)^2}{66{,}5} + \frac{\left(\left|81 - 66{,}5\right| - 0{,}5\right)^2}{66{,}5} +$$

$$+ \frac{\left(\left|83 - 66{,}5\right| - 0{,}5\right)^2}{66{,}5} + \frac{\left(\left|52 - 66{,}5\right| - 0{,}5\right)^2}{66{,}5} = 13{,}5940$$

En donde 66,5 = 266 × 0.25. Y la probabilidad:

$$F_{[13{,}5940;\,4\text{-}1]} = Y_0 \int_0^{13{,}5940} (13{,}5940)^{\frac{1}{2}(4-1)} e^{-\frac{1}{2}13{,}5940}\, d\chi = 0{,}0035$$

indica que al menos una de los estratos muestra una proporción diferente. Al estudio interesa el Nivel 1 con un porcentaje de 18% de la muestra.

Para verificar las diferencias se puede proceder a contrastar los números por pares de niveles usando la misma prueba de χ^2. Para el contraste del nivel N1 con el N2.

Nivel Económico	Frecuencia		Chi-Cudrada Parcial
	Observada	Esperada	
1	50	66,5	3,8496
2	81	66,5	2,9474
3	83	66,5	3,8496
4	52	66,5	2,9474
Sumas	266	266,0	13,5940
	Probabilidad Chi-cuadrada		0,0035

$$\chi^2_{2-1} = \frac{\left(\left|50 - 65{,}5\right| - 0{,}5\right)^2}{65{,}5} + \frac{\left(\left|80 - 65{,}5\right| - 0{,}5\right)^2}{65{,}5} = 6{,}8702$$

En donde 65,5 es el promedio de 50 + 80. La probabilidad de la χ^2 es:

$$F_{[6{,}8702;\,4\text{-}1]} = Y_0 \int_0^{6{,}8702} (6{,}8702)^{\frac{1}{2}(2-1)} e^{-\frac{1}{2}6{,}8702^2}\, d\chi^2 = 0{,}0088$$

Resultado que indica que el número, proporción o porcentaje del nivel económico N1 es menor que el del nivel N2 con probabilidad más que altamente significativa

Contrastando N1 con N3.

Los resultados son similares a los obtenidos en el contraste anterior. El número de pacientes de nivel N1 es significativamente inferior al número de mamás de la clase de nivel económico N3.

Contrastando N1 con N4.

Contraste	Frecuencia		Chi-Cudrada Parcial
	Observada	Esperada	
N1	50	65,5	3,4351
N2	81	65,5	3,4351
Suma Chi-cuadrada			6,8702
Probabilidad			0,0088

Nivel Económico	Frecuencia		Chi-Cudrada Parcial
	Observada	Esperada	
N1	50	66,5	4,0692
N3	83	66,5	4,0692
Suma Chi-cuadrada			8,1384
Probabilidad			0,0043

La prueba de χ^2 indica que no hay diferencia entre el número de mamás de la clase N1 y la clase N4. La siguiente comparación útil sería contrastar las clases N2 y N3 cuyo resultado se muestra en el

siguiente cuadro, en donde χ^2 = 0,0220 determina una probabilidad de 88,21%, que lleva a aceptar la hipótesis nula, esto es: no hay diferencias entre el número de pacientes de la clase N2 y N3.

Nivel Económico	Frecuencia Observada	Frecuencia Esperada	Chi-Cudrada Parcial
N1	50	51	0,0177
N4	52	51	0,0177
Suma Chi-cuadrada			0,0354
Probabilidad			0,8508

Nivel Económico	Frecuencia Observada	Frecuencia Esperada	Chi-Cudrada Parcial
N1	81	82	0,0110
N4	83	82	0,0110
Suma Chi-cuadrada			0,0220
Probabilidad			0,8821

8.33 Prueba de variables continuas.

Para efectuar la prueba de variables continuas se utiliza el Análisis de la Varianza (ANDEVA) y alguna prueba de promedios, se usará la de "t" en la versión de Diferencia Mínima Significativa. El ANDEVA de talla no indicó efectos importantes.

ANÁLISIS DE VARIANZA DE LA TALLA DE BEBÉS.

Origen de las variaciones	*Suma de cuadrados*	*Grados de e libertad*	*Promedio cuadrados*	*Estadístico F*	*Probabilidad F*	*Valores crítico para F 0,05*	0,01
Entre grupos	15,4565	3	5,1522	2,1787	0,0909	2,6391	3,8572
Dentro de los	619,5811	262	2,3648				
Total	635,0376	265					

1. ***Para efectuar el ANDEVA se procede de la siguiente manera.***
2. ***Copie los datos de las tallas acomodándolos por columnas una a la par de otra con las columnas tituladas;***
3. ***En el menú General solicite Herramientas /Análisis de Datos / Análisis de varianza de un factor. Obtendrá dos cuadros similares a los mostrados en la HE.***
4. ***Bajo el encabezado Probabilidad de F podrá observar el criterio para la prueba. Si la probabilidad es igual o inferior a 0,05 se rechazará la hipótesis nula que dice: Ho; $\mu_{T1} = \mu_{T2} = \mu_{T3} = \mu_{T4}$;***
5. ***Otra Forma de valorar la hipótesis es comparando el estadístico F = 2,1787 con los valores denominados tabulares de la distribución de $F_{(3;\ 262;\ 0,05)}$ = 2,6391 o $F_{(3;\ 262;\ 0,01)}$ = 3.8572. Sí el valor del estadístico es igual o menor al valor tabular elegido para la valoración se declararán Diferencias Significativas si se compara con el primer valor tabular y Diferencias Altamente Significativas sí se compara con el segundo valor tabular. Antes del advenimiento de programas de computadora, esta la manera de valorar las hipótesis estadísticas.***

El ANDEVA para el peso de los bebés indicó diferencias más que altamente significativas entre los niveles económicos.

ANÁLISIS DE VARIANZA DEL PESO DE BEBÉS.

Origen de las variaciones	*Suma de cuadrados*	*Grados de e libertad*	*Promedio cuadrados*	*Estadístico F*	*Probabilidad F*	*Valores crítico para F 0,05*	0,01
Entre grupos	61,1425	3	20,3808	337,0087	0,0000	2,6391	3,8572
Dentro de los	15,8446	262	0,0605				
Total	76,9871	265					

Los contrastes entre promedios indicaron diferencias significativas entre todos. El Nivel 1 mostró los pesos más bajos.

Contraste	Promedios 1	Promedios 2	Diferencia	Tamaño Muestra n1	Tamaño Muestra n2	Estadístico t	Probabilidad t	Criterios 0,05	Criterios 0,01	Resumen
N1 vs N2	2,6950	3,2672	-0,5722	50	81	12,937	0,0000	1,969	2,595	**
N1 vs N3	2,6950	3,6217	-0,9266	50	83	21,049	0,0000	1,969	2,595	**
N1 vs N4	2,6950	4,1771	-1,4821	50	52	30,428	0,0000	1,969	2,595	**
N2 vs N3	3,2672	3,6217	-0,3545	81	83	9,229	0,0000	1,969	2,595	**
N2 vs N4	3,2672	4,1771	-0,9099	83	52	20,921	0,0000	1,969	2,595	**
N3 vs N4	3,6217	4,1771	-0,5554	83	52	12,771	0,0000	1,969	2,595	**

8.34 Aclaración sobre el ejemplo.

Deberá entenderse que está utilizando un ejemplo dinámico para ejemplificar, y que los datos aun cuando se generan partiendo de los resultados originales no son reales, por esto, todos los registros de la población tienen resultados de la encuesta aplicada a las familias.

Se confía que el estudiante tiene la capacidad para discernir que en la realidad, la información que se utiliza a continuación sólo se recabó en una muestra de familias seleccionadas.

Puesto que se ejemplifican categorías en las que se encasilla a las familias en respuestas concretas, el tratamiento estadístico utiliza métodos para *Variables Cualitativas* en: números, proporciones o porcentajes.

8.35 Objetivo de los cuadros de resultados.

El Objetivo de los Cuadros de Resultados es ofrecer al lector una apreciación sintética de los resultados de la exploración, sea en números o en porcentajes, siendo más universales los porcentajes.

Estos Cuadros de Resultados, como se sabe, definidos de previo para responder a una o más *hipótesis* del estudio, se obtienen mediante clasificación de la información capturada y se presenta en tablas de dos o más entradas en las que concurre la información de dos o más variables *Cualitativas o Cuantitativas;* en el ejemplo:

- Nivel Económico en 4 clases: bajo, Medio Bajo; Medio Alto y Alto.
- Hábitos Higiénicos: Deficiente, Suficiente y Eficiente.
- Hábitos Nutricionales: Deficiente, Suficiente y Eficiente.

8.36 Cuadros de tres entradas.

En Números y en porcentajes.

Los cuadros de tres o más entradas se elaboran clasificando de manera anidada iniciando con las clases del margen izquierdo del cuadro hasta el encabezado.

Otra herramienta que permite elaborar cuadros es la de Cuadros y Gráficos

Nivel Económico	Hábitos Higiénicos	Hábitos Nutricionales: Deficiente	Hábitos Nutricionales: Suficiente	Hábitos Nutricionales: Eficiente	Suma H. Higiénicos
Bajo	Deficiente	3	0	0	3
	Suficiente	12	17	11	40
	Eficiente	0	6	1	7
Suma Nivel Bajo		15	23	12	50
Medio Bajo	Deficiente	8	0	0	8
	Suficiente	22	25	13	60
	Eficiente	0	9	4	13
Suma Nivel Medio Bajo		30	34	17	81
Medio Alto	Deficiente	0	0	11	11
	Suficiente	16	34	9	59
	Eficiente	0	9	4	13
Suma Nivel Medio Alto		16	43	24	83
Alto	Deficiente	8	0	0	8
	Suficiente	8	17	7	32
	Eficiente	0	8	4	12
Suma Nivel Alto		16	25	11	52
Suma H. Nutricionales		77	125	64	266
Suma H Higiçenicos		30	191	45	

Dinámico, otra más se refiere a los filtros. Aquí se usa la clasificación anidada por ser la más permanente y que menos recursos consume.

8.37 Cuadro de resultados general.

La Información del *Cuadro de Resultados* anterior, considera todas las variables de la encuesta, sirve para estudiar los números, proporciones o porcentajes que se presentan, en cada una de las categorías que se forman al separar la información en las diferentes clases que se crean al combinar las variables.

Por ejemplo:

El *Nivel Económico* bajo muestra únicamente *Hábitos Higiénicos deficientes y Hábitos Nutricionales* deficientes.

Los *Niveles Medios* muestran los mejores *Hábitos Higiénicos y Nutricionales.*

Cuadro de tres entradas. Porcentajes

Nivel Económico	Hábitos Higiénicos	Hábitos Nutricionales Deficiente	Hábitos Nutricionales Suficiente	Hábitos Nutricionales Eficiente	Suma H. Higiénicos
Bajo	Deficiente	1,13			1,13
	Suficiente	4,51	6,39	4,14	15,04
	Eficiente		2,26	0,38	2,63
Suma Nivel Bajo		5,64	8,65	4,51	18,80
Medio Bajo	Deficiente	3,01	0,00	0,00	3,01
	Suficiente	8,27	9,40	4,89	22,56
	Eficiente		3,38	1,50	4,89
Suma Nivel Medio Bajo		11,28	12,78	6,39	30,45
Medio Alto	Deficiente			4,14	4,14
	Suficiente	6,02	12,78	3,38	22,18
	Eficiente		3,38	1,50	4,89
Suma Nivel Medio Alto		6,02	16,17	9,02	31,20
Alto	Deficiente	3,01			3,01
	Suficiente	3,01	6,39	2,63	12,03
	Eficiente		3,01	1,50	4,51
Suma Nivel Alto		6,02	9,40	4,14	19,55
Suma H. Nutricionales		28,95	46,99	24,06	100,00
Suma H Higiénicos		11,28	71,80	16,92	

8.38 Cuadros de dos entradas.

Para obtener una visione más precisa sobre las relaciones que guardan entre sí las variables de tipo cualitativo, se acostumbra ir condensando la información en *Cuadros de Orden Inferior* hasta llegar a *Cuadros de Dos Entradas.*

En este ejemplo con tres variables el nivel subsecuente es la combinación de dos variables resultando en los siguientes *Cuadros de Resultados*.

- *Nivel Económico con Hábitos Higiénicos;*
- *Nivel Económico con Hábitos Nutricionales;*
- *Y Hábitos Higiénicos con Hábitos Nutricionales.*

Habrá notado que se usan, para los mismos cuadros los nombres de *Cuadros de n Entradas* y *Cuadros de Resultados*. *Cuadros de n Entradas* se refiere al proceso de acomodar los datos en una forma determinada y *Cuadros de Resultados* a la presentación de los resultados en respuesta a hipótesis del estudio.

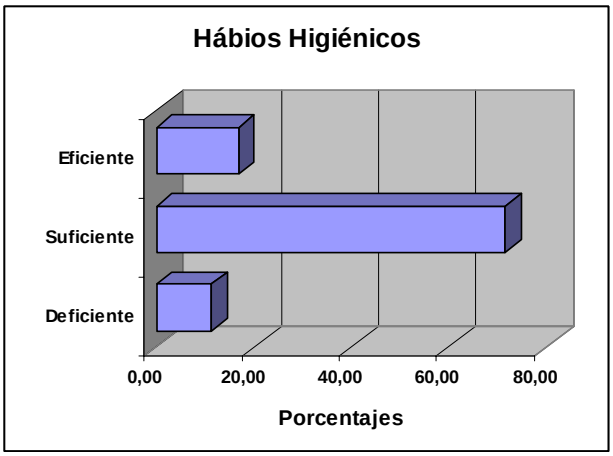
8.39 Nivel económico con hábitos higiénicos.

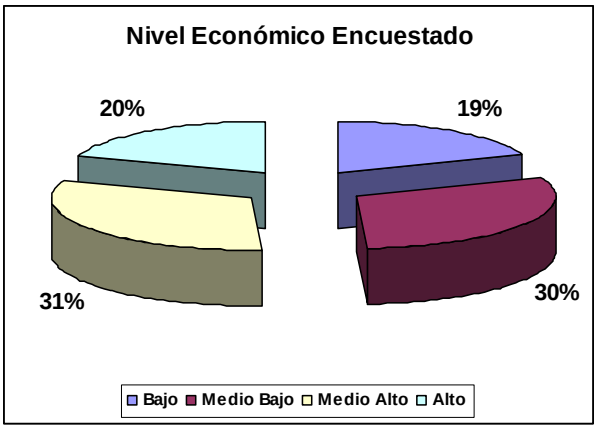
D_39. Número de Individuos. Nivel económico con Hábitos Higiénicos.

Nivel Económico	Hábitos Higiénicos Deficiente	Hábitos Higiénicos Suficiente	Hábitos Higiénicos Eficiente	Suma N. Económico
Bajo	3	40	7	50
Medio Bajo	8	60	13	81
Medio Alto	11	59	13	83
Alto	8	32	12	52
S. H. Higiénicos	30	191	45	266

D_39. Porcentajes. Nivele económico con Hábitos Higiénicos.

Nivel Económico	Hábitos Higiénicos			Suma N. Económico
	Deficiente	Suficiente	Eficiente	
Bajo	1,13	15,04	2,63	18,80
Medio Bajo	3,01	22,56	4,89	30,45
Medio Alto	4,14	22,18	4,89	31,20
Alto	3,01	12,03	4,51	19,55
S. H. Higiénicos	11,28	71,80	16,92	100,00





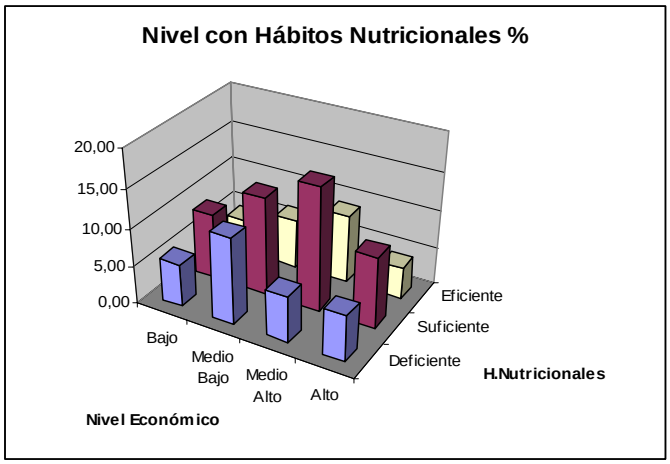
En la presentación de resultados muchas veces es conveniente utilizar gráficos, en este caso porcentuales, acompañando el cuadro de números de individuos. Los gráficos suelen ser muy explícitos.

Los gráficos suelen acompañarse con información sobre alguna prueba estadística. Es evidente la diferencia en las clases internas en el cuadro y en las marginales. En este ejemplo, las pruebas se verán más adelante.

8.40 Nivel económico con hábitos nutricionales.

Nivel Económico	Hábitos Nutricionales			S. Nivel Económico
	Deficiente	Suficiente	Eficiente	
Bajo	15	23	12	50
Medio Bajo	30	34	17	81
Medio Alto	16	43	24	83
Alto	16	25	11	52
S. H. Nutricionale	77	125	64	266

Nivel Económico	Hábitos Nutricionales			S. Nivel Económico
	Deficiente	Suficiente	Eficiente	
Bajo	5,64	8,65	4,51	18,80
Medio Bajo	11,28	12,78	6,39	30,45
Medio Alto	6,02	16,17	9,02	31,20
Alto	6,02	9,40	4,14	19,55
S. H. Nutricionale	28,95	46,99	24,06	100,00

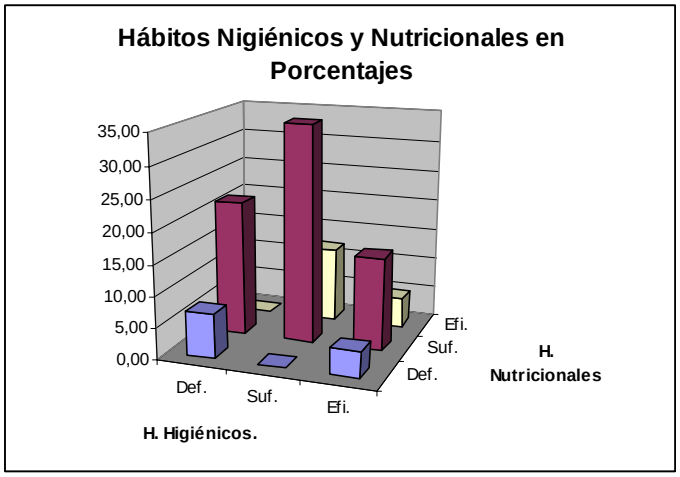


En ocasiones, el comportamiento de los niveles de una variable presenta magnitud y tendencias diferentes en presencia de los niveles de otras variables. Esta condición se conoce como *Interacción*, fenómeno que muchas veces interesa a los investigadores y puede ser valorado mediante pruebas de χ^2 en tablas de contingencia, como se verá más adelante.

8.41 Hábitos Higiénicos Con Hábitos Nutricionales.

Hábitos Higiénicos	Hábitos Nutricionales			S. Hábitos Higiénicos
	Deficiente	Suficiente	Eficiente	
Deficiente	19	0	11	30
Suficiente	58	93	40	191
Eficiente	0	32	13	45
S. H. Notricionales	77	125	64	266

Hábitos Higiénicos	Hábitos Nutricionales			S. Hábitos Higiénicos
	Def.	Suf.	Efi.	
Def.	7,14	0,00	4,14	11,28
Suf.	21,80	34,96	15,04	71,80
Efi.	0,00	12,03	4,89	16,92
S. H. Notricionales	28,95	46,99	24,06	100,00



En el último *Cuadro de Resultados* queda por mostrar el comportamiento de la segunda y tercera variables y las relaciones entre los niveles de ambas que aparenta una *interacción de dirección*: Los hábitos higiénicos muestran un comportamiento parabólico Con concavidad superior en el eficiente e inferior en los otros.

8.42 Pruebas de interacción.

En muchas ocasiones, los investigadores están interesados en valorar la interacción entre grupos de clasificación. Estas pueden ser de posición, esto es, que una o más clases (celdas) muestran valores de magnitud tal que resultan en pruebas significativas sin cambiar la dirección general (Nivel Económico con Hábitos Nutricionales); o interacciones de dirección en donde la tendencia general de una clase dentro de la alterna muestra dirección opuesta (Hábitos Higiénicos con Hábitos Nutricionales). La prueba recomendada es de χ^2 definida por:

$$\chi^2_{(h-1)(c-1)} = \sum_{j=1}^{h}\sum_{i=1}^{c} \frac{\left(\left|fo_{ij} - fe_{ij}\right| - 0{,}5\right)^2}{fe_{ij}}; \text{i} = 1,2,...,\text{c} = \text{Columnas}; \text{j} = 1,2,...,\text{h} = \text{Hileras}$$

En donde:

$$fe_{ij} = \frac{n_{i.} \times n_{.j}}{n_{..}}$$

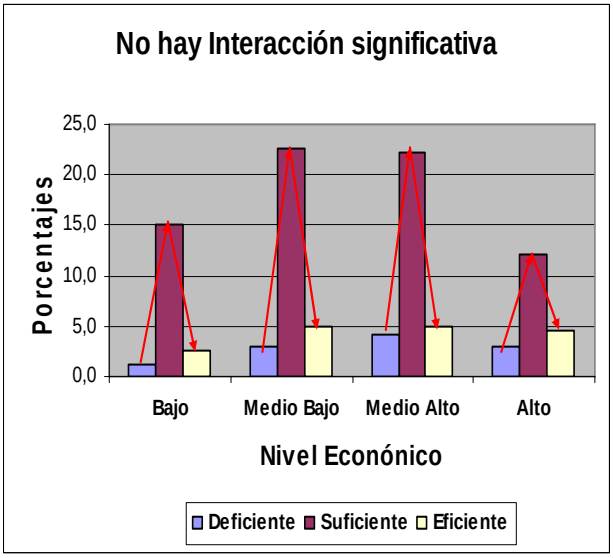
8.43 ¿Los hábitos higiénicos son independientes del nivel económico?

La prueba de χ^2 del cuadro indica que las diferencias se deben al azar y debe aceptarse la hipótesis nula.

No se refleja la interacción con una magnitud tal que la prueba lo indique. Esto se puede corroborar en el gráfico adjunto en donde para los Cuatro Niveles Económicos la tendencia es similar aun cuando hay diferencias en magnitud. $\chi^2_{(4-1)(3-1)} = \frac{\left(\left|3-5{,}6\right|-0{,}5\right)^2}{5{,}6} + \frac{\left(\left|40-35{,}9\right|-0{,}5\right)^2}{35{,}9} + ... + \frac{\left(\left|12-8{,}8\right|-0{,}5\right)^2}{8{,}8} = 3{,}4317$

En donde, como ejemplo, la frecuencia esperada de un hábito higiénico deficiente con novel económico bajo se calcula mediante:

Nivel Económico	Hábitos Higiénicos			Suma N. Económico
	Deficiente	Suficiente	Eficiente	
Bajo	3	40	7	50
Esperados	5,6	35,9	8,5	
X² parciales	0,8114	0,3605	0,1086	1,2806
Medio Bajo	8	60	13	81
Esperados	9,1	58,2	13,7	
X² parciales	0,0442	0,0308	0,0030	0,0780
Medio Alto	11	59	13	83
Esperados	9,4	59,6	14,0	
X² parciales	0,1386	0,0002	0,0209	0,1596
Alto	8	32	12	52
Esperados	5,9	37,3	8,8	
X² parciales	0,4560	0,6270	0,8305	1,9135
S. H. Higiénicos	30	191	45	266
Suma Chi-cuadradas parciales				**3,4317**
Probablida de la chi-cuadrada				**0,7530**



$$fe_{11} = \frac{n_{1.} \times n_{.1}}{n_{..}} = \frac{50 \times 30}{266} = 5{,}6$$

La probabilidad que determina el estadístico es:

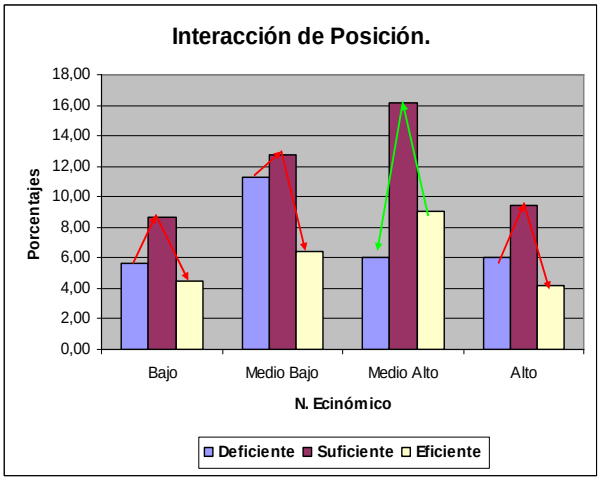
$$F_{[3,4317;\,(4-1)(3-1)]} = Y_0 \int_0^{3,4317} (3,4317)^{\frac{1}{2}(6-1)} e^{-\frac{1}{2}3,4317^2} d\chi^2 = 0,7530$$

Probabilidad que indica que la diferencia entre clases está dentro de la zona de aceptación de la hipótesis nula.

La conclusión sería: Los hábitos higiénicos son independientes del nivel económico.

8.44 ¿Ho; Los hábitos nutricionales son independientes del nivel económico?

Nivel Económico	Hábitos Nutricionales			S. Nivel Económico
	Deficiente	Suficiente	Eficiente	
Bajo	15	23	12	50
Esperado	14,5	23,5	12,0	
Chi-parcial	0,0000	0,0000	0,0184	0,0184
Medio Bajo	30	34	17	81
Esperado	23,4	38,1	19,5	
Chi-parcial	1,5624	0,3337	0,2029	2,0990
Medio Alto	16	43	24	83
Esperado	24,0	39,0	20,0	
Chi-parcial	2,3576	0,3134	0,6240	3,2950
Alto	16	25	11	52
Esperado	15,1	24,4	12,5	
Chi-parcial	0,0133	0,0002	0,0817	0,0952
S. H. Nutricionales	77	125	64	266
		Suma de chi-parciales		**5,4125**
		Probabilidad de la prueba		**0,0000**



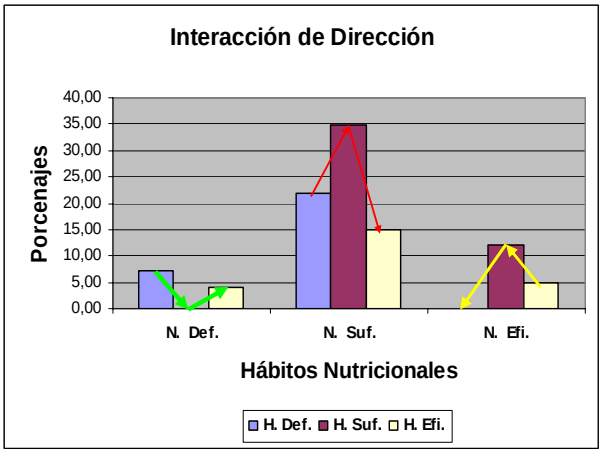
Aun cuando no es correcto graficar con líneas las variables cualitativas, facilitan la explicación de este tipo de interacción: Para los niveles Bajo, Medio Bajo, Medio Alto y Alto las tendencias van de valores medios para Hábitos Nutricionales deficientes, valores altos para niveles suficientes y bajos para Nivel Nutricional deficiente, señalados con líneas de color rojo (flechas hacia la derecha)

mientras que para el Nivel Económico Alto la dirección es inversa señalada con una línea de color verde brillante (flechas hacia la izquierda). Las diferencias son más que altamente significativas.

8.45 ¿Ho; Los hábitos higiénicos son independientes de los hábitos nutricionales?

Con Hábitos de Nutrición Deficientes la mayor proporción de familias muestra Hábitos Higiénicos deficientes y suficientes, no aparecen familias con Hábitos Suficientes (línea verde brillante); con un Nivel Nutricional Suficiente, la mayoría de las familias muestran Hábitos Higiénicos Suficientes seguidas de familias con hábitos Higiénicos deficientes y meno eficientes (línea roja); las familias con Hábitos Nutricionales Eficientes, no hay familias con Hábitos Higiénicos deficientes (línea roja). El estudiante habrá notado, además, una interacción de posición. Las diferencias son más que altamente significativas.

Hábitos Higiénicos	Hábitos Nutricionales			S. Hábitos Higiénicos
	Deficiente	Suficiente	Eficiente	
Deficiente	19	0	11	30
	8,6842	14,0977	7,2180	
	11,0948	13,1155	1,4923	25,7026
Suficiente	58	93	40	191
	55,2895	89,7556	45,9549	
	0,0884	0,0839	0,6475	0,8198
Eficiente	0	32	13	45
	13,0263	21,1466	10,8271	
	12,0455	5,0690	0,2585	17,3730
S. H. Nutricionales	77	125	64	266
		Suma de chi-parciales		**43,8954**
		Probabilidad de la prueba		**0,0000**



8.46 Opciones inductivas.

Aun cuando no se incluyó en el proyecto original, es posible conseguir *Opciones Inductivas*. Esto es, estimar algunas variables, por ejemplo *El Peso al Nacimiento* de los recién nacidos, partiendo de la información *cualitativa* recolectada en la encuesta. Esto, se intentará *calificando* en una escala de 0% a 100% las variables *Nivel Nutricional* con *25%, 50%, 75% y 100%; Hábitos Higiénicos y Hábitos Nutricionales* con 33,33%, 66,66% y 99,99%.

De esta manera, las variables *cualitativas* se transforman en variables *numéricas*, facultando entonces, la posibilidad de utilizar *Técnicas de Inducción* como *La Regresión* y *La Correlación.*

8.47 Regresión sobre el Peso.

Resumen

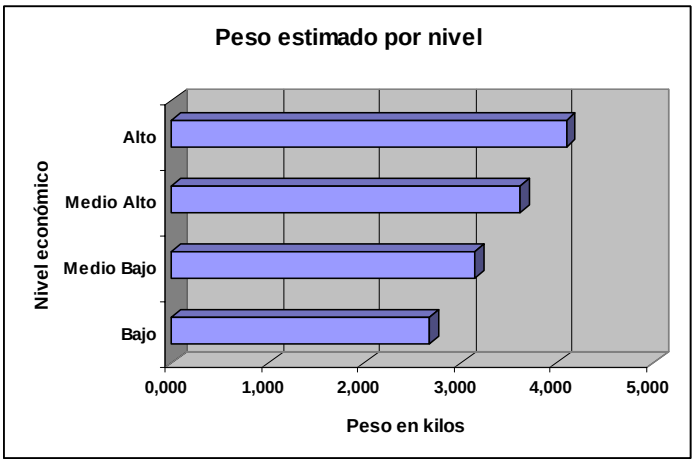
Estadísticas de la regresión	
Coeficiente de correlación mú	0,8865
Coeficiente de determinación	0,7860
R^2 ajustado	0,7835
Error típico	0,2508
Observaciones	266

ANÁLISIS DE VARIANZA

Fuente de La variación	*Grados de Libertad*	*Suma de Cuadrados*	*Promedio de Cuadrados*	*Estadístico F*	*Probabilida F*	*Valores Críticos 0,05*	*0,01*
Regresión	3	60,5096	20,1699	320,7094	0,0000	2,6391	3,8572
Residuos	262	16,4775	0,0629				
Total	265	76,9871					

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*
Intercepción	2,2072	0,0747	29,5315	0,0000	2,0601	2,3544
N. Económico	0,0189	0,0006	31,0119	0,0000	0,0177	0,0201
H. Higiénico	0,0008	0,0010	0,7955	0,4271	-0,0012	0,0027
H. Nutricional	0,0000	0,0007	-0,0432	0,9656	-0,0015	0,0014

8.48 Interpretación de la regresión.



Debe recordarse que la *Técnica de Regresión* cuando no se usa en experimentación planificada, generalmente refleja relacione fortuitas. Por esto, en estudios de muestreo, en donde no se tiene control del material experimental la relación nunca deberá considerarse *Causal.*

La finalidad de la técnica en conseguir un modelo lineal que permita proyectar los resultados de una combinación de factores. El modelo de regresión significativo de la diapositiva anterior únicamente refiere al Nivel Económico como factor de predicción mediante:

$$\hat{y}_i = 2{,}2176 + 0{,}0189(N_i)$$

Esto es, cada aumento de una unidad en el nivel económico se refleja en 0,0189 kilos de peso del recién nacido.

Usando el modelo es posible estimar los pesos según el nivel económico:

Es muy probable que se presente un efecto cuadrático que el estudiante podrá explorar.

NOTA 8,48.

Nivel	Porcentaje Relativo	Peso en kg. Estimado
Bajo	25	2,680
Medio Bajo	50	3,153
Medio Alto	75	3,626
Alto	100	4,099

8.49 La Correlación.

La *Correlación* es *La Técnica Estadística* que facilita la obtención de relaciones, casuales o causales, entre conjuntos de variables aleatorias. Este coeficiente va desde –1 cuando una variable *decrece una unidad* cuando la otra *se incrementa en una unidad*; pasando por 0 cuando no hay relación; hasta +1 cuando una variable *crece una unidad* cuando la otra se *incrementa en una unidad*. Los resultados se muestran en el cuadro en donde la única relación importante es el nivel económico con el peso. Y tal vez, los hábitos nutricionales con el peso.

Probabilidades de la Correlación.

	Peso	*N. Económico*	*H. Higiénico*	*H. Nutricional*	*Sexo*	*Talla*
Peso	1,0000					
N. Económico	0,7357	1				
H. Higiénico	0,0540	0,2410	1			
H. Nutricional	0,0000	0,9080	0,0605	1		
Sexo	0,8202	0,0892	0,8868	0,8432	1	
Talla	0,9516	0,8788	0,5116	0,8015	0,0000	1

Correlación

	Peso	***Sexo***	***Talla***	***N. Económico***	***H. Higiénico***	***H. Nutricional***
Peso	1,0000					
Sexo	-0,0208	1,0000				
Talla	0,1183	0,0721	1,0000			
N. Económico	0,8862	0,0071	0,1152	1,0000		
H. Higiénico	0,0140	0,1044	-0,0088	-0,0122	1,0000	
H. Nutricional	-0,0037	0,0094	-0,0404	-0,0155	0,4472	1,000

8.50 Otras técnicas de muestreo.

El muestreo completamente aleatorio en la elección de la muestra, es, por definición probabilístico y representativo. Sin embargo es, uno de los más costosos.

Los investigadores de la ciencia estadística han desarrollado técnica de muestreo, algunas probabilísticas con mucha seguridad de que sean a la vez representativas. Otras que no son probabilísticas pero sí representativas y otras que combinan ambos objetivos. Las familias de estas son:

- Muestreo de Unidades Accesibles;
- Muestreo Sistemático;
- Muestreo por Estratos;
- Muestreo por Etapas;
- Muestreo de Razón y Regresión.

8.51 Muestreo de unidades accesibles.

Recuerde que en un muestreo completo al azar se requiere una enumeración de cada unidad que puede ser objeto de selección. En muchas ocasiones esto no es posible recurriendo a esta técnica de muestreo.

- La técnica aprovecha proceso de mezcla y revoltura que sufren las unidades de muestreo previo a la elección de la muestra.
- Estibas de grano de café que llegan a una torrefactoría;
- Furgones de carbón que entran a una industria geotérmica;
- Apilados de minerales en un empresa fabricante de fertilizantes químicos;
- Cajas de verduras que entran a un centro de acopio;

No es estrictamente probabilístico, pero si es representativo del fenómeno, sobre todo económico y con buenos resultados.

Un ejemplo de este tipo de muestreo lo encuentra en el ejemplo 1 de la unidad 7. Un doctor oftalmólogo que valora dos técnicas quirúrgicas y utiliza la información que tiene a la mano.

8.52 Muestreo Sistemático.

Este tipo de muestreo se utiliza mucho en el control de la calidad en los procesos de líneas de fabricación en donde se opera sin interrupciones.

No es un muestreo probabilístico pero si representativo del proceso de producción de la unidades mustrales, generalmente más acucioso que un muestreo CA (Completo al Azar) y sobre todo, más económico.

Ejemplos de este tipo de muestreo lo encuentra en el problema 4 de la unidad 7, de una empresa que exporta en envases de cartón y en casi todas las unidades en que aborda el control de la calidad.

8.53 Muestreo por estratos.

Esta técnica de muestreo, es probabilística y específicamente representativa de grupos bien diferenciados.

Aprovecha la facilidad que ofrece el Análisis de la Varianza (ver capítulo 10) para reducir la variación. Entonces, por definición, éste tipo de muestreo es más eficiente.

La restricción es, que cada grupo constituye una población independiente y como tal deben ser tratadas. Esto significa que los resultados y conclusiones de un estrato únicamente son válidos para ese estrato.

La principal ventaja consiste en reducir enormemente el costo del muestreo al reducir drásticamente el tamaño de la muestra.

8.54 Tamaño de muestra en Muestreo por Estratos.

Suponga que tiene información de un muestro precio, por ejemplo en este caso. La diapositiva 33 muestra el ANDEVA en donde la varianza o Cuadrado Medio del Error alcanza un valor de:

$$S_E^2 = 0{,}0605$$

Qué, bajo las mismas condiciones determina un tamaño de muestra de:

$$n = \frac{z^2 S_E^2}{d^2} = \frac{-2{,}5758^2 \times 0{,}0605}{0{,}086^2} \equiv 54$$

Una reducción muy significante del tamaño de la muestra comparada con las 266 anteriores.

8.55 Asignación proporcional al tamaño del estrato.

La manera inmediata de asignar las 54 muestras es considerando el tamaño relativo del estrato. Recuerde que el estrato lo define el nivel económico.

La asignación proporcional se hace mediante:

$$n_i = n\frac{n_i}{N}; \text{para N1}; n_1 = 54\frac{50}{266} = 11$$

Cada una de las muestras se redondea al entero superior, por esto se deberán muestrear 56 unidades.

Estratos	Cuenta	Proporción	Muestras
N1	50	0,1880	11
N2	81	0,3045	17
N3	83	0,3120	17
N4	52	0,1955	11
Sumas	266	1,0000	56

8.56 Asignación eficiente de la muestra.

Otra forma de asignar la muestra a los estratos es utilizar un valor pesado mediante el número de individuos en un estrato y la variación intrínseca del mismo. Así, si el estrato es más grande y tiene menos varianza, es posible que la muestra sea menor si el estrato muestra menos varianza relativa.

Esto se puede conseguir multiplicando el coeficiente de variación por el tamaño del estrato, sumarlo y distribuir proporcionalmente la muestra. Esto es:

$$n_i = n_T\left(\frac{N_i \times (c.v)_i}{\sum_{i=1}^{E} N_i}\right); \text{ para el estrato N1}: n_1 = 56\frac{50\left(\frac{0,3218}{2,6950}\right)}{18,9250} = 18$$

Grupos	Cuenta	Promedio	D. Estándar	n * c.v.	Muestras
N1	50	2,6950	0,3218	5,9711	18
N2	81	3,2672	0,1869	4,6336	14
N3	83	3,6217	0,1960	4,4912	14
N4	52	4,1771	0,3076	3,8291	12
			Suma	18,9250	58

8.57 Elección de la muestra.

El estudiante habrá comprendido que en esta técnica de muestreo es necesario contar con un listado que identifique a la unidad muestral y el estrato al que pertenece.

A continuación se obtienen números aleatorios para cada estrato. La selección se hará de la misma manera en que se seleccionó la muestra en el diseño CA.

Puede auxiliarse del filtro de la HE para separar el estrato de la Base de Datos.

En la HE se ofrece la muestra para el primer estrato. De aquí en adelante, se procede como se hizo con el muestreo CA.

Deberá abrir el archivo E08_Muestreo_Irresticto_X02.xls y:

1) ***Obtener un conjunto de números aleatorios;***
2) ***Puede utilizar la opción Datos / Filtro / Autofiltro o copiar la base de datos, ordenar, renumerar cada estrato y hacer la elección de las muestras.***

8.58 Muestreo por etapas o anidado.

En ésta técnica de muestreo su supone que se suceden una serie de estaciones en el proceso de muestreo. Supóngase que se quiere explorar las enfermedades que afectan a los cafetos de Costa Rica. Obtener una muestra de todos los caficultores, ubicación y estado del cafeto puede ser complicado. Pero si se cuenta con la ubicación de las zonas cafetaleras del país por provincia cantón y distrito.

Supóngase que se obtiene una muestra al azar de 300 cantones productores, dentro de estos cantones, al azar también 3 distritos cafetaleros, dentro de estos distritos se revisarán 4 fincas que se eligen por su accesibilidad, dentro de la finca se eligen al azar 2 cultivares, dentro de estos, también al azar 2 calles, y dentro de las calles, dos matas también al azar. De esta manera se tendrá una muestra de:

$$n = 10 \times 3 \times 4 \times 2 \times 2 \times 2 = 960$$

Muestras. Con una etapa que no se hace al azar.

En el ejemplo 5 del capítulo de Análisis de la Varianza tiene un ejemplo de un muestreo por etapas.

8.59 Muestreo de Razón y Regresión.

Este tipo de muestreo se aplica a variables que están relacionadas sea de manera casual o de manera causal. El objetivo es, fundamentalmente, reducir el costo del muestreo utilizando, nuevamente, la propiedad de eficiencia estadística de los estimadores.

Esto quiere decir que vuelve a entrar en juego la técnica del Análisis de la varianza en la modalidad de regresión.

Únicamente como ejemplo, suponiendo que se utiliza la relación del nivel económico con el peso de la diapositiva 8,47 para obtener una muestra. La varianza o cuadrado medio del error para el peso del los bebés fue $S_E^2 = 0{,}0629$ entonces, el tamaño de muestra sería:

$$n = \frac{z^2 S_E^2}{d^2} = \frac{-2{,}5758^2 \times 0{,}0629}{0{,}086^2} \equiv 57$$

Cantidad muy parecida a la que se obtendría del muestreo estratificado. Cuando se analiza un factor como es el caso, el ANDEVA y la regresión suelen ser idénticos. La diferencia se debe a la concurrencia de más factores en la regresión múltiple.

8.60 Tamaño de muestra para variables cuantitativas.

Usando la HE puede calcular diferentes tamaños de muestra modificando los valores que se solicitan en verde brillante.

Para población finita y ejemplificando para una confiabilidad del 5% y una precisión del 5% para el peso promedio del ejemplo:

$$n = \frac{z^2 s^2}{d^2} = \frac{1{,}96^2 \times 0{,}291}{(3{,}448 \times 0{,}05)^2} = 38$$

Para una población finita:

$$n' = \frac{n}{1 + \frac{n}{N}} = \frac{38}{1 + \frac{38}{4.591}} = 38$$

8.61 Tamaño de muestra para variables cualitativas.

Usando la HE puede calcular diferentes tamaños de muestra modificando los valores que se solicitan en verde brillante.

Para población finita y ejemplificando para una confiabilidad del 5% y una precisión del 5% para la máxima varianza:

$$n = \frac{z^2 s^2}{d^2} = \frac{1{,}96^2 \times 0{,}5 \times 0{,}5}{(0{,}5 \times 0{,}05)^2} = 1.537$$

Para una población finita:

$$n' = \frac{n}{1 + \frac{n}{N}} = \frac{1.537}{1 + \frac{1.537}{4.591}} = 1.152$$

REFERENCIAS SELECTAS:

1. Abad, Adela, Servín, Luis A. Introducción al Muestreo. Capítulo 5. Editorial Limusa, S. A., México DF, 1987.
2. Azorín, Francisco. Curso de Muestreo y Aplicaciones. Capítulo 7. Ediciones Aguilar, S. A., Madrid España, 1972.
3. Box, George. Hunter, William. Hunter, Stuart. Estadística para Investigadores. 1ª edición. Capítulo 3. Editorial Reverte, S. A., España 1993.
4. Charles A., y Donald P. La Encuesta por Muestreo, Teoría y Práctica. Capítulo 1 y 2. Traducido de la 5ª edición en inglés. Editorial Continental, S. A. De C. V. 1985.
5. Cohran, William G. Técnicas de Muestreo. 3era edición. Capítulo 2. Editorial Continental, S. A., México DF, 1977.
6. Dimarco, Eugenio. Análisis Estadístico. 1ª Edición. Capítulo 8. Editorial Iberiaméricana, S. A., México DF. 1972.
7. Kish, Leslie. Muestreo de Encuestas. Capítulo 2. Offset Universal, S. A. 1982, México DF.
8. Murria R. Teoría y Problemas de Estadística. Traducido de la primera edición en inglés. Capítulo 4. Mc Graw-Hill 1970.
9. Newbold, Paul. Estadística para los Negocios y la Economía. Capítulo 6. Prentice may, Madrid España, 1998.
10. Poch, Azorín. Curso de Muestreo y Aplicaciones. Capítulo 2. Halar, 1972, Madrid España.
11. Pulido, A. Estadística y Técnicas de Investigación Social. Caplitulos 3, 4, 5 y 6. Editorial Pirámide 1978, Madrid España.
12. Quintana, C. Estadística Elemental. Capítulo 1. Universidad de Costa Rica, 1990, San José Costa Rica.
13. Richard, L., y Lacava, J. Estadística en los Negocios ¿por qué y cuándo?. Capítulo 5. Mc Graw-Hill Latinoamérica, Bogotá Colombia 1980.
14. Roberth D., y James H. Bioestadística: Principios y Procedimientos. Traducido de la 2ª edición en inglés. . Capítulo 1. 1980 McGraw-Hill México DF.
15. Scheater Richard L., Mendenhull William, y otros. Elementos de Muestreo. Capítulo 2, 4 y 5. Editorial Iberoamérica, 1987, México DF.
16. Shao, S. P. Estadística para Economista y Administradores de Empresas. 10ª edición. Capítulo 12. Herrero Hermanos Sucs., S. A., 1976.
17. Snédecor, George W, Cochran, William G. Métodos Estadísticos. Capítulo 1 y 17. Editorial Continental, S. A., México DF, 1982.

9 Análisis de la Varianza.

En esta sección se utilizan los archivos:
E09_ANDEVA_P01.pps
E09_ANDEVA_W01.doc
E09_ANDEVA_X01.xls

9.1 Portada.

Menú:

El caso del Agrónomo:

9.2 El caso del agrónomo.

Especialista en café se ha orientado a valorar nuevas variedades de cafetos para recomendar a los cafetaleros de Costa Rica.

Para cumplir los objetivos del estudio debe valorar entre otras cosas: rendimiento, exigencias nutricionales, resistencias a plagas y enfermedades, capacidades de reproducción. Todo esto en las diferentes zonas cafeteras del país y en al menos dos ciclos de cuatro años y una poda.

Para ofrecer sus recomendaciones planifica sus experiencias usando modelos lineales para el análisis de resultados.

9.3 El caso del economista.

Del departamento de comercio exterior que explora el aporte de productos básicos como: leche, huevos, arroz, maíz, azúcar; para obtener prospectivas del Producto Interno Bruto de los países desarrollados y del propio País.

En el análisis de los datos utiliza modelos lineales para conseguir estimaciones útiles que le faciliten estimar el indicador nacional del PIB.

9.4 El caso del ingeniero industrial.

Que tiene que decidir qué y cuánto de cada producto debe fabricar en los próximos seis meses considerando: el costo de la materia prima, el consumo de combustibles, el tiempo de proceso, las presentaciones, el gasto en mano de obra.

Considerando el ingreso neto de cada producto y las restricciones operativas de la fábrica y el mercado.

Para formar criterio analiza el conjunto de ecuaciones lineales que representan la operación de la empresa.

9.5 El caso del especialista en marketing.

Que debe entregar un estudio de mercado para una nueva presentación de un producto con mucha competencia.

En su estudio de factibilidad incluye un análisis de Fortalezas y Oportunidades, Debilidades y Amenazas conocido por sus siglas FODA modificado, en el que incluye calificaciones y valores métricos.

Para auxiliarse en sus recomendaciones utiliza modelos lineales para determinar los "puntos calientes" del estudio.

9.6 El caso del administrador médico.

Que debe programar las compras de equipo, materiales y medicinas para las clínicas de la Seguridad Social de país considerando: precios, calidades, presentaciones y otras variables relacionadas con los equipos, materiales y medicinas.

Debe además, considerar los presupuestos de compras de cada una de las clínicas, los presupuestos de gastos y prioridades de uso.

Para las prospectivas utiliza un modelo de simulación que incluye ecuaciones lineales.

9.7 El caso del biólogo.

Que debe estudiar el impacto que tendría abrir una zona franca en una región ecológicamente rica para no convertirla en una "zona caliente".

Por una parte debe estudiar el hábitat de las especies endémicas. Y por otra, los análisis de factibilidad que los especialistas en desarrollo urbano, económico e industrial han considerado para promover la apertura de una zona franca.

Para el análisis utilizará ecuaciones lineales.

9.8 En resumen.

En los ejemplos anteriores se presenta una constante:

Un conjunto de variables que pueden interactuar o no en su relación sobre una o varias variables que reflejan el resultado sobre el objetivo de estudio, en donde es posible:

- Al establecer un ordenamiento específico:
- Del lado izquierdo de una igualdad las variables consecuentes;
- Del lado derecho las variables influyentes separadas por un signo de suma que indica que los efectos son aditivos.

Lo esencial para un sistema de Ecuaciones Lineales.

9.9 El Modelo Lineal.

El modelo lineal toma la forma genérica:

$$y_i = a_0 + a_1 x_{1i} + a_2 x_{2i} + \ldots + a_p x_{pi} + \varepsilon_i$$

Es evidente que el comportamiento de la variable ***Y*** se verá influido en tasas de incremento o decremento a_i para cada variable ***X***.

Para el curso interesa que la variable ***Y*** sea de naturaleza aleatoria, por esto se incluye el error ε_i, mientras que las variables ***X***:

- Pueden ser de naturaleza Aleatoria;
- Pueden ser de naturaleza Factorial;
- pueden estar *Determinadas por la Investigación*.

9.10 El Problema.

El problema de los modelo lineales se centra en determinar cuál o cuáles tasas a_i son determinantes en el resultado de la variable ***Y*** utilizada para valorar el comportamiento del Sujeto Experimentado bajo la influencia de la correspondiente variable ***X***.

La Teoría Estadística ha desarrollado una técnica de valoración probabilística de *Hipótesis Nulas* establecidas sobre los coeficientes a_i en un modelo modificado como:

$$\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2=b_1^2\sum_{i=1}^{n}\left(X_{1i}-\overline{X}_1\right)^2+b_2^2\sum_{i=1}^{n}\left(X_{2i}-\overline{X}_2\right)^2+\ldots+b_p^2\sum_{i=1}^{n}\left(X_{pi}-\overline{X}_p\right)^2+\sum_{i=1}^{n}\varepsilon_i^2$$

Conocido como *Igualdad de la Suma de Cuadrados*, en la que la parte aleatoria $\sum_{i=1}^{n}\varepsilon_i^2$ ó *Suma de Cuadrados del Error* proporciona el criterio para las decisiones.

9.11 El ANDEVA.

Son las siglas para el Análisis de la Varianza, es además el concepto de una técnica para separar cada uno de los componentes de la *varianza total en términos de sumas de cuadrados* o sea:

$$\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2$$

y valorar la *Hipótesis Nula* del coeficiente a_j asociado a cada una de las variables $\boldsymbol{X}_j$.

Es evidente que las decisiones serán más sencillas y acertadas si únicamente aquellas *Variables Objetivo* ***X*** que realmente hagan evidente un cambio en la variable ***Y***, usada para medir el comportamiento del conjunto de sujetos experimentados o explorados.

9.12 El Cuadro del ANDEVA.

También, bajo el título de ANDEVA se conoce al cuadro resumen que compendia la información de las fuentes de variación en un formato simple y universal mostrada en un cuadro cuya forma puede cambiar de una a otra Técnica *de Análisis de los Modelos Lineales* pero con el mismo contenido: Las Fuentes de variación y el Criterio para valorar la hipótesis nula:

Ho; EL MODELO NO EXPLICA EL COMPORTAMIENTO DE ***Y***.

Fuente de la variación	*Grados de libertad*	*Sumas de cuadrados*	*Cuadrados medios*	*Criterio de F calculada*	*Probabilidad de F*
Del Modelo	***a - 1***	**SCR**	$CMR=\frac{SCR}{a-1}$	$Fc=\frac{CMR}{CME}$	$P_F\left[Fc;a-1;n-a\right]$
No Explicada	***n - a***	**SCE**	$CME=\frac{SCR}{n-a}$		
Total	***n - 1***	**SCT**			

9.13 El objetivo del tema.

El Objetivo de esta charla es introducir al estudiante en el uso de los *Modelos Lineales* en:

- Técnicas de *Regresión*:
- Técnicas de *Experimentación Planificada*;
- Y Técnicas de *Muestreo*.
- Todas relacionadas con la incertidumbre aleatoria.

Otra sección muy importante de la estadística se refiere al uso de los Modelos Lineales en la Optimización de Recursos, área de estudio más relacionada con la Investigación de Operaciones que se ve en otra unidad de estudio.

9.14 La Hoja Electrónica (HE).

La Mecánica de la charla incorpora, en este momento, el uso de un Libro Electrónico E09_ANDEVA_X01.xls, herramienta que permitirá mostrar a los estudiantes que la teoría no está en la cabeza del profesor o plasmada en los libros.

Se mostrará, que el avance de la ciencia, en este caso la *Estadística* con todas las complicaciones que se le achacan, puede tener un uso práctico sí, se es capaz, de aplicar razonamientos simples transformados en instrucciones para el *Programa Gestor de la Hoja Electrónica*, instrumento que responderá con *resultados* de operaciones, en ocasiones muy complejas que se resuelven mediante algoritmos internos.

Se verá que la aplicación teórica a los resultados no se desvirtúa por lo que implica hacer cálculos paso a paso, trabajo que se le deja a la HE quién los resuelve sin que el estudiante se entere.

9.15 ANDEVA EN LA REGRESIÓN.

Definiciones.

El modelo lineal más simple está definido mediante:

$$y_i = b_0 + b_1 X_i + \varepsilon_i$$

O mediante:

$$y_i = \bar{y} + b_1\left(X_i - \bar{X}\right) + \varepsilon_i$$

Que se obtiene resolviendo las Ecuaciones Normales:

$$b_0 = \bar{y} - b_1\bar{X}$$

Y

$$b_1 = \frac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)_2}$$

9.16 Inducción.

Para iniciar la discusión del ANDEVA es conveniente utilizar La Técnica de Regresión Simple cuyo modelo tiene una única variable *inductora* ***X*** (dominio de la función) para:

- Deducir el comportamiento de ***Y*** la variable *reflectora* (rango de la función) cuando se utiliza para explorar poblaciones;
- Para inducir acciones cuando el comportamiento de la variable ***Y*** refleja a acción de *Factores* en *Técnica de experimentación planificada*.

En todos casos, usualmente se utilizará la inferencia que resulte de una muestra.

9.17 Problema en la Regresión.

La hipótesis estadística que debe valorarse en la *Regresión Simple* es:

Determinar sí *El Modelo Lineal* aproxima convenientemente el comportamiento de la variable ***Y***, *sujeto del análisis*.

Considerando la forma alternativa del modelo lineal:

$$y_i = \bar{y} + b_1\left(X_i - \overline{X}\right) + \varepsilon_i$$

La hipótesis se plantea desde el punto de vista del coeficiente b_1. Esto es:

$$Ho; B_1 = 0$$

9.18 La solución.

Adecuando la Ecuación de *Sumas de Cuadrados de la Diapositiva* 10 al caso:

$$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = b_1^2\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 + \sum_{i=1}^{n}\left(y_i - \left[b_0 + b_1 x_i\right]\right)^2$$

O en su forma *explícita*:

SCT = SCR + SCE

O en su forma *proporcional*:

$$1 = r^2 + (1 - r^2)$$

Es evidente que entre mayor sea $\boldsymbol{r}^2$ más será la aproximación del Modelo Lineal al conjunto de datos.

En donde r^2 se conoce como Coeficiente de Determinación y mide la proporción de la variación que está siendo explicada por el modelo lineal.

9.19 El criterio.

Es evidente que entre mayor sea $\boldsymbol{r}^2$, la proporción de variación explicada es mayor el argumento para rechazar la hipótesis nula *Ho*; $B_1 = 0$.

La Teoría Estadística ha desarrollado, para estos casos, la prueba de ***F*** (de Fisher). Que consiste en ubicar en la ***Distribución de Probabilidad de F***, la cantidad de probabilidad desde el punto que determina la variable resultante de dividir dos varianzas, hasta infinito. La cantidad de probabilidad así medida se llama nivel de *significación* o nivel α y a la probabilidad complementaria $1 - \alpha$ nivel de *seguridad o nivel de confianza*, que viene desde 0 hasta el punto usado como criterio, ambas con respecto a la hipótesis nula.

9.20 El Cociente de F.

En el modelo de regresión hay dos parámetros, b_0 y b_1, por tanto, la varianza de regresión tiene 1 grado de libertad, esto es:

$$CMR = \frac{SCR}{1} = r^2 \times SCT$$

En la *varianza del error* intervienen ***n*** observaciones de las que se restan los dos coeficientes de regresión o restricciones paramétricas, por tanto tendrá ***n*** – 2 grados de libertad.

$$CME = \frac{SCE}{n-2} = \frac{\left(1 - r^2\right)SCT}{n-2}$$

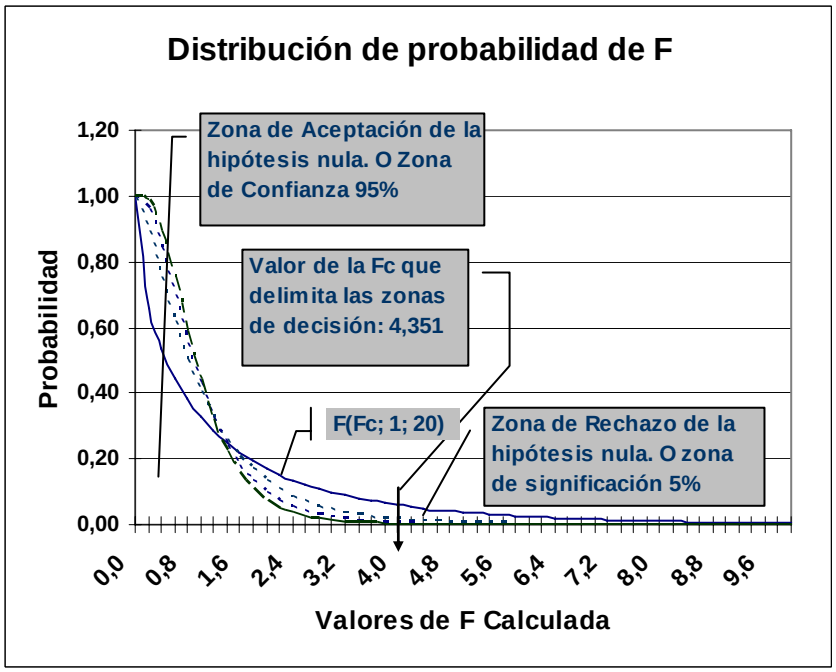
El cociente de ***F*** calculado es por tanto:

$$Fc = \frac{CMR}{CME} = \frac{(n-2)r^2}{1 - r^2}$$

9.21 La Regla de Decisión.

Todo estudio estadístico lleva implícito, que los responsables han analizado las consecuencias de *Aceptar* o *Rechazar* una *Hipótesis Nula* y con esta base, definido el nivel de confiabilidad que se usará en las pruebas de significación.

El Gráfico esquematiza una prueba de ***F*** con un nivel de significación 5%. La regla de decisión es simple: Si la función de la HE indica un nivel inferior o igual a 0,05 se rechaza la hipótesis. O bien, si el valor calculado de ***F*** es menor a 4,351 rechace *Ho*. Para una prueba con 1 y 20 grados de libertad en el cociente varianzas.



9.22 Problema 9.1. De Regresión Lineal.

En la escuela de Zootecnia de la Universidad de Costa Rica se efectuó un experimento planificado en una granja avícola en pollos de raza para carne. Uno de los objetivos era:

Obtener un promedio de libras de carne por libra de alimento consumido.

Los resultados son promedios de 22 jaulas que tenían 10 pollos cada una.

La hipótesis nula dice que el peso de los pollos no está relacionado con el consumo de alimento, esto es:

Ho; $B_1 = 0$

Con un nivel de significación de 0,05 o 5%, para las pruebas.

9.23 El ANDEVA promedios y sumas de productos cruzados.

Para construir un ANDEVA de Regresión Simple paso a paso debe obtener los siguientes estadísticos:

El promedio de ***Y***:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{1{,}425 + 2{,}603 + \ldots + 1{,}945}{22} = 2{,}391$$

El promedio de ***X***.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{37{,}233 + 43{,}136 + \ldots + 47{,}077}{22} = 43{,}666$$

La *Suma de Cuadrados de* ***Y***.

$$SCY = \sum_{i=1}^{n} (y_i - \bar{y})^2)(1{,}425 - 2{,}391)^2 + (2{,}603 - 2{,}391)^2 + \ldots + (1{,}945 - 2{,}392)^2 = 5{,}670$$

9.24 Sumas de Cuadrados de X, Productos Cruzados y Coeficiente b_1.

La *Suma se Cuadrados de* ***X***.

$$SCX = \sum_{i=1}^{n}(x_i - \bar{x})^2)(37{,}233 - 43{,}666) + (43{,}136 - 43{,}666)^2 + ... + (47{,}077 - 43{,}666)^2$$
$$= 264{,}866$$

La *Suma de Productos Cruzados* ***XY***:

$$SCXY = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) =$$
$$= (37{,}233 - 43{,}666)(1{,}425 - 2{,}391)^2 + ... + (47{,}077 - 43{,}666)(1{,}945 - 2{,}392)^2 = 27{,}709$$

El *Coeficiente de la Pendiente* $\boldsymbol{b_1}$.

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{SCXY}{SCX} = \frac{27{,}709}{264{,}866} = 0{,}1046$$

9.25 Coeficiente b_o Cuadrado Medio del Error y Desviación Estándar.

El *Coeficiente de la Interceptada.*

$b_0 = \bar{y} - b_1\bar{x} = 2{,}391 - 0{,}1046 \times 43{,}666 = -2{,}177$

Una vez que se obtiene el modelo lineal se puede calcula la *Suma de Cuadrados del Error*:

$$SCE = \sum_{i=1}^{n}(y_i - [b_0 + b_1 x_i])^2$$
$$= (1{,}425 - [-2{,}177 + 0{,}1046 \times 37{,}233])^2 + ... + (1{,}945 - [-2{,}177 + 0{,}1046 \times 47{,}077])^2 =$$
$$= 2{,}772$$

En el ANDEVA de regresión se acostumbra ofrecer pruebas sobre los coeficientes de regresión. Para esto debe considerarse La Desviación Estándar del Modelo:

$$S_E = \sqrt{\frac{SCE}{n-2}} = \sqrt{\frac{2{,}772}{20}} = 0{,}372$$

9.26 Suma de Cuadrados de Regresión, F calculada y Probabilidad de F.

El criterio para valorar la hipótesis se obtiene con la variable de ***F*** calculada, para esto se debe obtener la *Suma de Cuadrados de la Regresión*:

$$SCR = b_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 = 0{,}1046^2 \times 264{,}866 = 2{,}899$$

Estadístico con el que ya se puede obtener el cociente de ***F*** Calculada:

$$Fc = \frac{CMR}{CME} = \frac{2{,}899}{0{,}139} = 20{,}918$$

Y valorar la probabilidad que determina en la distribución de ***F*** mediante la rutina de la HE:

$$F(Fc; 1; 20) = \int_0^{20{,}918} f(20{,}918; 1; 20)dx = \text{DISTR.F}(20{,}918; 1; 20) = 0{,}00018$$

9.27 Prueba de t para la Intersectada.

Los estimadores de los coeficientes del modelo de regresión tienen una distribución normal alrededor de los parámetros. Esto significa que se pueden valorar las hipótesis al comparar variables estandarizadas con la *Distribución Normal Estándar* o usando la *Distribución de "t"*. Para la intersectada:

$$t_{b0} = \frac{b_0}{S_{b0}} = \frac{b_0}{S_E\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]} = \frac{-2{,}177}{0{,}372\left(\frac{1}{22} + \frac{43{,}666^2}{264{,}866}\right)} = -2{,}1730$$

Valorado con la rutina de la HE:

$$F(|-2{,}173|; 20; 2) = Y_0\int_0^{2{,}173} f\left(1 + \frac{t^2}{20}\right)^{\frac{20+1}{2}} dt = 0{,}04197$$

Implica rechazar la hipótesis con una probabilidad de 4,20%.

9.28 Prueba de t para la Pendiente.

La variable *t* para la pendiente b_1:

$$t_{b1} = \frac{b_1}{S_{b1}} = \frac{b_1}{\frac{S_E}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} = \frac{0{,}1046}{\frac{0{,}372}{\sqrt{264{,}866}}} = 4{,}5736$$

Valorado con la rutina de la HE:

$$F(|4{,}5736|; 20; 2) = Y_0\int_0^{4{,}5736} f\left(1 + \frac{t^2}{20}\right)^{\frac{20+1}{2}} dt = 0{,}00018$$

En el caso de la *Regresión Simple*, las probabilidades de ***F*** y de ***t*** son exactamente iguales. Nada extraño si se recuerda que la distribución de ***F*** es una Distribución de Probabilidad de variables $\mathbf{t}^2$.

9.29 El Cuadro del ANDEVA.

ANDEVA para regresión simple

Fuente de Variación	*Grados de Libertad*	*Suma de Cuadrados*	*Cuadrados Medios*	*Cocinete de F. Calculada*	*Probabilidad Significante*	*Límites Significantes P(0,05)*	*P(0,01)*
Regresión	1	2,899	2,8988	20,9180	0,00018	4,3512	8,0960
Error	20	2,772	0,1386				
Total	21	5,670					

Promedio de Y	2,3909
Desviación Estándar	0,3723
Coeficiente de variación	15,57%
Coeficiente de Correlación	71,50%
Coeficiente de Determinación	51,12%

Información sobre los coeficientes de regresión

Variable	*Coeficiente Regresión*	*Error Típico*	*Estadístico t*	*Probabilidad Coeficiente*	*Límites 95% Inferior*	*Superior*
Intersectada	-2,1773	1,0019	-2,1730	0,04197	-4,2673	-0,0872
X	0,1046	0,022873581	4,5736	0,00018	0,0569	0,1523

Todos los cálculos se acomodan en el cuadro del ANDEVA para regresión tomado como base el de la diapositiva 12.

9.30 Conclusiones del ANDEVA.

La primera conclusión que indica el ANDEVA es que debe *rechazarse la Hipótesis Nula* con un nivel de significación de 0,00018 o 0,018%. Dicho de otra forma, el Modelo Lineal:

$$\hat{y}_i = -2{,}1773 + 0{,}1046x_i$$

Aproxima los datos con un nivel de confianza de 99,98%.

La intersectada se estima entre:

$$\Pr\left\{b_0 - t_{(20;0,05)}S_{b0} \geq \beta_0 \leq b_0 + t_{(20;0,05)}S_{b0}\right\} = 95\%$$

$$\Pr\left\{-4{,}2673 \geq \beta_0 \leq -0{,}0872\right\} = 95\%$$

Evidentemente, no pasa por el origen como se dedujo con anterioridad.

La pendiente puede fluctuar entre:

$$\Pr\left\{b_1 - t_{(20;0,05)}S_{b1} \geq \beta_1 \leq b_1 + t_{(20;0,05)}S_{b1}\right\} = 95\%$$

$$\Pr\left\{0{,}0569 \geq \beta_1 \leq 0{,}1523\right\} = 95\%$$

La conclusión de la regresión sería: Los pollos pesan más entre más alimento consumen. Una libra de alimento implica un aumento de peso que puede ir de 0,060 a 0,150 libras con una confianza del 95%.

9.31 El ANDEVA que ofrece la HE.

Ya que se ha dado un repaso a la teoría para reafirmar conceptos, se está en posición de interpretar el resumen del ANDEVA que ofrece la rutina de la HE. Con los mismos valores que el elaborado paso a paso.

Resumen

Estadísticas de la regresión	
Coeficiente de correlación	0,7150
Coeficiente de determinac	0,5112
R^2 ajustado	0,4868
Error típico	0,3723
Observaciones	22

ANÁLISIS DE VARIANZA

Fuente de Variación	*Grados de Libertad*	*Suma de Cuadrados*	*Cuadrados Medios*	*Cocinete de F. Calculada*	*Probabilidad Significante*	*Límites Significantes P(0,05)*	*P(0,01)*
Regresión	1	2,8988	2,8988	20,9180	0,00018	4,3512	8,0960
Residuos	20	2,7716	0,1386				
Total	21	5,6703					

Variable	*Coeficiente Regresión*	*Error Típico*	*Estadístico t*	*Probabilidad Coeficiente*	*Límites 95% Inferior*	*Superior*
Intercepción	-2,1773	1,0019	-2,1730	0,0420	-4,2673	-0,0872
X	0,1046	0,0229	4,5736	0,0002	0,0569	0,1523

9.32 El ANDEVA en Experimentos Planificados con Modelo Lineal de un Factor.

El experimento planificado.

El Modelo Lineal más simple que se utiliza en experimentos planificados es el denominado *Diseño Completo al Azar*. En donde el Objeto de la experiencia es un *Factor* con más de un Nivel de Aplicación o Tratamientos.

Es también, el tránsito entre el Modelo de Regresión Simple y el Modelo Experimental, pues como de verá, el ANDEVA se puede realizar con la rutina de Regresión de la HE.

Aun cuando Los Modelo utilizados en la Experimentación planificada son un refinamiento del Modelo de Regresión tienen, al menos, una forma tradicional de ser presentados que requiere un poco de discusión.

9.33 El Modelo Lineal en Experimentos Planificados.

El Modelo Lineal básico para los experimentos se define como:

$$y_{ij} = \bar{y}.. + T_i + E_{ij}$$

Se usa indicar la observación con la notación y_{ij} para hacer ver que debe considerarse como una repetición *j* dentro del tratamiento *i*.

La expresión $\boldsymbol{T}_i$ indica el efecto de un Factor dentro del modelo. Pero en este, se incluye una *Regresión Polinomial* de grado $t - 1$, donde *t* es número de niveles del factor o Tratamientos.

Y la expresión $\boldsymbol{E}_{ij}$ indica que la aproximación que se hace mediante el Modelo Experimental a una variable aleatoria, no podrá ser exacta, pues hay fuentes de variación inherentes a la unidad experimentada *ij* que no podrán ser explicadas por el Modelo usado en la aproximación.

9.34 El ejemplo de ANDEVA con un factor.

Una empresa que fabrica estructuras de lámina galvanizada, quiere incursionar en estructuras preformadas para tejados.

El departamento de control de la calidad está efectuando pruebas de resistencia tratando de ubicar el punto óptimo de carbono agregado a la colada del hierro.

	Porcentaje de carbono		
Repetición	0,1	0,2	0,3
1	25	40	34
2	28	31	37
3	25	27	37
4	22	40	39
5	23	43	32
6	31	35	38
Sumas T	154	216	217

Uno de los ensayos proporcionó los resultados mostrados en el cuadro como unidades de resistencia a la tensión.

9.35 Suma de Cuados Total y de Tratamientos.

Son pocos los cálculos para llevar a cabo un *Análisis de la Varianza* paso a paso de la manera tradicional adaptada a las facilidades de la HE. Lo primero a considerar son:

La cantidad de tratamientos o niveles del factor: $t = 3$

Las veces que se repite cada tratamiento: $r = 6$

La Suma de Cuadrados Total.

$$SCY = (t \times r - 1)S_Y^2 = (3 \times 6 - 1)41{,}8987 = 712{,}2778$$

La Suma de Cuadrados de Tratamientos o del Factor.

$$SCT = \frac{(t-1)}{r} S_T^2 = \frac{3-1}{6} 1302{,}3333 = 434{,}1111$$

9.36 Cuadrado Medio de Tratamientos, del Error; F calculada.

El Cuadrado Medio de Tratamientos.

$$CMT = \frac{SCT}{t-1} = \frac{434{,}1111}{3-1} = 217{,}0556$$

La *Suma de Cuadrados del Error* se obtiene por diferencia.

$$SCE = SCY - SCT = 712{,}2778 - 434{,}1111 = 278{,}1667$$

El Cuadrado Medio del Error.

$$CME = \frac{SCE}{n-t} = \frac{278{,}1667}{18-3} = 18{,}5444$$

El Cociente de F.

$$Fc = \frac{CMT}{CME} = \frac{217{,}0556}{18{,}5444} = 11{,}7046$$

9.37 La prueba de F.

Finalmente de valora la ***F*** calculada mediante la rutina de la HE:

$$F(Fc;2;15) = \int_0^{11{,}7046} f(11{,}7046;2;15)dx = \text{DISTR.F}(11{,}7046;2;15) = 0{,}0009$$

Con lo que se concluye el ANDEVA.

El resultado indica que la probabilidad de que las diferencias de la resistencia a la tensión de las estructuras de metal se deban al azar es de 0,0009 0 0,09%, o nivel de significación.

Se puede interpretar considerando el nivel de confianza diciendo:

Se concluye: El porcentaje de carbón agregado al metal produce estructuras de diferente dureza con una confiabilidad de 99,91%

9.38 El ANDEVA de la Hoja Electrónica.

La HE proporciona una rutina para el cálculo del ANDEVA para un Factor con la siguiente salida.

Análisis de varianza de un factor

RESUMEN

Grupos	*Cuenta*	*Suma*	*Promedio*	*Varianza*
0,1	6	154	25,6667	11,0667
0,2	6	216	36,0000	37,6000
0,3	6	217	36,1667	6,9667

ANÁLISIS DE VARIANZA

Fuente de Variación	*Suma de Cuadrados*	*Grados de Libertad*	*Cuadrados Medios*	*Cocinete de F. Calculada*	*Probabilidad Significante*	*Límites Significantes P(0,05)*	*P(0,01)*
Entre grupos	434,1111	2	217,0556	11,7046	0,0009	3,6823	6,3589
Dentro de los grupos	278,1667	15	18,5444				
Total	712,2778	17					

Queda el problema de decidir ¿cuál nivel de carbón es el adecuado?

9.39 Los Tratamientos: un polinomio de grado t – 1.

El Modelo de la diapositiva 35 puede escribirse:

$$\hat{\bar{y}}_i = \bar{y}.. + b_1 X1_i + b_2 X2_i + E_{ij}$$

En donde:

$$T_i = b_1 X1_i + b_2 X2_i + E_{ij}$$

Y las Sumas de Cuadrados como:

$$\sum_{i=1}^{t}\sum_{j=1}^{r}\left(y_{ij} - \bar{y}\right)^2 = \left\{ SCT = b_1^2 r \sum_{i=1}^{t} X1_i^2 + b_2^2 r \sum_{i=1}^{t} X2_i^2 \right\} + SCE$$

En donde $\boldsymbol{X}1 = [-1; 0; 1]$ y $\boldsymbol{X}2 = [1; -2; 1]$, polinomios mínimos transformados de los niveles de los tratamientos para efectuar un análisis que se conoce Por Contrastes Ortogonales.

9.40 *Transformando los niveles a polinomios mínimos.*

Los Tratamientos aplicado son: 0,10, 0,20 y 0,30 por ciento de carbono. Para obtener un polinomio mínimo debe buscarse la cantidad:

$$X1_i = \eta\left(t_i - \bar{t}\right)$$

Tal que la suma de $\boldsymbol{X1}_i = 0$ y la suma de $(\boldsymbol{X1}_i)^2$ sea mínima. Para esto, se tiene que encontrar el escalar η que haga mínimo al polinomio:

$$\eta[(0{,}1-0{,}2);(0{,}2-0{,}2);(0{,}3-0{,}2)]$$

Estas cantidades son:

$$\frac{1}{0{,}10} = [-0{,}10; 0; 0{,}10] = [-1; 0; 1] \text{ y } 3\left[1-\frac{2}{3}; 0-\frac{2}{3}; 1-\frac{2}{3}\right] = [1; -2; 1]$$

Para $\boldsymbol{X}1_i$ y $\boldsymbol{X}2_i$ respectivamente.

9.41 *Los Contrastes.*

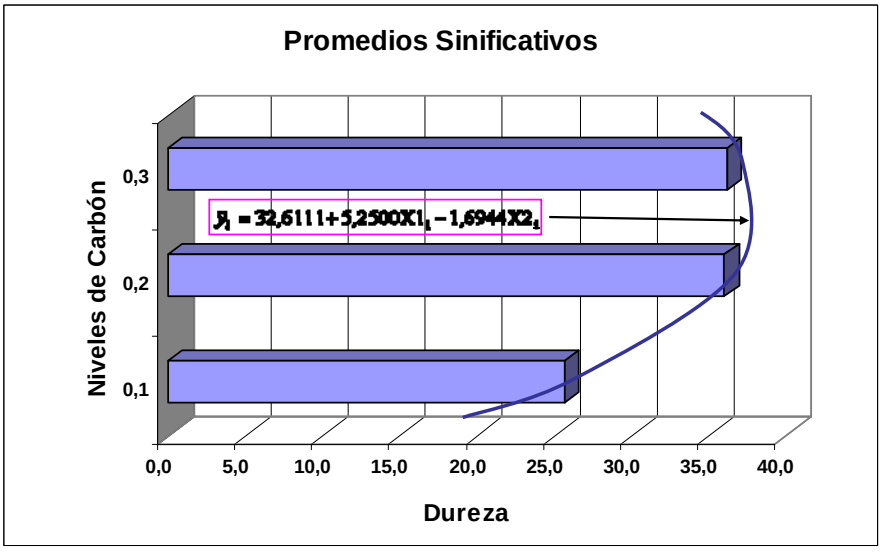
Cuando se aplican los polinomios a los promedios de los tratamientos se efectúan los siguientes contrastes o comparaciones:

Contraste	Promedios de Tratamientos 25,6667	36,0000	36,1667	Suma de Yi * Xij	S.Cuadrados X	S.Cuadrdos Contraste	Cociente F	Probabilidad F	Coeficentes Regresión
T1 vs T2	-1	0	1	10,50	2	330,7500	17,8355	0,0007	5,2500
2(T2) vs T1+T2	1	-2	1	-10,17	6	103,3611	5,5737	0,0322	-1,6944

En el gráfico se observa el modelo de regresión significativo.

Es evidente que la dureza aumenta a medida que se aumenta el contenido de carbón en la colada, pero hasta cierto nivel u óptimo. Éste será deducido más adelante.

Los coeficientes de regresión para los contrastes se obtienen mediante:



$$b_1 = \frac{\sum_{i=1}^{3} X1_i \times \bar{y}_i}{SCX1} = \frac{-1\times 25{,}6667 + 0\times 36{,}0000 + 1\times 36{,}1667}{2} = \frac{10{,}50}{2} = 5{,}2500$$

$$b_2 = \frac{\sum_{i=1}^{3} X2_i \times \bar{y}_i}{SCX2} = \frac{1\times 25{,}6667 - 2\times 36{,}0000 + 1\times 36{,}1667}{6} = \frac{-10{,}17}{6} = -1{,}6944$$

Las sumas de cuadrados para los efectos lineal y cuadrático:

$$SCC1 = \frac{r\left(\sum_{i=1}^{3} X1_i \bar{y}_i\right)^2}{SCX1} = \frac{6\times 10{,}5000^2}{2} = 330{,}7500$$

$$SCC2 = \frac{r\left(\sum_{i=1}^{3} X2_i \bar{y}_i\right)^2}{SCX2} = \frac{6\times -10{,}17^2}{6} = 103{,}3611$$

Los resultados se corroboran sí la suma de cuadrados de los contrastes es igual a la suma de cuadrados de tratamientos:

$$SCT = SCC1 + SCC2 = 330{,}7500 + 103{,}3611 = 434{,}1111$$

Los cocientes de F.

$$Fc = \frac{SCC1}{CME} = \frac{330{,}7500}{18{,}5444} = 17{,}8355$$

$$Fc = \frac{SCC2}{CME} = \frac{103{,}3611}{18{,}5444} = 5{,}5737$$

La probabilidad de las F:

$$F(Fc;1;15) = \int_0^{17{,}8355} f(17{,}8355;1;15)dx = \text{DISTR.F}(17{,}8355;1;15) = 0{,}0007$$

$$F(Fc;1;15) = \int_0^{5{,}5737} f(5{,}5737;1;15)dx = \text{DISTR.F}(5{,}5737;1;15) = 0{,}0322$$

Conclusión:

Rechazar Ho; $b_1 = 0$. Por tanto, el efecto lineal positivo indica que la dureza aumenta a medida que se aumenta el contenido de carbón en la colada.

Rechazar Ho; $b_2 = 0$. Por tanto, el efecto cuadrático negativo indica que existe un máximo. Recordar el teorema de la segunda derivada: si es positiva, se tendrá un mínimo; si es negativa se tendrá un máximo; si es 0, puede estar indeterminada.

El modelo de regresión significativo es:

$$\bar{y}_i = 32{,}6111 + 5{,}2500\text{X1}_\text{i} - 1{,}6944\text{X2}_\text{i}$$

Con el que se obtienen los promedios estimados:

$$\bar{y}_{0,1} = 32{,}6111 + (5{,}2500\times -1) - (1{,}6944\times 1) = 25{,}7$$

$$\bar{y}_{0,2} = 32{,}6111 + (5{,}2500\times 0) - (1{,}6944\times -2) = 36{,}0$$

$$\bar{y}_{0,3} = 32{,}6111 + (5{,}2500\times -1) - (1{,}6944\times 1) = 25{,}7$$

Resistencia	POLINOMIOS	
Y	X1	X2
25	-1	1
28	-1	1
25	-1	1
22	-1	1
23	-1	1
31	-1	1
40	0	-2
31	0	-2
27	0	-2
40	0	-2
43	0	-2
35	0	-2
34	1	1
37	1	1
37	1	1
39	1	1
32	1	1
38	1	1

Notará que el modelo contiene toda la información puesto que los promedios reales y los esperados alcanzan el mismo valor.

Con los datos del cuadro aledaño en donde se asocia cada valor del polinomio con el tratamiento y repetición es posible utilizar la función de regresión múltiple de la HE para conseguir el ANDEVA y los valores de los coeficientes de regresión tal como se ve en la salida de la HE.

En las técnicas de experimentación los resultados se obtienen en términos de los polinomios ortogonales mínimos. Asociando cada observación con el polinomio respectivo tal como se muestra en el cuadro aledaño, es posible obtener El ANDEVA del modelo lineal utilizando la rutina de regresión múltiple que proporciona la HE.

9.42 Coeficientes de regresión Calculados por Regresión de la HE.

Resumen

Estadísticas de la regresión	
Coeficiente de correlación	0,7807
Coeficiente de determinac	0,6095
R^2 ajustado	0,5574
Error típico	4,3063
Observaciones	18

ANÁLISIS DE VARIANZA

Fuente de Variación	*Grados de Libertad*	*Suma de Cuadrados*	*Cuadrados Medios*	*Cocinete de F. Calculada*	*Probabilidad Significante*	*Límites Significantes P(0,05)*	*P(0,01)*
Regresión	2	434,1111	217,0556	11,7046	0,0009	3,6823	6,3589
Residuos	15	278,1667	18,5444				
Total	17	712,2778					

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*
Intercepción	**32,6111**	**1,0150**	**32,1288**	**0,0000**	30,4477	34,7746
X1	**5,2500**	**1,2431**	**4,2232**	**0,0007**	2,6003	7,8997
X2	**-1,6944**	**0,7177**	**-2,3609**	**0,0322**	-3,2242	-0,1647

Nº	% Carbono X	X²	Resistencia Y
1	0,1	0,01	25
2	0,1	0,01	28
3	0,1	0,01	25
4	0,1	0,01	22
5	0,1	0,01	23
6	0,1	0,01	31
7	0,2	0,04	40
8	0,2	0,04	31
9	0,2	0,04	27
10	0,2	0,04	40
11	0,2	0,04	43
12	0,2	0,04	35
13	0,3	0,09	34
14	0,3	0,09	37
15	0,3	0,09	37
16	0,3	0,09	39
17	0,3	0,09	32
18	0,3	0,09	38

Estadísticas de la regresión	
Coeficiente de correlación	0,78068
Coeficiente de determinac	0,60947
R^2 ajustado	0,55740
Error típico	4,30633
Observaciones	18

ANÁLISIS DE VARIANZA

Fuente de Variación	*Grados de Libertad*	*Suma de Cuadrados*	*Cuadrados Medios*	*Cocinete de F. Calculada*	*Probabilidad Significante*	*Límites Significantes P(0,05)*	*P(0,01)*
Regresión	2	434,1111	217,0556	11,7046	0,0009	3,6823	6,3589
Residuos	15	278,1667	18,5444				
Total	17	712,2778					

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*Inferior 95,0%*	*Superior 95,0%*
Intercepción	5,1667	7,6632	0,6742	0,5104	-11,1670	21,5003	-11,1670	21,5003
X	255,8333	87,0190	2,9400	0,0101	70,3565	441,3102	70,3565	441,3102
X²	-508,3333	215,3163	-2,3609	0,0322	-967,2695	-49,3972	-967,2695	-49,3972

Si los datos se arreglan tal como se muestra de la derecha, es posible efectuar un análisis aún más explícito en términos de niveles de tratamientos y resultados. En este caso los niveles X1 = 0,1; 0,2 y 0,3 por ciento de carbono. Para función cuadrática basta elevar estos al cuadrado, $X2_1 = 0,1^2 = 0,01$; $X2_2 = 0,02^2 = 0,04$ y $X2_3 = 0,3^2 = 0,09$. Se introducen los datos pertinentes en la rutina de regresión de la HE para obtener el ANDEVA de regresión que se muestra a continuación.

Aquí interesan las probabilidades de los coeficientes. La probabilidad para la interceptada es 0,5104 valor que indica no rechazar la hipótesis Ho; $b_o = 0$; esto es, la línea de regresión estimada

pasa por el origen, en otras palabras si no se agrega carbón a colada la dureza del material andará muy cercana a 0.

La probabilidad para el coeficiente lineal es 0,0101, que significativa e implica que por cada unidad porcentual de carbón que se agregue a la colada, la dureza aumenta en b_1 = 255,8333 unidades.

La probabilidad para el coeficiente cuadrático es 0,0322 para el coeficiente b_2 = −508,3333. Resultado que indica un crecimiento parabólico positivo, es decir, habrá un máximo en el rango de la función (Y) dado un valor del dominio (X).

9.43 Interpretación.

El modelo significativo de los valores sin transforma es:

$$\hat{\hat{y}}_{ij} = 5{,}1667 + 255{,}8333T_i - 508{,}333T_i^2$$

Usando el criterio de la segunda derivada es posible encontrar un óptimo mediante:

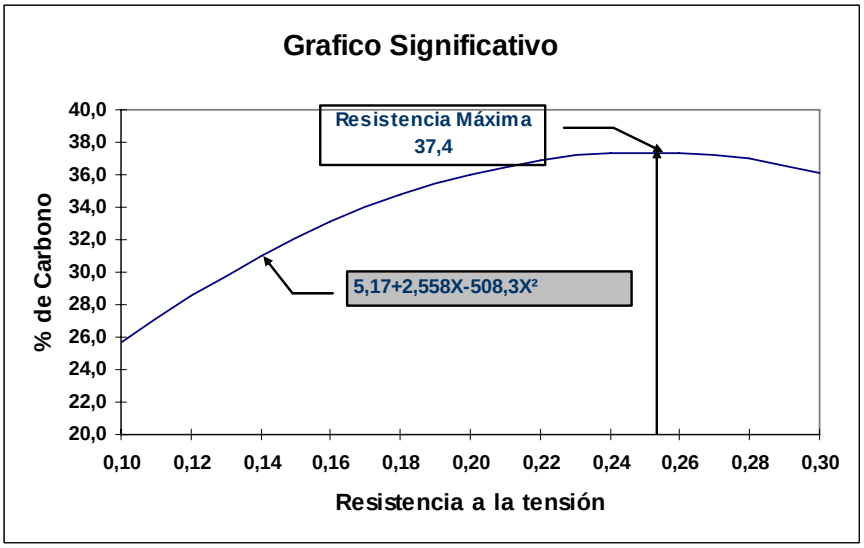
$$T_o = \frac{b_1}{-2b_2} = \frac{255{,}833}{-(2\times -508{,}333)} = 0{,}25$$

Por ciento de carbono con lo que se estima una resistencia de:

$$\hat{\hat{y}}_{ij} = 5{,}1667 + 255{,}8333(0{,}25) - 508{,}333(0{,}25^2) = 37{,}4$$

9.44 Presentación de Resultados.

Para presentar los resultados es suficiente un gráfico, en este caso que muestre que ***Y*** es una variable continua. Indicando además los resultados determinantes.



Es importante usar el gráfico apropiado al tipo de variable y análisis. En este caso, se busca que la resistencia a la tensión sea máxima. El gráfico hace evidente que la tensión aumenta a medida que se le agrega carbono a la colada hasta llegar a 0,25% de donde empieza a disminuir. Un posible siguiente ensayo partirá de este punto central explorando hacia arriba y abajo cantidades más próximas hasta que el resultado se consolide.

9.45 El ANDEVA en experiencias planificadas en modelos lineales de dos factores.

Tema:
Consideraciones.

Posiblemente, el Modelo Lineal más utilizado en la Experimentación Planificada es el esquema en el que uno de los Factores es Operativo y el otro, obviamente Objeto de la Experimentación. Este

esquema se conoce como Diseño en Bloques Completamente al Azar, reminiscencias del pasado agrícola del modelo, en el que el bloque correspondía a una parcela de tierra con características diferenciadas. Ahora podría asociarse el nombre de Bloque con la función que desempeña: una fuente de variación homogénea a su interior y heterogénea con otras de la misma naturaleza, que regularmente no interesa analizar pero que sin esta no podría realizarse el ensayo, además, podría afectar el resultado.

9.46 Descripción del Modelo.

El Modelo Lineal para experimentos con dos Factores se define como:

$$y_{ij} = \bar{y}.. + T_i + B_j + E_{ij}$$

Una ecuación ampliada del modelo completo al azar en el factor $\boldsymbol{B}_j$ representante de los bloques. Usualmente, la inclusión de un factor reduce la cantidad de los efectos no explicados del error. Esto se traduce en diseños más *eficientes.*

La diferencia con el proceso del diseño completo al azar es que los tratamientos o niveles del Factor Objetivo deben asignarse al azar a las repeticiones de cada bloque. Y es más conveniente, tener la misma cantidad de repeticiones por bloque.

9.47 El ejemplo de ANDEVA con dos factores.

Un fabricante que se inicia en el mercado de las botanas mediante un producto innovador que incorpora tubérculos y cormos tropicales de camote, ñáme, dos variedades de tiquisque y yuca. El producto se ofrece en 4 tipo de empaques, 30, 60, 120 y 240 gramos.

Las cantidades que se agregan de cada producto es similar, esto es 1/5 para cada producto.

Los dueños, solicitaron a la Escuela de Ingeniería Agropecuaria Administrativa del Instituto Tecnológico de Costa Rica, se les hiciera una valoración de la calidad del producto.

La primera pregunta que les hizo el estudiante fue:

Si se puede saber ¿qué porcentaje se agrega de cada producto?

La respuesta fue, el mismo para cada producto, esto es: el 20%.

9.48 Planeamiento y Método.

Factor Operativo:

La presentación, con todo y que se empaque por las mismas máquinas puede considerase como factor de variación sin el cual no podría llevarse a cabo la experiencia.

Factor Objetivo:

El porcentaje de cada producto en los envases con 5 niveles con un valor de 20% para cada producto.

El Método de Análisis:

Se usará diseño de dos Factores en Bloques Aleatorios.

Nivel de Confianza:

El 95% o un coeficiente de significación de 5%.

Hipótesis Nula:

Todos los tratamientos son iguales: *Ho*; $T_i = 0$.

9.49 El modelo específico.

De acuerdo al objetivo de la experiencia, habrá una composición homogénea del producto si cada envase contiene un 20% de cada una de las hojuelas. En otras palabras, se espera aceptar la hipótesis nula. El Modelo Lineal que se va a usar para valorar la experiencia es:

$$y_{ij} = \bar{y}.. + T_i + B_j + E_{ij}$$

En donde i = 1,2,...,5 = t productos; j = 1,2,...,6 = b repeticiones por cada producto en los 4 tipo de envases.

El proceso de selección de los productos se realizó adquiriendo 10 bolsas para cada uno de los 4 tipos de envases en seis de los grandes supermercados elegidos al azar en la zona metropolitana de la capital, para obtener el promedio de cada producto.

Los datos individuales, tal como se capturaron en los supermercados se presenta en la HE. Puede elaborar el cuadro siguiente utilizando la Herramienta de manejo de datos de Generador de Tablas Dinámicas para obtener el resumen de la siguiente diapositiva.

El estudiante podrá visualizar que la experiencia podría tratarse como un muestreo.

9.50 Datos de campo y SCY.

Tradicionalmente se usa un cuadro de dos entradas totalizado por factores llamado Datos de Campo como resumen informativo y para efectuar el ANDEVA.

	PRODUCTOS					
Presentación	Camote	Ñame	Tiquisque 1	Tiquisque 2	Yuca	Suma
30	20,5	16,8	23,6	19,1	17,9	97,9
60	22,1	18,1	14,9	13	27,7	95,8
120	18,1	15	18,5	23,7	17,7	93,0
240	22,4	24,6	18,9	16,8	19,8	102,5
Suma	83,1	74,5	75,9	72,6	83,1	389,2

Número de tratamientos: 5 = t;
Número de repeticiones: 4 = r.

La Suma de cuadrados Total.

$$SCY = (t \times r - 1)S_Y^2 = (5 \times 4 - 1)13{,}3373 = 253{,}4080$$

El estudiante puede elaborar el cuadro sumando las cantidades correspondientes o utilizando la rutina de Datos / Informe de tablas y gráficos dinámicos.

9.51 SC de Bloques, Tratamientos y Error.

La Suma de Cuadrados de Presentaciones (Bloques):

$$SCB = \frac{b-1}{t} S_B^2 = \frac{3}{4} 16{,}0467 = 9{,}6280$$

La Suma de Cuadrados de Productos (Tratamientos):

$$SCT = \frac{t-1}{b} S_T^2 = \frac{4}{4} 24{,}4280 = 24{,}4280$$

La Suma de Cuadrados del Error (Residual o no explicada).

$$SCE = SCY - SCB - SCT = 253{,}4080 - 9{,}6280 - 24{,}4280 = 219{,}3520$$

Resultados que se resumen en el cuadro del ANDEVA en la siguiente diapositiva.

9.52 Resumen del ANDEVA.

De acuerdo a los resultados del ANDEVA se concluye que no hay evidencias para declarar diferencias entre presentaciones o entre tratamientos.

No obstante, hay que estar completamente seguros que en presentaciones o en tratamientos hay algún efecto escondido. Para esto se considera el Modelo Lineal completo:

$$\hat{\bar{y}}_{ij} = \bar{y}.. + b_1 B_j^1 + b_{21} B_j^2 + b_3 B_j^3 + b_4 T_j^1 + b_5 T_j^2 + b_6 T_j^3 + b_7 T_j^4 + E_{ij}$$

En el que B y T son los polinomios ortogonales para presentaciones y tratamientos.

ANDEVA de Bloques Aleatorios.

Fuente de Variación	*Grados de Libertad*	*Suma de Cuadrados*	*Cuadrados Medios*	*Cocinete de F. Calculada*	*Probabilidad Significante*	*Límites Significantes P(0,05)*	*P(0,01)*
Total	19	253,4080					
Presentaciones	3	9,6280	3,2093	0,1756	0,9109	3,4903	5,9525
Productos	4	24,4280	6,1070	0,3341	0,8498		
Error	12	219,3520	18,2793				

9.53 Sumas de Cuadrados de los Contrastes.

Para el Contraste lineal de Envases:

$$(30g + 60g) \text{ vs } (120g + 240g) = b_1^2 \sum_{i=1}^{r} B_{1i}^1 = 0{,}1100^2 \times 100 = 1{,}2100$$

Para el Contraste Cuadrático de Envases:

$$(30g + 240g) \text{ vs } (60g + 120g) = b_2^2 \sum_{i=1}^{r} B_{2i} = 0{,}5800^2 \times 20 = 6{,}7280$$

Para el Contraste Cúbico de Envases:

$$(30g + 120g) \text{ vs } (60g + 240g) = b_3^2 \sum_{i=1}^{r} B_{3i} = 0{,}1300^2 \times 100 = 1{,}690$$

La suma de cuadrados de los contrastes es igual a la suma de cuadrados de los Envases, Presentaciones o Bloques:

$$SCB = 1{,}2100 + 6{,}7280 + 1{,}6900 = 9{,}6280$$

Procediendo de la misma forma para Tratamientos se llega al ANDEVA de la siguiente diapositiva.

En donde el coeficiente de regresión se puede obtener usando la función de la HE:

$$b_1 = \frac{\sum_{i=1}^{18} B_i^1 y_{ij}}{SCX} = \text{PENDIENTE}(\$C\$649:\$C\$668; D649:D668) = 0{,}1100$$

O tomándolos directamente del ANDEVA de regresión, puesto que son idénticos.

9.54 El ANDEVA con los contrastes.

ANÁLISIS DE VARIANZA

Fuente de Variación	*Grados de Libertad*	*Suma de Cuadrados*	*Cuadrados Medios*	*Cocinete de F. Calculada*	*Probabilidad Significante*	*Límites Significantes P(0,05)*	*P(0,01)*
Regresión	7	34,056	4,8651	0,2662	0,9559	2,9134	4,6395
Empaques	3	9,6280	3,2093	0,1756	0,9109	3,4903	5,9525
E. Lineal	1	1,2100	1,2100	0,0662	0,8013	4,7472	9,3302
E: Cuadrático	1	6,7280	6,7280	0,3681	0,5554	4,7472	9,3302
E. Cúbico	1	1,6900	1,6900	0,0925	0,7663	4,7472	9,3302
Productos	4	24,4280	6,1070	0,3341	0,8498	3,2592	5,4120
P. Lineal	1	0,09025	0,0903	0,0049	0,9451	4,7472	9,3302
P. Cuadrático	1	20,0402	20,0402	1,0963	0,3157	4,7472	9,3302
P. Cúbico	1	0,3610	0,3610	0,0197	0,8906	4,7472	9,3302
P. Cuarto	1	3,9366	3,9366	0,2154	0,6509	4,7472	9,3302
Residuos	12	219,3520	18,2793				
Total	19	253,4080					

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*
Intercepción	19,4600	0,9560	20,3553	0,0000	17,3770	21,5430
B1	0,1100	0,4275	0,2573	0,8013	-0,8215	1,0415
B2	0,5800	0,9560	0,6067	0,5554	-1,5030	2,6630
B3	0,1300	0,4275	0,3041	0,7663	-0,8015	1,0615
T1	-0,0475	0,6760	-0,0703	0,9451	-1,5204	1,4254
T2	0,5982	0,5713	1,0471	0,3157	-0,6466	1,8430
T3	0,0950	0,6760	0,1405	0,8906	-1,3779	1,5679
T4	0,1186	0,2555	0,4641	0,6509	-0,4381	0,6753

9.55 Conclusión.

EN el ANDEVA para el Modelo Completo no indicó diferencias importantes para los factores. Por tanto, debe concluirse que el porcentaje de hojuelas de cada materia prima en el producto terminado es la misma, de 20% o 1/5.

Dos técnicas adicionales se pueden utilizar:

El ANDEVA solicitado directamente a la HE como: *Herramientas / Análisis de datos / Análisis de la varianza para dos factores con una repetición por grupo*. Cuyo resultado se muestra en el cuadro anterior,

ANÁLISIS DE VARIANZA

Fuente de Variación	*Suma de Cuadrados*	*Grados de Libertad*	*Cuadrados Medios*	*Cocinete de F. Calculada*	*Probabilidad Significante*	*Límites Significantes P(0,05)*	*P(0,01)*
Filas	9,6280	3	3,2093	0,1756	0,9109	3,4903	5,9525
Columnas	24,4280	4	6,1070	0,3341	0,8498	3,2592	5,9525
Error	219,3520	12	18,2793				
Total	253,4080	19					

El ANDEVA de observación por observación.

En todo lo caso, se acepta la hipótesis nula.

9.56 Recomendación.

La segunda técnica consiste en elaborar un ANDEVA con el modelo completo sin promediar por muestra. Con más observaciones se aumenta la confiabilidad de las conclusiones. Como se puede corroborar en la HE.

Es conveniente mencionar que una alternativa apropiada para valorar la hipótesis es analizar las diferencias de los porcentajes con la distribución binomial mediante la comparación de las frecuencias observadas con las frecuencias esperadas.

También es conveniente indicar al estudiante, que situaciones de control de calidad pueden analizarse mediante las cartas de control. Ver Control de la calidad.

ANÁLISIS DE VARIANZA

Fuente de Variación	Grados de Libertad	Suma de Cuadrados	Cuadrados Medios	Cocinete de F. Calculada	Probabilidad Significante	Límites Significantes P(0,05)	P(0,01)
Regresión	7	340,56	48,6514	0,4019	0,9003	2,0575	2,7336
Empaques	3	96,2800	32,0933	0,2651	0,8505	2,6516	3,8852
E. Lineal	1	12,1000	12,1000	0,1000	0,7522	3,8903	6,7687
E: Cuadrático	1	67,2800	67,2800	0,5558	0,4569	3,8903	6,7687
E. Cúbico	1	16,9000	16,9000	0,1396	0,7091	3,8903	6,7687
Prodictos	4	244,2800	61,0700	0,5045	0,7325	2,4187	3,4184
P. Lineal	1	0,9025	0,9025	0,0075	0,9313	3,8903	6,7687
P. Cuadrático	1	200,4018	200,4018	1,6556	0,1998	3,8903	6,7687
P. Cúbico	1	3,6100	3,6100	0,0298	0,8631	3,8903	6,7687
P. Cuarto	1	39,3657	39,3657	0,3252	0,5692	3,8903	6,7687
Residuos	192	23241,12	121,0475				
Total	199	23581,68					

9.57 El ANDEVA en Técnicas de Muestreo.

El Modelo.

El Modelo Lineal para Técnicas de Muestreo puede esquematizarse mediante:

$$y_{ij} = \bar{y}.. + G_i + E_{ij}$$

En donde G_i representa un efecto del agrupamiento, sea:
En muestreos Estratificados;
En muestreos por Conglomerados;
O en muestreos de Razón y Regresión.

9.58 El ejemplo de ANDEVA en muestreo.

Una empresa que vende semen de toros mantiene un proyecto permanente de valoración de la fertilidad del producto. Por lo menos dos veces al año ofrece a una universidad un trabajo para pasantes de veterinaria con el objeto de que realicen el estudio. Este consiste en tomar *n* número de toros en *r* fincas y de los registros de vida, localizar las vacas que se hayan inseminado en los últimos seis meses calificando la preñez con la palpación exitosa del producto.

El Análisis se hace con un nivel de confianza de 95%. Se espera, que no haya diferencias por toro y una fertilidad superior a 62%.

9.59 El conjunto de datos.

Los datos se pueden clasificar por Finca y por Toro, interesa el ordenamiento por toro tal como se muestra en el siguiente cuadro:

	T O R O S								
Nº	1	2	3	4	5	6	7	8	9
1	42,34	39,41	74,48	28,11	74,65	28,74	31,62	63,16	83,62
2	86,49	70,80	67,68	83,61	60,84	69,43	50,76	67,73	48,92
3	70,07	74,77	62,03	30,10	66,46	54,90	60,23	86,74	30,47
4	34,72	38,35	77,48	82,88	86,55	76,57	80,38	34,50	40,23
5	53,43	62,20	76,68	50,72	90,18	69,05	62,71	28,20	87,49
6	48,07	76,05	86,21	58,64	62,78	75,74	80,84	54,21	43,76
7	83,23	38,40	33,32	32,88	36,60	87,31	41,96	47,84	58,28
8	37,47	58,88	59,18	60,73	77,09	69,92	52,99	84,41	42,34
9	80,23	31,29	75,07	62,19	75,43	89,43	69,88	40,87	44,40
10	61,15	43,19	70,37	30,86	55,71	43,47	73,88	55,84	82,86
11	33,98	46,10	53,94	45,67	49,47	42,80	56,48	88,58	81,60
12	54,75	79,03	88,91	91,42	56,01	75,79	85,84	54,28	82,48
13	63,94	48,14	81,17	59,49	35,30			78,93	50,76
14	84,71	62,05	64,14	80,82	75,57			48,31	36,18
15	76,09	46,98	44,09	82,35				65,45	58,53
16	40,01	92,55						62,25	91,92
17	46,70	44,84						62,03	
18	63,48	71,04						50,59	
19		94,68						45,39	
20		93,99						58,37	
21		65,77						97,47	
22								34,88	
23								61,59	
24								82,45	
25								40,80	

9.60 Las Estadísticas Descriptivas.

En estudios de poblaciones por muestreo siempre es importante mostrar las estadísticas descriptivas de la población o los grupos, se esperen o no diferencias entre ellos. Se muestran la varianza y el total en negrillas por el interés que tienen en el cálculo del ANDEVA paso a paso.

Estadístico	*Toro 1*	*Toro 2*	*Toro 3*	*Toro 4*	*Toro 5*	*Toro 6*	*Toro 7*	*Toro 8*	*Toro 9*
Media	58,94	60,88	67,65	58,70	64,47	65,26	62,30	59,79	60,24
Error típico	4,2608	4,2874	3,9664	5,6780	4,4940	5,4305	4,7907	3,6977	5,2791
Mediana	57,95	62,05	70,37	59,49	64,62	69,675	61,47	58,37	54,52
Moda	----	----	----	----	----	----	----	----	----
Desviación estándar	18,0770	19,6475	15,3617	21,9906	16,8150	18,8119	16,5953	18,4887	21,1164
Varianza de la muestra	**326,7779**	**386,0252**	**235,9827**	**483,5880**	**282,7451**	**353,8894**	**275,4049**	**341,8320**	**445,9041**
Curtosis	-1,3671	-1,0158	0,4070	-1,3971	-0,5807	-0,3971	-0,6329	-0,6034	-1,6459
Coeficiente de asimetría	0,1607	0,3256	-0,8193	-0,0369	-0,3342	-0,6995	-0,2985	0,3366	0,2614
Rango	52,51	63,39	55,59	63,31	54,88	60,69	54,22	69,27	61,45
Mínimo	33,98	31,29	33,32	28,11	35,3	28,74	31,62	28,2	30,47
Máximo	86,49	94,68	88,91	91,42	90,18	89,43	85,84	97,47	91,92
Suma	1060,86	1278,51	1014,75	880,47	902,64	783,15	747,57	1494,87	963,84
Cuenta	**18**	**21**	**15**	**15**	**14**	**12**	**12**	**25**	**16**

9.61 Sumas de cuadrados.

Es lógico pensar que un "buen" estimador de la varianza entre grupos los sea el promedio ponderado de las varianzas. En términos de *Sumas de Cuadrados* para usar la igualdad: *SCY = SCG + SCE.*

$$SCE = \sum_{i=1}^{9} (n_i - 1)(S_i^2) =$$

$$= (18-1)326,7779 + (21-1)386,0252 + \ldots + (16-1)445,9041 = 48.840,1723$$

Con la Suma de Cuadrados Total se puede obtener por diferencia la Suma de Cuadrados Entre Grupos:

$$SCY = \left[\left(\sum_{i=1}^{9} n_i\right) - 1\right] S_Y^2 = [(18+21+...+16)-1]340{,}4553 = 50.046{,}9289$$

Finalmente:

$$SCG = SCY - SCE = 20.046{,}9289 - 48.840{,}1723 = 1.206{,}7566$$

9.62 *Resumen del ANDEVA.*

Fuente de Variación	*Grados de Libertad*	*Suma de Cuadrados*	*Cuadrados Medios*	*Cocinete de F. Calculada*	*Probabilidad Significante*	*Límites Significantes P(0,05)*	*P(0,01)*
Total	147	50.046,9289					
Entre Toros	8	1.206,7566	150,8446	0,4293	0,9019	2,0056	2,6417
Dentro de Toros	139	48.840,1723	351,3681				
Promedio	61,67						
Desviación Estándar	18,7448						
Coeficiente de Variación	30,40%						

El ANDEVA indica que no hay motivos para dudar que todos los toros presentar porcentajes de fertilidad similares sobre el 61,67%. Algo más bajo que lo esperado. Esto es, 1,6 ampollas de semen por fecundación exitosa a la palpación aproximadamente entre 2 y 3 meses después de la inseminación.

Puede comprobar los resultados operando la rutina de la HE.

El ANDEVA se puede obtener directamente con la HE con los resultados que se muestran a continuación.

Análisis de varianza de un factor

RESUMEN

Grupos	*Cuenta*	*Suma*	*Promedio*	*Varianza*
1	18	1.060,86	58,94	326,7779
2	21	1.278,51	60,88	386,0252
3	15	1.014,75	67,65	235,9827
4	15	880,47	58,70	483,5880
5	14	902,64	64,47	282,7451
6	12	783,15	65,26	353,8894
7	12	747,57	62,30	275,4049
8	25	1.494,87	59,79	341,8320
9	16	963,84	60,24	445,9041

ANÁLISIS DE VARIANZA

Fuente de Variación	*Suma de Cuadrados*	*Grados de Libertad*	*Cuadrados Medios*	*Cocinete de F. Calculada*	*Probabilidad Significante*	*Límites Significantes P(0,05)*	*P(0,01)*
Entre grupos	1.206,7566	8	150,8446	0,4293	0,9019	2,0056	2,6417
Dentro de los gı	48.840,1723	139	351,3681				
Total	50.046,9289	147					

9.63 *El ANDEVA Anidado Completo.*

En la toma de datos se tienen dos niveles o etapas de muestreo: primero se llega a la finca y dentro de la finca se llega al toro. Este esquema se estudia detalladamente en diseños de muestreo, en este caso, los cálculos se muestran el la HE.

Cada anidamiento provoca un error que debe considerarse en la prueba de ***F*** en donde el cuadrado medio del error de una clase es el cuadrado medio de la clase siguiente.

Del ANDEVA se concluye que no hay efecto *Entre Fincas dentro de toros* P(0,1487); ni efecto entre toros dentro de repeticiones P(0,3841).

Fuente de Variación	*Grados de Libertad*	*Suma de Cuadrados*	*Cuadrados Medios*	*Cocinete de F. Calculada*	*Probabilidad Significante*	*Límites Significantes*	
						P(0,05)	*P(0,01)*
Entre Fincas	5	3.015,9737	603,1947	1,7357	0,1487	2,4495	3,5138
Entre toros dentro de fincas	40	13.900,5884	347,5147	1,0699	0,3841	1,5128	1,7931
Dentro de toros	102	33.130,3668	324,8075				
Total	147	50.046,9289					

Los cálculos para conseguir el ANDEVA rebasan los objetivos del curso, no obstante se pueden entender mejor si se observa el modelo lineal y el equivalente en grados de libertad:

$$Y_{ijk} = \overline{Y}... + F_i + TF_{i(j)} + RTF_{ij(k)}$$

En donde el paréntesis indica el efecto del anidamiento.

$$\sum_{i=1}^{f}\sum_{j=1}^{t}\sum_{k=1}^{r} y_{ijk}^2 - C = \sum_{i=1}^{f}\frac{\left(\sum_{j=1}^{t}\sum_{k=1}^{r} y_{jk}\right)^2}{r_i + t_j} - C + \sum_{i}^{f}\sum_{j}^{t}\frac{\left(\sum_{k=1}^{r} y_{..k}\right)^2}{r_k} - \sum_{i=1}^{f}\frac{\left(\sum_{j=1}^{t}\sum_{k=1}^{r} y_{jk}\right)^2}{r_i + t_j} + SCE$$

$$(n-1) = (f-1) + (m-f) + n - f - m$$

Suma de cuadrados total o de Y se puede obtener de la manera simple:

$$SCY = (n-1)S_Y^2 = (148-1)340{,}4553 = 50.046{,}9289$$

O en la forma tradicional:

$$\sum_{i=1}^{f}\sum_{j=1}^{t}\sum_{k=1}^{r} y_{ijk}^2 - C = 42{,}34^2 + 86{,}49^2 + ... + 91{,}92^2 - \frac{9.126{,}66^2}{n} =$$

$$= 612.857{,}22 - 562.810{,}29 = 50.046{,}93$$

La suma de cuadrados de fincas. Para esta, se suman las observaciones de una finca; la suma se eleva al cuadrado y se divide por el número de observaciones sumadas y finalmente se suman los aportes cuadráticos de cada finca; a esta suma se le resta el corrector:

$$SCF = \sum_{i=1}^{f}\frac{\left(\sum_{j=1}^{t}\sum_{k=1}^{r} y_{jk}\right)^2}{r_i + t_j} - C =$$

$$= \frac{(42{,}34 + 86{,}49 + ... + 40{,}23)^2}{13} + ... + \frac{(84{,}71 + 76{,}09 + ... + 91{,}92)^2}{44} - C =$$

$$= \frac{466.953{,}56}{13} + ... + \frac{8.302.926{,}99}{44} - C = 35.919{,}50 + ... + 188.702{,}89 - C =$$

$$= 565.826{,}26 - 562.810{,}29 = 3.015{,}9737$$

$$GLF = 6 - 1 = 5$$

Suma de cuadrados de toros dentro de las fincas.

$$SCT = \sum_{i}^{f}\sum_{j}^{t}\frac{\left(\sum_{k=1}^{r} y_{..k}\right)^2}{r_k} - CF = \frac{(42{,}34+86{,}49)^2}{2} + \frac{(39{,}41)^2}{1} + ... + \frac{(5076+...91{,}92)^2}{4} - CF =$$

$$= \frac{128{,}83^2}{2} + 39{,}41^2 + ... + \frac{237{,}39^2}{4} - CF = 8.298{,}58 + 1.553{,}15 + ... + 14.088{,}50 - CF =$$

$$= 579.726{,}85 - 565.826{,}26 = 13.900{,}5884$$

$$GLT = 46 - 6 = 40$$

El diseño anidado considera las particularidades de cada finca como son manejo, alimentación, prácticas higiénicas etc. Por esto, debe sustraerse de la variación que tienen los toros la que ocurre en las fincas.

La suma de cuadrados entre repeticiones dentro de toros y fincas se obtiene por diferencia:

$$SCE = SCY - SCF - SCT = 50.046{,}9289 - 3.015{,}9737 - 13.900{,}5884 = 33.130{,}3668$$

$$GLE = 148 - 40 - 5 = 102$$

9.64 Conclusión.

Se han tratado aspectos relevantes del Análisis de Poblaciones mediante el uso de Modelos Lineales y Las Sumas de Cuadrados que implican en un método denominado Análisis de la Varianza, conocida universalmente por sus siglas en los diferentes idiomas como ANDEVA.

Desde el más simple de la Regresión Lineal, pasando por modelos de uso en la Investigación Planificada, para un Factor y para dos Factores, uno Operativo y él otro Objetivo.

Se introdujo el concepto del Polinomio Mínimo y el contraste de Tratamientos mediante Polinomios Ortogonales imbuidos en los niveles de los factores. Y de cómo complementan al ANDEVA.

Finalmente, se mostró el uso del ANDEVA en las técnicas del muestreo.

Cada una de las áreas abordadas, integran apartados que la Teoría Estadística trata por separado. Por tanto, la herramienta del ANDEVA se puede estudiar específicamente en cada subárea.

REFERENCIAS SELECTAS.

1. Buwker, A y Liberman, G. Estadística para ingenieros. Capítulo: Análisis de la Varianza. Prentice Hall, México, 1985.
2. Chou Ya-Lun. Análisis Estadístico. Capítulo 11. Editorial Interamericana S.A. México. 1975.
3. Cristensen, Howard. Estadística Paso A Paso. Capítulo 10. Editorial Trillas, México D.F., 1983.
4. Di Marco, Luis Eugenio. Análisis Estadístico. Capítulo 13. Editorial Interamericana S.A. México. 1969.
5. Jonson R. De Probabilidad y Estadística Para Ingenieros de Millar y Freud. Capítulo 12. Prentice hall Hispanoamericana, S. A., 1997.
6. Kazmier, Leonard. Estadística Aplicada a la Administración y a la Economía. Capítulo 13. Editorial McGraw-Hill, México D. F., 1986.
7. Kempthorne, O. "The Design and Analysis of Experiments". Capítulos 1,2 y 3. Robert E. Krieger Publishing Company, Huntington, N. Y., 1973.
8. Koosis Donad. Elementos de Inferencia Estadística. Capítulo 6. Editorial LIMUSA 1974.
9. Levin Richard. Estadística para Administradores. Capítulo 10. Prentice-Hall, México 1978.
10. Ostle. Bernard. Estadística Aplicada, Técnicas de la Estadística Moderna, ¿cuándo y dónde aplicarlas?. Capítulos 10, 11, 12 y 13. Editorial Limusa, México 1977.

11. Mason, Robet. Douglas, Lind. Estadística Para Administración y Economía. Capítulo 12. Editorial Alfaomega. México. 1992.
12. Mendelhall, William. Introducción a la Probabilidad y Estadística. Grupo Editorial Iberoamericana, S. A. De C. V. México 1987.
13. Miller, I y Freud, J. E. Probabilidad y Estadística Para Ingenieros. Capítulo 13. Editorial Reverte Mexicana, S. A., 1973.
14. Mewbold, Paul. Estadísticas Para Los Negocios y la Economía. Prentice-Hall, Madrid España, 1998.
15. Richartds, L. Estadística En Los Negocios ¿Por Qué y Cuando?. Capítulo 12. McGraw-Hill Latinoamericana, S. A. Bogotá Colombia, 1980.
16. Ríos, Sixto. Ejercicios De Estadística. Capítulo 3. Ediciones ICE, Madrid España, 1977.
17. Snedecor, G. Cochran, W. Métodos Estadísticos. Capítulo 10. Editorial Continental, S.A. Madrid España, 1967.
18. Spiegel, Murray R. Probabilidad y Estadística. Capítulo 2. Editorial McGraw-Hill, México D.F., 1991.
19. Steel, Robert; Torrier, James. "Bioestadística, Principios y Procedimientos" Capítulo 7. Editorial McGraw-Hill, México D. F., 1985.
20. Toranzos Fausto. Teoría Estadística y Aplicaciones. Capítulos 22. Buenos Aires 1971.
21. Walpole, R. E., Myers, R. H. Probabilidad y Estadística para Ingenieros. Capítulo 9. McGraw-Hill Interamericana de México, S. A., México 1991.

10 Regresión Lineal y Correlación.

Los archivos de esta sección son:
E01_RLineal_P01.pps;
E01_RLineal_X01.xls.

10.1 Menú de distribución.

Introducción;
El Modelo de Regresión Lineal Simple;
Ejemplo 1: Caso de Ambas variables Aleatorias;
La Línea de Regresión Estimada;
Las Hipótesis y la Prueba;
La Correlación.: Ejemplo 2;
Ejemplo 3: Caso en que X es un Factor;
Análisis de la Varianza Completo;
Problema 4: Caso en que X se determina (Anualidades);
Regresión con variables Indexadas.

10.2 Objetivos.

La concomitancia de variables siempre ha intrigado al investigador, llevándolo a buscar una explicación lógica a esta asociación entre variables, que muchas veces es fortuita.

- Siempre ha sido la emulación matemática de fenómenos una manera lógica y muy aceptada para explicar la naturaleza de las relaciones.
- Para estudiar un fenómeno de concomitancia, interesa conocer la función matemática que explique la relación.
- Y también, saber con qué precisión y confiabilidad se puede predecir el valor de una variable.
- Los Métodos de Regresión y Los Métodos de Correlación son las Técnicas Estadísticas para lograr los dos objetivos anteriores.

10.3 Método de la Regresión.

Es la Técnica Estadística desarrollada para estudiar la "mejor" relación funcional.
Suele aplicarse en:

- Técnicas de Muestreo;
- Análisis de Experimentos;
- Simples exploraciones estadísticas.

Sólo es indispensable que la variable de Interés sea de naturaleza aleatoria.

10.4 Método de correlación.

Es la *Técnica Estadística* para medir el grado y la naturaleza de la asociación de las distintas variables.

Suele aplicarse en:

- Técnicas de Muestreo;
- Análisis de Experimentos;
- simples exploraciones estadísticas.

Para esto, es insoslayable que *todas las variables* sean de naturaleza aleatoria.

Usualmente, en los estudios por muestreo tanto la variable *Dependiente Y* (rango de la función) como la variable *Independiente X* (dominio de la función) son de naturaleza aleatoria. Es estos casos, *Los Métodos de Regresión y de Correlación* son alternativas de análisis válidas.

10.5 Problema de regresión según Galton.

Galton bautizó la técnica de predicción mediante el modelo lineal simple haciendo una análisis de algo más de mil observaciones de las estaturas de los padres variable *X*, y de los hijos, variable *Y*. En teoría la relación debería ser muy próxima a 1; en el análisis, obtuvo una pendiente de 0,516, concluyendo. "La estatura de los hijos sufre una "*Regresión a la Mediocridad*" con respecto a la estatura de los hijos. De este resultado se adoptó el nombre.

Con este nombre se ha desarrollado toda una técnica estadística que se utiliza en exploraciones mediante el muestreo, en la deducción del efecto de factores en la experimentación planificada y el proyección de futuros con gran éxito.

Sir Francis Galton (16 de febrero de 1822 – 17 de enero de 1911), explorador y científico británico con un amplio espectro de intereses.

No tuvo cátedras universitarias y realizó la mayoría de sus investigaciones por su cuenta. Sus múltiples contribuciones recibieron reconocimiento formal cuando, a la edad de 87 años, se le concedió el título de Sir o caballero del Reino.

De intereses muy variados, Galton contribuyó a diferentes áreas de la ciencia como la psicología, la biología, la tecnología, la geografía, la estadística o meteorología. A menudo sus investigaciones fueron continuadas dando lugar a nuevas disciplinas.

Primo de Charles Darwin, aplicó sus principios a numerosos campos, principalmente al estudio del ser humano y de las diferencias individuales.

10.6 El modelo de regresión básico.

La relación funcional lineal más simple se puede aproximar mediante el modelo:

$$Y_i = b_0 + b_1 X_i$$

En este:

Y_i es la el valor de variable de interés en la í-sima observación, obligadamente de naturaleza aleatoria. En matemáticas el *Rango de la Función.*;

b_0 ó *Interceptada* y b_1 o *Pendiente*, son los parámetros que definen la función lineal;

X_i es el valor de la variable concomitante de la í-sima observación. No necesariamente será de naturaleza aleatoria. En matemáticas el *Dominio de la función.*

10.7 Caso de ambas variables aleatorias.

Usualmente, en los estudios por muestreo tanto la variable *Dependiente Y* como la variable *Independiente X* son de naturaleza aleatoria. Es estos casos, *Los Métodos de Regresión y de Correlación* son alternativas de análisis válidas.

Problema:

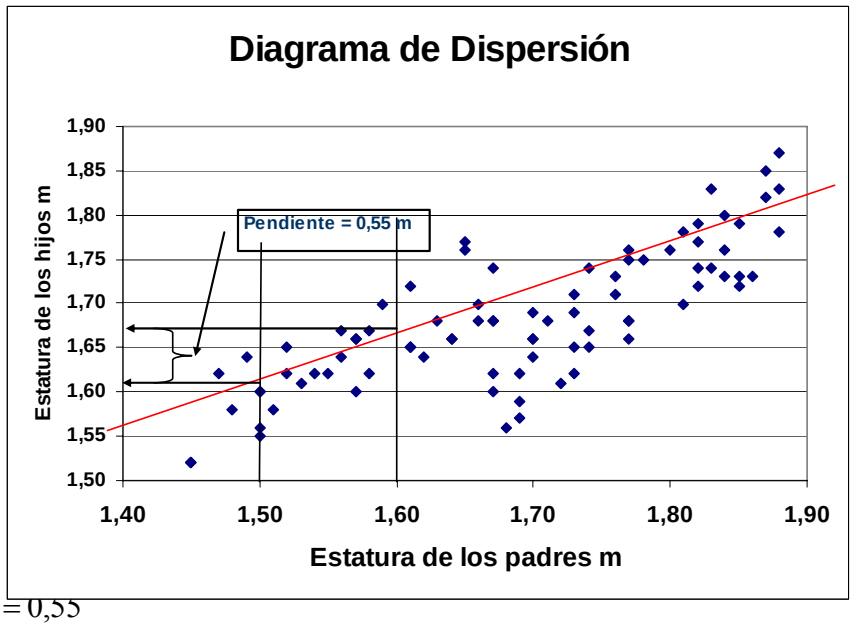
Galton bautizó la técnica de predicción mediante el modelo lineal simple haciendo una análisis de algo más de mil observaciones de las estaturas de los padres variable *X*, y de los hijos, variable *Y*. En teoría la relación debería ser muy próxima a 1; en el análisis, obtuvo una pendiente de 0,516, concluyendo. "La estatura de los hijos sufre una "*Regresión a la Mediocridad*" con respecto a la estatura de los hijos. De este resultado se adoptó el nombre.

Un profesor pidió a sus estudiantes recopilaran muestras de estaturas de padres e hijos varones para emular el estudio de regresión de Galton.

10.8 El diagrama de dispersión.

El punteo de datos en un plano cartesiano se conoce como *Diagrama de Dispersión.*

Ordenando los datos ascendentemente tomado como base la estatura de los padres. Usando el Gráfico *X, Y* de la HE y acomode a mano alzada una línea que ajuste de manera aproximada a los datos. Obtendrá un diagrama parecido al siguiente.



La pendiente se puede calcular a partir del gráfico usando la fórmula:

$$b_1 = \frac{y_{i+1} - y_i}{x_{i+1} - x_i} = \frac{1{,}670 - 1{,}615}{1{,}60 - 1{,}50} = 0{,}55$$

10.9 La Pendiente.

La ***Pendiente*** es sin duda, el indicador más importante de la relación funcional entre dos variables.

Indica el *Incremento o Decremento* de la variable *Y* (rango) a un *Incremento o Decremento* unitario de la variable *X* (dominio).

El estimado de la pendiente $b_1 = 0{,}55$ indicará al investigador que en promedio, un metro en el aumento de la estatura de los padres se reflejará en los hijos en promedio 0,55 metros.

En este ejemplo, un incremento en la estatura de los padres se refleja en un incremento en la estatura de los hijos, por tanto, la relación es positiva.

Habrá otros problemas en los que un incremento en la variable *X* se traduzca en decrementos de la variable *Y*, entonces la relación será negativa.

10.10 Obteniendo el modelo con la Hoja Electrónica.

Definitivamente, para aproximar resultados es preferible utilizar una estructura firme, por ejemplo el modelo matemático de la *Regresión Lineal Simple*:

$$y_i = b_0 + b_1 x_i$$

La Interceptada es b_0; *La Pendiente* es b_1; y_i la estatura del i-ésimo hijo; x_i la estatura del i-ésimo padre. Que la HE calcula directamente para cada uno de los parámetros dando por resultado el modelo:

$$y_i = 0{,}8581 + 0{,}4885 x_i$$

Siendo las ecuaciones normales:

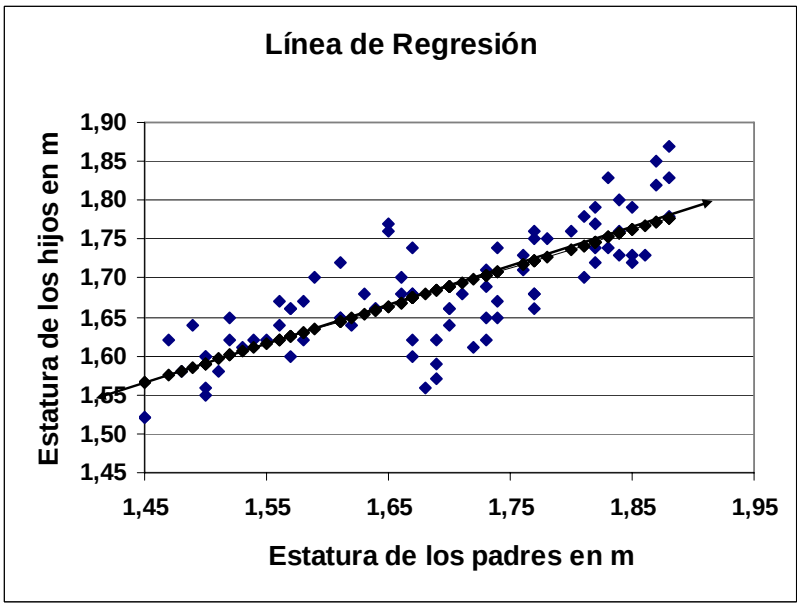
$$b_0 = \bar{y} - b_1\bar{x} = \text{INTERSECCION.EJE(Rango Y : Rango X)} = 0{,}8581$$

$$b_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \text{PENDIENTE(Rango Y; Rango X)} = 0{,}4885$$

Nota: Si incluyen las fórmulas de cálculo para que el estudiante se empiece a acostumbrar a relacionar la notación matemática y las instrucciones de la HE para efectuar un cálculo que puede leer directamente en la celda que ofrece la cantidad y aparece como fórmula o función.

10.11 La línea de regresión estimada.

Con tal modelo se estima la línea de regresión qué se acomoda de maneta tal que aproxima al diagrama de dispersión de los a los datos. Esta se puede observar en el gráfico como una línea sobre una serie de puntos continuos.Esta línea que *Ajusta Los Datos* posee cualidades muy deseables en un modelo de aproximación y predicción que se irán desvelando a medida que se avance en el tema.



10.12 ¿Es la línea de mejor ajuste?

En el gráfico anterior es notorio que hay diferencias entre la línea estimada y los puntos observados. Si se supone que la línea es un promedio de las estaturas de los hijos para cada una de las estaturas de los padres muestreadas, debe cumplirse que la suma de las diferencias de los valores observados menos los esperados es cero, esto es:

$$\sum_{i=1}^{n} d_i = \left[y_i - \left(b_0 + b_1 x_i\right)\right] = 0$$

Por ejemplo para d_1:

$$d_1 = 1{,}52 - \left(0{,}8581 + 0{,}4885 \times 1{,}45\right) = -0{,}0464$$

Proposición que se cumple según la prueba práctica de la HE. Si esto ocurre, la suma de cuadrados de las diferencias será mínima, con un valor de:

$$\sum_{i=1}^{n} d_i^2 = -0{,}0464^2 + -0{,}0464^2 + ... + 0{,}0935^2 = 0{,}0022 + 0{,}0022 + ... + 0{,}0087 = 0{,}1866$$

10.13 Estadística descriptiva.

Ambas variables son de naturaleza: continua y además aleatorias, por tanto, se puede interpretar la estadística descriptiva en ambas.

En este momento, de especial interés es la suma de cuadrados agregada. Se requiere para efectuar comparaciones de variación importantes para comprender la justificación teórica del potencial de la regresión en la estimación estadística.

Estadísticos	Padres	Hijos
Media	1,6866	1,6820
Error típico	0,0014	0,0009
Mediana	1,69	1,68
Moda	1,67	1,62
Desviación estándar	0,1225	0,0762
Varianza de la muestra	0,0150	0,0058
Suma de Cuadrados	1,2595	0,4872
Curtosis	-1,0378	1,6820
Coeficiente de asimetría	-0,19	0,21
Rango	0,43	0,35
Mínimo	1,45	1,52
Máximo	1,88	1,87
Suma	143,36	142,97
Cuenta	85	85

Recuerde que:

$$S_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{SCY}{n-1}$$ ***Por tanto:***

$$SCY = (n-1)S_y^2$$

Cantidad que se conoce como Suma de Cuadrados de Y. Entenderá que la fórmula es genérica aplicable a cualquier variable aún cuando esta no sea aleatoria.

10.14 Las Sumas de Cuadrados.

De la misma variable *Y* se tienen dos sumas de cuadrados o varianzas. Las usuales de las estadísticas descriptivas que llamaremos *Suma de Cuadrados* de *Y* o Suma de *Cuadrados Total*, definida por:

$$SCY = (n-1)S_Y^2 = (85-1)0{,}0058 = 0{,}4872$$

Y la obtenida de las diferencias cuadráticas de los valores observados con respecto a los valores esperados que se identificarán como *Suma de Cuadrados del Error*:

$$SCE = \sum_{i=1}^{n} d_i^2 = \left[y_i - (b_0 + b_1 x_i)\right]^2 = 0{,}1866$$

¡No son iguales! Entonces ¿en donde está la suma de cuadrados faltante?

10.15 La Suma de Cuadrados de la Regresión.

Se considerará que existe una fuente de variación adicional y se supondrá que proviene de la relación entre la variable ***Y*** con la variable ***X***. De esta manera se puede establecer la siguiente relación:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = SCR + \sum_{i=1}^{n}\left[y_i - (b_0 + b_1 x_i)\right]^2 = \{SCT = SCR + SCE\}$$

De la ecuación anterior denominada *Ecuación de las Sumas de Cuadrados* y resolviendo para *SCR* se obtiene que:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = SCR + \sum_{i=1}^{n}\left[y_i - (b_0 + b_1 x_i)\right]^2 =$$

Resolviendo para SCR:

$$SCR = \sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}[y_i - (b_0 + b_1 x_i)]^2 =$$

Sustituyendo b_0 para obtenerlo en términos de b_1:

$$SCR = \sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}[y_i - (\{\bar{y} - b_1\bar{x}\} + b_1 x_i)]^2 =$$

$$= \sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}[(y_i - \bar{y}) + b_1(x_i - \bar{x})]^2 =$$

Se puede probar que los términos dentro del paréntesis cuadrado del segundo elemento del lado izquierdo no interactúan, en términos de sumas de cuadrados se tendría:

$$SCR = \sum_{i=1}^{n}(y_i - \bar{y})^2 - \left[\sum_{i=1}^{n}(y_i - \bar{y})^2 + b_1^2\sum_{i=1}^{n}(x_i - \bar{x})^2\right]$$

Finalmente:

$$SCR = b_1^2\sum_{i=1}^{n}(x_i - \bar{x})^2 = 0{,}4885^2 \times 1{,}2595 = 0{,}3006$$

La Suma de Cuadrados de Regresión resuelve La Ecuación de la Suma de Cuadrados fundamental, esto es:

$$SCT = SCR + SCE;$$

$$0{,}4872 = 0{,}3006 + 0{,}1866$$

10.16 Las varianzas o Cuadrados Medios.

Cada una de las sumas de cuadrados divididos por sus respectivos grados de libertad son respectivamente: la *Varianza Total*, *La Varianza de Regresión* y *La Varianza del Error*, llamados *Cuadrados Medios*.

Por razones que se verán más adelante, la atención debe ponerse en los siguientes cuadrados medios:

- *El Cuadrado Medio de Regresión* cuyo origen de variación es conocido, en el ejemplo la parte genética del aporte de los padres a la estatura de los hijos;
- Y *El Cuadrado Medio del Error* cuya fuente es desconocida o no interesa discernir.

10.17 La hipótesis y la prueba.

Interesa conocer sí las dos componentes de variación total son iguales. Esto es:

$$Ho; \mathrm{CM}_{(\text{Regresión})} = CM_{(Error)};\ \text{ó}\ Ho; S_R^2 = S_E^2$$

La Teoría Estadística desarrolló una prueba para valorar dos varianzas mediante cociente llamado de *F*. Esto es:

$$F_{(x;\,GLR;\,GLE)} = \frac{S_R^2}{S_E^2} = \frac{SCR/(c-1)}{SCE/(n-2)}$$

Evidentemente, sí $F = 1$ las varianzas serán iguales, cuando el número de observaciones es alto. Para compensar esto, se ha desarrollado la función de densidad de ***F*** que considera los grados de libertad. En el ejemplo el valor calculado de *F*:

$$F_{(x;1\,;83)} = \frac{0{,}3006/1}{0{,}1866/83} = \frac{0{,}3006}{0{,}0022} = 133{,}6997$$

La función de densidad de $F_{(133{,}7;1;83)} = 0{,}000..$, indica la probabilidad de que ambas variables sean iguales.

Esquemáticamente:

$$F(133,6997;1;83)=\int_0^{133,6997} f\left(\frac{0,3006}{0,0022};1;83\right)df=5,5777\text{E-}19$$

10.18 El Cuadro del ANDEVA.

Se ha venido trabajando con la finalidad de mostrar en un cuadro sinóptico el resumen de las fuentes de variación que componen el comportamiento de una variable aleatoria bajo la influencia de un *Factor*.

Otra manera de expresarlo sería:

El ANDEVA es un resumen de las variaciones involucradas en el análisis de poblaciones cuando se utiliza un modelo lineal. En todo caso, el sistema se utiliza para probar hipótesis estadísticas.

La probabilidad indica cuánto se aproxima la regresión (el coeficiente b_1) a 0. Ó si ***F*** es mayor a alguno de los criterios debe rechazarse la hipótesis.

10.19 Cálculos para el ANDEVA: Cálculo de estadísticos.

Aún cuando es poco probable que tenga que efectuar los cálculos para el análisis de la varianza sin un equipo de computación, es conveniente que sepa que cálculos debe considerar para efectuar un análisis manual.

	De Y	De X	Productos XY
Número	85		
Suma	142,97	143,36	
Promedio	1,6820	1,6866	
Corrector por la media	240,4755	241,7893	241,1315
Sumas Cuadraticas	240,9627	243,0488	241,7468
Sumas de Cuadrados	0,4872	1,2595	0,6153
Pendiente	0,4885		
Interceptada	0,8581		

Número de Observaciones: ***85***

Sumas totales:

$$\sum_{i=1}^{85} y_i = 1,64+1,62+\ldots+1,60=142,97$$

$$\sum_{i=1}^{85} x_i = 1,49+1,52+\ldots+1,50=143,36$$

Promedios:

$$\bar{y}=\frac{\sum_{i=1}^{85} y_i}{n}=\frac{142,97}{85}=1,6820$$

$$\bar{x}=\frac{\sum_{i=1}^{85} x_i}{n}=\frac{143,36}{85}=1,6866$$

Los Correctores por los promedios:

$$C_y = \frac{\left(\sum_{i=1}^{85} y_i\right)^2}{85} = \frac{(142{,}97)^2}{85} = 240{,}4755$$

$$C_x = \frac{\left(\sum_{i=1}^{85} x_i\right)}{85} = \frac{(143{,}36)^2}{85} = 241{,}7893$$

$$C_{xy} = \frac{\left(\sum_{i=1}^{85} x_i\right)\left(\sum_{i=1}^{85} y_i\right)}{85} = \frac{(142{,}97)(143{,}36))}{85} = 241{,}1315$$

Sumas de cuadrados y productos cruzados sin corregir.

$$\sum_{i=1}^{85} y_i^2 = 1{,}64^2 + 1{,}62^2 + \ldots + 1{,}60^2 = 240{,}9627$$

$$\sum_{i=1}^{85} x_i^2 = 1{,}49^2 + 1{,}52^2 + \ldots + 1{,}50^2 = 243{,}0488$$

$$\sum_{i=1}^{85} x_i y_i = 1{,}49 \times 1{,}64 + 1{,}52 \times 1{,}62 + \ldots + 1{,}50 \times 1{,}60 = 241{,}1315$$

Sumas de Cuadrados y Productos cruzados.

$$SCY = \sum_{i=1}^{85} y_i^2 - C_y = 240{,}9627 - 240{,}4755 = 0{,}4872$$

$$SCX = \sum_{i=1}^{85} x_i^2 - C_x = 243{,}0488 - 241{,}7893 = 1{,}2595$$

$$SCXY = \sum_{i=1}^{85} x_i y_i - C_{xy} = 241{,}7468 - 241{,}1315 = 0{,}6153$$

Coeficiente de la pendiente b_1:

$$b_1 = \frac{\sum_{i=1}^{85} x_1 y_i - C_{xy}}{\sum_{i=1}^{85} x_i^2 - C_x} = \frac{SCXY}{SCX} = \frac{0{,}6153}{1{,}2595} = 0{,}4885$$

Coeficiente de la intersectada b_0:

$$b_0 = \bar{y} - b_1 \bar{x} = 1{,}6820 - 0{,}4885 \times 1{,}6866 = 0{,}8581$$

Estos cálculos se resumen en el cuadro anterior.

10.20 Cuadro de la varianza o ANDEVA

Cuadro de ANDEVA completo

Fuente de la Variación	*Grados de Libertad*	*Sumas de Cuadrados*	*Cuadrados Medios*	*Cociente de Fc.*	*Probabilidad de F*	*Valores Críticos de F*	
						P < 0,05	*P < 0,01*
Regresión	1	0,3006	0,3006	133,6997	5,578E-19	3,9560	6,9504
Error (o Residual)	83	0,1866	0,0022				
Total	84	0,4872					
Promedio de Y	1,6820						
Desvío Estándar	0,0474						
Coeficiente Variación	2,82%						
Coeficiente Determinación	61,70%						

Como se mencionó con anterioridad, el investigador decide la probabilidad de significación para la prueba. Si la probabilidad del cociente de ***F*** es igual o menor al nivel de significación o si el valor del estadístico ***F*** es igual o mayor al criterio de ***F*** elegido deberá rechazarse la hipótesis nula que dice que no hay relación entre las variables *X* e *Y*. Se ofrece un ANDEVA con más información.

Se puntualizan los cálculos para: El Desvío Estándar:

$$S_e = \sqrt{CME} = \sqrt{0{,}4872} = 0{,}0474$$

El Coeficiente de Variación:

$$CV = \frac{S_e}{\bar{y}}\% = \frac{0{,}0474}{1{,}6820} = 2{,}82\%$$

El Coeficiente de Determinación:

$$r^2 = \frac{SCR}{SCT}\% = \frac{0{,}3006}{0{,}4872}\% = 61{,}70\%$$

10.21 Prueba de hipótesis sobre la Interceptada.

En ocasiones, el coeficiente de la *Interceptada* tiene poca utilidad deductiva e inductiva, como es el caso, pues indica que cuando el padre mida 0 metros el hijo medirá 0,8581 metros. Siempre se reporta en el ANDEVA de regresión los elementos para valorar la hipótesis nula:

$Ho; \beta_0 = 0$ contra $Ha: \beta_0 \neq 0$

En palabras: establecer que la línea de regresión pasa por 0 cuando la variable independiente tiene valor 0. Los estimadores de la regresión se aglomeran sobre los parámetros con una distribución normal. Usualmente, se utiliza la Distribución de "t" de *Student* para aproximarse considerando los grados de libertad para la prueba. El estadístico *tc* se obtiene con la siguiente fórmula:

$$t_c = \frac{b_0}{S_{bo}} = \frac{b_0}{S_e\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SCX}}} = \frac{0{,}8581}{0{,}0474\sqrt{\frac{1}{85} + \frac{1{,}6866^2}{1{,}2595}}} = 12{,}0113$$

La probabilidad que determina el estadístico tc se ejemplifica como:

$$F(12{,}0113;83) = Y_0\int_0^{\infty}\left(1 + \frac{12{,}0113^2}{83}\right)^{\frac{83+1}{2}} dt = 7{,}622E-20$$

La probabilidad de la intersectada β_0 pase por el cruce de las coordenadas $x = 0$ e $y = 0$ es prácticamente cero.

Es lógico, puesto que la estatura de un humano nunca es 0.

10.22 Intervalo de confianza para la intersectada.

Como una ampliación de la prueba de "*t*" se obtiene el intervalo de confianza para la interceptada.

$$\Pr\{b_0 - t(S_{b0}) \geq \beta_0 \leq b_0 + t(S_{b0})\} = 1 - \alpha$$

$$\Pr\{0{,}8581 - 1{,}9890(0{,}0714) \geq \beta_0 \leq 0{,}8581 + 1{,}9890(0{,}0714)\} = 95\%$$

$$\Pr\{0{,}7160 \geq \beta_0 \leq 1{,}0002\} = 95\%$$

Se reitera que en este caso, la interpretación el valor de la *Interceptada* no siempre permite conclusiones lógicas. No obstante, siempre se ofrecen los límites confiables como marcos de referencia para estudios similares.

En el ejemplo se espera que el Parámetro β_0 se encuentre entre 0,72 y 1,00 metros en 19 de cada 20 ensayos, cuándo los padres tengan altura 0. Galton encontró 0,84 metros, por tanto, los resultados son consistentes.

El valor de t para una probabilidad de 0,05 y 83 grados de libertad se obtiene directamente de la HE o consultando las Tablas Estadísticas. Se esquematiza mediante:

$$T_{(0,05;83)} = \text{DISTR.T.INV}(0,05;83) = 1,9890$$

Notará que se aproxima mucho al valor de $z_{(0,025\ o\ 0,975)}$ = 1,96

10.23 Prueba de hipótesis sobre la Pendiente.

Sin duda, el parámetro importante en la regresión es la pendiente. Para una sola variable ***X*** la probabilidad de *Fc* del ANDEVA es idéntica a la que se obtiene usando una prueba de "*t*" para valorar la hipótesis:

$$Ho; \beta_1 = 0 \text{ contra } Ha: \beta_1 \neq 0$$

En Palabras: probar que no hay relación entre *Y* y *X*. La t_c se obtiene mediante:

$$t_c = \frac{b_1}{S_{b1}} = \frac{b_1}{\frac{S_e}{\sqrt{SCX}}} = \frac{0,4885}{\frac{0,0474}{\sqrt{1,2595}}} = \frac{0,4885}{0,0422} = 11,5629$$

Valor que determina una probabilidad de 0,000.. De que $B1$ sea cero. Existe una relación entre la estatura de los padres y la estatura de los hijos.

Puede comprobar que elevando al cuadrado esta t obtiene el valor de la F.

Elevando al cuadrado este valor de t se obtiene el valor del estadístico F:

"tc"^{2} = 11,5629^2 = 133,6997 = Fc.

La probabilidad que determina el estadístico tc se ejemplifica como:

$$F(11,5629;83) = Y_0 \int_0^{\infty} \left(1 + \frac{11,5629^2}{83}\right)^{\frac{83+1}{2}} dt = 5,577E-19$$

10.24 Intervalo de confianza para la Pendiente.

En un análisis de *Regresión*, es insoslayable que se presente el intervalo confiable para la *Pendiente* β_1. En este caso, para ejemplificar se usará un intervalo confiable de 99%

$$\Pr\{b_1 - t(S_{b1}) \geq \beta_1 \leq b_1 + t(S_{b1})\} = 1 - \alpha$$

$$\Pr\{0,4885 - 2,6364 \times 0,0422 \geq \beta_1 \leq 0,4885 + 2,6364 \times 0,0422\} = 95\%$$

$$\Pr\{0,3771 \geq \beta_1 \leq 0,5999\} = 95\%$$

Se espera que el verdadero parámetro de la *Pendiente* se encuentre entre valores que van desde 0,38 hasta 0,60 metros por cada metro en la estatura de los padres con una probabilidad del 99%. Galton encontró, 0,516 unidades de estatura en los hijos por unidad paterna. Estos resultados son consistentes.

10.25 Bandas de Confianza.

Dado que la línea de regresión cubre una infinidad de puntos y_i asociados con cada x_i los intervalos de confianza se estiman en todo el recorrido de la línea de regresión. En el plano cartesiano parecen bandas a ambos lados de la línea estimada.

Se acostumbra presentar dos tipos de intervalos confiables:

- Para promedios, que el *Teorema Central del Límite* asegura que siempre serán válidos si se cumple que la variable *Y* se distribuya normal o se trabaja con promedios;
- Y para observaciones, válidos únicamente si la distribución Y_i (rango) es normal en cada punto X_i (del dominio).

En las siguientes diapositivas se ofrecen las fórmulas y resultados.

10.26 Bandas de confianza para promedios.

Los intervalos de confianza para promedios, estiman como se espera, promedios de la variable $\boldsymbol{Y_i}$ —promedio de estaturas de los hijos —en cada punto $\boldsymbol{X_i}$ —estatura de los padres. —Se acostumbra presentarlos en todo el recorrido de $\boldsymbol{X}$, esto es desde la estatura más baja de 1,45 hasta la mayor de 1,88 a espacios regulares. Se obtiene aplicando la siguiente fórmula en cada punto.

$$\Pr\left\{\hat{y}_i - t_{(\alpha;\text{n-2})}S_{\bar{y}_i} \geq \overline{Y}_i \leq \hat{y}_i + t_{(\alpha;\text{n-2})}S_{\bar{y}_i}\right\} = 1-\alpha$$

Y en donde la Desviación Típica en cada punto x_i es:

$$S_{\bar{y}_i} = S_e\sqrt{\frac{1}{n}+\frac{(x_i-\bar{x})^2}{SCX}} = 0{,}0474\sqrt{\frac{1}{85}+\frac{(1{,}45-1{,}6866)^2}{1{,}2595}} = 0{,}0112$$

La estimación se presentará más adelante. Se ejemplifica con la estatura del padre de 1,45:

$$\Pr\left\{1{,}56 \geq \overline{Y}_i \leq 1{,}58\right\} = 1-\alpha$$

10.27 Bandas de confianza para observaciones.

Los intervalos de confianza para observaciones, estiman valores de individuos de la variable $\boldsymbol{Y}_i$ —estatura de los hijos— en cada punto $\boldsymbol{X}_i$ —estatura de los padres—. Se procede igual que con los promedios, en el recorrido de $\boldsymbol{X}$. Debe hacerse hincapié que la extrapolación, esto es, ir más allá del recorrido de $\boldsymbol{X}$, debe tomarse con cuidado en ambas bandas de confianza.

$$\Pr\left\{\hat{y}_i - t_{(\alpha;\text{n-2})}S_{y_i} \geq Y_i \leq \hat{y}_i + t_{(\alpha;\text{n-2})}S_{y_i}\right\} = 1-\alpha$$

Y en donde la Desviación Típica en cada punto x_i es:

$$S_{y_i} = S_e\sqrt{1+\frac{1}{n}+\frac{(x_i-\bar{x})^2}{SCX}} = 0{,}0474\sqrt{1+\frac{1}{85}+\frac{(1{,}45-1{,}6866)^2}{1{,}2595}} = 0{,}0487$$

La estimación se presentará más adelante. Se ejemplifica con la estatura del padre de 1,45:

$$\Pr\left\{1{,}57 - 1{,}989\times 0{,}0487 \geq \text{Y}_{(1{,}45)} < 1{,}57 + 1{,}989\times 0{,}0487\right\} = 95\%$$

Recordar:

$$\hat{y}_i = 0{,}8581 + 0{,}4885x_i$$

10.28 Tabla de bandas de confianza.

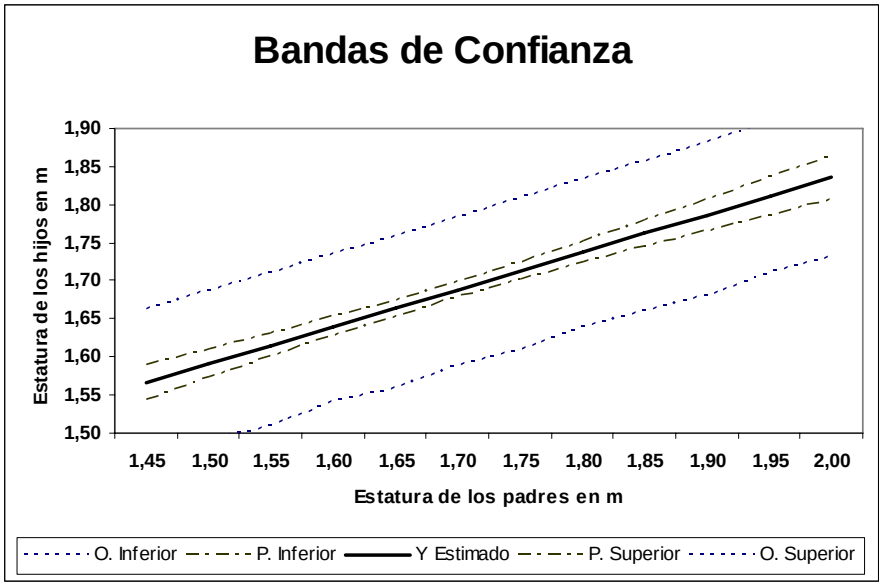
La tabla de valores estimados muestra que los intervalos de confianza para los promedios se estiman más próximos a la línea de regresión mientras los intervalos confiables para las observaciones son más amplios. Puesto que las estaturas se

Nº	Valor de X	O. Inferior	P. Inferior	Y Estimado	P. Superior	O. Superior
1	1,45	1,46	1,54	1,57	1,59	1,66
2	1,50	1,49	1,57	1,59	1,61	1,69
3	1,55	1,51	1,60	1,62	1,63	1,71
4	1,60	1,54	1,63	1,64	1,65	1,73
5	1,65	1,56	1,65	1,66	1,67	1,76
6	1,70	1,59	1,68	1,69	1,70	1,78
7	1,75	1,61	1,70	1,71	1,72	1,81
8	1,80	1,64	1,72	1,74	1,75	1,83
9	1,85	1,66	1,74	1,76	1,78	1,86
10	1,90	1,68	1,77	1,79	1,81	1,88
11	1,95	1,71	1,79	1,81	1,84	1,91
12	2,00	1,73	1,81	1,84	1,86	1,93

distribuyen normales, los valores extremos serán menos frecuentes, situación que se refleja en bandas de confianza más amplias en los extremos.

Muchas tablas de uso como parangones de estaturas, pesos y otras constantes fisiológicas, físicas, químicas etc., se elaboran mediante modelos de regresión e intervalos de confianza en donde se refieren mínimos y máximos de situaciones estándar.

10.29 Gráfico de las bandas de confianza.



Generalmente, el Análisis de Regresión Concluye con este gráfico.

La línea central corresponde a los valores estimados. Sobre estos se calculan los límites de confianza para cada x_i.

Los más próximos corresponden a los intervalos de confianza para los promedios. Los más alejados serán los intervalos de confianza para los valores individuales. Así, el promedio de estatura de los hijos varones de una familia se estima con las bandas más estrechas. Una estatura individual se estimará con las bandas más amplias.

10.30 La Correlación.

Para poder ofrecer conclusiones mediante análisis de correlación, es insoslayable que las variables involucradas sean de naturaleza aleatoria. Esto no impide, por ejemplo, calcule la correlación entre niveles de fertilizante aplicados a una planta y el rendimiento, o las ventas y el costo del producto de las mismas de una empresa en una serie de años. Será incorrecto que reporte por ejemplo que la correlación entre ventas y años es significativa, puesto que los años no son una variable aleatoria, para estos casos se usa la regresión.

Ambos coeficientes, él de correlación y él de regresión están íntimamente relacionados, pero su uso es diferente.

10.31 El ejemplo de correlación.

El ejemplo consiste en agregar un nuevo conjunto de 85 observaciones que incluye el peso de padres e hijos varones.

Todas las variables *X*1 = *La Estatura de los Padres*; *X*2 = *El Peso de los Padres*; *X*3 = *La Estatura de los hijos*; *X*4 = *El peso de los hijos* son de naturaleza aleatoria.

El Objetivo del análisis:

Conocer las relaciones entre las variables mencionadas en un estudio genético de peso y estatura entre hijos y padres varones.

10.32 Definición de correlación.

La correlación mide la relación concomitante entre dos variables aleatorias. Esta definida por la ecuación:

$$\rho = \frac{S_{xy}}{\sqrt{S_x S_y}} = \frac{\dfrac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{n-1}}{\sqrt{\dfrac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}\sqrt{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}}$$

Es el cociente de la coovarianza dividida por las desviaciones estándar de las varianzas. Un indicador que corre de –1 cuando una variable disminuye mientras la otra aumenta hasta +1 cuando una variable aumenta mientras la otra también lo hace y 0 cuando no hay relación entre las variables.

10.33 El cálculo y la prueba.

Los coeficientes de correlación se pueden obtener aplicando, para cada par de variables la fórmula de la diapositiva anterior, solicitándolo a la HE por cada par de variables la opción CORRELACION(Y;X) u operando el algoritmo de *Coeficiente de Correlación* de la HE. Por ser más común y simple, se usará este último con los resultados mostrados en el cuadro.

	X1 = E. Padre	X2 = P. Padre	X3 = E. Hijos	X4 = P.Hijos
X1 = E. Pa	1,0000			
X2 = P. Pa	**0,4250**	1,0000		
X3 = E. Hij	**0,7807**	**0,25662**	1,0000	
X4 = P.Hijc	**0,3335**	0,08213	0,40568	1,0000

10.34 Estadísticos para la prueba.

Según la definición, entre más se acerque el coeficiente de correlación a 1 o –1 mayor será la relación funcional ente las variables. Sin embargo, es necesario usar una prueba estadística de significación. Por la relación que tiene con el coeficiente de regresión, el estimador muestral del coeficiente de correlación se distribuye alrededor del parámetro con una distribución normal o como una "*t*" de *Students* con $n - 2$ grados de libertad. Ejemplificando con la estatura y peso del padre:

$$t_{(n-2)} = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{|0{,}4250|\sqrt{85-2}}{\sqrt{1-0{,}4250^2}} = 4{,}2770$$

O como una $\boldsymbol{F_{(1;\ n-2)}}$ en donde el coeficiente de correlación $\boldsymbol{r}$ se transforma en el coeficiente de determinación $\boldsymbol{r}^2$.

$$F_{(1;\,n\text{-}2)} = \frac{r^2(n-2)}{1-r^2} = \frac{0{,}4250^2(85-2)}{1-0{,}4250^2} = 18{,}2927$$

10.35 Valorando la hipótesis.

La hipótesis que se valora para el coeficiente de determinación es:

$Ho; \rho = 0 : \text{Contra } Ha : \rho \neq 0$

Los estadísticos "*t*" o *F* se valoran directamente utilizando la función de densidad de la HE que indicará la probabilidad que va del punto determinado por el valor del estadístico a mas infinito. Dicho de otra forma, la probabilidad de la zona de rechazo. Para la t:

$$F(4{,}2770;83) = Y_0 \int_0^{\infty} \left(1 + \frac{4{,}2770^2}{83}\right)^{\frac{83+1}{2}} dt = 0{,}0001$$

O para F:

$$F(18{,}2927;1;83) = \int_0^{18{,}2927} f\left(\frac{14{,}989}{0{,}819};1;83\right) df = 0{,}0001$$

Se puede usar el método antiguo, de cuando no se contaba con la facilidad de la HE comparando el valor del estadístico contra el valor que determina una probabilidad de significación definida, por ejemplo 0,05. Para t = 1,9890 y para $\boldsymbol{F}$ = 3,9560 y tomado la decisión de rechazar la hipótesis si los estadísticos son mayores o iguales a los criterios.

10.36 Los cálculos y las pruebas.

La prueba completa indica relaciones significantes en las correlaciones señaladas con negrilla.

El cuadro de la parte superior muestra los coeficientes de correlación

El cuadro central los estadísticos de F.

El cuadro inferior las probabilidades de la relación.

D_34. Cálculo de los Coeficientes de Correlación

	X1 = E. Padres	X2 = P. Padres	X3 = E. Hijos	X4 = P.Hijos
X1 = E. Padres	1,0000			
X2 = P. Padres	**0,4250**	1,0000		
X3 = E. Hijos	**0,7807**	**0,25662**	1,0000	
X4 = P.Hijos	**0,3335**	0,08213	**0,40568**	1,0000

D_35. Estadísticos de F para la prueba de Ho; r = 0

	X1 = E. Padres	X2 = P. Padres	X3 = E. Hijos	X4 = P.Hijos
X1 = E. Padres	0,0000			
X2 = P. Padres	18,2927	0,0000		
X3 = E. Hijos	129,5702	5,8513	0,0000	
X4 = P.Hijos	10,3895	0,5637	16,3505	0,0000

D_36. Valorando las pruebas; Probabilidad para la hipótesis r = 0.

	X1 = E. Padres	X2 = P. Padres	X3 = E. Hijos	X4 = P.Hijos
X1 = E. Padres	1,0000			
X2 = P. Padres	**0,0001**	1,0000		
X3 = E. Hijos	**0,0000**	**0,0178**	1,0000	
X4 = P.Hijos	**0,0018**	0,4549	**0,0001**	1,0000

10.37 Interpretación.

La interpretación de los resultados es simple:

La talla de los padres se relaciona significativamente en:

r = 51,69% con el peso de los mismos;

r = 78,31% con la talla de los hijos y;

r = 33,35% con el peso de los hijos.

El peso de los padres con:

r = 44,35% con la talla de los hijos;

Pero no con el peso de los hijos.

Y finalmente, la estatura de los hijos;

r = 40,57% con el peso de ellos mismos.

Notará que todas las relaciones son positivas, esto significa que al aumentar una aumenta concomitantemente la otra.

10.38 Relación con el Coeficiente de Regresión.

Si se estuviera interesado en obtener un *Modelo de Regresión* bastaría conocer promedio y la desviación estándar de la o las variables de interés. Por ejemplo: estimar la estatura con el peso de los padres. Aplicando la ecuación:

$$b_1 = r\frac{S_y}{S_x} = 0{,}4250\frac{0{,}1256}{12{,}5115} = 0{,}0043$$

Para estimar la estatura de los padres conociendo el peso y utilizando el modelo alternativo.

$y_i = \bar{y} + b_1(x_i - \bar{x}) = 1{,}6735 + 0{,}0043(x_i - 71{,}8094)$ La relación viceversa es de uso menos frecuente:

$$r = b_1\frac{S_x}{S_y} = 0{,}0043\frac{12{,}5115}{0{,}1256} = 0{,}4250$$

10.39 Caso en que X es un Factor.

Planteamiento del problema en donde X es un factor.

En la *Experimentación Planificada*, el investigador mantiene un control estricto sobre los tratamientos aplicados a "*unidades experimentales*", la relación funcional del *Factor X* (dominio de la función) con la variable *Objetivo Y* (rango de la función) se espera que sea de naturaleza causal. Esto es, ¡qué produzca cambios en los sujetos experimentados por efecto de los tratamientos.

Problema:

Una investigación estaba interesada en valorar la densidad de siembra sobre el rendimiento de una variedad de tomatillo (*Phisalis*). Se analizaron 5 distancias entre surcos sobre 12 repeticiones.

La hipótesis nula se planteó de la siguiente manera:

Ho; La densidad de siembra no afecta el rendimiento del tomatillo.

El interés en la investigación es determinar la relación entre rendimiento o variable *Y* a diferentes densidades de siembra *X*, un *Factor* aplicado a 5 niveles predeterminados 45, 60, 75, 90, 105 centímetros entre surcos, que se considera, causará diferencias en el rendimiento del sujeto experimentado, El *Tomatillo*.

10.40 Propiedad del los Polinomios Mínimos.

Cuando las facilidades de cálculo eran restringidas a reglas de cálculo y sumadoras manuales, se idearon los polinomios ortogonales para solucionar problemas de regresión. Estos polinomios tienen la característica de que su suma es cero y la suma del producto entre dos polinomios también deberá ser cero.

El polinomio mínimo para los tratamientos aplicados en el ejemplo sería:

$$X1_1 = \frac{45-75}{15} = -2;\ X1_2 = \frac{60-75}{15} = -1;\ X1_3 = \frac{75-75}{15} = 0;$$

$$X1_4 = \frac{90-75}{12} = 1;\ X1_5 = \frac{105-75}{15} = 2$$

Los cálculos se facilitan mucho.

Pero la importancia de esta manera de transformar los niveles de los factores se hace importante en el análisis de experimentos mediante modelos lineales. Por el momento se obtendrá la regresión del Rendimiento sobre el polinomio usando en algoritmo para la regresión de las HE.

El polinomio de segundo grado se obtiene:

$$X2_1 = X1_{1_1}^2 - 2 = 2; X2_2 = X1_2^2 - 2 = -1; X2_3 = X1_3^2 - 2 = -2;$$
$$X2_4 = X1_4^2 - 2 = +1; X2_5 = X1_5^2 = +1$$

10.41 Los Polinomios de Grado Superior.

Para un *factor* de 5 niveles o tratamientos se requiere un *polinomio de grado* 4 para recorrer el espacio muestral mediante una línea sinusoide que permita pasar por los diferentes puntos que se puedan crear. Un polinomio de grado 4 significa un modelo de la forma:

ratamiento ivel de Fact	Lineal X1	Cuadrático X2	Cúbico X3	Cuártico X4
45	-2	2	-1	1
60	-1	-1	2	-4
75	0	-2	0	6
90	1	-1	-2	-4
105	2	2	1	1
SX²	10	14	10	70

$$\bar{y}_i = \bar{y} + b_1\lambda_i^1 + b_2\lambda_i^2 + b_3\lambda_i^3 + b_4\lambda_i^4 + \varepsilon_i$$

En donde cada λ es un polinomio mínimo de la potencia de ***X***. Estos polinomios se obtienen de manera similar a la mostrada en la diapositiva anterior. Para no preocuparse por minucias se han desarrollado tablas de polinomios mínimos como la mostrada en la HE_Tablas de donde se extracta el cuado aledaño.

10.42 Las Sumas de Cuadrados por Coeficiente.

Acomodándolos apropiadamente a los datos, basta operar la función de *fx* = PENDIENTE() de la HE para obtener los coeficientes de regresión y la función *fx* = COEFICIENTE.R2() para obtener los coeficientes de determinación. Finalmente se obtienen las sumas de cuadrados correspondientes a cada polinomio multiplicando la *Suma de Cuadrados Total* por el *Coeficiente de Determinación*. La suma de cuadrados del cada polinomio es:

Efecto Lineal: $SCX1 = r_1^2 \times SCT = 0{,}15254 \times 6.560{,}79 = 1.000{,}76$

Efecto Cuadrático: $SCX2 = r_2^2 \times SCT = 0{,}71462 \times 6.560{,}79 = 4.688{,}46$

Efecto Cúbico: $SCX3 = r_3^2 \times SCT = 0{,}00121 \times 6.560{,}79 = 7{,}92$

Efecto Cuártico: $SCX4 = r_4^2 \times SCT = 0{,}01194 \times 6.560{,}79 = 78{,}31$

Es evidente que la suma de los efectos independientes hace la suma de cuadrados de los tratamientos del factor Densidad de Siembra.

$$SCF = SCT\sum_{i=1}^{t-1} r_i^2 = 1.00{,}76 + 4.668{,}46 + 7{,}92 + 78{,}31 = 5.775{,}44$$

10.43 El Análisis de la Varianza Completo.

Fuente de la Variación	*Grados de Libertad*	*Suma de Cuadrados*	*Promedio de los Cuadrados*	*Cociente de F*	*Probabilidad de F*	*Valores Críticos*	
						0,05	*0,01*
Total	59	6.560,79					
Efectos:							
Lineal	1	1.000,76	1.000,76	70,0856	0,00000	4,0162	7,1194
Cuadrátic	1	4.688,46	4.688,46	328,3454	0,00000	""	""
Cúbico	1	7,92	7,92	0,5544	0,45969	""	""
Cuártico	1	78,31	78,31	5,4842	0,02284	""	""
Tratamient	4	5.775,44	1.443,86	26,252015	0,00000	2,5397	3,6809
Error	55	785,35	14,28				

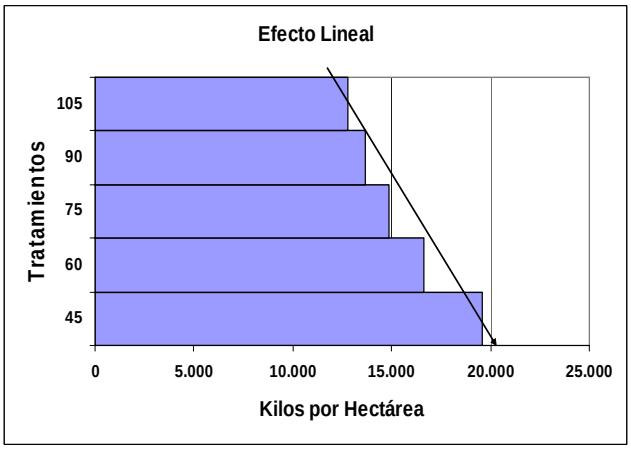
Notará que los efectos significativos se refieren al efecto lineal, al efecto cuadrático y al efecto cuártico. El efecto cúbico no mostró efectos importantes, el modelo de regresión que estima los promedios significativos quedará definido por:

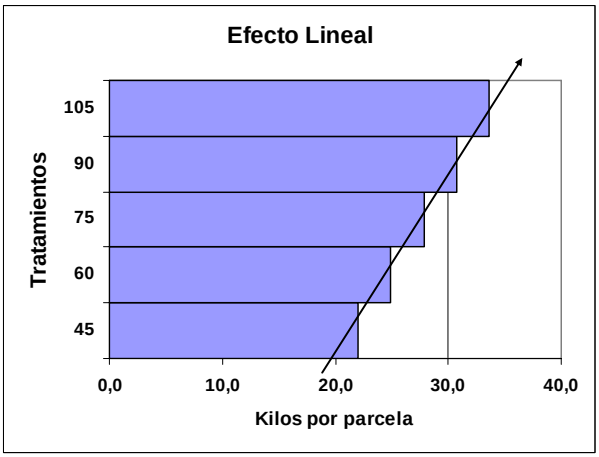
$$\bar{y}_i = 27{,}824 + 2{,}8878X1_i + \text{-}5{,}2828X2_i + 0{,}3053X3_i$$

Se analizará efecto por efecto.

10.44 Análisis del Efecto Lineal.

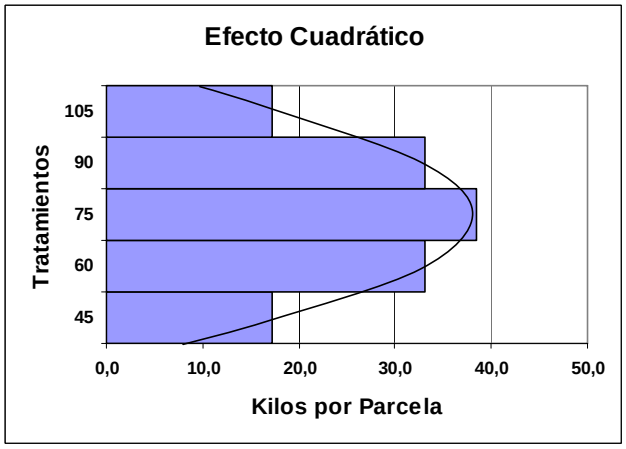
Es importante que el estudiante entienda la respuesta del sujeto experimental bajo los efectos que integran el modelo significativo. El primero corresponde al efecto lineal: En el rendimiento por parcela es ascendente, aumenta de 45 a 105; pero al llevarlo a Ha el efecto se invierte y curva por efecto de la densidad (Recuerde que interesa el rendimiento por Ha).

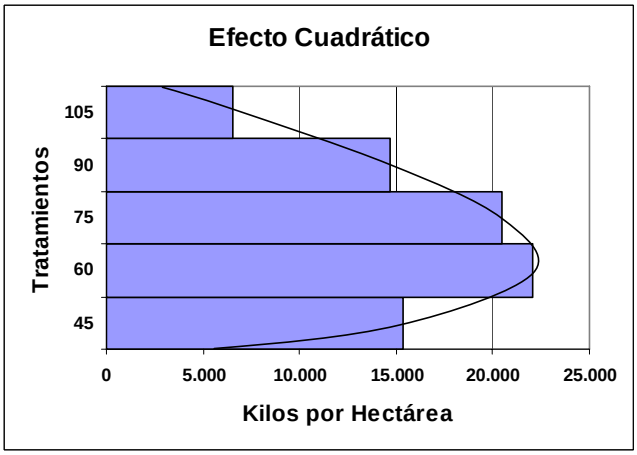




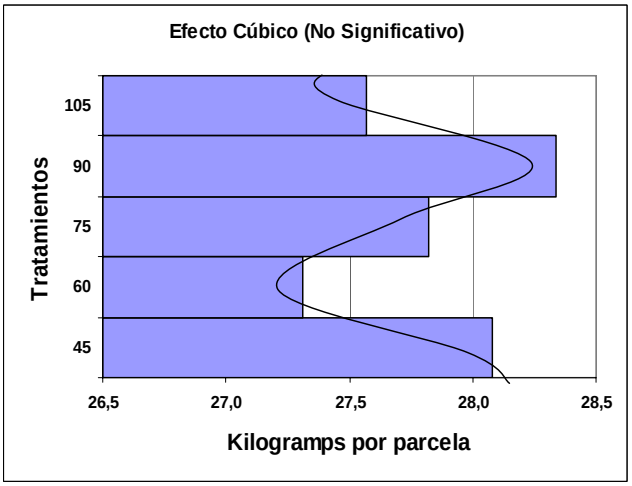
10.45 El Efecto Cuadrático.

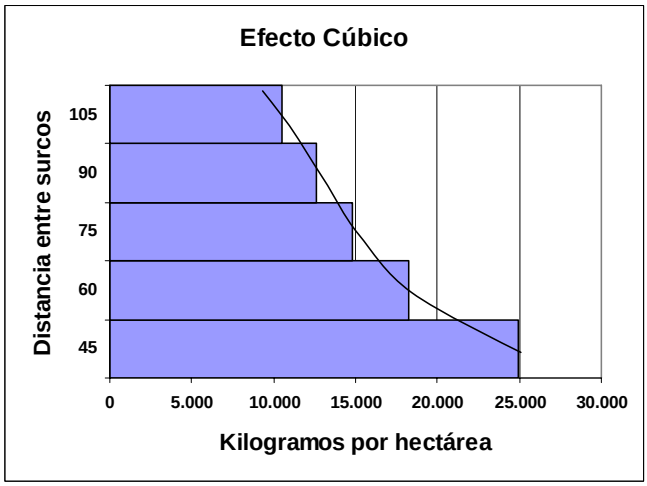
El Efecto cuadrático indica que el rendimiento tiende a ser mayor hacia el tratamiento de una distancia entre surcos de 75 cm., al extrapolar a rendimiento por hectárea, el rendimiento mayor se desplaza hacia una distancia entre surcos de 60 centímetros.





10.46 Efecto Cúbico.

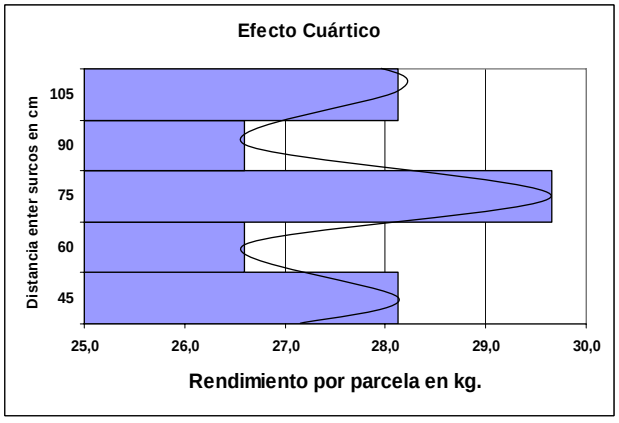


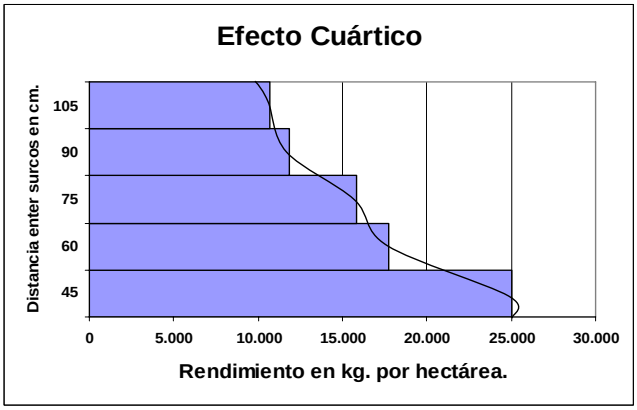


Aun cuando el efecto cúbico no se mostró a niveles significativos es conveniente que el estudiante observe la tendencia de éste y piense en la interrogante: ¿la variable de rendimiento por hectárea está determinada por la densidad de siembra?

10.47 El Efecto Cuártico.

El efecto cuártico compara los valores intermedios del efecto cuadrático. En el rendimiento por





parcela el valor más alto se obtiene con el tratamiento de 90 cm., de distancia entre surcos. Al llevar el rendimiento a hectáreas, el rendimiento se desplaza a valores de mayor densidad siendo entonces, el tratamiento de 45 cm., entre surcos el más recomendable.

10.48 Los Promedios Significativos Integrados.

Cada efecto independiente aporta su efecto al modelo integrado. Las estimaciones de los promedios de rendimiento por parcela par para cada tratamiento mediante el modelo integrado serán:Para 45 cm: $\bar{y}_1 = 27{,}824 + 2{,}8878(-2) - 5{,}2828(+2) + 0{,}3053(+1) = 11{,}8$

Para 60 cm: $y_2 = 27{,}824 + 2{,}8878(1) - 5{,}2828(-1) + 0{,}3053(+4) = 29{,}0$

Para 75 cm: $y_3 = 27{,}824 + 2{,}8878(+0) - 5{,}2828(-2) + 0{,}3053(+6) = 40{,}2$

Para 90 cm: $y_4 = 27{,}824 + 2{,}8878(+1) - 5{,}2828(-1) + 0{,}3053(-4) = 34{,}8$

Para 105 cm: $y_5 = 27{,}824 + 2{,}8878(+2) - 5{,}2828(+2) + 0{,}3053(+1) = 23{,}3$

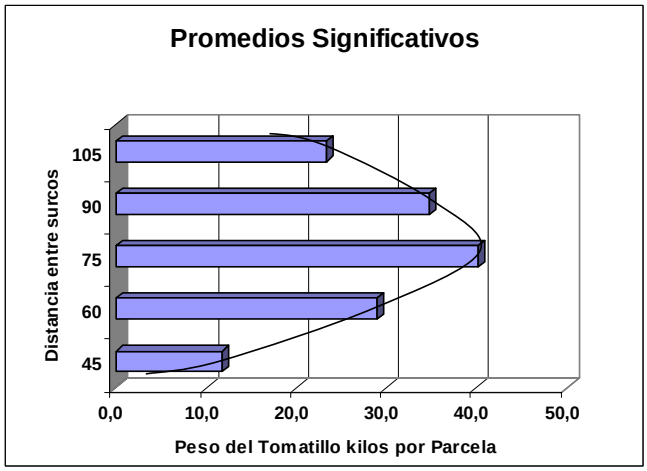
Estos valores deberán calcularse para rendimiento por hectárea mediante la fórmula para una parcela de 25 metros de largo:

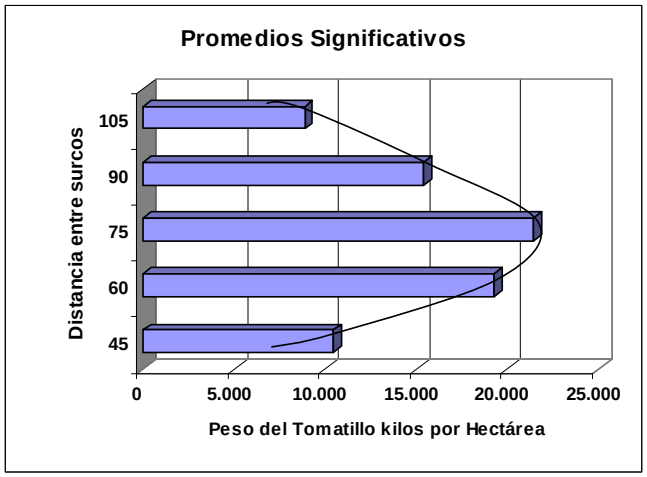
$$R_i = \frac{10.000 \times r_i \times 4}{\text{Distancia\_Entre\_Surco}}$$

R_i es el rendimiento por hectárea y r_i el rendimiento por parcela.

10.49 La representación gráfica.

La representación gráfica de la prueba de hipótesis utilizando el modelo significativo elaborado a partir del *Análisis de la Varianza para Modelos de Regresión* mediante polinomios ortogonales es muy ilustrativa utilizando el gráfico apropiado y los datos convenientes. El gráfico de la Izquierda muestra los promedios de kilos de parcela en surcos de 25 metros, el de la derecha los mismos datos llevados a hectáreas. Recuerde que se está trabajando con densidad de siembra. Para obtener el óptimo el Tomatillo debe sembrarse en surcos de 75 y 90 centímetros entre ellos, dando un poco de más peso a los datos por parcela.





10.50 Conclusión importante.

Paso a paso se ha llegado a una conclusión importante para los diseños experimentales que usan modelos lineales:

Con la Regresión de Polinomios Ortogonales se pueden encontrar todos los efectos de un Factor.

Esto, es trascendental en el análisis de Experiencias Planificadas, pues basta aplicar el modelo de regresión significativo para determinar sin ambigüedad, cuál o cuáles tratamientos han demostrado un efecto importante sobre el sujeto de la experimentación, en este ejemplo, El Tomatillo.

10.51 La Rutina de Cálculo Directo en la HE.

La HE tiene una herramienta que permite calcular regresiones múltiples hasta de 16 variables. Puede comprobar que los resultados son idénticos, fije su atención en las probabilidades de ***F*** en el anterior y de "***t***" en este.

Resumen

Estadísticas de la regresión	
Coeficiente de correlación múltip	0,9382
Coeficiente de determinación R^2	0,8803
R^2 ajustado	0,8716
Error típico	3,7788
Observaciones	60

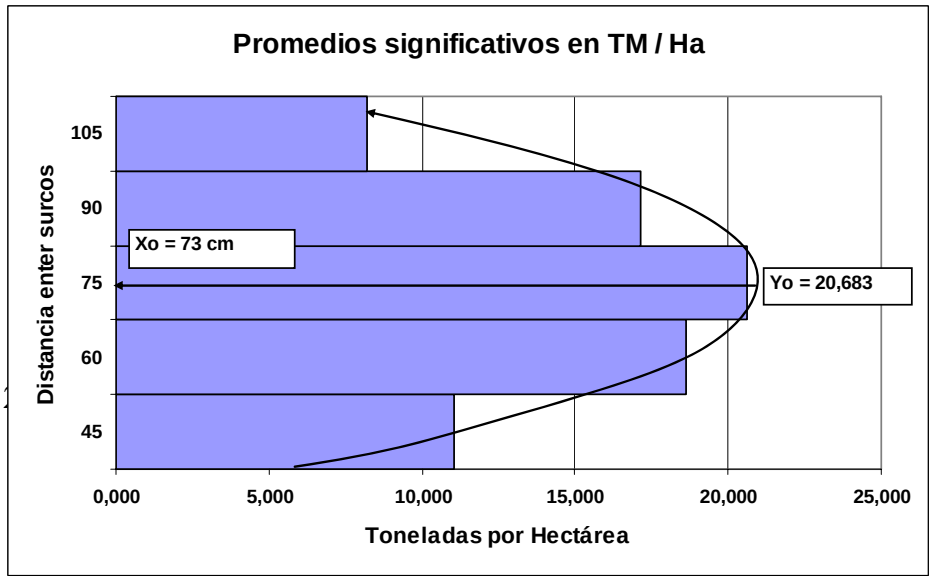
ANÁLISIS DE VARIANZA

Fuente de la Variación	*Grados de Libertad*	*Suma de Cuadrados*	*Promedio de los Cuadrados*	*Cociente de F*	*Probabilidad de F*	*Valores Críticos*	
						0,05	*0,01*
Regresión	4	5.775,44	1.443,86	101,12	1,1206E-24	2,54	3,68
Residuos	55	785,35	14,28				
Total	59	6.560,79					

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*
Intercepción	27,8241	0,4878	57,0357	0,0000	26,8465	28,8017455
X1	2,8878	0,3450	8,3717	0,0000	2,1965	3,57914142
X2	-5,2828	0,2915	-18,1203	0,0000	-5,8670	-4,69850102
X3	-0,2568	0,3450	-0,7446	0,4597	-0,9481	0,43444976
X4	0,3053	0,1304	2,3418	0,0228	0,0440	0,56661532

10.52 Análisis de las Toneladas por Hectárea.

Se ha venido expandiendo el rendimiento por parcela a rendimiento por hectárea. Es más conveniente expandir cada observación y analizarla para obtener un modelo significativo para la variable.



$$\bar{y}_i = 15{,}125 - 0{,}7133X1_i - 1$$

Igualando a 0 y derivando se puede obtener el punto óptimo:

$$\lambda o = \frac{-b_1}{2b_2} = \frac{-(-0{,}7133)}{2(-2{,}7554)} = -0{,}1294$$

Llevándolo a valores de X para obtener 73 cm, con un óptimo de 20,682 TM.

$$Xo = \lambda \times 15 + 75 = -0{,}1294 \times 15 + 75 = 73$$

Calculando para obtener un promedio máximo de 20,682 TM.

$$\lambda 2o = \lambda_1^2 - 2 = (-0.1294)^2 - 2 = -1{,}983$$

$$\bar{y}_o = 15{,}125 - 0{,}7133(-0{,}1294) - 2{,}7554(-1{,}983) = 20{,}682$$

Valor de λ^2 (Lambda cuadrático).

$$\lambda_o = X_o^2 - 2 = \text{-}0{,}1294^2 - 2 = -1{,}983$$

$$\bar{y}_o = 15{,}125 \text{ -}0{,}7133(\text{-}0{,}1294) \text{-}2{,}7554(\text{-}1{,}983) = 20{,}682$$

10.53 Regresión en donde X se determina.

Variable independiente X, se hace convencional.

En trabajos de economía es muy frecuente utilizar convencionalmente, directa o indirectamente, los años como variables ***X***. Definitivamente, los años están dados, lo que sí es aleatorio es el resultado de las variables *Y* consideradas en los análisis.

En este tipo de estudios los resultados no pueden achacarse estrictamente al año, sino qué, más bien el año refleja una serie de acontecimientos que ocurren, y que afectan a las variables *Y,* pero que no interesa descubrir por sí solos. Entonces, se toma como referencia la variable determinada como año, que puede ser natural de enero a diciembre o eventualmente fiscal de octubre a septiembre como se usa en Costa Rica.

10.54 El problema: Regresión lineal con X determinada.

El departamento de planeamiento preparó para la junta directiva un programa de inversiones para los siguientes cuatro años. Para mostrar lo que se podría esperar de inversión gubernamental. Decidieron estudiar el comportamiento del gasto en construcción pública *Y2* y el monto de la inversión privada *Y1* en la construcción de vivienda no subvencionada por el gobierno de los últimos 30 años.

Decidieron hacer el estudio mediante *Regresión Lineal Simple* de las inversiones sobre los años.

10.55 La Correlación.

Aún cuando las correlaciones de los gastos con el año se manifiestan positivas y altamente significativas, estrictamente no se puede hacer referencia a estas pues el año no es una variable aleatoria. Por tanto, la única correlación a la que se puede hacer mención es entre gastos que es de 0,7139 o 71,39% altamente significativa.

	Año	*I. Vivienda*	*G. Público*
Año	1,0000		
I. Vivienda	**0,8305**	1,0000	
G. Público	**0,8159**	**0,7139**	1,0000
Probabilidad de F			
Año	1,0000		
I. Vivienda	0,0000	1,0000	
G. Público	0,0000	0,0000	1,0000

En donde se introduce la ecuación directamente en la función de probabilidad:

$$F(29{,}1013;1;28) = \int_0^{29{,}1013} f\left(\frac{14{,}2700}{0{,}4904};1;28\right)df = 0{,}000009\,.$$

10.56 Las tasas de crecimiento.

A partir de los coeficientes de correlación y las estadísticas descriptivas de promedio y varianza se obtienen las tasas de

Estadístico	Años	Inversión en Vivienda	Gasto cons. Gobierno
Número de años del estudio	30		
Promedios	1984,50	55,21	504,80
Desviaciones Estándar	8,8034	37,4567	360,6812

crecimiento:

La Inversión en vivienda crece a una tasa de 3,43 miles por año y la inversión del gobierno en 33,43 por año.

El gasto de una unidad moneda por el gobierno se traduce en una inversión privada en vivienda en 0,074 miles de unidades moneda por año.

Las diferentes tasas de crecimiento, por su magnitud no permiten deducir, por ejemplo ¿cuál crece más por año?

10.57 Las tasas estandarizadas.

La estandarización permite la comparación de cualquier conjunto de variables puesto que transforman los estadísticos a una unidad común, además en un número puro. Esta se consigue dividiendo el coeficiente de regresión por la desviación estándar de la variable que se estudia, *Y*.

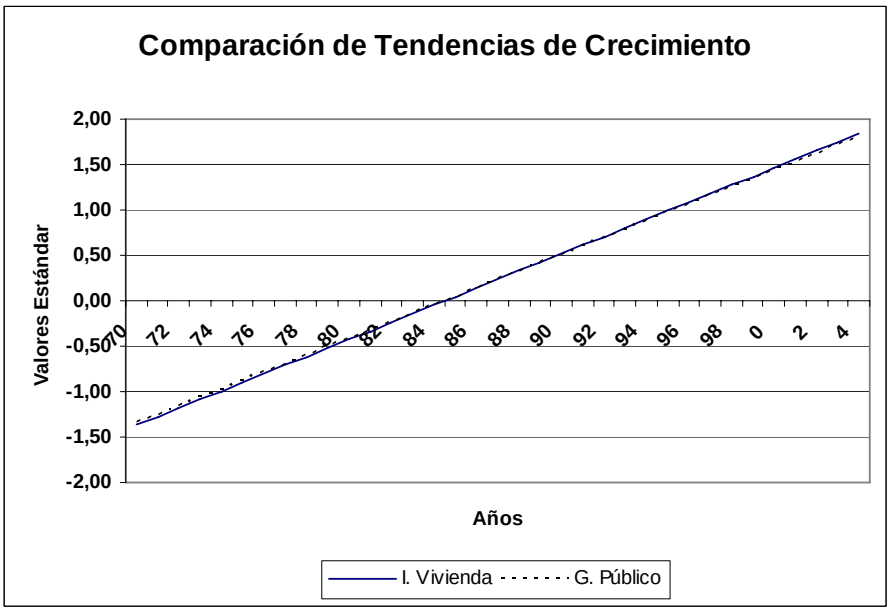
La estandarización de la tasa de la inversión particular en vivienda es:

$$b_{Z1} = \frac{b_{Y1}}{S_{Y1}} = \frac{3,5336}{37,4567} = 0,0943$$ La estandarización de la tasa de gasto de gobierno es:

$$b_{Z2} = \frac{b_{Y2}}{S_{Y2}} = \frac{33,4291}{360,6812} = 0,0927$$

Ha crecido más la inversión particular que el gasto del Gobierno.

10.58 Representación gráfica.



Las líneas del gráfico muestran que prácticamente no hay diferencias entre las tendencias de crecimiento entre las variables.

10.59 Indexando las Variables.

Observando la tabla de valores estandarizados estimados se aprecia que las tendencias cambian de signo en 1985. Si se toma el índice medio desde 1983 hasta 1987. Para la Inversión en vivienda de:

$I1_{(1983-1987)} = \frac{1}{5}(25{,}00 + 23{,}80 + 7{,}68 + 15{,}45 + 60{,}25) = 26{,}44$ Para el gasto del gobierno en construcción: $I2_{(1983-1987)} = \frac{1}{5}(843 + 550 + 275 + 191 + 113) = 394{,}40$ Dividiendo el índice medio respectivo por cada valor desde 1986 se consiguen los modelos de regresión:

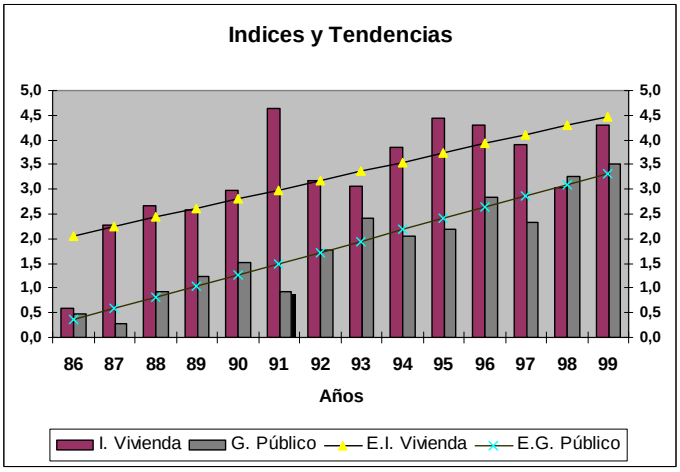
10.60 Los Modelos de los índices.

La tasa de inversión desde 1986 de la inversión privada en vivienda ha sido inferior y más errática que la tasa de gasto en construcción del gobierno.

Si la empresa depende de sus fondos o de préstamos para invertir en vivienda corre más riesgo que ofrecer sus servicios para el gobierno.

Aun cuando la tasa de inversión privada indica que ha gastado más que en 1966 que lo que gastó el gobierno; es probable que haya aumentado la competencia por contratos gubernamentales.

Año	Inversión en Privada en I. Vivienda	Gasto en Construcción G. Público
1983	25,00	843,00
1984	23,80	550,00
1985	7,68	275,00
1986	15,45	191,00
1987	60,25	113,00
Promedio	26,44	394,40



10.61 Conclusión.

Se ha estudiado la *Regresión Lineal* en sus tres modalidades:

Cuando Ambas Variables *X e Y* en la relación son de naturaleza aleatoria, usualmente obtenidas mediante técnicas de muestreo;

Cuando La Variable Independiente *X* es un *Factor* aplicado en una experiencia planificada con modelos lineales;

Cuando La Variable Independiente *X* se determina a conveniencia de la investigación.

Cualquiera que sea la modalidad, es evidente que la *Regresión Lineal* es una poderosa herramienta de análisis que permite una síntesis precisa.

10.62 Recomendación.

Para el estudiante que desee profundizar en el estudio de la *Regresión Lineal* y aprender sobre las posibilidades de inducción que poseen los modelos lineales debe estudiar las dos secciones avanzadas relacionadas con la *Regresión:*

La *Regresión de Modelos no Lineales* pero linealizables, en donde entran la familia de los modelos logarítmicos y exponenciales y los modelos de potencias;

Y la *Regresión Lineal Múltiple*, una herramienta de elección muy útil en el estudio de relaciones entre variables.

REFERENCIAS SELECTAS:

1. Di Marco, Luis Eugenio. Análisis Estadístico. Capítulo 19. Editorial Interamericana S.A. México. 1969.
2. Hillier Frederick S., y Lieberman Gerard j., *Introducción a la Investigación de Operaciones*. Capítulo 19. Segunda edición en español traducida de la cuarta edición en ingles. McGraw-Hill Interamericana de México, S. A. De C. V., 1990.
3. Koosis, Donald. Elementos de Inferencia Estadística. Capítulo 2. Editorial Limusa, S. A. México, 1974.
4. Levin, Richard. Estadística Para Administración. Capítulo 11. Editorial Prentice-Hall Hispanoamérica S.A. México. 1987.
5. Mason, Robet. Douglas, Lind. Estadística Para Administración y Economía. Capítulo 15. Editorial Alfaomega. México. 1992.
6. Mendelhall, William Reinmuth, james. Estadística para Administración y Economía. Capitulo 11. Editorial Iberoamérica. México. 1981.
7. Miller Irwin, Freund John E., Johnson Richard A: *Probabilidad y Estadística para Ingenieros*. Capítulo 11. Traducido de la cuarta edición en ingles; Prentice-Hall Hispanoamericana, S. A. 1992.
8. Murray R. Spiegel: *Serie de compendios Schaum, Teoría y Problemas de Estadística.* Capítulos 13,14, 15. Primera edición en español, traducido de la primera edición en ingles; Libros McGraw-Hill de México, S. A. De C. V., 1973.
9. Ostle Bernard: *Estadística Aplicada*. Capítulos 8 y 9. Primera edición en español traducida de la primera edición en ingles. Editorial Limusa, S. A., 1977.
10. Snedecor George W., y Cochran William G: *Statistical Methods*. Capítulos 6 y 7. Sexta edición; The Iowa State University, 1974.
11. Steel Robert G. D., Torrie James H: *Principles and Procedures of Statistics*. Capítulos 9, 10. Primera edición; McGraw-Hill Book Company, Inc, 1960.
12. Ya-Lun Chou. Análisis Estadístico. Capítulo 17. Editorial Interamericana S.A. México. 1975.

Este libro se imprimió en:
LOAIZA IMPRESIÓN DIGITAL
Teléfono 551-6580. Fax. 552-3844
E-mail. loaizaimpresion@gmail.com
ENERO 2007.