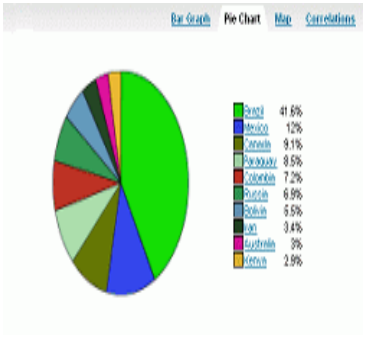
MINISTERIO DE DESARROLLO AGROPECUARIO

DIRECCIÓN DE PLANIFICACIÓN SECTORIAL

DEPARTAMENTO DE ESTADÍSTICAS E INFORMACIÓN



Terminología Estadística Común y sus Usos

Como toda profesión, también los estadísticos tienen sus propias palabras claves y frases para facilitar una comunicación precisa. Sin embargo, uno debe interpretar los resultados de cualquier toma de decisión en un lenguaje que sea fácil de entender para a los tomadores de decisiones. Si no, el/ella no creerá en lo que usted recomienda, y por lo tanto no entrara a la fase de implementación. Esta carencia de comunicación entre los estadísticos y gerentes es la barrera principal para usar la estadística.

Población: Una población es cualquier colección entera de personas, animales, plantas o cosas de las cuales podríamos recolectar datos. Es el grupo entero que nos interesa, el cual deseamos describir o sobre cuál deseamos establecer conclusiones. En la figura anterior la vida de las bombillas de luz fabricadas, digamos por GE, es la población en cuestión.

Variables Cualitativas y Cuantitativas: Cualquier objeto o acontecimiento, que pueda variar en observaciones sucesivas ya sea en cantidad o cualidad se llama "variable." Las variables se clasifican por consiguiente como cuantitativas o cualitativas. Una variable cualitativa, a diferencia de una variable cuantitativa no varía en magnitud en observaciones sucesivas. Los valores de variables cuantitativas y cualitativas se llaman “valores” y “cualidades o atributos”, respectivamente.

Variable: Una característica o fenómeno, que pueden tomar diversos valores tales como peso o género, ya que los mismos son diferentes entre individuos.

Aleatoriedad: La aleatoriedad significa algo impredecible. El hecho fascinador sobre estadística deductiva es que, aunque cada observación aleatoria podría no ser predecible cuando es tomada sola, colectivamente siguen un patrón confiable llamado

función de distribución. Por ejemplo, es un hecho de que la distribución promedio de una muestra sigue una distribución normal para una muestra mayor a 30. Es decir, un valor exagerado de la media de la muestra es más certero que un valor exagerado de algunos pocos valores de datos.

Muestra: Un subconjunto de una población o universo.

Un Experimento: Un experimento es un proceso mediante el cual el no se sabe con certeza cual será el resultado por adelantado.

Experimento Estadístico: Un experimento en general es una operación en la cual una elige los valores de algunas variables y mide los valores de otras variables, como en la física. Un experimento estadístico, en contraste es una operación en la cual uno toma una muestra aleatoria de una población e infiere los valores de algunas variables. Por ejemplo, en una encuesta, "examinamos" es decir, "observamos" la situación sin intentar cambiarla, tal como en una encuesta de opiniones políticas. Una muestra aleatoria de una población relevante proporciona la información sobre las intenciones de votación.

Para hacer cualquier generalización sobre una población, una muestra escogida al azar de la población entera, que se considere representativa de la población, es frecuentemente estudiada. Para cada población, hay muchas muestras posibles. Una muestra estadística da información sobre los parámetros poblacionales correspondiente. Por ejemplo, la media de la muestra para un conjunto de datos daría información sobre la media m correspondiente a toda la población.

Es importante que el investigador defina total y cuidadosamente a la población antes de recolectar la muestra, incluyendo una descripción de los miembros.

Ejemplo: La población para un estudio de la salud infantil podría ser todos los niños nacidos en los Chile durante los años 80. La muestra podría ser todos los bebés nacidos el 7 de mayo en cualquiera de los años.

Un experimento es cualquier proceso o estudio en el cual los resultados obtenidos en la recolección de datos eran anteriormente desconocidos. En estadística, el término se restringe generalmente a las situaciones en las cuales el investigador tiene control sobre algunas de las condiciones bajo las cuales el experimento ocurre.

Ejemplo: Antes de introducir un nuevo tratamiento medico con el uso de una nueva droga para reducir la alta tensión arterial, los

fabricantes de la misma realizan un experimento para comparar la eficacia de la nueva droga con la prescrita actualmente. Pacientes recientemente diagnosticados son seleccionados de un grupo para las prácticas generales. La mitad de ellos son elegidos al azar para recibir la nueva droga, el resto recibe la droga actual. De esta manera, el investigador tiene control sobre los pacientes seleccionados y de la manera en la cual el tratamiento es asignado.

Diseño de Experimentos: Es una herramienta para incrementar el índice de adquirir nuevos conocimientos. El conocimiento alternativamente se puede utilizar para ganar ventajas competitivas, para acortar el ciclo de desarrollo de productos, y para producir nuevos productos y procesos que satisfagan y excedan las expectativas de sus clientes.

Datos Primarios y Conjunto de Datos Secundarios: Si los datos son obtenidos de un experimento planificado el cual es relevante y relacionado al objetivo (s) de la investigación estadística, son recolectados directamente por el analista, se llaman datos primarios. Sin embargo, si algunos registros resumidos son dados al analista, se llama conjunto de datos secundarios.

Variable aleatoria: Una variable aleatoria (escogida al azar) es una función (se llama “variable”, pero en realidad es una función) que asigna un valor numérico a cada evento simple. Por ejemplo, en el muestreo para el control de calidad, un artículo podría ser defectuoso o no defectuoso, por lo tanto, se podría asignar X =1, y X =0 para un artículo defectuoso y no defectuoso respectivamente. Se podrían asignar cualquier otros dos valores de números reales distintos; sin embargo, es más fácil trabajar con números enteros no negativos para variables aleatorias. Estas son necesarias porque no se pueden realizar operaciones aritméticas con palabras. Las variables aleatorias nos permiten realizar cálculos estadísticos, tal como promedio varianza. Cualquier variable aleatoria tiene una distribución de probabilidad asociada.

Probabilidad: La probabilidad (es decir, sondeando sobre lo desconocido) es la herramienta usada para anticipar como una distribución de datos debería ser representada bajo un modelo dado. Fenómenos aleatorios no son casuales: exhiben un orden que se desarrolla solamente a largo y que es descrita por una distribución. La descripción matemática de la variación es básica para la estadística. La probabilidad requerida para la inferencia estadística no es principalmente axiomática o combinatoria, sino que se orienta hacia la descripción de las distribuciones de los datos .

Unidad de Muestreo: Una unidad es una persona, un animal, una planta o una cosa que son estudiadas por un investigador; son los objetos básicos sobre los cuales se ejecuta el estudio o el experimento. Por ejemplo, una persona; una muestra de suelo; un pote de semillas; un área de código postal; el área de especialización de un medico.

Parámetro: Un parámetro es un valor desconocido, y por lo tanto tiene que ser estimado. Los parámetros se utilizan para representar una determinada característica de la población. Por ejemplo, la media poblacional m es un parámetro que normalmente se utiliza para indicar el valor promedio medio de una cantidad.

Dentro de una población, un parámetro es un valor fijo que no varía. Cada muestra tomada de la población tiene su propio valor de cualquier estadística que se utilice para estimar este parámetro. Por ejemplo, la media de los datos en una muestra es utilizada para dar información sobre la media de la población total m de la cual esa muestra fue tomada.

Estadístico: Un estadístico es una cantidad calculada de una muestra de datos. Se utiliza para dar información sobre valores desconocidos correspondientes a la población. Por ejemplo, el promedio de los datos en una muestra se utiliza para dar información sobre el promedio total de la población de la cual esa muestra fue tomada.

Un estadístico es una función de una muestra aleatoria observable. Por lo tanto es en sí, una variable aleatoria observable. Note que, mientras que un estadístico es una "función" de observaciones, desafortunadamente, es comúnmente llamado una “variable” aleatoria, no una función.

Es posible obtener más de una muestra de la misma población, y el valor del estadístico en general variara entre muestra y muestra. Por ejemplo, el valor promedio de una muestra es un estadístico. Los valores promedios en más de una muestra, obtenidos de la misma población, no serán necesariamente iguales.

Estadísticos se les asignan normalmente letras romanas (por ejemplo x̄y s), mientras que los valores equivalentes desconocidos de la población (parámetros) se asignan las letras griegas (por ejemplo μ, s).

La palabra estimación significa estimar, esto significa darle un valor a algo. Una estimación estadística es una indicación de valor de una cantidad desconocida basada en datos observados.

Más formalmente, una estimación es el valor particular de un estimador que es obtenido de una muestra particular de datos y que es utilizado para indicar el valor de un parámetro.

Ejemplo: Suponga que el gerente de una tienda deseó saber el valor de m ,el gasto promedio por cliente de su tienda durante el año pasado. Ella podría calcular el gasto promedio de los centenares (o quizás de los miles) de clientes que compraron mercancías en su tienda; es decir, la media poblacional m . En lugar de esto, ella podría utilizar una estimación de la media poblacional m calculando la media de una muestra representativa de clientes. Si se encontrara que el valor fuera $25, estos $25 serían su estimación.

Existen dos amplias subdivisiones de la estadística: Estadística descriptiva y estadística deductiva, tal y como se describirá a continuación.

Estadística Descriptiva: Los datos numéricos estadísticos deben ser presentados de manera clara, consistente, y de manera tal que los tomadores de decisiones puedan obtener rápidamente las características esenciales de los datos e incorporarlos en proceso de.

La principal cantidad descriptiva derivada de datos de la muestra es la media (x̄), la cual es la media aritmética de los datos de la muestra. Esta sirve como la más confiable medida de valor de un miembro típico de la muestra. Si la muestra contiene algunos valores que son demasiado grandes o demasiado pequeños los cuales pudieran generar un efecto distorsionador en el valor de la media, la muestra es representada con mayor exactitud por la mediana, el cual es el valor donde la mitad de los valores de la muestra se ubican por debajo y la otra mitad por arriba de la misma.

Las cantidades comúnmente usadas para medir la dispersión de los valores con respecto a su media son la varianza s^2 y su raíz cuadrada, la desviación estándar s. La varianza es calculada determinando la media, luego restándole dicha media a cada uno de los valores de la muestra (que generan la desviación de las muestras), y después haciendo un promedio de los cuadrados de estas desviaciones. La media y la desviación estándar de la muestra se utiliza como estimadores de las características correspondientes de todo el grupo del cual la muestra fue obtenida. Ellos en general, ***no*** describen totalmente la distribución (F_x) de los valores dentro de la muestra o del grupo del relacionado; de hecho, diversas distribuciones pueden tener la misma media y distribución estándar. Sin embargo, ellos si proporcionan una descripción completa de la distribución normal, en la cual las desviaciones positivas y negativas con respecto a la

media son igualmente comunes, y pequeñas desviaciones pequeñas son mucho más comunes que las grandes. Para un sistema de valores normalmente distribuido, un gráfico que demuestre la dependencia de la frecuencia de las desviaciones sobre sus magnitudes tiene una curva acampanada. Cerca de 68 por ciento de los valores diferirán con respecto al valor de la media por menos que el valor de la desviación estándar, y casi 100 por ciento diferenciarán por menos de tres veces el valor de la desviación estándar.

Estadística Deductiva (inferencial): La estadística deductiva se refiere al hecho de hacer inferencias sobre las poblaciones basándose en muestras que han sido extraídas de ellas. Es decir, si encontramos una diferencia entre dos muestras, nos gustaría saber si estas son diferencias “reales” (es decir, que están presentes en la población) o quizás una diferencia de "oportunidad" (es decir, que podrían ser el resultado de un error de la muestra aleatoria). Eso es a lo que las pruebas de significancia estadística se refieren. Cualquier conclusión deducida de los datos de la muestra y que se refieran a la población de los cuales fueron obtenidos, deben ser expresados en términos probabilísticos. La probabilidad es el lenguaje y la herramienta que mide la incertidumbre en nuestras conclusiones estadísticas.

La estadística deductiva se podía utilizar para explicar un fenómeno o para comprobar la validez de una proposición. En este caso, la estadística deductiva es llamada análisis exploratorio de datos o análisis confirmativo de datos, respectivamente.

Inferencia Estadística: La inferencia estadística esta referida a ampliar sus conocimientos obtenidos de una muestra escogida al azar de la población entera y aplicarla para población entera. Esto es conocido en matemáticas razonamiento inductivo, es decir, el conocimiento del todo proveniente de un detalle particular. Su uso principal es la prueba de hipótesis en una población dada. La inferencia estadística dirige la selección de los modelos estadísticos apropiados. Los modelos y los datos interactúan recíprocamente en trabajo estadístico. La inferencia con base en los datos puede ser pensada como el proceso de seleccionar un modelo razonable, incluyendo una proposición en lenguaje probabilístico de cuan confiable se puede estar sobre la selección hecha.

Condición de la Distribución Normal: La distribución normal o distribución de Gauss es una distribución simétrica y continua que sigue una curva de forma acampanada. Una de sus características más notable es que la media y la varianza de manera única e independiente determinan la distribución. Se ha

observado empíricamente que muchas variables de medición tienen distribuciones aproximadamente normales. Incluso cuando una distribución es no normal, la distribución de la media de muchas observaciones independientes de la misma distribución

se convierten arbitrariamente a una distribución similar a la normal, a medida que el número de observaciones crece. Muchas pruebas estadísticas frecuentemente usadas tienen la condición de que los datos provengan de una distribución normal.

Estimación y Prueba de Hipótesis: Las inferencias en estadística son de dos tipos. La primera es la valoración o estimación, la cual implica la determinación, con la posibilidad de error debido al muestreo, de un valor desconocido de alguna característica de la población, tal como la proporción que tiene una cualidad específica o el valor de la media μ en ciertas medidas numéricas. Para expresar la exactitud de las estimaciones sobre las características de la población, se debe calcular también el error estándar de las estimaciones. El segundo tipo de inferencia es el contraste o prueba de hipótesis. Esto implica la definición de una hipótesis como un sistema de valores posibles para la población y una alternativa, para valores diferentes. Existen muchos procedimientos estadísticos para determinar, con relación a una muestra, si las verdaderas características de la población pertenecen al sistema de valores en la hipótesis o en la alternativa.

El concepto de inferencia estadística esta inmerso en el de la probabilidad, son conceptos idealizados del grupo que esta sujeto a estudio, llamados población y muestra. Los estadísticos podrían ver a la población como un grupo de bolas de las cuales la muestra se selecciona al azar, es decir, de una manera tal que cada bola tenga la misma oportunidad de ser seleccionada para la muestra.

Note que para poder estimar los parámetros de la población, el tamaño de la muestra n debe ser mayor que uno (1). Por ejemplo, con un tamaño de muestra uno, la variación (s^2) dentro de la muestra es 0/1 = 0. Una estimación para la variación (s^2) dentro de la población sería 0/0, que es cantidad indeterminada, lo cual es imposible.

Letras Griegas Comúnmente Usadas como Anotaciones Estadísticas

En estadística, al igual que en otras áreas de la ciencia, se utilizas las letras griegas como anotaciones científicas. Esto, para hacer honor a nuestros ancestros filósofos Griegos que inventaron la

ciencia y el pensamiento científico. Antes de Sócrates, en el siglo VI AC, Tales y Pitágoras entre otros, aplicaron conceptos geométricos a la aritmética, mientras que Sócrates en su época inventó el razonamiento dialéctico. El renacimiento del

pensamiento científico (iniciado por los trabajos de Newton) fue valorado y por lo tanto reapareció casi 2000 años más tarde.

Letras Griegas Comúnmente Usadas como Anotaciones Estadísticas										
alpha	beta	Ki al cuadrado	delta	mu	nu	pi	rho	sigma	tau	theta
a	B	c^2	d	m	n	p	r	s	t	q

Nota: Ki al cuadrado (o Chi-cuadrado) c^2, no es el cuadrado de algo en particular, su nombre simplemente implica Chi al cuadrado. Ki no tiene ningún significado en estadística.

Me alegra que usted poco a poco este venciendo todas las confusiones que existen cuando se aprende estadística.

Tipo de Datos y Niveles de Medición

En estadística, la información puede ser recolectada usando datos cualitativos o cuantitativos. Los datos cualitativos, tal como el color del ojo de un grupo de individuos, no pueden ser medidos por relaciones aritméticas. Existen ciertas particularidades que orientan en cuales categorías o clases debe ubicarse un individuo, objeto, o proceso. Estas son llamadas variables categóricas.

El conjunto de datos cuantitativos que consiste en las medidas que toman valores numéricos, en cuales descripciones tales como la media y la desviación estándar tienen sentido. Pueden ser puestos en un orden y ser subdivididos en dos grupos: datos discretos o datos continuos.

Los datos discretos son datos contables y recolectados por **conteo**, por ejemplo, el número de los artículos defectuosos producidos durante un día de producción.

Los datos continuos son recolectados por **medición** y expresados en una escala continua. Por ejemplo, midiendo la altura de una persona o la extensión de una parcela.

Entre las primeras actividades del análisis estadístico se encuentran contar o medir: La teoría de Conteo / medición se

refiere a la conexión entre los datos y la realidad. Un sistema de datos es una representación (es decir, un modelo) de la realidad basada en escalas numéricas y mensurables. Los datos son llamados de “tipo primario” si el analista ha estado envuelto directamente en la recolección de datos relevantes para su investigación. Si no, son llamados datos de “tipo secundario”.

Los datos vienen en forma **N**ominal, **O**rdinal, de **I**ntervalo, y **C**ociente. Los datos pueden ser continuos o discretos.

Niveles de Medición

	Nominal	**Ordinal**	**Intervalo/Cociente**
Posición	no	si	si
Diferencia Numérica	no	no	si

Tanto el punto cero como las unidades de medida son arbitrarios en la escala de Intervalo. Mientras que la unidad de medida es arbitraria en la escala de Cocientes, el punto cero es un atributo natural. La variable categórica es medida en una escala ordinal o nominal.

La teoría de Conteo / medición se refiere a la conexión entre los datos y la realidad. Ambas, la teoría estadística y la teoría de conteo y medición son necesarias hacer inferencias sobre realidad.

Puesto que los estadísticos viven para la precisión, prefieren niveles de Intervalo / Cociente de medición.

Para una buena aplicación en negocios de variables aleatorias discretas, visite Calculadora para la Cadena de Markov , Calculadora para Cadenas Grandes de Markov y Juegos Suma Cero.

¿Por qué el Muestreo Estadístico?

Muestreo es la selección de una parte de un agregado o totalidad conocida como Población, de las cuales se basan las decisiones con respecto a la población.

Las siguientes, son ventajas y /o necesidades para el muestreo en la toma de decisiones estadísticas:

1. Costos: El costo es uno de los principales argumentos a favor del muestreo, básicamente porque una muestra puede proveer datos de suficiente exactitud y con mucho menor costo que un censo.
2. Exactitud: En el muestreo, a diferencia que en un censo, existe un mayor control sobre los errores en la recolección porque una muestra es una agrupación a menor escala.
3. Menor tiempo: Otra ventaja de la muestra sobre el censo es que provee resultados e información más rápida. Esto es
4. importante para una toma de decisión sujeta a un tiempo limitado.
5. Cantidad de información: Información mas detallada puede ser mejor obtenida una muestra que en de un censo, porque la muestra toma menos tiempo, es menos costosa y nos permite tener mas cuidado en las etapas de procesamiento de los datos.
6. Pruebas deductivas: Cuando una prueba envuelve la deducción de un objeto en estudio, el muestreo tiene que ser usado. La determinación del muestreo estadístico puede ser usado para encontrar el tamaño optimo de la muestra a un costo aceptable.

Métodos de Muestreo

Desde la comida que usted come hasta la televisión que usted ve, desde las elecciones políticas hasta el consejo disciplinario del colegio, muchos aspectos de su vida están controlados y regulados por encuestas sobre muestras.

Una muestra es un grupo de unidades seleccionadas de un grupo mayor (población). Mediante el estudio de una muestra, se espera que proporcione conclusiones validas sobre el grupo mayor.

La muestra es generalmente seleccionada para ser el objeto de estudio ya que las poblaciones son muy largas para estudiarlas en su totalidad. La muestra debería ser representativa de la población. Esto es normalmente mejor alcanzado mediante el muestreo aleatorio. Adicionalmente, antes de recolectar la muestra, es importante que la población sea definida cuidadosa y completamente, incluyendo una descripción de los miembros que la conformaran.

Un problema común en la toma de decisión estadística de negocios se presenta cuando necesitamos la información en referencia a una población, pero encontramos que el costo de

obtenerla es exagerado. Por ejemplo, suponga que necesitamos saber el tiempo promedio de vida del inventario actual. Si el inventario es grande, el costo de comprobar los registros de cada uno de los artículos podría cancelar el beneficio de tener la información. Por otra parte, la intuición acerca del posible tiempo promedio de vida del inventario podría no ser suficiente para el propósito de toma de decisiones. Esto significa que debemos abordar la situación que implique el seleccionar un número pequeño de artículos y calcular su average de vida útil dentro del inventario, como una estimación del tiempo promedio de vida del

inventario total. Esto es un compromiso, puesto que las medidas para la muestra del inventario producirán solo una estimación del valor que deseamos, pero con ahorros substanciales. Lo que quisiéramos saber es que tan "buena" es la estimación y cuánto mas costara para hacerla "mejor". La información de este tipo esta directamente relacionada con las técnicas de muestreo. Esta sección proporciona una discusión corta sobre los métodos comunes de muestreo estadístico de negocios.

Muestreo de Grupos se puede utilizar siempre que la población sea homogénea, pero que a su vez puede ser particionada. En muchos casos las particiones son resultados de distancias físicas. Por ejemplo, en la industria de seguros, existen "grupos" pequeños de empleados en oficinas del mismo ramo o especialización, las cuales están dispersadas alrededor de todo el país. En este caso, un muestreo aleatorio de los hábitos de trabajo del empleado no requeriría el viajar a muchos de estos "grupos" o campos de trabajo con el objetivo de recolectar los datos. El muestreo total de cada uno de los contados grupos elegidos podría reducir mucho el costo asociado a los requerimiento de datos por parte de la gerencia.

Muestreo Estratificado puede ser utilizado siempre que la población pueda ser particionada en sub poblaciones más pequeñas, cada uno de las cuales es homogénea según las características particulares de interés. Si existen k sub poblaciones y dejamos que N_i denote el tamaño de la sub población i, N denote el tamaño de la población total, y dejamos que n represente el tamaño de la muestra, y deje n denotar el tamaño de muestra, entonces seleccionamos una muestra estratificada siempre que escogemos:

$n_i = n(N_i/N)$
unidades aleatorias de la sub población i, donde i = 1,2, …. ,k.

El estimador es:

$\bar{x}_s$ = S W_t. $\bar{x}_t$, sobre 1 , 2, .L (estratificado), y $\bar{x}_t$ es SX_{it}/n_t.

Su varianza es:

$SW^2_t /(N_t-n_t)S^2_t/[n_t(N_t-1)]$

La población total T es estimada por N. $\bar{x}_s$; su varianza es:

$SN^2_t(N_t-n_t)S^2_t/[n_t(N_t-1)]$.

Muestreo Aleatorio es probablemente el método de muestreo más usado en la toma de decisiones de hoy en día. Muchas decisiones, por lo tanto, son escogiendo un número dentro de un sombrero o un grano de un barril, estos dos métodos son intentos para alcanzar una selección aleatoria de un conjunto de elementos. Pero, un verdadero muestreo aleatorio debe ser alcanzado con la ayuda de una computadora o de una tabla de números aleatorios de los cuales sus valores son generados por generadores de números aleatorios.

Un muestreo aleatorio de tamaño n es obtenido de una población de tamaño N. La estimación balanceada para la varianza de $\bar{x}$es:

Var($\bar{x}$) = $S^2(1-n/N)/n$,
donde n /N la fracción de la muestra con respecto a la población. Para proporción de muestra menor a 10%, el factor de corrección para una población finita es (N-n)/ (N-1), el cual es casi 1.

El T total es estimado por N ´ $\bar{x}$, su varianza es N^2Var($\bar{x}$).

Para variables tipo 0, 1 (binarias), variación en la proporción estimada p es:

S^2 = p(1-p) ´ (1-n/N)/(n-1).

Para el cociente r = $Sx_i/Sy_i=\bar{x}$ / $\bar{y}$, la variación para r es:

$[(N-n)(r^2S^2_x + S^2_y -2 r Cov(x, y)]/[n(N-1)\bar{x}^2]$.

Determinación del tamaño de la muestra (n) con referencia a datos binarios: Los integradores mas pequeños que sean mas grandes o iguales a:

$[t^2 N p(1-p)] / [t^2 p(1-p) + a^2 (N-1)]$,
de donde N es el tamaño total de números de casos, n el tamaño de la muestra, a el error esperado, t el valor obtenido de la distribución t correspondiente a un cierto intervalo de confianza, y p la probabilidad de un evento.

Muestreo de Selección Cruzada: La selección cruzada estudia las observaciones de una población definida un momento o intervalo de tiempo determinado. Muestras y resultados son calculados al mismo tiempo.

¿Qué es un Instrumento Estadístico? Un instrumento estadístico es cualquier proceso que tiene como objetivo describir los fenómenos usando cualquier instrumento o dispositivo. No obstante, los resultados se pueden utilizados como herramientas del control. Ejemplos de instrumentos estadísticos son los cuestionario y muestreos por encuestas.

¿Cuál es la Técnica de Muestreo por Captura? Esta técnica consiste en tomar una muestra relativamente pequeña por un período del tiempo muy corto, donde los resultados son obtenidos generalmente de manera instantánea. Sin embargo, el **muestreo pasivo** es una técnica donde un instrumento de muestreo se utiliza por un periodo de tiempo mas largo y manteniendo condiciones similares. Dependiendo de la investigación estadística deseable, el muestreo pasivo puede ser una alternativa útil o aún más apropiado que el muestreo por captura. Sin embargo, una técnica de muestreo pasiva necesita ser desarrollada y ser probada en el campo. No obstante, la técnica de muestreo pasivo necesita ser desarrollada y probada directamente en el campo de estudio.

Sumario de Estadísticos

Representativo de una Muestra: Sumario de Medidas de Tendencia Central

¿Cómo describiría el “promedio” o un pedazo de información “típica” de un conjunto de datos? Diversos procedimientos se utilizan para resumir la información más representativa de acuerdo al tipo de pregunta y a la naturaleza de los datos que son resumidos.

Las medidas de ubicación dan la información sobre el **lugar** hacia donde existe la tendencia central dentro de un grupo de números. Las medidas de ubicación presentadas en esta unidad para datos no agrupados son la media, la mediana, y la moda.

Media: La media aritmética (o el promedio, media simple) es calculada sumando todos los números de un conjunto de números (x_i) y después dividiéndolos por el número de observaciones (n) del conjunto.

Media = $\bar{x} = S\ X_i\ /n$, la suma incluye todos los i's.

La media utiliza todas las observaciones, y cada observación afecta la media. Aunque la media es sensible a los valores extremos; es decir, los datos extremadamente grandes o pequeños pueden causar que la media se ubique o más cerca de uno de los datos extremos; A pesar de esto, la media sigue siendo la medida lo más usada para medir la localización. Esto se debe a que la media posee valiosas propiedades matemáticas que la hacen conveniente para el uso en el análisis estadístico de inferencia o deductivo. Por ejemplo, la suma de las desviaciones entre los números de un conjunto de datos con respecto a la media es cero, y la suma de las desviaciones elevadas al cuadrado entre los números en un conjunto de datos con respecto a la media es el valor mínimo.

A usted podría gustarle usar Applets de Estadística Descriptiva para calcular la media.

Media Ponderada: en algunos casos, los datos de una muestra o población no deberían ser ponderados de la misma manera, es preferible ponderarlos de acuerdo a su importancia.

Mediana: La mediana es el valor medio de una grupo **ordenado** de observaciones. Si existe un número par de observaciones correspondientes al grupo, la mediana es el **average** de los dos números ubicados en el medio del grupo. Si existe un número impar de observaciones correspondientes al grupo, la mediana es el número en el **medio** del grupo.

La mediana es normalmente utilizada resumir los resultados de una distribución. Si la distribución es oblicua o sesgada, la mediana y el rango inter cuartíl (RIC), serían los mejores indicadores de medida para saber donde los datos observados se encuentran concentrados.

Generalmente, la mediana proporciona una mejor medida mejor de localización que la media cuando hay algunas observaciones extremadamente grandes o pequeñas; es decir, cuando los datos se sesgan a la derecha o a la izquierda. Por esta razón, la mediana de la renta se utiliza como la medida de ubicación para la renta por hogar en los Estados Unidos. Observe que si el valor de la mediana es **menor que** que el de la media, los datos están sesgados a la derecha. Si el valor de la mediana es **mayor que** que el de la media, los datos están sesgados a la izquierda. Para una población normal, la mediana de la muestra se distribuye normalmente con media = m = y error estándar de la mediana de $(p/2)^{½}$ veces con respecto a la media.

La media tiene dos ventajas distintas sobre la mediana. Es más estable, y uno puede calcular la media basada de dos muestras combinando las dos medios de las mismas.

Moda: La moda es el valor lo más con frecuencia posible que ocurre de un sistema de observaciones. ¿Por qué utilizar la moda? El ejemplo clásico es el fabricante de zapatos/ camisas que desea decidir a qué tallas introducir en el mercado. Los datos pueden tener dos modas. En este caso, decimos que los datos son **bimodales**, y los grupos de observaciones con más de dos modos están referidos como **multimodales**. Observe que la moda no es una medida útil de ubicación, porque puede haber más de una moda o quizás ninguna.

Cuando la media y la mediana son conocidas, es posible estimar la moda para la distribución unimodal usando los otros dos promedios como se muestra a continuación:

Moda » 3(medianas) - 2(medias)
Esta estimación es aplicable a ambos, conjuntos agrupado y no agrupado de datos.

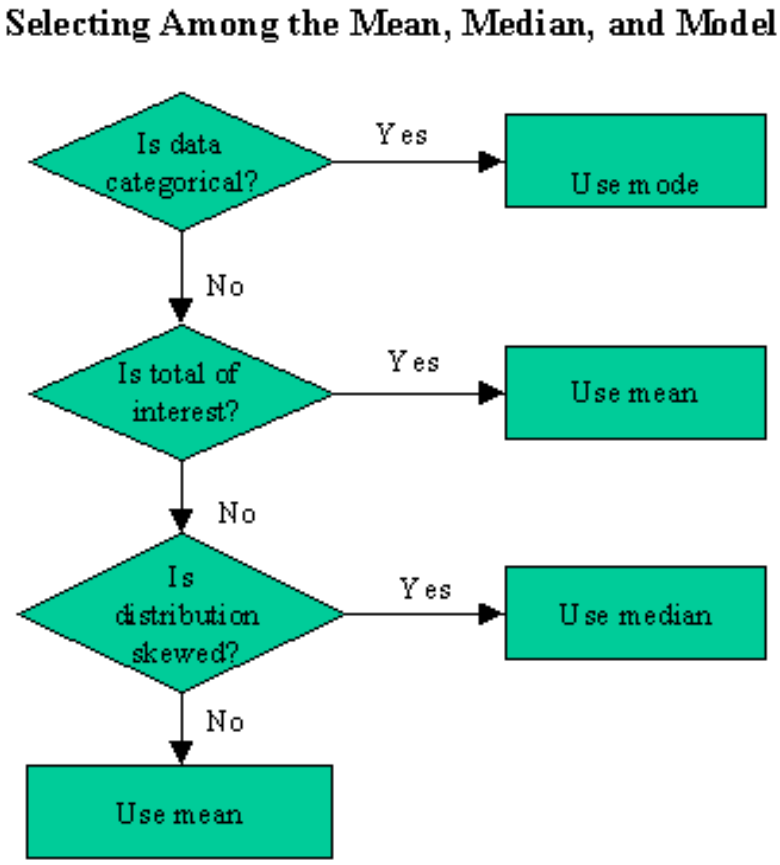
Siempre que exista más de una moda, la población de la cual la muestra es obtenida es una mezcla de más de una población. Sin embargo, note que una distribución Uniforme tiene un incontable número de modas que tienen igual valor de densidad; por lo tanto se considera como población homogénea.

Casi todos los análisis estadísticos estándar se condicionan en la asunción de que la población es homogénea.

Note que Excel tiene una capacidad estadística muy limitada. Por ejemplo, exhibe solamente una moda, la primera. Desafortunadamente, esto es muy engañoso. Sin embargo, usted puede descubrir si existen otras modas mediante el método de inspección, como sigue: Cree una distribución de frecuencia, invoque la secuencia del menú: Herramientas, análisis de datos, frecuencia y sigua las instrucciones en la pantalla. Usted verá la distribución de frecuencia y después encontrará la moda visualmente. Desafortunadamente, Excel no proporciona diagramas de árbol. Todo el software disponible comercialmente, tal como el SAS y SPSS, exhiben diagramas de árbol, el cual es una distribución de frecuencia de un grupo dado de datos.

Seleccionando Entre la Media (Mean), Mediana (Median) y Moda (Mode)

Es un error común el especificar el índice equivocado para la tendencia central.



La primera consideración es el tipo de data, si la variable es categórica, la moda es la medida más simple que mejor describe los datos.

La segunda consideración para seleccionar el índice es preguntarse si el total de las observaciones tiene algún interés. Si la respuesta es si, entonces la media es el índice apropiado para la tendencia central.

Si el total no interesa, dependerá entonces si el histograma es simétrico o sesgado, y se deberá utilizar la media o la mediana respectivamente.

En todo los casos, el histograma debe ser unimodal. Sin embrago, note que por ejemplo una distribución uniforme tiene un número incontable de modas con igual valor de densidad, por lo tanto es considerada como una población homogénea. Adicionalmente note que:

|Media - Mediana| £s

Las características principales de estos tres estadísticos son tabuladas a continuación:

Principales Características de la Moda, Mediana y Media Hechos	**Moda**	**Mediana**	**Media**
1	Es el valor mas frecuente en la distribución. Es el punto de mas alto densidad.	Es el valor del punto medio de la selección (no del rango), tal que la mitad de los datos están por arriba y por debajo de ella.	Es el valor en algún agregado, el cual se obtendría si todos los valores fueran iguales.
2	Su valor es establecido por la frecuencia predominante, no por los valores en la distribución.	El valor de la media es fijado por su posición en la selección, y no refleja valores individuales.	La suma de las desviaciones en cualquier lado de la media son iguales; por lo tanto la suma algebraica de sus desviaciones es cero.
3	Este es el valor mas probable, por lo tanto el mas común.	La distancia agregada entre la mediana y cualquier otro punto de la muestra es menor que en cualquier otro punto.	Esta refleja la magnitud de cada valor.
4	Una distribución puede tener mas de 2 modas, pero no existe moda en una distribución rectangular.	Cada selección tiene solo una mediana.	Una muestra tiene solo una media.
5	No puede ser manipulada algebraicamente. Modas de subgrupos no pueden ser ponderadas o combinadas.	No puede ser manipulada algebraicamente. Medianas de subgrupos no pueden ser ponderadas o combinadas.	Pueden ser manipuladas algebraicamente. Medias de subgrupos pueden ser combinadas cuando son ponderadas apropiadamente.
6	Es inestable, puede ser influenciada en el proceso de agrupación.	Es estable en cuanto a que procedimientos para agrupar no afecta su	Es estable en cuanto a que procedimientos para agrupar no afecta su

		apreciación.	apreciación.
7	La moda no refleja el grado de modalidad.	No es aplicable para datos cualitativos.	Podría ser calcula igualmente cuando los valores individuales son desconocidos, si se posee la suma de los valores y el tamaño de la muestra.
8	Puede ser calculada cuando los extremos de los valores de los grupos son abiertos.	Puede ser calculado cuando los valores extremos son abiertos.	No puede ser calculado de una tabla de frecuencia cuando sus valores extremos son abiertos.
9	Valores deben ser ordenados para su cálculo.	Valores deben ser ordenados y agrupados para su cálculo.	Los valores no necesitan ser ordenados para su cálculo.

Para la Estadística Descriptiva, JavaScript proporciona un conjunto completo de información que usted podría necesitar. A usted le podría gustar usarlo para realizar algunas experimentaciones numéricas que validan las aserciones anteriores para un entendimiento mas profundo.

Promedios Especializados: La Media Geométrica y la Media Armónica

La Media Geométrica: La media geométrica (G) de n valores no negativos es la enésima raíz del producto de los n valores.

Si algunos valores son muy grandes en magnitud y otros muy pequeños, la media geométrica proporciona una mejor representación de los datos que un simple promedio. In una “serie geométrica”, el average mas significativo es la media geométrica (G). La media aritmética es muy favorecida por valores grandes de la serie.

Una aplicación: Suponga que las ventas de un determinado producto incrementan en 110% en el primer año y en 150% en el segundo. Por simplicidad, asuma que usted inicialmente vendió 100 unidades. Entonces el número de unidades vendidas en el primer año fueron 110 y en el segundo fueron 150% x110= 165. Usando la media aritmética de 110% y 150% que es 130%, estimaríamos incorrectamente las unidades vendidas en el primer año de 130 y las del segundo año de 169. Mediante la media geométrica de 110% y

150% obtendríamos G = $(1,65)^{1/2}$ la cual es la estimación correcta, por lo cual venderíamos 100 $(G)^2$ = 165 unidades en el segundo año.

La Media Armónica: La media armónica otro average especializado, el cual es útil para calcular promedios de variables expresadas en proporciones de unidades por tiempo, tales como kilómetros por hora, número de unidades de producción por día. La media armónica (G) de n valores no cero x(i) es: H = n/[S (1/x(i)].

Una aplicación: Suponga que cuatro maquinas en un taller son usadas para producir la misma pieza. Pero, cada una de las maquinas se toma 2,5, 2, 1,5 y 6 minutos para realizar dicha pieza. ¿ Cuál es la velocidad promedio de producción?

La media armónica es: H = 4/[(1/2,5) + (1/2,0) + 1/(1,5) + (1/6,0)] = 2,31 minutos.

Si todas las maquinas trabajaran por una hora, ¿cuántas unidades serian producidas? Porque cuatro maquinas trabajando por una hora representan 240 minutos de operación, se obtiene que: 240 / 2,31 = 104 piezas serán producidas.

El Orden Entre las Tres Medias: Si todas las tres medias existen, la media aritmética nunca es menor que las otras dos, además, la media armónica nunca es mayor que las otras.

A usted podría gustarle usar el JavaScript de Las Otras Medias en Javasript para realizar algunos experimentos numéricos que validan las aserciones anteriores para un entendimiento mas profundo.

Histogramas: Analizando la Homogeneidad de la Población

Un histograma es una representación gráfica de una estimación para la densidad (para variables aleatorias continuas) o la función de probabilidad total (para variables aleatorias discretas) de la población.

Las características geométricas del histograma nos permiten descubrir información útil sobre los datos, por ejemplo:

1. La localización del “centro” de los datos.
2. El grado de dispersión.
3. La sección a la cual se sesga, es decir, cuando no cae simétricamente en ambos lados del pico.
4. El grado de agudeza del pico. Cómo se levanta y baja la pendiente.

La moda es el valor más frecuente que ocurre en un grupo de observaciones. Los datos pueden tener dos modas. En este caso, decimos que los datos son

bimodales, y los grupos de observaciones con más de dos modas están referidos como **multimodales**. Siempre que exista más de una moda, la población de la cual la muestra es obtenida es una mezcla de más de una población. Casi todos los análisis estadísticos estándares se condicionan en la asunción que la población es homogénea, lo que significa que su densidad (para variables aleatorias continuas) o la función total de la probabilidad (para variables aleatorias discretas) es unimodal. Sin embargo, note que, por ejemplo, una Uniforme tiene un número incontable de modas que tienen igual valor de densidad, por lo tanto se considera como población homogénea.

Para comprobar el unimodalidad de los datos de la muestra, se podría utilizar el proceso de creación de histogramas.

número de intervalos de clase en un histograma: Antes de que poder construir nuestra distribución de frecuencia debemos determinar cuántas clases debemos utilizar. Esto es puramente arbitrario, pero demasiadas o pocas clases no proporcionarán una clara visión de la distribución a la que se obtendría con un número de clases cercanas al óptimo. Una relación empírica (es decir, observada), conocida como la regla de Sturge, se puede utilizar como guía útil para determinar el número óptimo de clases (k), el cual es dado por el entero mas pequeño mayor o igual a:

Mínimo de { $n^{1/2}$, 10 Log(n) }, $n \geq 30$,

de donde k es el número de clases, Log es en base a 10, y n es el número total de los valores numéricos que abarcan los datos.

Por lo tanto, la anchura de la clase es:

(Valor mas alto – valor mas bajo) / k

El siguiente Javascript genera un histograma basado en esta regla: Prueba de homogeneidad para una población.

Para lograr un “óptimo” se necesitan ciertas medidas de calidad, probablemente en este caso, esta sea la “mejor” manera de exhibir cualquier información disponible de los datos. El tamaño de muestra contribuye a esto; las pautas generalmente deben utilizar entre 5 y 15 clases, con más clases si se tiene una muestra más grande. Usted debe considerar la preferencia por anchuras ordenadas de la clase, preferiblemente un múltiplo de 5 o 10, la cual la haría más fácil de entender.

Más allá de aquí, esto se convierte en una cuestión de juicio. Pruebe varios rangos de anchura de las clases, y elija el que trabaje lo mejor posible. Esto asume que usted tiene una computadora y que puede generar histogramas alternativos fáciles de leer.

A menudo existen también problemas de gerencia que se unen al juego. Por ejemplo, si sus datos van a ser comparados a datos similares, tales como de

estudios anteriores, o de otros países, sus parámetros se restringen a los intervalos a usados en estos.

Si el histograma es muy sesgado, clases desiguales deben ser consideradas. Utilice clases estrechas donde las frecuencias de clase sean altas, y anchas donde estas sean bajas.

Los acercamientos siguientes son comunes:

Deje que n sea el tamaño de la muestra, después el número de intervalos de clase podría ser:

Min $\{n^{½}, 10 \text{ Log}(n)\}$.

El logaritmo en base 10. De esta forma, para 200 observaciones usted utilizaría 14 intervalos pero para 2000 utilizara 33.

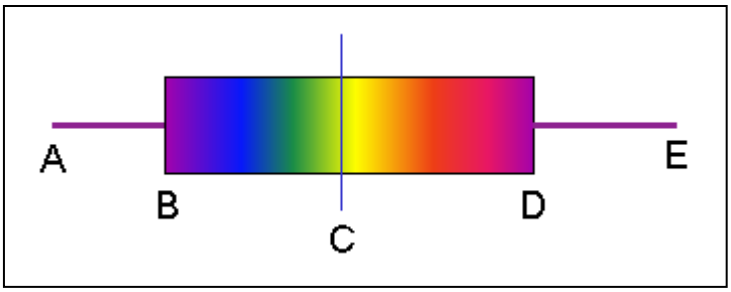
Alternativamente,

1. Encuentre el rango (Valor más alto - el valor más bajo).
2. Divida el rango por un tamaño razonable de intervalos: 2, 3, 5, 10 o un múltiplo de 10.
3. Pruebe intervalos no menores de 5 no mayores de 15.

Uno de los usos principales de los histogramas es para la Prueba la Homogeneidad de una Población. El unimodalidad del histograma es una condición necesaria para la homogeneidad de la población, con el objetivo de hacer cualquier análisis estadístico significativo. Sin embargo, note que una distribución Uniforme tiene incontable cantidad de modas que tienen igual valor de densidad, por lo tanto es considerada como población homogénea.

Cómo Construir un BoxPlot

Un BoxPlot es una exhibición gráfica que tiene muchas características. Incluye la presencia de posibles outliers. Ilustra los rangos de los datos. Muestra una medida de dispersión tal como el cuartíl superior, cuartíl inferior y los rangos intercuartiles (RIC) de un conjunto de datos, así como también a la mediana como medida central ubicación, el cual es útil para comparar grupos de datos. También indica acerca de la simetría o de la oblicuidad de la distribución. La razón principal del renombre de boxplots es porque ofrecen mucha información de una manera compacta.



Pasos para Construir un Boxplot:

1. Líneas horizontales son obtenidas de las observaciones mas pequeñas (A), en el cuartíl mas bajo, y otro para el cuartíl mas alto (D), de observaciones mas largas (E). Las líneas verticales que producen la caja, se unen con las líneas horizontales en los puntos B y D.
2. La línea vertical es dibujada en el punto medio (C), como es mostrado en la figura anterior.

Para un entendimiento mas profundo, usted podría utilizar papel para gráficos, y el JavaScript de muestreo de estadística descriptiva para construir boxplots para un conjunto de datos, por ejemplo, de su libro de texto.

Midiendo la Calidad de la Muestra

El promedio por sí mismo no es una buena indicación de la calidad. Usted necesita conocer la varianza para cualquier evaluación educada. Esto nos recuerda el dilema del estadístico que media dos metros de alto y que se ahogó en una corriente que tenía un metro de profundidad.

Las mediciones estadísticas son normalmente utilizadas para describir la naturaleza y el grado de diferencias entre la información de la distribución. Una medida de variabilidad es generalmente expresada junto con una medida de tendencia central.

Las mediciones estadísticas de variación son valores numéricos que indican la **variabilidad** inherente en un grupo de mediciones de datos. Observe que un valor pequeño para la medida de dispersión indica que los datos están concentrados alrededor de la media; por lo tanto, la media es una buena representación de los datos. Por otra parte, una medida grande de dispersión indica que la media no es una buena representación de los datos. Adicionalmente, las medidas de dispersión pueden ser utilizadas cuando deseamos comparar las distribuciones de dos o más conjuntos de datos. *La calidad de un conjunto de datos es medida por su variabilidad: variabilidad grande indica baja calidad.* Esta es la razón del porque gerentes se preocupan cuando encuentran grandes variaciones. Su trabajo, como estadístico, es medir la variación, y si es demasiado alto e inaceptable, entonces es trabajo del personal técnico, tal como ingenieros, en ajustar el proceso.

Situaciones de decisión con la carencia absoluta de conocimiento, conocida como **incertidumbre plena**, tienen el riesgo más grande. Para simplificar, considere el caso cuando hay solamente dos resultados, uno con la probabilidad de p. Entonces, la variación en los resultados es p(1-p). Esta variación es la más grande si fijamos p = 50%. Es decir, igual oportunidad para cada resultado. En este caso, la calidad de la información está en su nivel más bajo.

Recuerde, **calidad en la información y variación están relacionadas inversamente**. Cuanto más grande es la variación en los datos, más baja es la calidad de los datos (información): **el Diablo está en las Desviaciones.**

Las cuatro medidas de variación más comunes son: **el rango**, **varianza**, **desviación estándar**, y **el coeficiente de variación**.

Rango: El rango de un grupo de observaciones es el valor absoluto de la diferencia entre el valor más grande y más pequeño del conjunto de datos. Mide el tamaño del intervalo inmediato de números reales más pequeño que abarcan todos los valores de los datos. No es útil cuando existen valores extremos. Se basa solamente en dos valores, no en la totalidad de los datos. Adicionalmente, no puede ser definido en distribuciones de extremos abiertos tales como la distribución normal.

Note que, al trabajar con observaciones aleatorias discretas, algunos autores definen el rango como:
Rango = Valor más grande - valor más pequeño + 1.

Una distribución normal no tiene rango. Un estudiante dijo, “porque las colas de una función de densidad normal nunca toca el eje de las x, y porque para que una observación contribuya a la creación de dicha curva, muchos valores negativos y positivos deben existir”, pero estos valores remotos siempre tienen la posibilidad de existir, pero cada vez son más improbable. Esto encapsula muy bien el comportamiento asintótico de la densidad normal. Por lo tanto, a pesar de este comportamiento, es útil y aplicable a una amplia gama de las situaciones de toma de decisión.

Cuartiles: Cuando requerimos los datos, por ejemplo en orden ascendente, podemos dividir los datos en cuartos, Q1... Q4, conocidos como cuartiles. El primer cuartíl (Q1) es el valor donde están 25% de los valores mas pequeños y en el otro 75% los más grandes. El segundo cuartíl (Q2) es el valor donde están 50% de los valores mas pequeños y en el otro 50% los más grandes. En el tercer cuartíl (Q3) es el valor donde están 75% de los valores mas pequeños y en el otro 25% los más grandes.

Porcentajes: Los porcentajes tienen un concepto similar y por lo tanto, están relacionados; por ejemplo, el 25 porciento corresponde al primer cuartíl Q1, etc. La ventaja de los porcentajes es que pueden ser subdivididos en 100 porciones. Los porcentajes y los cuartiles son más convenientes de leer cuando son tomados de una función de distribución acumulativa.

Rango entre: El rango intercuartíl (RIC) describe el grado de dispersión o acumulación del 50% de las observaciones ubicadas en el medio de la distribución. Es la distancia entre el primero y tercer cuartíl:

RIC = Q3 - Q1,

el cual es dos veces la Desviación Cuartíl. Para datos que están sesgados, la **dispersión relativa**, similar to the coefficient of variation (C.V.) similar al

coeficiente de variación (CV) es dada (provisto de numerador no-cero) por el Coeficiente de Variación Cuartíl:

CVC = (Q3-Q1) / (Q3 + Q1).

Note que casi todos los estadísticos que hemos cubierto hasta ahora pueden ser obtenidos y entendidos con mayor profundidad por métodos gráficos usando la Función de Distribución Empírica (observada) Acumulativa (FDEA) en Javascript. Sin embargo, el JavaScript numérico de Estadística Descriptiva proporciona un conjunto completo de información de todos los estadísticos que usted podría necesitar.

La Dualidad entre la FDEA y el Histograma: Note que la función de distribución empírica(observada) acumulativa (FDEA) indicada por la su altura en un punto particular de la curva, es numéricamente igual al área en el histograma correspondiente al lado izquierdo de ese punto. Por lo tanto, cualquiera o ambos se podían utilizar dependiendo de los usos previstos.

Media de desviación absoluta (MDA): Una simple medida de variabilidad es la media de desviación absoluta:

$MDA = S\ |(x_i - \bar{x})| / n.$

La media de desviación absoluta es ampliamente utilizada como medida de funcionamiento para determinar la calidad del modelo, tales como las técnicas de predicción. Sin embargo, el MDA no se presta para el cálculo de inferencias; por otra parte, igualmente en los estudios de análisis de error, la varianza es preferida, porque las varianzas de errores independientes (o sin correlación) son aditivas; Sin embargo, la MDA no tiene tan elegantes presentaciones.

La MDA es una simple medida de variabilidad, que a diferencia del rango y de la desviación cuartíl, toma en cuenta cada objeto de la muestra, y es más simple y menos afectada por desviaciones extremas. Por lo tanto se utiliza a menudo en las muestras pequeñas que incluyen valores extremos.

La media de desviación absoluta teóricamente debe ser medida con respecto a la mediana porque esta representa su mínimo; sin embargo, es más conveniente medir las desviaciones con respecto a la media.

Como ejemplo numérico, considere el precio (en $) del mismo artículo en 5 diversos almacenes: $4,75, $5,00, $4,65, $6,10, y $6,30. La media de la desviación absoluta con respecto a la media es $0,67, mientras que con respecto a la mediana es $0,60, el cual es una mejor representación de la desviación entre los precios.

Varianza: Es una importante medida de variabilidad. La varianza es el **promedio** de las **desviaciones estándar** elevadas al cuadrado de cada una de las observaciones con respecto a la media.

$Varianza = S\ (x_i - \bar{x})^2 / (n - 1),$ de donde n por lo menos 2.

La varianza es una medida de dispersión entre valores de los datos. Por lo tanto, mientras más grande sea la varianza, menor será la calidad de los datos.

La varianza **no es expresada en las mismas unidades que las observaciones.** Es decir, la varianza es difícil de entender porque las desviaciones con respecto a la media están elevadas al cuadrado, haciéndola demasiado grande para explicaciones lógicas. Este problema puede ser solucionado trabajando con la **raíz cuadrada** de la varianza, lo cual se conoce como la **desviación estándar.**

Desviación Estándar: Ambas, la varianza y la desviación estándar proporcionan la misma información; **una siempre puede ser obtenida de la otra**. Es decir, el proceso de cálculo de la desviación estándar siempre implica el cálculo de la varianza. Puesto que la desviación estándar es la raíz cuadrada de la varianza, esta siempre es expresada en **las mismas unidades** que el conjunto de datos:

Desviación estándar= S = (Varianza) ^{½}

Para conjunto de datos grandes (digamos más de 30), aproximadamente el 68% de los datos están contenidos dentro de una desviación estándar con respecto a la media, 95% de los datos caen dentro de dos desviaciones estándar. 97,7% (o casi 100%) de los datos se encuentran dentro de tres desviaciones estándar (S) con respecto a la media.

Usted puede utilizar el JavaScript de Estadística Descriptiva para calcular la media, y la desviación estándar.

La Media de los Errores al Cuadrado (MEC) de una estimación es la varianza de la estimación más el cuadrado de su desviaciones; por lo tanto, si una estimación es imparcial, entonces su MEC es igual a su varianza, como es el caso de la tabla de ANOVA.

Coeficiente de Variación: El coeficiente de variación (CV) es la *desviación relativa absoluta* con respecto al tamaño , siempre que sea cero, expresado en porcentaje:

CV =100 |S/| %

El CV es independiente de las unidades de medida. En la estimación de un parámetro, cuando su CV es menos del 10%, la estimación se asume aceptable. En el caso contrario, digamos, 1/CV se llama el **Cociente de señal de ruido**.

El coeficiente de variación se utiliza para representar la relación de la desviación estándar hacia la media, diciendo cuan representativa es la media de los números de los cuales fue calculada. Esta expresa la desviación estándar como porcentaje de la media; es decir, refleja la variación de una distribución con respecto a la media. Sin embargo, los intervalos de la confianza para el coeficiente de variación generalmente no son expresados.

Una de las razones es que el cálculo exacto del intervalo de confianza para el coeficiente de variación es tedioso de obtener.

Observe que, para un conjunto de datos agrupados o sesgados, el **coeficiente de variación cuartíl** es:

$V_Q = 100(Q_3 - Q_1)/(Q_3 + Q_1)\%$
es mas útil que el CV.

Usted puede utilizar el JavaScript de Estadística Descriptiva para calcular la media, la desviación estándar y el coeficiente de variación.

Cociente de Variación para Datos Cualitativos: Puesto que la moda es la medida mas usada para la tendencia central de variables cualitativas, la variabilidad es medida con respecto a la moda. El estadístico que describe la variabilidad de datos cuantitativos es el cociente de variación (VR):

$VR = 1 - f_m/n,$

de donde f_m es la frecuencia de la moda, y n es el número total de cálculos en la distribución.

Score Z: cuántas desviaciones estándar en un punto dado (es decir, observación) están por debajo a arriba de la media. Es decir, valor **Z** representa el número de las desviaciones estándar que una observación (x) está *arriba o debajo* de la media. *Cuanto más grande sea el valor de Z, más lejos estará el valor de la media*. Observe que valores más allá de tres desviaciones estándar son bastante raros. Si un score Z es negativo, la observación (x) está debajo de la media. Si el score Z es positivo, la observación (x) está por arriba de la media. El score Z se obtiene por:

Z = (x -) / Desviación Estándar de X

El score Z es una medida del número de desviaciones estándar en la que una observación está por arriba o por debajo de la media. Puesto que la desviación estándar nunca es negativa, un valor Z positiva indica que la observación está por arriba de la media, una score Z negativa indica que la observación está por debajo de la media. Note que Z es un valor sin dimensiones, y por lo tanto es una medida útil para comparar valores de datos de dos poblaciones distintas, incluso cuando sean medidas por unidades distintas.

Transformación -Z: Aplicando la fórmula z = (X - m) / s siempre se producirá una variable transformada con media de cero y desviación estándar uno. Sin embargo, la forma de la distribución no será afectada por la transformación. Si X no es normal, entonces la distribución transformada tampoco será normal.

Una de las características interesantes de la Transformación-Z es que la distribución resultante de los datos transformados tiene una **forma idéntica** pero con media cero, y desviación estándar igual a 1.

Se podría generalizar esta transformación de los datos para obtener cualquier media y desviación estándar deseable diferentes de 0 y 1, respectivamente. Suponga que deseamos que los datos transformados tengan media M y desviación estándar D, respectivamente. Por ejemplo, en los resultados de una prueba para ingresar a la escuela de leyes, se fijan en M = 500, y D =100. La transformación siguiente debe ser aplicada:

Z = (estándar Z) ´ D + M

Suponga que usted tiene dos grupos de datos con escalas muy diferentes (por ejemplo, una tiene valores muy bajos y la otra valores muy altos). Si usted deseara comparar estos dos grupos, debido a las diferencias en las escalas respectivas, los estadísticos que se generarían no serian comparables. Seria una buena idea utilizar la transformación-Z de ambos datos originales y después hacer cualquier comparación.

Usted ha oído los términos valor z, la prueba z, la transformación z, y el score Z . ¿Todos estos términos significan lo mismo? Ciertamente no:

El valor z refiere al valor crítico (un punto en los ejes horizontales) de una Función de Densidad Normal (0, 1) para un área dada a la izquierda de ese valor z.

La prueba z se refiere a los procedimientos para probar la igualdad de la media(s) de un (o dos) población (es).

El score Z de una observación x dada, en una muestra del tamaño n, el cual es simplemente (x - promedio de la muestra) dividida por la desviación estándar de la muestra. Se debe tener cuidado de no confundir los valores Z con los valores estándares.

La transformación - z de un sistema de observaciones de tamaño n es simplemente (cada observación - promedio de todas las observaciones) dividida por la desviación estándar entre todas las observaciones. El objetivo es producir datos transformados con una media cero y desviación estándar uno. Esto hace de los datos transformados sin dimensiones y manejable con respecto a sus magnitudes. Se utiliza también en comparar varios grupos de datos que han medidos usando diversas escalas de medición.

Pearson recalcó el término "desviación estándar" en algún momento durante los años 1900’s. La idea de usar desviaciones al cuadrado va mucho mas atrás con Laplace a comienzo de los 1800's.

Finalmente, note de nuevo, que transformando los datos originales a valor Z no normalizan los datos.

Cálculo de Estadísticos Descriptivos para Datos Agrupados: Una de las maneras más comunes de describir una sola variable es con una distribución de frecuencia. Un histograma es una representación gráfica de una estimación para la distribución de frecuencia de la población. Dependiendo de las variables

particulares, todos los valores de los datos podrían ser representados, o se podrían agrupar los valores primero por categorías (por ejemplo, por edad). Generalmente, no sería sensible determinar las frecuencias para cada valor. Preferiblemente, los valores deberían ser agrupados en rangos, y luego determinar la frecuencia. Las distribuciones de frecuencia se pueden representar de dos maneras: como tablas o como gráficos, los cuales a menudo se refieren a histogramas o gráfico de barras. Los gráficos de barras son normalmente utilizados para mostrar la relación entre dos variables categóricas.

Los datos agrupados son derivados de informaciones ordinarias, y consisten en frecuencias (cálculo de valores ordinarios) tabulados con las clases en las cuales ocurren. Los límites de las clases representan los valores más pequeños (inferiores) y más grandes (superior) que la clase contendrá. Las fórmulas para los estadísticos descriptivos son mucho más simples para los datos agrupados, así como se muestra en las siguientes formulas para la media, varianza, y la desviación estándar, respectivamente, de donde f representa la frecuencia de cada clase, y n es la frecuencia total:

$$\bar{X} = \frac{\Sigma fX}{n}$$

$$s^2 = \frac{\Sigma fX^2 - \frac{(\Sigma fX)^2}{n}}{n-1}$$

$$s = \sqrt{\frac{\Sigma fX^2 - \frac{(\Sigma fX)^2}{n}}{n-1}}$$

Seleccionando entre Desviación Cuartíl, Media de Desviación Absoluta y Desviación Estándar

Una guía general para seleccionar el estadístico adecuado para describir la dispersión de la población, incluye la consideración de los siguientes factores:

1. El concepto de dispersión que el problema requiere. ¿Es un simple par de valores adecuado, tal como los dos extremos o los dos cuartiles (rango o Q)?
2. El tipo de datos disponibles. Si son pocos en números, o contiene valores extremos, evite la desviación estándar. Si se encuentran sesgados, evite la media de desviación absoluta. Si existen brechas entre los cuartiles, la desviación cuartíl se debería evitar.
3. La peculiaridad de la dispersión que los mide. Estos son resumidos en el cuadro de “las Características Principales de la Desviación Cuartíl, la Media de Desviación Absoluta y la Desviación Estándar”, que se muestra a continuación.

Características Principales de la Desviación Cuartíl, la Media de Desviación Absoluta y la Desviación Estándar **Hechos**	**La Desviación Cuartíl**	**La Media de Desviación Absoluta**	**La Desviación Estándar**
1	La desviación cuartíl es fácil de calcular y entender. Sin embargo, esta es inconsistente si existen brechas entre los datos alrededor de los cuartiles.	La Media de Desviación Absoluta tiene la ventaja de dar igual peso a la desviación de cada valor con respecto a la media o la mediana.	La Desviación Estándar es normalmente mas útil y mejor adaptable a análisis mas profundos que lo que es La Media de Desviación Absoluta.
2	Solo depende de dos valores, los cuales incluyen la mitad central de los mismos.	Es una medida de dispersión más sensitiva que cualquiera de las descritas anteriormente, y normalmente tiene errores de muestreo más pequeños.	Es más adaptable como estimador de la dispersión de la población que cualquier otra medición, haciendo que la distribución sea normal.
3	Es normalmente superior al rango como una medida cruda de dispersión.	Es más fácil de calcular y entender, además es menos sensible que la desviación estándar a valores extremos.	Es la más amplia medida de dispersión usada, y la más fácil de manejar algebraicamente.
4	Esta podría ser determinada en una distribución abierta en los extremos, o en una en la cual los datos pueden ser seleccionados pero no medidos cuantitativamente.	Desafortunadamente, es muy difícil de manejar algebraicamente, dado que el signo negativo debe ser ignorado cuando se calcula.	En comparación con los demás, esta es mas difícil de calcular y de entender.
5	Es muy útil en distribuciones muy sesgadas, o en	Su aplicación principal es la precisa elección de modelos en	Es normalmente afectada por valores extremos,

	aquellas en las cuales otras medidas de dispersión serian deformadas por valores extremos.	técnicas de predicciones comparativas.	los cuales podrían ocasionar el sesgamiento de los datos.

A usted podría gustarle utilizar el JavaScript Muestreo Estadístico Descriptivo en Javasript y realizar algunos experimentos numéricos para validar las aserciones anteriores y tener entendimiento mas profundo de los mismos.

Forma de la Función de Distribución: Tabla de Oblicuidad-Kurtosis

El par de medidas estadísticas, oblicuidad y kurtosis, son herramientas de medición, las cuales son usadas para seleccionar la distribución(es) que satisfaga los datos determinados. Para hacer una inferencia con respecto a la distribución de la población, usted primero podría calcular la oblicuidad y kurtosis de su muestra aleatoria de la población entera. Luego, localizar un punto con las coordinadas encontradas en la ampliamente utilizada Tabla de Oblicuidad-Kurtosis, hacer conjetura acerca de las posibles distribuciones que satisfagan los datos. Finalmente, se podrían utilizar la prueba de calidad de ajuste para que rigurosamente obtenga el mejor candidato que satisface los datos. Quitando un outliers se mejora la exactitud de la oblicuidad y kurtosis.

Oblicuidad: La oblicuidad es una medida del grado al cual la muestra de la población se desvía de la simetría con la media ubicada en el centro.

Oblicuidad = $S (x_i - \bar{x})^3 / [(n - 1) S^3]$, n es por lo menos 2.

La oblicuidad adquirirá un valor de cero cuando la distribución es una curva simétrica. Un valor positivo indica que las observaciones están concentradas más a la izquierda de la media con la mayoría de los valores extremos a la derecha de la media. Una oblicuidad negativa indica observaciones concentradas a la derecha. En este caso tenemos: Media £ Mediana £ Moda. El orden reverso se cumple para observaciones con oblicuidad positiva.

Kurtosis: La kurtosis es una medida del apuntamiento relativo de la curva definida por la distribución de las observaciones.

Kurtosis = $S (x_i - \bar{x})^4 / [(n - 1) S^4]$, n es por lo menos 2.

La distribución normal estándar tiene kurtosis de +3. Una kurtosis mayor a 3 indica que la distribución es más elevada que la distribución normal estándar.

Coeficiente de exceso de kurtosis = kurtosis – 3.

Un valor menor a 3 para la kurtosis indica que la distribución es mas plana que la distribución normal estándar.

Se puede demostrado que,

Kurtosis - Oblicuidad ^{2} es mayor o igual que 1, y Kurtosis es menor o igual al tamaño de la muestral n..

Estas desigualdades se mantienen para cualquier distribución de probabilidad que tiene oblicuidad y kurtosis finitos.

En la **Tabla de Oblicuidad-Kurtosis** , se pueden notar dos familias útiles de distribuciones, las familias beta y gammas.

La Función de Densidad tipo Beta: Puesto que la densidad beta tiene parámetros de forma y de escala, esta describe muchos fenómenos aleatorios que hacen que la variable aleatoria se encuentra entre [0, 1]. Por ejemplo, cuando ambos parámetros son números enteros con variables aleatorias el resultado es la función de probabilidad binomial.

Aplicaciones: Una distribución básica de estadísticos para variables limitadas en ambos lados; por ejemplo x entre [0, 1]. La densidad beta es útil para problemas aplicados y teóricos de muchas áreas. Los ejemplos incluyen la distribución de la proporción de la población localizada en el medio del valor más bajo y más alto de una muestra; la distribución del porcentaje diario de en un proceso de producción; la descripción de etapas transcurridas en la terminación de la tarea (PERT). También existe una relación entre las distribuciones beta y normal. El cálculo convencional es que dado un PERT beta con el valor más alto b, el mas bajo a, y muy probablemente como m, la distribución normal equivalente tiene una media y una moda de (a + 4M + b)/6 y una desviación estándar de (b - a)/6.

Comentarios: Distribuciones uniformes, de triangulo rectángulo, y parabólicas son casos especiales. Para generar beta, cree dos valores aleatorios de una gamma, g_1, g_2. El cociente $g_1/(g_1 + g_2)$ se distribuye como una distribución beta. La distribución beta también se puede pensar como la distribución de X1 dado (X1+ X2), cuando X1 y X2 son variables aleatorias gammas independientes.

La Función de Densidad tipo Gamma: Algunas variables son siempre no negativas. La función de densidad asociada a estas variables aleatorias es modelada acorde a una función de densidad tipo gamma. La función de densidad tipo gamma tiene parámetros de forma y de escala ambos iguales a 1, lo cual resulta en función de densidad exponencial. La Chi-cuadrado es también un caso especial de la función de densidad gamma con parámetros de forma igual a 2.

Aplicaciones: Una distribución básica de estadística para variables limitadas en un lado; por ejemplo x mayor o igual a cero. La densidad gamma da a la distribución el tiempo requerido para que exactamente k exactamente eventos independientes ocurran, suponiendo que los eventos toman lugar a una tasa

constante. Es utilizada con frecuencia en teoría de alineación, confiabilidad, y otros usos industriales. Los ejemplos incluyen distribución de tiempo entre reajuste de instrumentos que necesitan ser reajustados después de k veces utilizados; tiempo entre la reposición de inventarios, tiempo de falla de un sistema con componentes inactivos.

Comentarios: Las distribuciones de Erlangian, exponenciales, y Chi-cuadrado son casos especiales. La binomial negativa es análoga a la distribución gamma con variable aleatorias discretas.

¿Cuál es la distribución del producto de las observaciones de una muestra aleatoria uniforme (0, 1)? Como muchos problemas con productos, esto se transforma en un problema familiar cuando se convierte en un problema de sumas. Si X es uniforme (para simplificar la notación haga U(0,1)), Y =-log(X) es exponencialmente distribuida, tal que el producto de X1, X2..., Xn es la suma de Y1, Y2..., Yn, el cual tiene una distribución gamma (Chi-cuadrado a escala). De esta forma, es una densidad gamma con parámetro de forma n y escala 1.

La Función Normal de Densidad Logarítmica: Permite la representación de una variable aleatoria de la cual su logaritmo sigue una distribución normal. El cociente de dos variables aleatorias logarítmicas normal es también logarítmica normal.

Aplicaciones: Modelo para un proceso creciente de pequeños errores multiplicativos. Apropiado cuando el valor de una variable observada es una proporción aleatoria del valor previamente observado.

Aplicaciones: Los ejemplos incluyen el tamaño de la distribución de un proceso de quiebra; el tamaño de la distribución de la renta, herencias y depósitos bancarios; distribución de fenómenos biológicos; distribución de la vida de algunos tipos de transistores, etc.

La distribución logarítmica normal es extensamente utilizada en situaciones donde los valores son sesgados positivamente (donde la distribución tiene una cola larga hacia la derecha; las distribuciones sesgadas negativamente tienen una cola larga hacia la izquierda; una distribución normal no tiene ninguna oblicuidad). Ejemplos de datos que se “ajustan” a una distribución logarítmica

normal incluyen valuaciones de la seguridad financiera o valuaciones de propiedades inmobiliarias. Analistas financieros han observado que los precios de acciones bursátiles generalmente se muestran sesgados positivamente, en vez de estar normalmente (simétricamente) distribuidos. Los precios de las acciones en la bolsa de valores muestran esta tendencia porque dichos precios no puedes bajar del límite de cero valor, pero pueden aumentar sin límite a cualquier precio. De manera semejante, los costos de salud publica ilustran oblicuidad positiva puesto que los costos unitarios no pueden ser negativos. Por ejemplo, no puede haber costos negativos para un contrato de servicios capitalización. Esta distribución describe exactamente la mayoría de los datos de salud pública..

En el caso donde los datos son logarítmicos normalmente distribuidos, la Media Geométrica describe mejor de los datos que la media. Mientras mas cerca los datos sigan a una distribución logarítmica normal, más cerca estará la media geométrica a la mediana, puesto que la reexpresión logarítmica produce una distribución simétrica.

Ejemplo Numérico y Discusiones

Un ejemplo numérico: Dado el siguiente grupo pequeño de datos (n =4), calcule los estadísticos descriptivos: $x_1 = 1$, $x_2 = 2$, $x_3 = 3$, y $x_4 = 6$.

i	x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^3$	$(x_i - \bar{x})^4$
1	1	-2	4	-8	16
2	2	-1	1	-1	1
3	3	0	0	0	0
4	6	3	9	27	81
Sum	12	0	14	18	98

LA media $\bar{x}$ is 12 / 4 = 3; la varianza es s^2 = 14 / 3 = 4,67; la desviación estándar = $(14/3)^{0.5}$ = 2,16; la oblicuidad es 18 / [3 $(2,16)^3$] = 0,5952, y finalmente, la Kurtosis es = 98 / [3 $(2,16)^4$] = 1,5.

A usted podría interesarle usar el JavaScript de Estadística Descriptiva para comprobar sus cálculos manuales.

Una Pequeña Discusión Acerca de la Estadística Descriptiva:

Las desviaciones con respecto a la media m de una distribución son la base para la mayoría de las pruebas estadísticas que aprenderemos. Puesto que estamos midiendo cuánto se dispersa un sistema de valores con respecto a la media m , estamos midiendo **variabilidad**. Podemos calcular las desviaciones con respecto a la media m y expresarlas como varianza s^2 o como desviación estándar s. **Es muy importante tener un conocimiento firme de este concepto porque será una noción fundamental a través de su curso de estadística**.

Tanto la varianza s^2 y la desviación estándar s miden la variabilidad dentro de una distribución. La desviación estándar s es un número que indica cuánto en promedio cada uno de los valores en la distribución se desvía de la media m (o del centro) de la distribución. Tenga presente que la varianza s^2 mide lo mismo que la desviación estándar s (dispersión de valores en una distribución). Sin embargo, la varianza s^2 corresponde al average al cuadrado de las desviaciones con respecto a la media. Así, la varianza m . Por lo tanto, la varianza s^2 es el cuadrado de la desviación estándar s.

El valor esperado y la varianza del $\bar{x}$ son m y s^2/n, respectivamente.

El valor esperado y la varianza del estadístico S^2 son s^2 y $2s^4$ / (n-1), respectivamente.

$\bar{x}$ y S^2 son los mejores estimadores para m y s^2. Estos son imparciales (usted puede actualizar su estimación); Eficientes (tienen la varianza más pequeña entre otros estimadores); Consistente (incrementos en el tamaño de la muestra proporciona una mejor estimación); y suficiente (no se necesita tener el grupo entero de datos; todo lo que se necesita es Sx_i y Sx_i^2 para las estimaciones). Adicionalmente, observe que la varianza anterior S^2 se justificada solamente en el caso donde la distribución de la población tiende a ser normal, de otra manera se podrían utilizar técnicas de enlace.

En general, se cree que el patrón de la moda, la mediana y la media van de menor a mayor oblicuidad positiva con respecto a los datos, y apenas el patrón opuesto en datos sesgados negativamente. Sin embargo, por ejemplo, en los 23 números siguientes, la media = 2,87 y la mediana = 3, pero los datos están sesgados positivamente:

4, 2, 7, 6, 4, 3, 5, 3, 1, 3, 1, 2, 4, 3, 1, 2, 1, 1, 5, 2, 2, 3, 1

por otro lado, los siguientes 10 números tienen media = mediana = moda = 4, pero los datos están sesgados hacia la izquierda (negativamente):

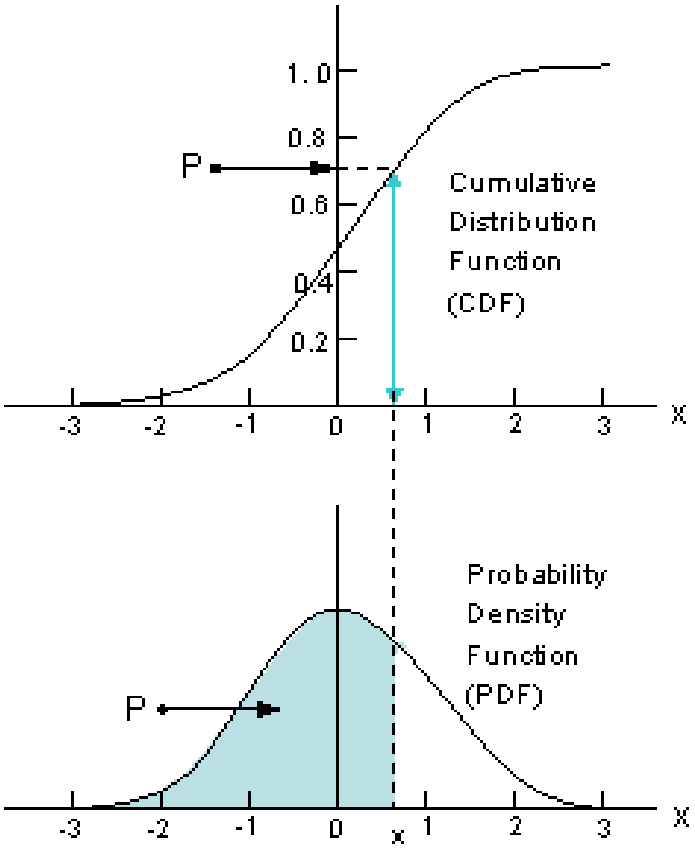
1, 2, 3, 4, 4, 4, 5, 5, 6, 6.

Adicionalmente, note que los software más comercial no calculan correctamente la Oblicuidad y Kurtosis. No existe manera fácil de determinar intervalos de confianza sobre un valor calculado de la oblicuidad o kurtosis de una muestra pequeña a media. Las literaturas dan tablas basadas en métodos asintóticos para sistemas de muestras mayores a 100 y solo para distribuciones normales.

Se podría notar que usando el ejemplo numérico anterior en algunos paquetes estadísticos de computadora tales como SPSS, la oblicuidad y la kurtosis son diferentes a las que hemos calculado. Por ejemplo, los resultados del SPSS para la oblicuidad es 1,190. Sin embargo, para muestras n mas grandes, el resultados es idéntico.

Las Dos Representaciones Estadísticas de la Población

La siguiente figura representa una relación típica entre la función de distribución acumulativa (fda) y la de densidad (para variables aleatorias) contínuas,



Relations Between Two Different Typical Representations of a Population

Todas las características de la población están bien descritas por cualquiera de estas dos funciones. La figura también ilustra sus aplicaciones para determinar la medición del percentil (más bajo) denotado por **P**:

P = P[X £ x] = Probabilidad de que la variable aleatoria X sea menor o igual a un número dado x is less than or equal to a given number x,

entre otras informaciones útiles. Note que la probabilidad P es el área bajo la curva de la función de densidad, mientras que es numéricamente igual a la altura de la curva fdc en el punto x.

Ambas funciones pueden ser estimadas *suavizando la función empírica (observada) acumulativa, y suavizando el histograma* construido de la muestra.

Función de Distribución Empírica (observada) Acumulativa

La función de distribución empírica acumulativa (FDEA), también conocida como **ojiva**, se utiliza para graficar frecuencias acumulativas.

La ojiva es el estimador para la función de distribución acumulativa de la población, la cual contiene todas las características de la población. La distribución empírica es una función de escalonada con la localización aleatoria de los puntos. El tamaño de la cada escalera para cada punto depende de la frecuencia del valor de ese punto, y es igual a la **frecuencia /n** donde n es el tamaño de la muestra. El tamaño de muestra es la suma de todas las frecuencias.

Note que todos los estadísticos cubiertos hasta ahora, pueden ser obtenidos y entendidos más profundamente en papel para graficar usando la Función de Distribución Empírica en Javascript. A usted podría gustarle usar este Javascript para ejecutar ciertas experimentaciones numéricas y tener una comprensión o más profundamente.

Otros modelos de decisión extensamente utilizados, los cuales estas basados en la función de distribución empírica acumulativa (FDEA) como herramienta de medición y procedimiento de decisiones son la Clasificación ABC de Inventarios, Análisis de Inventarios en Periodos Simples (modelo de Newsboy), y el de determinación del Mejor Momento para Remplazar Equipos. Para otras decisiones acerca de inventarios, visite el sitio Web Modelos de Control de Inventario.

Introducción

Modelamiento de un Conjunto de Datos: Las familias de distribuciones paramétricas son ampliamente utilizadas para resumir enormes grupos de datos, para obtener predicciones, determinan la calidad de ajuste, estimar funciones de datos que no son fácil de derivar directamente, o para alcanzar efectos aleatorios manejables. La credibilidad de los resultados obtenidos dependerá de la generalidad de la distribución de las familias empleadas.

Inferencia Inductiva: Esta extensión de nuestro conocimiento proveniente de una muestra particular escogida al azar de una población se llama inferencia inductiva. La función principal de la estadística de negocios es de proveer las técnicas para hacer inferencia inductiva y para medir el grado de incertidumbre de tal inferencia. La incertidumbre es medida en términos de probabilidad y ésa es la razón por la cual necesitamos aprender la lengua de la incertidumbre y su herramienta de medición llamada probabilidad.

En contraste con la inferencia inductiva, las matemáticas normalmente utilizan inferencia deductiva para probar teoremas,

mientras que en ciencia empírica, tal como la estadística, la inferencia inductiva es utilizada para ampliar o encontrar nuevo conocimiento.

Probabilidad, Chance, Oportunidad, y Posibilidad

El concepto de probabilidad ocupa un lugar importante en el proceso de toma de decisión bajo incertidumbre, no importa si el problema es enfrentado en el campo de negocios, del gobierno, en las ciencias sociales, o simplemente en nuestras vidas diarias. En muy pocas situaciones de toma de decisión la información perfecta esta disponible --todos los factores u hechos necesarios--. La mayoría de las decisiones se toman encarando la incertidumbre. La probabilidad entra en el proceso desempeñando el papel de substituto para la certeza, substituto para el completo conocimiento.

La Probabilidad es especialmente significativa en el área de la inferencia estadística. Aquí la preocupación principal de los estadísticos es obtener conclusiones o hacer inferencias provenientes de experimentos que implican incertidumbre. El concepto de la probabilidad permite al estadístico generalizar de la información obtenida de lo sabido (muestra) a lo desconocido (población), y agregar un alto grado de confianza en estas generalizaciones. Por lo tanto, la probabilidad es una de las herramientas más importantes de la inferencia estadística.

La probabilidad tiene un significado técnico exacto (bueno, de hecho tiene varios, y todavía existen discusiones de cual término debería ser utilizado). Sin embargo, para la mayoría de los acontecimientos para los cuales la probabilidad se calcula fácilmente; por ejemplo, la probabilidad de tirar un dado y conseguir cuatro [::], casi todos están de acuerdo en que el valor es (1/6), y no es una interpretación filosófica. Una probabilidad es siempre un número entre 0 y 1. Cero no significa “exactamente” lo mismo que imposibilidad. Es posible que “si” una moneda fuera

lanzada muchas veces, nunca mostrara la “cruz", pero la probabilidad de que se obtengan “caras” infinitamente es 0. Estos conceptos no significan “exactamente” lo mismo, pero son bastante cercanos.

La palabra “chance” o “chances” son frecuentemente utilizadas como sinónimos aproximados de “probabilidad”, ya sea por variedad o por ahorrar sílabas. Sería mejor si dejamos la palabra “chance” para uso informal, y la palabra “probabilidad” para definir lo que significa realmente. En otras oportunidades se podrían encontrar los términos “posibilidad” y “ocasión”, sin embargo,

estos términos se utilizan ocasionalmente como sinónimos para lo "probable" y la "probabilidad".

Oportunidad es un concepto probabilística relacionado con la probabilidad. Es el cociente de la probabilidad (p) de un evento con respecto a la probabilidad (1-p) de que no sucede: p/(1-p). Se puede expresar como cociente, o como número entero como en los “Oportunidad” de 1 a 5 en el ejemplo anterior del dado, pero para fines técnicos la división se pueden realizar para alcanzar un número real positivo (aquí 0,2). Oportunidad son el cociente de no-ocurrencia ningún de un evento a un evento. Si el cociente de ocurrencia de una enfermedad es 0,1 (10%), el cociente de no-ocurrencia es 0,9 y por lo tanto sus probabilidades son 9:1.

Otra manera de comparar probabilidades y Oportunidad es utilizando el “pensamiento parte-entera” con un binario (dicotómico) partido en un grupo. Una probabilidad es un cociente de una parte a un conjunto; por ejemplo, el cociente entre [aquellos que sobrevivieron 5 años después de haber sido diagnosticados con una enfermedad] al conjunto de [todos los que fueron diagnosticadas con la enfermedad]. Oportunidad son normalmente un cociente de una parte a otra parte, por ejemplo, las Oportunidad de los que estaban en contra de morir son el cociente de la parte que tuvo éxito [los que sobrevivieron 5 años después de ser diagnosticado con la enfermedad] a la parte que “falló” [los que no sobrevivieron 5 años después de ser diagnosticado con la enfermedad].

Aparte de su valor en apuestas, las Oportunidad permiten especificar una probabilidad pequeña (cerca de cero) o una probabilidad grande (cerca de uno) usando números enteros grandes (1.000 a 1 o un millón a uno). Las Oportunidad magnifican probabilidades pequeñas (o probabilidades grandes) con el objetivo de hacer las diferencias relativas visibles. Considere dos probabilidades: 0,01 y 0,005. Ambas son pequeñas. Un observador inexperto podría no darse cuenta que una es el doble de la otra. Pero si esta se encuentra expresada como Oportunidad (99 a 1 contra 199 a 1) podrían ser más fácil de comparar las dos situaciones centrándose en los números

enteros grandes (199 contra 99) en vez de los cocientes pequeños o fracciones.

Como Asignar Probabilidades

La probabilidad es una herramienta para medir la posibilidad de ocurrencia de un evento. Existen 5 aproximaciones para asignar probabilidad: Aproximación Clásica, Aproximación de la Frecuencia Relativa, Aproximación Subjetiva, Anclaje, y la técnica de Delphi:

1. Aproximación Clásica : La probabilidad clásica se basa en la condición de que los resultados de un experimento son igualmente probables suceder. La probabilidad clásica utiliza la idea de que la carencia del conocimiento implica que todas las posibilidades son igualmente probables. La probabilidad clásica es aplicada cuando los acontecimientos tienen la misma oportunidad de ocurrencia (llamado eventos igualmente probables), y los grupos de eventos son mutuamente excluyentes y colectivamente exhaustivo. La probabilidad clásica se define como:

   P (X) = Número de resultados favorables / Número total de posibles resultados.

2. Aproximación de la Frecuencia Relativa: La probabilidad relativa se basa datos históricos o experimentales acumulados. Probabilidad basada en frecuencia se define como:

   P (X) = Número de veces que un evento ocurre / Número total de oportunidades de ocurrencia del evento.

   Note que la probabilidad relativa se basa en la idea de que lo que ha ocurrido en el pasado se mantendra.

3. Aproximación Subjetiva: La probabilidad subjetiva se basa en juicios y experiencias personales. Por ejemplo, los médicos algunas veces asignan probabilidad subjetiva al periodo de vida de una persona diagnosticada con cáncer.
4. Anclaje: Es la practica de asignar un valor obtenido de una experiencia previa y ajustando el valor en consideración a las circunstancias y expectativas del momento.
5. Técnica de Delphi: Consiste en una serie de cuestionarios. Cada serie es un “círculo”. Las respuestas del primer “círculo” se recolectan y se convierten en la base para las preguntas y realimento del segundo “circulo”. El proceso generalmente se repite para un número predeterminado de “círculos” o hasta que las respuestas se ajustan al patrón observado. Este proceso permite que la opinión de los expertos sea circulada a todos los miembros del grupo y elimine el efecto de distraer la opinión de la mayoría.

El análisis de Delphi se utiliza en el proceso de toma de decisión, particularmente en pronósticos. Varios “expertos” se sientan a discutir e intentan comprometerse en algo sobre el cual no pueden convenir.

Leyes Generales de la Probabilidad

1. Ley general de la adición: Cuando dos o más eventos ocurren al mismo tiempo, y los eventos **no son** mutuamente excluyentes, se tiene que:

   P (X ó Y) = P (X) + P (Y) - P (X e Y)

   Note que, la ecuación P (X ó Y) = P (X) + P (Y) - P (X e Y), contiene eventos especiales: un evento (X e Y), el cual es la intersección del grupo de eventos X e Y, y otro evento (X o Y), el cual es la unión de los grupos X e Y. A pesar de que esta formula es bastante sencilla, dice relativamente poco acerca de cómo un evento X in fluencia al evento Y, y viceversa. Si P (X e Y) es 0, indica que los eventos X e Y no se interceptan (son mutuamente excluyentes), por lo tanto tenemos P (X ó Y) = P (X) + P (Y). Por otro lado, si P (X e Y) es 0, existe una intercepción entre los eventos X e Y. Generalmente, esto podría ser una iteración física entre ellos. Esto hace que la relación P (X ó Y) = P (X) + P (Y) - P (X e Y) sea no lineal porque el termino P (X e Y) es sustraído el cual influencia al resultado.

   Esta ley es también conocida como la Formula de Inclusión-Exclusión. Esta puede ser extendida para mas de dos eventos. Por ejemplo, para A, B, y C, esta se convierte en:

   P (A ó B ó C) = P(A) + P(B) + P(C) - P(A y B) - P(A y C) - P(B y C) + P(A y B y C)

2. Ley Especial de Adición: Cuando dos o mas eventos ocurren al mismo tiempo, y los eventos **son** mutuamente excluyentes, se obtiene:

   P(X ó Y) = P(X) + P(Y)

3. Ley General de la Multiplicación: Cuando dos o mas eventos ocurren al mismo tiempo, y los **son** dependientes, la Ley General de la Multiplicación es usada para obtener la probabilidad conjunta:

   P(X e Y) = P(Y) ´ P(X | Y), ´ P(X|Y),

   de donde P(X | Y) es la probabilidad condicional.

4. Ley Multiplicativa: Cuando dos o mas eventos ocurren al mismo tiempo, y los eventos **son** independientes, la regla especial de la ley multiplicativa es usada para obtener la probabilidad:

   P(X e Y) = P(X) ´ P(Y)´ P(Y)

5. Ley de la Probabilidad Condicional: Una probabilidad condicional es denotada por P(X|Y). Esta frase se lee: La probabilidad de que X ocurra conociendo como **dada que** la probabilidad de Y haya ocurrido.

   Probabilidades condicionales se basan en el conocimiento de una de las variables. La probabilidad condicional de un evento, tal que X ocurra sujeto a que el evento Y ha ocurrido, es expresada como:

   P(X|Y) = P(X|Y) = P(X e Y) ¸ P(Y),

   Provisto de que P(Y) no es cero. Note que cuando se una la ley de probabilidad condicional, siempre se divide la probabilidad conjunta entre la probabilidad de un evento después de la palabra dada. Por lo tanto, para obtener P(X dada Y), se divide la probabilidad conjunta de X e Y entre la probabilidad incondicional de Y. En otras palabras, la ecuación anterior es usada para encontrar la probabilidad condicional para dos eventos **dependientes** cualquiera.

   La versión mas simple del teorema de Bayes es:

   P(X|Y) = P(Y|X) ´ P(X) ¸ P(Y)

   Si dos eventos, tales como X e Y, son **independientes** entonces:

   P(X|Y) = P(X),

   y

   P(Y|X) = P(Y)

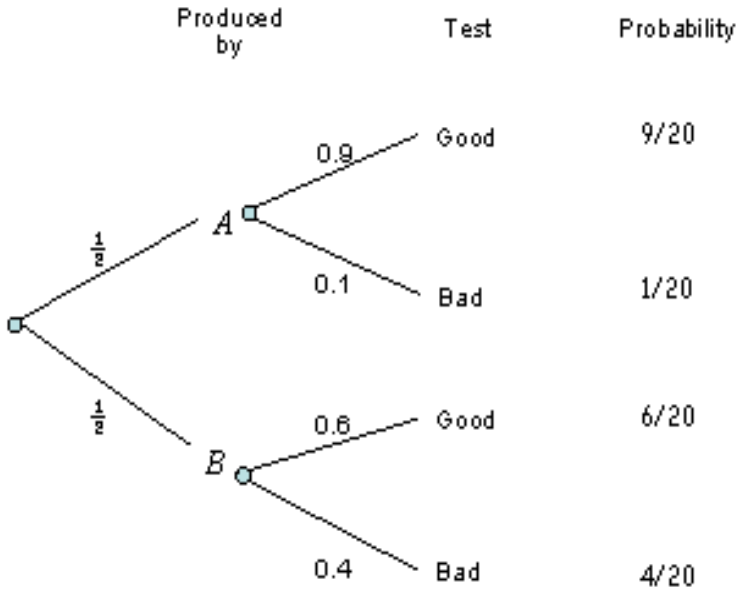
6. **La ley de Bayes:**

   P(X|Y) = [P(X) ´ P(Y|X)] ¸ [P(X) ´P(Y|X) + P(no X) ´ P(Y| no X)]

   La ley de Bayes proporciona una probabilidad posterior [por ejemplo, P(X|Y)] agudizando la probabilidad anterior [P(X)] por la disponibilidad de mejorar y relevar información en términos probabilísticos.

Una Aplicación: Suponga que dos maquinas, A y B, producen partes idénticas. La maquina A tiene una probabilidad de 0,1 de producción defectuosa cada vez que se utiliza, mientras que la maquina B tiene una probabilidad de 0,4 de producción defectuosa cada vez que se usa. Cada maquina produce una parte a la vez. Una de estas partes es selecciona al azar, probada, y se encuentra que es defectuosa. ¿Cuál es la probabilidad de que esa parte fue producida por la maquina B?

Probabilidad de Diagramas de Árbol: representa eventos o secuencias de eventos como rama de árboles. El Diagrama de árbol es una visualización útil de probabilidades condicionales:



Conditional Probabilities

Las probabilidades al final de cada rama son las probabilidades de que eventos dirigidos al final de cada rama ocurrirán simultáneamente. El diagrama de árbol anterior indica que la probabilidad de las partes probadas como buenas es 9/20 + 6/20 = 3/4, por lo tanto, la probabilidad de partes defectuosas es 1/4. esto significa que la P(sea hecha por B | esta es defectuosa) = (4/20) / (1/4) = 4/5.

Ahora, usando la Ley de Bayes podemos obtener información útil, como por ejemplo:

P(esta es defectuosa | hecha por B) = 1/4(4/5) / [1/4(4/5) + 3/4(2/5)] = 2/5.

Equivalentemente, usando la probabilidad condicional anterior, se obtiene que:

P(esta es defectuosa | sea hecha por B) = P(esta es defectuosa y sea hecha por B)/P(hecha por B) = (4/20)/(1/2) = 2/5.

A usted le gustaría utilizar la Probabilidad Revisada de Bayes en JavaScript.

Mutuamente Excluyente contra Eventos Independientes

Mutuamente Excluyente (ME): Los eventos A y B son mutuamente excluyentes si los dos no pueden ocurrir al mismo tiempo. Esto es P[A y B] =0.

Independencia: Los eventos A y B son independiente si, cuando se tiene la información de que B ha ocurrido y esto no altera la probabilidad de que A ocurra. Esto es P[A dado B] = P[A].

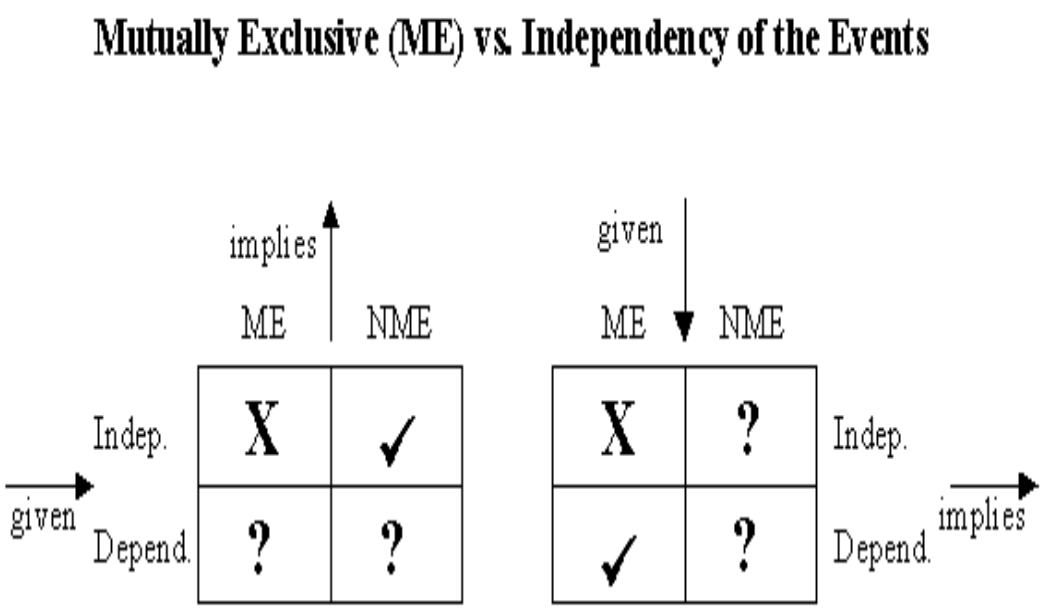
Si dos eventos son ME, ellos también son dependientes: P(A dado B)= P[A y B] ¸ P[B], y porque P[A y B] =0 (por ME), entonces P[A dado B] = 0. Parecidamente,

Si dos eventos son independientes implica que también son ME.

Si dos eventos son dependientes, implica que ellos podrían o no ser ME.

Si dos eventos no son ME, implica que ellos podrían o no ser independientes.

La siguiente figura muestra todas las posibilidades. Las notaciones usadas en esta tabla son las siguientes: **X** significa ausencia de implicación, signo de interrogación **?** significa que podría o no implicar, mientras que la **marca de chequeo** significa que si implica.



Note que la independencia probabilística y independencia mutual para un grupo de eventos $A_1,..., A_n$ son dos nociones diferentes.

¿Qué es tan Importante de la Distribución Normal?

El término “normal” posiblemente se presentó debido a los varios intentos para establecer esta distribución como la ley subyacente que

gobierna todas las variables continuas. Estos intentos se basaron en premisas falsas y por lo tanto en fracasos. No obstante, la distribución normal ocupa un lugar preeminente en el campo de la probabilidad. Además de retratar las distribuciones de muchos tipos de fenómenos naturales y físicos (tales como la altura del hombre, los diámetros de piezas hechas por máquinas, etc.), también sirve como una aproximación conveniente de muchas otras distribuciones que sean menos manejables. Lo más importante, esta distribución describe la manera en la cual ciertos estimadores de características de la población varían de muestra a la muestra. De forma tal, que sirven como fundamento de inferencia estadística de muestras aleatorias escogidas de una población.

La curva de Distribución Normal (llamada también Gaussiana), la cual tiene un aspecto acampanado (algunas veces es referida como “curvas acampanadas”) son muy importantes en el análisis estadístico. En cualquier distribución normal sus observaciones se distribuyen simétricamente alrededor de la media, el 68% de todos los valores bajo la curva se encuentran dentro de una desviación estándar con respecto a la media, y el 95% dentro de dos desviaciones estándar.

Existen muchas razones por la cual proviene su popularidad. Las siguientes, son las razones más importantes de su:

7. Una razón por la cual la distribución normal es importante es que una amplia variedad de variables aleatorias de *ocurrencia natural* tales como la altura y el peso de todas las criaturas, se encuentran distribuidas uniformemente alrededor de un valor central, de un promedio, o de una norma (de aquí el nombre de distribución normal). Aunque las distribuciones son solo aproximadamente normales, están generalmente bastante cerca de su forma.

   Siempre que existan demasiados factores que influencian el resultado de un resultado aleatorio, se considera a la distribución subyacente como aproximadamente normal. Por ejemplo, la altura de un árbol es determinada por la "suma" de los factores tales como la lluvia, la calidad del suelo, el sol, las enfermedades, etc.

   Tal y como Francis Galton escribió en 1889, "siempre que una muestra grande de elementos caóticos se tomen y se arreglen de acuerdo al orden de sus magnitudes, la más insospechada y hermosa forma de regularidad demuestra haber estado latiendo en cada uno de ellos."

8. Casi todas las **tablas de estadísticas están limitadas** al tamaño de sus parámetros. Sin embargo, cuando estos parámetros son suficientemente grandes se puede utilizar la distribución normal para calcular los valores críticos para estas tablas. Por ejemplo, la

   distribución F-estadística se relaciona con la Z-estadística normal estándar de la siguiente forma: $F = z^2$, de donde F tiene ($gl_1 = 1$, y gl_2 es el valor mas grande de la tabla F). Si requiere mas información, visite las Relaciones entre Distribuciones Comunes.

   Aproximación de la Binomial: Por ejemplo, la distribución normal proporciona una buena aproximación de la binomial cuando n es grande y p está cerca del 1/2. Incluso si n es pequeña y p no está extremadamente cerca de 0 o a 1, la aproximación es adecuada. De hecho, la aproximación normal de la binomial será satisfactoria para la mayoría de los propósitos si se cumple que $np > 5$ y $nq > 5$.

   Esta es la forma que la aproximación se realiza. Primero, fije $m = np$ y $s^2 = npq$. Para tener en cuenta el hecho que el binomial es una distribución discreta, utilizamos unfactor de corrección de continuidad de la unidad de 1/2 agregado o restada de X considerando que el valor discreto ($x = a$) debe corresponder a una escala continua de $(a - 1/2) < x < (a + 1/2)$. Luego se calcula el valor de la variable normal estándar por:

$z = [(a - 1/2) - m]/s$ ó $z = [(a + 1/2) - m]/s$

Desde este punto, se podría usar la tabla normal estándar para los valores numéricos.

Una Aplicación: La probabilidad de que un artículo defectuoso proveniente de cierta planta fabricación es $p = 0,25$. Una muestra de 400 artículos se selecciona de una porción grande de estos artículos. ¿Cuál es la probabilidad de que 90 artículos o menos sean defectuosos?

9. Si la media y la desviación estándar son conocidas, es fácil convertir hacia adelante y hacia atrás los **valores brutas a percentiles**.
10. Se ha probado que la distribución subyacente es normal si y solo si la media muestral es independiente de la varianza de la muestra, **esto caracteriza a la distribución normal.** Por lo tanto muchas **transformaciones** efectivas pueden ser aplicadas para convertir casi cualquier distribución a forma normal.
11. La razón más importante de popularidad la distribución normal es el **Teorema del Límite Central (TLC)**. La distribución de los promedios de la muestra de una gran cantidad de variables aleatorias será aproximadamente normal **sin importar** las distribuciones de las variables aleatorias individuales.
12. **La Distribución de Muestreo** para poblaciones normales proporciona más información que cualquier otra distribución. Por ejemplo, los siguientes errores estándar (es decir, que tienen la misma unidad que tienen los datos) son fácilmente disponibles:
    - Error Estándar de la Media = $(p/2n)^{½}S$.
    - Error Estándar de la Desviación Estándar = $S/(2n)^{½}$. Por lo tanto, la prueba estadística para la hipótesis nula $s = s_0$, es $Z = (2n)^{½} (S - s_0)/s_0$.
    - Error Estándar de la Varianza = $S^2[(2/(n-1)]^{½}$.
    - Error Estándar Intercuartíl de Mitad de Rango (Q) = $1,166Q/n^{½}$
    - Standard Error of the Skewness = $(6/n)^{½}$.
    - Error Estándar de la Oblicuidad = $(6/n)^{½}$

      Note que la oblicuidad en distribuciones de muestreo de la media, desaparecen rápidamente cuando n se hace mas.

    - Error Estándar de Kurtosis = $(24/n)^{½}$ = 2 veces el error estándar de la oblicuidad.
    - Error Estándar de la Correlación (r) = $[(1 - r^2)/(n-1)]^{½}$.

Por otra parte,

Desviación Cuartíl» 2S/3, y, Media Absoluta de desviación » 4S/5.

13. La otro razón del porque las distribuciones normales son tan importantes es que la other reason the normal distributions are so important is that the condición de normalidad es requerida por casi todas las clases de **pruebas estadísticas** paramétricas. El teorema del límite central es una herramienta útil cuando se está trabajando con una población de distribución desconocida. A menudo, usted podría analizar la media (o la suma) de una muestra de tamaño n. Por ejemplo en vez de analizar los pesos de elementos individuales se podría analizar el conjunto de tamaño n, es decir, cada paquete que contiene n elementos.

¿Qué es una Distribución de Muestreo?

Una distribución de muestreo describe las probabilidades asociadas a un estadístico cuando una muestra aleatoria es dibujada de la población entera.

La distribución de muestreo es la densidad (para un estadístico continua, tal como una media estimada), o función de probabilidad (para estadístico discreto, tal como una proporción estimada).

La derivación de la distribución de muestreo es el primer paso para calcular un intervalo de confianza o para realizar una prueba de hipótesis a un parámetro.

Ejemplo: Suponga que x1,.....,xn son valores de una muestra simple escogida al azar de una población normalmente distribuida con el valor esperado: m y varianza conocida s^2. Por lo tanto, la media muestral se distribuye normalmente con valor esperado m y varianza s^2/n.

La idea principal de la inferencia estadística es tomar una muestra escogida aleatoria de una población particular y después utilizar la información de la muestra para hacer inferencias sobre las características particulares de la población tales como la media m (medida de la tendencia central), la desviación estándar s (medida de dispersión, de dispersión), o de la proporción de las unidades en la población que tienen cierta característica. El muestreo ahorra el dinero, tiempo, y esfuerzo. Además, una muestra puede proporcionar, en ciertos casos, tanto o más exactitud que el correspondiente estudio que procure investigar a una población entera. La recolección minuciosa de datos de una muestra proporcionará a menudo mucho mejor información que un estudio menos minucioso y que intente mirar todo.

A menudo, se debe estudiar también el comportamiento de la media de los valores de la muestra tomados de diferentes poblaciones especificas; es decir, para propósitos comparativos.

Porque una muestra examina solamente a parte de una población, la media muestral no es exactamente igual a la media poblacional m correspondiente. Por lo tanto, una consideración importante para el planeamiento e interpretación de los resultados del muestreo es el grado en el cual las estimaciones de la muestra, tales como la medio muestral, convendrán con la característica de la población correspondiente.

En la práctica, generalmente solo una muestra es tomada. En algunos casos una “muestra piloto” se utiliza para probar los mecanismos de recolección de datos y para conseguir la información preliminar para planear el esquema principal de la muestreo principal. Sin embargo, para los propósitos de entender el grado al cual la media muestral convendrá con la correspondiente media poblacional m , sería útil considerar qué sucedería si 10, o 50, o 100 estudios separados del muestreo, del mismo tipo, fueran conducidos. ¿Cuan consistente serian los resultados a través de estos diversos estudios? Si podemos ver que los resultados de cada uno de las muestras son casi iguales (y honestamente corregibles!), entonces, tendríamos confianza de que la simple muestra será utilizada realmente. Por otra parte, viendo que las respuestas de las muestras repetidas son demasiado variables para la exactitud necesaria, sugeriría que un diverso plan de muestreo (quizás con un tamaño de muestra más grande) debería ser utilizado.

Una distribución de muestreo es utilizada para describir la distribución de los resultados que uno observaría de la réplica de un plan de muestreo particular.

Sepa que las estimaciones calculadas a partir de una muestra serán diferentes de las estimaciones que se obtendrían de los cálculos de otra muestra.

Entienda que las estimaciones se esperan que difieran de las características de la población (parámetros), las cuales son las que se intenta estimar, pero que las propiedades de las distribuciones de muestreo nos permiten que calculemos, basadas en probabilidad, y como ellos se diferenciarán.

Entienda que diferentes estadísticos tienen diferentes distribuciones de muestreo, con forma de la distribución dependiendo de (a) del estadístico específico, (b) el tamaño de la muestra, y (c) la distribución familiar.

Entienda la relación entre el tamaño de la muestra y la distribución de las estimaciones de la muestra.

Entienda que en muestras grandes, muchas distribuciones de muestreo pueden ser aproximadas a una distribución normal.

Vea que en muestras grandes, muchas distribuciones del muestreo se pueden aproximar con una distribución normal.

Distribución de Muestreo de la Media y la Varianza para Poblaciones Normales: Dado una variable aleatoria X que se distribuye normalmente con media m y desviación estándar s, para una muestra escogida al azar del tamaño n:

- La distribución de muestreo de [x̄ - m] ´ $n^{½}$ ¸ s, es la distribución normal estándar.
- La distribución de muestreo de [x̄ - m] ´ $n^{½}$ ¸ S, es una distribución T con parámetro gl = n-1.
- La distribución de muestreo de [S^2(n-1) ¸ s^2], es un c^2 es una distribución con parámetro gl = n-1.
- Para dos muestras independientes, la distribución de muestreo de [S_1^2 / S_2^2], la distribución de muestreo de $gl_1 = n_1$-1, y $gl_2 = n_2$-1.

¿Que es el Teorema del Límite Central?

El teorema de límite central (TLC) es un "límite" que es "central" para prácticas estadísticas. Para propósitos prácticos, la idea principal del TLC es que el promedio (centro de datos) de una muestra de observaciones dibujadas de alguna población está distribuido aproximadamente como una distribución normal si se resuelven ciertas condiciones. En estadística teórica hay varias versiones del teorema de límite central dependiendo de cómo se especifican estas condiciones. Éstos se refieren a los tipos de condiciones hechas sobre la distribución de la población parientes (población de la cual la muestra es dibujada) y del procedimiento actual de muestreo.

Una de las versiones más simples del teorema de límite central indicada por muchos libros de textos es: si tomamos una muestra aleatoria de tamaño (n) de la población entera, entonces, el medio de la muestra el cual es una variable aleatoria se definida por:

S x_i / n,

tiene un histograma que converge a la forma de una distribución normal si n es suficientemente grande. Equivalente, la distribución de la media muestral se acerca a la distribución normal mientras que el tamaño de muestra aumenta.

Algunos estudiantes que tienen dificultad al reconciliar de su propia comprensión del teorema de límite central con algunas de las declaraciones de los libros de textos. Algunos libros de textos no profundizan en los conceptos de **independencia, las muestras aleatorias de tamaño fijo** n (digamos más de 30).

La forma de las distribuciones de muestreo para la media – se convierte cada vez más normal a medida que el tamaño de la muestra n se hace más grande. El incremento del tamaño de la muestra es lo que hace que

la distribución se haga mas normal y que la condición de independencia proporcione la contracción Ön de la desviación estándar.

TLC para los datos de la proporción, tales como los binarios 0, 1, otra vez la distribución de muestreo -- mientras que se hace cada vez más grande “forma acampanado” se mantiene limitada al dominio [0, 1]. Este dominio representa una diferencia dramática con respecto a una distribución normal, el cual tiene un dominio ilimitado. Sin embargo, cuando n aumenta sin límite, la “anchura” de la campana llega a ser muy pequeña de modo que el TLC “todavía trabaja”.

Existen aplicaciones del teorema de límite central problemas prácticos en estadística inferencia, sin embargo, estamos más interesados en cómo la distribución aproximada de la media muestral sigue de cerca una distribución normal para tamaños de muestras finitas, que en la distribución limitadora en sí. El acuerdo suficientemente cercano con una distribución normal nos permite utilizar la teoría normal para hacer inferencias acerca de los parámetros de la población (tales como la media) usando la media muestral, independiente de la forma real de la población original.

Puede ser demostrado que, si la población original tiene media m y una desviación estándar s, finita, entonces la media de la distribución de la muestra tiene la misma m pero con desviación estándar s mas pequeña, la cuál es dividida por $n^{½}$.

Usted ahora sabe que, independientemente de como sea la población original, la variable estandardizada Z = (X - m)/s tendrá una distribución con media m = 0, y desviación estándar s =1 bajo un muestreo aleatorio. Por otra parte, **si** la población original es normal, Z se distribuye exactamente como la normal estándar. El teorema del límite central indica el resultado notable de que, igualmente cuando la población original no sea normal, la variable estandardizada es aproximadamente normal si el tamaño de la muestra es suficientemente bastante. Generalmente no es posible indicar cuales son las condiciones bajo las cuales la aproximación dada por el teorema del límite central funcione y qué tamaños de muestra son necesarios para que la aproximación llegue a ser bastante buena. Como pauta general, los estadísticos han utilizado la regla de que, si la distribución original es simétrica y de colas relativamente cortas, la media muestral se aproxima más de cerca de la normalidad para muestras **más** pequeñas a que si la población original es sesgada o de colas largas.

Bajo ciertas condiciones, en muestras grandes, la distribución de muestreo de la media muestral se puede aproximar a una distribución normal. El tamaño de muestra necesitada para que la aproximación sea adecuada depende en gran medida de la forma de la distribución original. La simetría (o carencia de eso) es particularmente importante.

Para una distribución original simétrica, igualmente si difiere a la forma de una distribución normal, una aproximación adecuada puede ser obtenido con muestras pequeñas (por ejemplo, 15 o más para la distribución uniforme). Para distribuciones simétricas, de distribuciones originales de colas cortas, la media muestral se aproxima más a la normal para tamaños de muestra más pequeños **que si** la población original es sesgada y de colas largas. En algunos casos extremos (como la binomial) tamaños de muestra que se excedan las pautas típicas (sobre 30) **son necesarias** para una aproximación adecuada. Para algunas distribuciones sin primer y segundo momentos (por ejemplo, una es conocida como la distribución de Cauchy el teorema del límite central no se sostiene.

Para algunas distribuciones, las muestras extremadamente grandes (imprácticas) serían requeridas para acercarse a una distribución normal. En la fabricación, por ejemplo, cuando los defectos ocurren a una tasa de menos de 100 unidades por millón, usando, una distribución beta proporcionaría un Intervalo de Confianza (IC) de defectos totales en la poblaciónsome distributions.

¿Qué son los Grados de Libertad (gl)?

Recuerde que al estimar la varianza de la población, utilizamos (n-1) en vez de n en el denominador. El factor (n-1) se llama los “grados de libertad.”

Estimación de la Varianza Poblacional: Varianza en una población se define como el promedio de desviaciones elevadas al cuadrado con respecto a la media de la población. Si dibujamos una muestra escogida al azar de n casos provenientes de una población de donde se sabe la media, podemos estimar la varianza de la población de una manera intuitiva. Sumamos las desviaciones de los valores con respecto a la media de la población y dividimos esa suma por n. Esta estimación se basa en n piezas independientes de información, y tienen n grados de libertad. Cada uno de las n observaciones, incluyendo la ultima son disipadas (es decir, “libres” de variar).

Cuando no sabemos la media de la población, podemos todavía estimar la varianza poblacional; pero, ahora calculamos desviaciones alrededor de la media muestral. Esto introduce un contraste importante porque la suma de las desviaciones alrededor de la media de la muestra es conocida como cero. Si sabemos el valor para las primeras (n-1) desviaciones, la ultima también es conocida. Existen solamente n-1 elementos independientes de información en esta estimación de la varianza.

Si usted estudia un sistema con n parámetros x_i, i =1..., n, usted puede representarlo en un espacio de dimensión n. Cualquier punto de este

espacio representará un estado potencial de su sistema. Si sus n parámetros pudieran variar independientemente, entonces su sistema sería completamente descrito en un hiper volumen de n-dimensiones (para n mayor a 3). Ahora, imagine que tiene una contracción entre parámetros (una ecuación con sus n parámetros), su sistema sería descrito por una hiper superficie de n-1 dimensiones (para n mayor a 3). Por ejemplo, en un espacio tridimensional, una relación lineal significa un plano que sea de 2 dimensiones.

En estadística, sus n parámetros son sus n datos. Para evaluar la varianza, usted primero necesita inferir la media m . De tal forma que cuando usted evalúa la varianza, usted tiene una limitación en su sistema (que es la expresión de la media), y esta se mantiene solo (n-1) grados de libertad a su sistema.

Por lo tanto, dividimos la suma de desviaciones al cuadrado por n-1, en vez de por n, cuando tenemos datos de la muestra. En promedio, las desviaciones alrededor de la media muestral son más pequeñas que desviaciones alrededor de la media poblacional. Esto es porque nuestra media muestral se encuentra siempre en el centro de nuestros valores muestrales; de hecho, la mínima suma posible de las desviaciones al cuadrado para cualquier muestra de números está alrededor de la media de esa muestra de números. Por lo tanto, si sumamos las desviaciones al cuadrado de la media muestral y la dividimos por n, tenemos una subestimación de la varianza en la población (que se basa en desviaciones alrededor de la media poblacional).

Si dividimos la suma de desviaciones al cuadrado por n-1 en vez de n, nuestra estimación es un poco mas larga, y se puede demostrar que este ajuste nos da una estimación imparcial de la varianza poblacional. Sin embargo, para n grande, por ejemplo, sobre 30, no hace existe mucha diferencia si dividimos por n o n-1.

Grados de Libertad en ANOVA: Usted también verá la palabra clave “grados de libertad” que aparece en la Tablas Análisis de las de Varianza (ANOVA). Si le preguntara por 4 números, pero que no me dijera cuáles son, el promedio podría ser cualquier cosa. Tengo 4 grados de libertad en el conjunto de datos. Si le dijera 3 de esos números, y a promedio, usted puede descifrar el cuarto número. El conjunto de datos, dado el promedio, tiene 3 grados de libertad. Si le digo el promedio y la desviación estándar de los números, le habría dado 2 piezas de información, y he reducido los grados de libertad de 4 a 2. Usted necesita conocer solamente 2 de los valores de los números para conjeturar los otros 2.

En una tabla ANOVA, el grado de la libertad (gl) es el divisor en (suma de desviaciones al cuadrado)/gl, el cual dará lugar a una estimación imparcial de la varianza de una población.

En general, un grado de libertad gl = N - k, donde N es el tamaño de la muestra, y k es un número pequeño, igual al número de "requerimientos", el número de "pedazos de información" ya "usada". Así como veremos en la sección de ANOVA, los grados de libertad son cantidades aditivas; cantidades totales de ellos pueden ser "particionados" en varios componentes. Por ejemplo, suponga que tenemos una muestra de tamaño 13 y calculamos su media, y desviaciones con respecto a la media; solamente 12 de las desviaciones son libres de variar. Una vez que se hayan encontrado 12 de las desviaciones, la decimotercera es determinada.

En situaciones de doble correlación o de regresión, k = 2. El cálculo de las medias de la muestra de cada variable "utiliza" dos piezas de información, dejando N - 2 piezas de la información independientes.

En un análisis de variación unidireccional (ANOVA) con g grupos, existen tres maneras de usar los datos para estimar la varianza de la población. Si todos los datos son reunidos, el SST/(n-1) convencional proporcionaría una estimación de la varianza de la población.

Si se consideran a los grupos del tratamiento por separado, las medias de la muestra se pueden también considerar como las estimaciones de la media de la población, y por lo tanto SSb/(g - 1) se puede utilizar como una estimación. La varianza ("dentro-grupo", "error") restante se puede estimar de SSw/(n - g). Este ejemplo demuestra el repartimiento de los gl:
gl total = n - 1 = gl (entre) + .(contenidos) = (g - 1) + (n - g).

Por lo tanto, una definición simple de trabajo de gl es el tamaño de la muestra menos el número de los parámetros estimados. Una respuesta más completa tendría que explicar porqué existen las situaciones en las cuales los grados de libertad no son un número entero. Después de haber dicho todo esto, la mejor explicación, es matemáticamente por la cual **usamos gl es para obtener una estimación imparcial**.

En resumen, el concepto de grados de libertad se utiliza para los siguiente dos diversos propósitos:

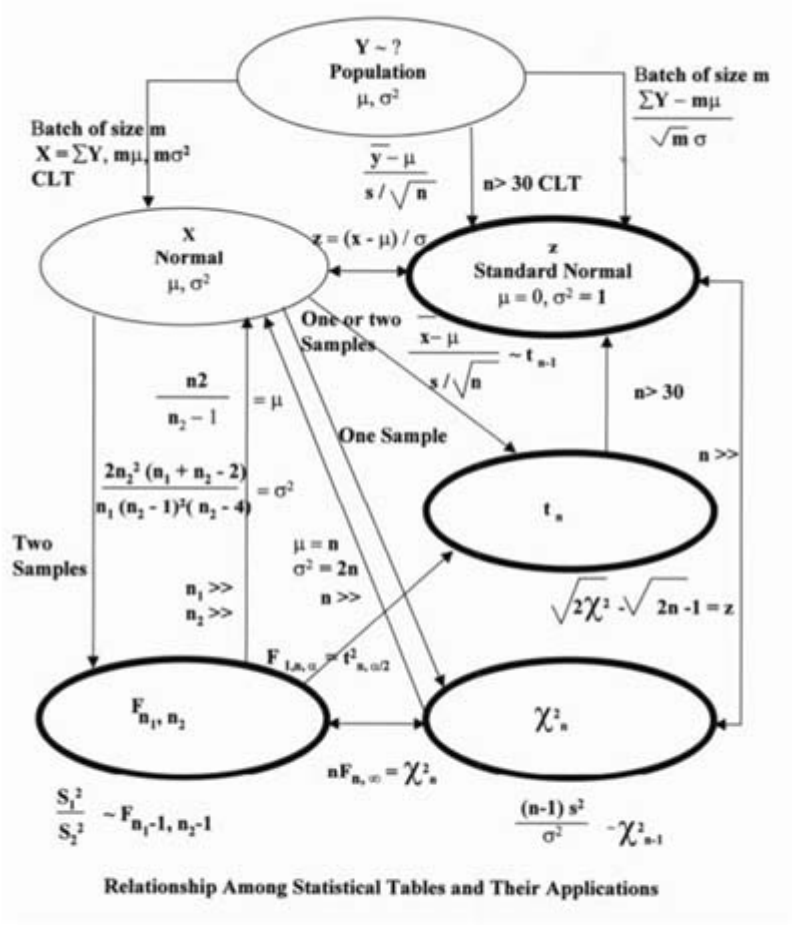
- Parámetro(s) de ciertas distribuciones, tales como F y distribución t, se llama grados de libertad.
- Lo más importantemente, los grados de libertad se utilizan para obtener estimaciones imparciales de los parámetros de la población.

Aplicaciones y Condiciones para usar Tablas Estadísticas

Uno de los problemas que casi todos los libros de textos en estadística tienen es que no proporcionan información suficiente para entender las

conexiones entre las tablas estadísticas. Los estudiantes se preguntan a menudo: ¿Por qué los valores de la tabla del t con 1 grado de libertad son mucho más grandes comparados con otros valores de diferentes grados de libertad?. Algunas tablas son limitadas, ¿Qué se debería hacer cuando el tamaño de la muestra es demasiado grande?, ¿Cómo se puede conseguir familiaridad con las tablas y sus diferencias?, ¿Existe algún tipo de integración entre las tablas?, ¿Existen algunas conexiones entre las pruebas de hipótesis y el intervalo de confianza bajo diversos panoramas?, Por ejemplo, pruebas con respecto a una, dos o más poblaciones. Etcétera.

La figura siguiente muestra algunas relaciones útiles entre las tablas estadísticas mas comunes:



Algunas aplicaciones ampliamente usadas de las tablas estadísticas más comunes, pueden ser resumidas a continuación:

Tabla -T:

20. Prueba m de para una Población Simple.
21. Prueba de μ's para Dos Poblaciones Independientes.
22. La Prueba de μ's Antes-y-Después .
23. Prueba Concerniente a Coeficientes de Regresión.
24. Prueba Concerniente a Correlación.

Condiciones para usar esta tabla: La prueba dealeatoriedad de los datos es necesaria antes de usar esta tabla. La prueba para la condición de normalidad de la distribución de la población también es necesaria si el tamaño de la muestra es pequeño, de otra forma no sería posible invocar el teorema de límite central.

Tabla -Z :

25. Pruebe para la Aleatoriedad.
26. Pruebas referentes a la μ para una población o dos poblaciones basadas en tamaños grandes de muestras aleatorias (digamos mayores que 30) para invocar el teorema de límite central. Esto incluye la prueba referente a **proporciones**, con tamaño grande, muestras aleatorias tamaño n (mayores que 30) para invocar resultados de convergencia en la distribución.
27. Para Comparar Dos Coeficientes De Correlación.

Notas: Como usted sabe, en la prueba de hipótesis referentes a m, y la construcción de su intervalo de la confianza, comenzamos con una s sabido, puesto que el valor crítico (y el valor-p) de la distribución de la Tabla- Z pueden ser utilizados. Considerando las situaciones más realistas, cuando no sabemos s, se utiliza la Tabla- T. En ambos casos, necesitamos verificar la condición de normalidad de la distribución de la población; sin embargo, si el tamaño de muestra n es muy grande, automáticamente podríamos utilizar de hecho la Tabla- Z en virtud del teorema de límite central. Para poblaciones perfectamente normales, la distribución-t corrige cualquier error introducido por la estimación de s con s, en el caso de realizar inferencia.

Observe también que, en la prueba de hipótesis referente a los parámetros de las distribuciones binomiales y de Poisson para los tamaños de muestra grandes, la desviación estándar se conoce bajo la hipótesis nulas. Esta es la razón por la cual usted puede utilizar las aproximaciones normales para estas dos distribuciones.

Condiciones para usar esta tabla: La prueba para la aleatoriedad de los datos es necesaria antes de usar esta tabla. La prueba para la condición de normalidad de la distribución de la población también se necesita si el tamaño de muestra es pequeño, o podría no ser posible invocar el Teorema de Límite Central.

Tabla Chi- Cuadrado:

28. Prueba para la Relación de Tablas Cruzadas.
29. Prueba de Poblaciones- Idénticas para Datos de Tablas Cruzadas.
30. Prueba para la Igualdad de varias Proporciones de la Población.
31. Prueba para la Igualdad de varios Medianas de la Población.
32. Prueba de Bondad de Ajuste para la Probabilidad de Funciones Masivas.
33. Compatibilidad de Conteos Múltiples.
34. Prueba Del Coeficiente de Correlación.
35. Condiciones Necesarias para la Aplicación de las Pruebas Anteriores.
36. ¿Prueba de la Varianza: ¿Es la Calidad Buena?.
37. ¿Prueba de la Igualdad Varianzas Múltiples?.

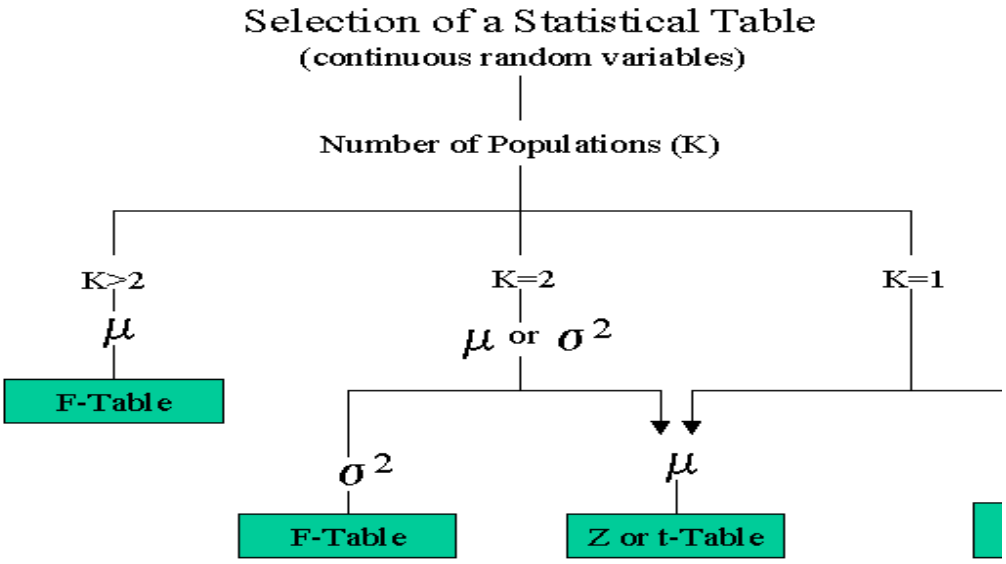
Condiciones para usar esta Tabla: Las condiciones necesarias para usar esta tabla para todas las pruebas antedichas, a excepción de la última, se pueden encontrar en las Condiciones para las Pruebas Basadas en la Chi-cuadrado. La última aplicación requiere de la normalidad (condición) de la distribución de la población.

Tabla-F:

38. Comparación de Medias Múltiples: Análisis de la Varianza (ANOVA).
39. Pruebas Referentes a Dos Varianzas.
40. Evaluación Total de los Modelos de Regresión.

Condiciones para usar esta Tabla: Las pruebas para la aleatoriedad de los datos y de la normalidad (condición) de las poblaciones son necesarias antes de usar esta tabla para ANOVA. Las mismas condiciones deben ser satisfechas para los residuos en análisis de regresión.

El cuadro siguiente resume las aplicaciones de las tablas estadísticas con respecto a la prueba de hipótesis y para la construcción de los intervalos de la confianza para media m y varianza s^2 en una población o en la comparación de dos o más.



Selección de una Tabla Estadística Apropiada

Usted podría gustarle u utilizar

Cálculos Estadísticos en Línea en la ejecución de la mayoría de estas pruebas. El sitio Web Valores-P para una Distribuciones Popular proporciona los valores-P útiles en importantes pruebas estadísticas. Los resultados son más exactos que los que se pueden obtener (por la interpolación) de tablas estadísticas o su libro de textos.

Función De Probabilidad Binomial

Una clase importante de los problemas de decisión bajo incertidumbre implica las situaciones para las cuales existen solo dos resultados aleatorios posibles.

La función de probabilidad binomial suministra la probabilidad exacta del número de "éxitos" en n pruebas independientes, cuando la probabilidad de éxito p en una sola prueba es una constante. Cada ensayo simple se llama Prueba de Bernoulli, la cual satisface las condiciones siguientes:

41. Cada ensayo da lugar a uno de dos posibles, mutuamente excluyentes, resultados. Uno de los posibles resultados se denota (arbitrariamente) como éxito, y el otro se denota un fracaso.
42. La probabilidad de éxitos, denotada por p, se mantiene constante de prueba a prueba. La probabilidad de fracaso, 1-p, es denotada por q.
43. Las pruebas o ensayos son independientes; es decir, el resultado de cualquier ensayo en particular no es afectado por el resultado de ningún otro ensayo.

Las formas conseguir r éxitos en n ensayos se obtiene mediante:

P (r éxitos en n pruebas) = ${}^{n}C_{r}$. p^{r} . $(1- p)^{(n-r)}$ = n! / [r!(n-r)!] . $[p^{r} . (1- p)^{(n-r)}]$.

La media y la varianza de la variable aleatoria r, son np y np(1-p) respectivamente, donde q = 1 - p. La oblicuidad y la kurtosis son (2q -1)/ $(npq)^{½}$, y (1 6pq)/(npq), respectivamente. De su oblicuidad, notamos que la distribución es simétrica para p = 1/2 y mas sesgada cuando p es 0 o 1.

Su moda está dentro del intervalo [(n+1)p -1, (n+1)p], por lo tanto si (n+1) p no es un número entero, la moda es un número entero dentro del intervalo. Sin embargo si (n+1)p es un número entero, su función de la probabilidad tiene dos modas pero adyacentes: (n+1)p -1, y (n+1)p.

Determinación de las Probabilidades para p Mayores a 0,5: Las tablas binomiales en algunos libros de textos se limitan a disuadir los valores de las probabilidades de p hasta 0,5. Sin embargo, estas tablas se pueden utilizar para valores de p mayores a 0,5. Modificando un problema en términos de p a 1- p, y fijando r a n-r, la probabilidad de obtener r éxitos en n ensayos para un valor dado p es igual a la probabilidad de obtener n-r fracasos en n ensayos con 1-p.

Una aplicación: Un envío grande de piezas compradas es recibido en un almacén, y una muestra de 10 porciones es revisada para saber su calidad. El fabricante establece que un máximo de 5% de las piezas podrían salir defectuosas. ¿Cuál es la probabilidad de que la muestra incluye una pieza defectuosa?

P (una defectuosa de diez) = $\{10! /[(1!)(9!)]\}(0,05)^{1}(0,95)^{9}$ = 32%.

Entienda que la distribución binomial satisface los cinco requisitos siguientes: (1) Cada ensayo puede tener solamente dos resultados o sus resultados se pueden reducir a dos categorías que se llamen éxito o fracaso, (2) Deben existir un número fijo de ensayos, (3)El resultado de cada ensayo o prueba debe ser independiente, (4) las probabilidades deben mantenerse constantes, (5) y el resultado de interés es el número de éxitos.

Aproximación Normal para Binomial: Todas las tablas binomiales son limitadas en su alcance; por lo tanto es necesario utilizar la distribución normal estándar para calcular las probabilidades binomiales. El siguiente ejemplo numérico ilustra cuan buena la aproximación podría ser. Este proporciona una indicación para aplicaciones reales cuando n está más allá de los valores dados en las tablas binomiales disponibles.

Ejemplo Numérico: Una muestra de 20 artículos es tomada aleatoriamente de un proceso de fabricación con probabilidad de artículos defectuosos p = 0,40. ¿Cuál es la probabilidad de obtener exactamente 5 artículos defectuosos?

P (5 de 20) = {20!/[(5!)(15!)]} ´ $(0,40)^5(0,6)^{15}$= 7,5%

Por que la media y la desviación estándar de la distribución son:

m = np = 8, y s = $(npq)^{1/2}$ = 2,19,

respectivamente; Por lo tanto, las observaciones estandarizadas para r = 5, mediante el uso del factor de continuidad (el cual siempre agranda) son:

z_1 = [(r-1/2) - m] / s = (4,5 -8)/2,19 = -1,60, y

z_2 = [(r+1/2) - m] / s = (5,5 -8)/2,19 = -1,14.

Como consecuencia, la P (5 de 20) aproximada es P (z estando dentro de los intervalos -1,60, -1,14). Ahora, mediante el uso de la tabla normal estándar, se obtiene:

P (5 de 20) = 0,44520 – 0,37286 = 7,2%

Comentarios: La aproximación para la distribución binomial se utiliza frecuentemente en procesos de control de calidad, confiabilidad, muestreo en censos, y otros problemas industriales.

A usted podría gustarle utilizar el JavaScript de Construcción de Intervalos de Confianza Exactos y la Prueba Hipótesis para Poblaciones Binomial , y el JavaScript de la Función de Probabilidad Binomial en Javascript para realizar algunos experimentos numéricos para validar las aserciones anteriores y proveerlo de un conocimiento mas profundo.

Función de Densidad Exponencial

Una parte importante de los problemas de decisión bajo incertidumbre se refiere a las duraciones aleatorias entre eventos. Por ejemplo, la longitud de tiempo entre las interrupciones de funcionamiento de una máquina que no excedan cierto intervalos, tal como la fotocopiadora en su oficina que no se haya dejado de funcionar durante esta semana.

La distribución exponencial da distribución de tiempo entre los acontecimientos independientes que ocurren a una tasa constante. Su función de densidad es:

f(t) = l exp(-lt),

donde l es el número promedio de eventos por unidad de tiempo, el cual es positivo.

La media y la varianza de la variable aleatoria t (tiempo entre los eventos) son 1/l, y $1/l^2$, respectivamente.

Entre las aplicaciones se incluyen la estimación probabilística del tiempo entre las llegadas de pacientes a la sala de emergencia de un hospital, o el tiempo entre de llegadas de barcos en un puerto particular.

Comentarios: Este es un casos especial de la distribución Gamma.

A usted podría gustarle utilizar el JavaScript de la Densidad Exponencial para realizar sus cálculos, y la Prueba de Lilliefors para la Exponencialidad para realizar calidad de ajustes en la prueba.

Función de Densidad F

La distribución F es la distribución del cociente de dos estimaciones de la varianza de dos muestreos independientes (de tamaño n_1, y n_2, respectivamente) con respecto a una distribución normas estándar. También es formada por el cociente de dos variables independientes Chi-cuadrado divididas por sus respectivos grados de libertad independiente.

Sus usos principales son en la prueba de igualdad de dos varianzas poblacionales independientes con respecto a dos muestras aleatorias independientes, ANOVA, y análisis de la regresión.

A usted podría gustarle utilizar la Función de Densidad F para obtener sus valores de P.

Función de Densidad Chi-cuadrado

La curva de la densidad de probabilidad de una distribución Chi-cuadrado es una curva asimétrica estirada sobre el lado positivo de la línea y que tiene una cola derecha larga. La forma de la curva depende del valor de un parámetro conocido como grado de libertad (gl).

El valor esperado del estadístico Chi-cuadrado es su gl, su varianza está dos veces de su gl, y su moda es igual a (gl - 2).

Relación de la distribución Chi-cuadrado con la Distribución Normal: La distribución Chi-cuadrado se relaciona con la distribución de muestreo en la varianza cuando la muestra viene de una distribución normal. La varianza de la muestra es la suma de los cuadrados de las variables normales estándares N (0, 1). Por lo tanto, del cuadrado de la variable aleatoria N (0.1) es una Chi-cuadrado con 1 gl.

Note que la Chi-cuadrado está relacionado con el estadístico F de la siguiente manera: F= Chi-cuadrado/ $gl._1$, donde F tiene (gl_1 = gl de la tabla Chi-cuadrado, y gl_2 es el valor disponible más grande de la tabla F).

De manera similar a las variables aleatorias normales, la Chi-cuadrado tiene la propiedad aditiva. Por ejemplo, para dos variables independientes Chi-cuadrado, su suma es también Chi-cuadrado con los grados de libertad iguales a la suma del los gl individuales. Por lo tanto, la varianza muestral **imparcial** para una muestra de tamaño n de N (0,1) es una suma de n-1 Chi-cuadrados, cada uno con gl = 1, es decir, Chi-cuadrado con gl = n-1.

Las aplicaciones mas comunes de la distribución Chi-cuadrado son:

La prueba Chi-cuadrado por asociación es una prueba no paramétrica; por lo tanto, puede ser utilizada también para datos nominales. Es una prueba de significancia estadística ampliamente utilizada de doble variación en el análisis tabular de asociación. Típicamente, la hipótesis es si o no dos poblaciones son diferentes en ciertas características o aspectos de su comportamiento basado en dos muestras escogidas al azar. El procedimiento de esta prueba es también conocido como la prueba Chi-cuadrado de Pearson.

La prueba de Bondad de Ajuste Chi-cuadrado se utiliza para probar si una distribución observada satisface a cualquier otra distribución particular. El cálculo de la prueba de bondad de ajuste se realiza mediante la comparación de datos observados con los datos esperados basados en una distribución particular.

A usted podría gustarle utilizar la Densidad Chi-cuadrado para encontrar sus valores de P.

Función de Probabilidad Multinomial

Unavariable aleatoria multinomial es una binomial extendida. Sin embargo, la diferencia está en que en el caso multinomial, existen s más de dos resultados posibles. Existe un número fijo de resultados independientes, con una probabilidad dada para cada resultado.

El Valor Esperado (es decir, promedio):

Valor Esperado = m = $S(X_i \acute{} P_i)$, la suma incluye todos los i's.

El valor esperado es otro nombre con el cual se puede llamar a la media y al average (aritmético.)

Este es un concepto estadístico importante porque, por ejemplo, sus clientes quieren saber que “esperar” de su producto/ servicio, ó usted como comprador necesita saber que esta comprando como “materia prima” para su producto/ servicio, es decir, lo que usted “espera” obtener de su compra.

Para entender el significado de la formula anterior, considere calcular el average de los siguientes datos:

2, 3, 2, 2, 0, 3

El average se obtiene sumando todos los números y dividiéndolos por el número de conteos (o unidades), es decir:

(2 + 3 + 2 + 2 + 0 + 3) / 6

Estos datos pueden ser agrupados y reescritos de la siguiente forma:

[2(3) + 3(2) + 0(1)] / 6 = 2(3/6) + 3(2/6) + 0(1/6)

lo cual el la suma de las multiplicaciones de cada observación particular por su probabilidad asociada. Alguna duda?

El valor esperado se conoce también como el Primer Momento, prestado de la física, porque este es el punto de balance donde los datos y las probabilidades son las distancias y los pesos, respectivamente.

La Varianza es:

Varianza = $s^2 = S\ [X_i^2 \acute{}\ P_i] - m^2$, la suma incluyetodos los i's.

La variación no se expresa en las mismas unidades que el valor esperado. Por lo tanto, la varianza es difícil de entender y de explicar como resultado del término al cuadrado en su cálculo. Esto se podría remediar si se trabaja con la raíz cuadrada de la varianza, el cual se llama la **Desviación Estándar (es decir, utilizando las mismas unidades que tienen los datos):** :

Desviación Estándar = s = (Varianza) ^{½}

La varianza y la desviación estándar proporcionan la misma información y, por lo tanto, una se puede obtener siempre de la otra. Es decir el proceso de calcular la desviación estándar implica siempre el cálculo de la varianza. Puesto que la desviación estándar es la raíz cuadrada de la varianza, siempre es expresada en las mismas unidades que el valor esperado.

Para el proceso dinámico, la Volatilidad como medida para el riesgo incluye el período de tiempo sobre el cual la desviación estándar se

calcula. La **medida de Volatilidad** se define como desviación estándar dividida por la raíz cuadrada de la duración del tiempo.

Coeficiente de variación: El coeficiente de variación (CV) es la *desviación absoluta relativa* con respecto al tamaño x̄provisto de que x̄ no es cero, expresado en porcentaje:

CV =100 |s/x̄| %

Note que el CV es independiente del valor esperado de la medida. El coeficiente de variación demuestra la relación entre la desviación estándar y el valor esperado, mediante el enunciado del riesgo como un porcentaje del valor esperado. El inverso del CV (es decir 1/CV) es llamado el **Cociente de Señal de Ruido**.

A usted podría gustarle utilizar el JavaScript Multinomial para revisar sus cálculos y realizar un experimento asistido por computadoras.

Una Aplicación: Considere dos alternativas de inversión, inversiones I y II con sus características respectivas descritas en la tabla siguiente:

- Dos Alternativas de Inversión -			
Inversión I		Inversión II	
Pagos %	Prob.	Pagos %	Prob.
1	0,25	3	0,33
7	0,50	5	0,33
12	0,25	8	0,34

Comportamiento de dos Inversiones

Para alinear estas dos inversiones bajo el *Acercamiento de Dominación Estándar en Finanzas*, primero debemos calcular la media y la desviación estándar y luego analizar los resultados. Usando el JavaScript Multinomial para los cálculos, notamos que la inversión I tiene media = 6,75% y desviación estándar = 3,9%, mientras que la segunda inversión tiene media = 5,36% y desviación estándar = 2,06%. Primero observe que bajo el análisis generadle media-varianza, estas dos inversiones no pueden ser alineadas. Esto se debe a que la primera inversión tiene una media mas grande; También tiene mayor desviación estándar; por lo tanto, el **Acercamiento de Dominación Estándar** no es una herramienta útil aquí. Tenemos que recurrir al coeficiente de variación (CV) como base sistemática de la comparación. El CV para la inversión I es 57,74% y para la inversión II es 38,43%. Por lo tanto, la inversión II tiene preferencia sobre la inversión I. Claramente, este acercamiento se puede utilizar para alinear cualquier número de inversiones alternativas. Note que mientras menor sea la variación en los retornos de inversión menor es el riesgo implícito.

A usted podría gustarle utilizar este Applet en la ejecución de algunos experimentos numéricos que:

44. Muestre: $E[aX + b] = aE(X) + b$.
45. Muestre: $V[aX + b] = a^2V(X)$.
46. Muestre: $E(X^2)= V(X) + (E(X))^2$.

Función de Densidad Normal

En la Sección de Estadística Descriptiva de este sitio del Web, nos hemos referido a cómo se distribuyen los valores empíricas y a cómo describir sus distribuciones de la mejor manera posible. Hemos discutido diversas medidas, pero la media m será la medida que utilizaremos para describir el centro de la distribución, y la desviación estándar s será la medida que utilizaremos para describir la extensión de la distribución. Saber estos dos hechos nos da una amplia información para hacer fundamentaciones sobre la probabilidad de observar cierto valor dentro de esa distribución. Si se sabe, por ejemplo, que el valor promedio del Coeficiente de Inteligencia (siglas IQ en Ingles) es 100 y que tiene una desviación estándar de s = 20, entonces sabemos que alguien con un índice de inteligencia de 140 es muy perspicaz. Esto se sabe porque 140 se desvía de la media m dos veces el valor promedio del resto de los valores de la distribución. Por lo tanto, es poco probable ver un valor tan extrema como 140 porque la mayoría de los valores de IQ se encuentran agrupados alrededor de 100 y se desvían solamente 20 puntos de la media m .

Muchos aplicaciones surgen del teorema del límite central (TLC). El TLC indica eso, el promedio de valores de n observaciones que se aproximan la distribución normal, independientes de la forma de la distribución original bajo condiciones generales. Por lo tanto, la distribución normal es un modelo apropiado para muchos, pero no todos, los fenómenos físicos, tales como distribución de medidas físicas en los organismos vivos, pruebas de inteligencia, dimensiones de productos, temperaturas medias, etcétera.

Sepa que la distribución normal debe satisfacer siete requisitos: (1) el gráfico debe ser de formada campana; (2) la media, la mediana y la moda son todas iguales; (3) la media, la mediana y la moda están situadas en el centro de la distribución; (4) tiene solamente una moda, (5) es simétrica con respecto a la media, (6) es una función continua; (6) nunca toca el eje de las x; y (7) el área bajo la curva es igual a 1.

Muchos métodos de análisis estadístico presumen la distribución normal.

Cuando sabemos la media y la varianza de una Normal estamos capacitados a encontrar probabilidades. Así pues, por ejemplo, si usted sabe algunas cosas sobre la altura media de las mujeres en su país,

incluyendo el hecho de que las alturas están distribuidas normalmente, usted podría medir a todas las mujeres de su familia y encontraría la altura promedio. Esto le permite determinar una probabilidad asociada a su resultado, si la probabilidad de conseguir su resultado dada el conocimiento de la estatura de las mujeres a nivel nacional es alta, entonces la altura femenina de su familia no podría ser diferente del promedio. Si esa probabilidad es baja, entonces su resultado es raro (dado el conocimiento de las alturas de las mujeres en toda la nación), y usted podría decir que su familia es diferente. Usted simplemente acaba de realizar una prueba de hipótesis de que la altura media de mujeres en su familia es diferente del promedio total.

El coeficiente de dos observaciones independientes con respecto a la normal estándar se distribuye como la distribución de Cauchy la cual tiene colas más gruesas que una distribución normal. Su función de la densidad es $f(x) = 1/[p(1+x^2)]$, para cualquier valor real de x.

A usted podría gustarle utilizar el JavaScript de la Normal Estándar en vez de usar valores tabulares de su libro de texto, y la muy conicidad Prueba de Normalidad de Lilliefors para determinar la calidad de ajuste.

Función de Probabilidad de Poisson

> La vida es buena solo por dos cosas, por descubrir las matemáticas y por enseñar las matemáticas.
> -- *Simeon Poisson*

Un tipo importante de problemas de decisión bajo incertidumbre es caracterizado por el pequeño chance de ocurrencia de un acontecimiento particular, tal como un accidente. La función de probabilidad de Poisson calcula la probabilidad de exactamente x ocurrencias independientes durante un período de tiempo dado, si los eventos ocurren independientemente y a una tasa constante. La función de la probabilidad de Poisson también representa el número de ocurrencias sobre áreas o volúmenes constantes:

Las probabilidades de Poisson se utilizan a menudo; por ejemplo en control de calidad, confiabilidad de software y hardware, reclamos de seguro, el número de llamadas telefónicas entrantes, y la teoría de alineación.

Una aplicación: Uno de los usos más útiles de la distribución de Poisson es en el campo de la teoría de alineación. En muchas situaciones donde ocurren colas, se ha demostrado que el número de la gente que se une a la misma en un período de tiempo dado, sigue el modelo de Poisson. Por ejemplo, si el índice de llegadas a una sala de emergencia es l por unidad de período de tiempo (1 hora), entonces:

$P(\text{n llegadas}) = l^n \ e^{-l} / n!$

La media y la varianza de la variable aleatoria n son ambas l . Sin embargo si la media y la varianza de una variable aleatoria tienen valores numéricos iguales, no necesariamente implica que su distribución es de Poisson. Su moda está dentro del intervalo [l -1, l].

Aplicaciones:

P (0 llegadas) = e^{-l}
P (1 llegada) = l e^{-l} / 1!
P (2 llegadas) = l^2 e^{-l}/ 2!

y así sucesivamente, en general:

P (n+1 llegadas) = l P (n llegadas) / n.

Aproximación Normal para Poisson: Todas las tablas de Poisson se limitan en su alcance; por lo tanto, es necesario utilizar la distribución normal estándar para calcular las probabilidades de Poisson. El siguiente ejemplo numérico ilustra cuan buena la aproximación podría ser.

Ejemplo Numérico: Los pacientes de la sala emergencia llegan a un hospital a una tasa de 0,033 por minuto. ¿Cuál es la probabilidad de que exactamente dos pacientes lleguen durante los próximos 30 minutos?

La tasa de llegada durante 30 minutos es l = (30)(0,033) = 1. Por lo tanto,

P (2 llegadas) = [1^2 /(2!)] e^{-1} = 18%

La media y la desviación estándar de la distribución son:

m = l = 1, and s = $l^{1/2}$ = 1,

respectivamente; por lo tanto, las observaciones estandarizadas para n = 2, mediante el uso del factor continuo (el cual siempre engrandece) son:

z_1 = [(r-1/2) - m] / s = (1,5 -1)/1 = 0,5, y

z_2 = [(r+1/2) - m] / s = (2,5 -1)/1 = 1,5.

Por lo tanto, la P (2 llegadas) aproximada es P (z estando entre los intervalos 0,5, 1,5). Ahora, mediante el uso de la tabla normal estándar, se obtiene:

P (2 llegadas) = 0,43319 – 0,19146 = 24%

Como se puede observar la aproximación se sobrestima levemente, por lo tanto el error está en el lado seguro. Para valores grandes de l,

digamos mayores a 20, se podría utilizar la aproximación normal para el cálculo de las probabilidades de Poisson.

Note que tomando la raíz cuadrada de una variable aleatoria de Poisson, la variable aleatoria, transformada es más simétrica. Esto es una transformación útil en el análisis de regresión de las observaciones de Poisson.

A usted podría gustarle utilizar el JavaScript de la Función de Probabilidad de Poisson para realizar sus cálculos, y la Prueba de Poisson para realizar la calidad de ajuste.

Función de Densidad T de Student

Las distribuciones t fueron descubiertas en 1908 por William Gosset, que era un químico y estadístico empleado por la compañía de elaboración de la cerveza Guinness. Él se consideraba como un estudiante todavía que aprendía estadística, y el firmaba sus trabajos bajo el seudónimo de "estudiante", o quizás él utilizó el seudónimo debido a las restricciones "secretas" de Guinness.

Observe que hay diversas distribuciones t; esta es una clase de distribuciones. Cuando hablamos de una distribución específica t, tenemos que especificar los grados de libertad. Las curvas de la densidad t son simétricas y acampanadas como la distribución normal y tienen su pico en 0. Sin embargo, la extensión es mayor que el de la distribución normal estándar. Mientras mas grandes sean los grados de libertad, más cercana se encuentra la densidad t de la densidad normal.

La forma de una distribución t depende del parámetro llamado "grado de libertad". Mientras el grado de libertad sea mas grande, distribución t se asemeja mas y mas a la distribución estándar normal.. Para propósitos prácticos, la distribución se maneja como una distribución normal estándar cuando los grados de libertad sean mayores a 30.

Suponga que tenemos dos variables aleatorias independientes, una es Z, distribuida como la distribución normal estándar, y la otra esta distribuida como la Chi-cuadrado con (n-1) gl; entonces la variable aleatoria:

$(n-1)Z / c^2$

tiene una distribución t con (n-1) gl, para tamaños de muestra grande (n mayor a 30), la nueva variable aleatoria tiene un valor esperado igual a cero, y su varianza es (n-1)/(n-3) la cuál se acerca a uno.

Note que la t estadística está relacionada con la F-estadística de la siguiente forma: $F = t^2$, donde F tiene ($gl_1 = 1$, y gl_2 = gl de la tabla t).

A usted podría gustarle utilizar la Densidad t de Student para obtener sus valores de P.

6. Función de Densidad Triangular

Función de Densidad Triangular

La distribución triangular muestra el número de éxitos cuando se saben el mínimo, el máximo, y los valores más probable. Por ejemplo, se podría describir el número de productos consumidos por semana cuando los últimos datos de consumo muestran el mínimo, el máximo, y el número más probable de los casos considerados. Esta representa una distribución de probabilidad.

Los parámetros para la distribución triangular son: Mínimo, máximo, y lo más probablemente posible. Existen tres condiciones subyacente a la distribución triangular:

- El número mínimo de artículos es fijo.
- El número máximo de artículos es fijo.
- El número más probable de artículos se encuentra entre los valores mínimos y máximos.

Estos tres parámetros forman una distribución triangular, la cual muestra que los valores cerca del mínimo y del máximo son menos probables de ocurrir que eso cercanos el valor más probable.

Función de Densidad Uniforme

La función de densidad uniforme proporciona la probabilidad de que una observación ocurrirá dentro de un intervalo particular [a, b] cuando la probabilidad de la ocurrencia dentro de ese intervalo es directamente proporcional a la longitud del intervalo. Su media y varianza son:

$m = (a+b)/2, \quad s^2 = (b-a)^2/12.$

Aplicaciones: usada para generar números aleatorios azar en muestreos y en la simulación de Monte Carlo.

Comentarios: Caso especial de la distribución beta.

A usted podría gustarle utilizar Prueba de Bondad de Ajuste uniforme y realizar algunos experimentos numéricos para una comprensión mas profunda de los conceptos.

Note que cualquier distribución uniforme tiene incontable número de modas que tienen igual valor de densidad; por lo tanto se considera como población homogénea.

Condiciones Necesarias para la toma de Decisiones Estadísticas

Introducción a las Condiciones Necesarias para el Análisis Deductivo de Datos: No aprenda simplemente fórmulas y combinaciones de números. Aprenda sobre las *condiciones* bajo las cuales los métodos de prueba estadística se aplican. Las condiciones siguientes son comunes para casi todas las pruebas estadísticas:

3. Cualquier outliers puede tener impacto importante y puede influenciar los resultados de casi toda la valoración y métodos de pruebas estadísticas.
4. Población homogénea. Es decir, no hay más de una moda. Realice la Prueba para la Homogeneidad de una Población
5. La muestra debe ser aleatoria. Realice la Prueba de Aleatoriedad .
6. Además del requisito de homogeneidad, cada población tiene una distribución normal. Realice la Prueba de Lilliefors para la Normalidad .
7. Homogeneidad de las varianzas. La variación en cada población es casi igual que la que ocurre en otras poblaciones. Realice la Prueba De Bartlett .

   Para dos poblaciones utilice la prueba F. Para 3 o más poblaciones existe una regla práctica conocida como la “regla de 2”. En esta regla, se divide la varianza más alta de una muestra por la varianza más baja de la otra muestra. Dado que los tamaños de las muestras similares, y el valor de la división es menor a 2, las variaciones de las poblaciones son casi iguales.

   Aviso: Esta importante condición en el análisis de la varianza (ANOVA y la prueba t para diferencias en las medias) es comúnmente evaluada por la prueba de Levene o su prueba modificada conocida como la prueba Brown-Forsythe. Interesante, ambas pruebas confían en la condición de homogeneidad de las varianzas!

   Estas condiciones son cruciales, no para el método de cálculo, sino para la prueba usando el estadístico resultante. De otra forma, podríamos hacer ANOVA y regresión sin ningún supuesto, y los números resultantes serían los mismos. Simples cálculos nos darían los ajustes de último cuadrado, particiones de la varianza, coeficientes de regresión, etcétera. Necesitamos las condiciones anteriores cuando la prueba de hipótesis es nuestra preocupación principal.

Medida de Extrañeza para la Detección del Resultados

Las técnicas estadísticas fuertes son necesarias hacer frente a cualquier outlier desapercibido; si no fuesen mas probables de invalidar las técnicas estadísticas de las condiciones subyacentes , , y podrían distorsionar seriamente las estimaciones y producir conclusiones engañosas de la prueba de hipótesis. Un acercamiento común consiste en asumir que los modelos contaminados, son diferentes a los que se generan el resto de los datos, generan los outliers (posibles).

Debido a una varianza potencialmente grande, los outliers podían ser los resultados de los errores de muestreo o de los errores administrativos tales como recolección de datos. Por lo tanto, usted debe ser muy cuidadoso y cauteloso. Antes de declarar una observación como "outlier", descubra porqué y cómo ocurrió tal observación. Esto incluso podría ser un error en la etapa que entraba de los datos mientras se usa cualquier paquete de la computadora.

En la práctica, cualquier observación con un valor estandardizado mayor de 2,5 en valor absoluto es un candidato a ser un outlier. En tal caso, es necesario primero investigar la fuente del dato. Si no hay duda sobre la exactitud o la veracidad de la observación, entonces debe ser quitada, y el modelo debe ser reinstalado.

8. Calcule la media ($\bar{x}$) y la desviación estándar (S) de la muestra entera.
9. Fije los límites para la media $\bar{x}$:

   $\bar{x}$- k ´ S, $\bar{x}$+ k ´ S.

   Un valor típico para k es 2,5

10. Remueva todos los valores de la muestra fuera de los límites.
11. Ahora, itere con el algoritmo, el grupo de la muestra debería reducirse después de quitar los outliers aplicando el paso 3.
12. En la mayoría de los casos, necesitamos iterar con este algoritmo varias veces hasta que todos los outliers sean removidos.

Una aplicación: Suponga usted pide que diez de sus compañeros de clase midan una longitud dada X. Los resultados (en el milímetro) son:

46, 48, 38, 45, 47, 58, 44, 45, 43, 44

¿Es 58 un outlier? Calculando la media y la varianza de las diez medidas usando el Javascript de Estadística Descriptiva de Muestreo se obtiene 45,8 y 5,1 respectivamente (después de los ajustes necesarios). El valor Z para 58 es Z (58) = 2,4. Puesto que las medidas, en general, siguen una distribución normal, por lo tanto,

la probabilidad [X tan grande como 2,4 veces la desviación estándar] = 0,008,

obtenida mediante el uso del Javascript Valor P Normal Estándar o de la tabla normal en su libro de textos.

De acuerdo a esta probabilidad, se espera que solamente 0,09 de las diez medidas sean tan malas como ésta. Esto es un acontecimiento muy raro, sin embargo, como esta probabilidad tan pequeña ha ocurrido, podría ser que sea un outlier.

La próxima medida mas sospechada es 38, ¿es este un outlier? Esta es una pregunta para usted.

Nota: La detección de outliers en una población simple no es demasiado difícil. Frecuentemente, sin embargo, se puede discutir que los outliers detectados no sean realmente outliers, sino una **formar de una segunda población** . Si éste es el caso, un acercamiento a la separación de datos necesita ser tomado.

A usted podría gustarle utilizar Identificación de Outliers en Javascript para la realización de algunas experimentaciones numéricas para validar y para obtener una comprensión más profunda de los conceptos.

Población Homogénea

Una población homogénea es una población estadística que tiene **una única moda**.

Note que, por ejemplo, una distribución Uniforme tiene incontable número de modas que tienen valor de densidad igual; por lo tanto se considera como población homogénea.

Para determinarse si una población dada es homogénea o no, construya el histograma de una muestra escogida al azar de la población entera. Si hay más de una moda, se tiene una mezcla de una o mas poblaciones diversas. Sepa que para realizar cualquier prueba estadística, usted necesita cerciorarse de que usted esté tratando con una población homogénea.

Uno de las aplicaciones principales de la histografía es Probar la Homogeneidad de una Población. La unimodalidad del histograma es una condición necesaria para la homogeneidad de una población con el objetivo de conducir cualquier análisis estadístico significativo. Sin embargo, note que, una distribución Uniforme tiene incontables cantidades de modas que tienen valor de densidad igual; por lo tanto se considera como población homogénea.

Prueba de Aleatoriedad: la Prueba de Corridas (Wald-Wolfowitz)

Una condición básica en casi toda la estadística deductiva es que un sistema de datos constituye una muestra escogida aleatoria de una población homogénea dada. La condición de la aleatoriedad es esencial para cerciorarse de que la muestra es verdaderamente representativa de la población. La prueba mas usada para la aleatoriedad es la Prueba de corridas (Wald-Wolfowitz).

Una “Corrida” es una sub secuencia máxima de elementos semejantes.

Considere la siguiente secuencia (D para artículos defectuosos, N para artículos No-defectuosos) de una cadena de producción: DDDNNDNDNDDD. El número de corridas es R = 7, con $n_1 = 8$, y $n_2 = 4$ los cuales son números de D’s y N’s.

Una secuencia es una secuencia aleatoria si, ni es “sobre mezclada” ni es “sub mezclada”. Un ejemplo de la secuencia sobre mezclada es DDDNDNDNDNDD, con R = 9 mientras que una sub mezclada luciría como DDDDDDDDNNNN con R = 2. Allí la secuencia antedicha parece ser una secuencia aleatoria.

Las Pruebas de Corridas, que también se conoce como Prueba de Wald-Wolfowitz, es diseñada para probar la aleatoriedad de una muestra dada a un nivel de confianza de 100(1- a)% Para conducir una Prueba de corridas en una muestra, realice los pasos siguientes:

Paso 1: calcule la media de la muestra.

Paso 2: pasando por la secuencia de la muestra, substituya cualquier observación con +, ó - dependiendo si está por debajo o por arriba de la media. Deseche cualquier lazo.

Paso 3: Calcule R, n_1, y n_2.

Paso 4: calcule la media y la varianza esperada de R, como sigue:

$a = 1 + 2n_1n_2/(n_1 + n_2)$.

$s^2 = 2n_1n_2(2n_1n_2-n_1- n_2)/[[n_1 + n_2)^2 (n_1 + n_2 -1)]$.

Paso 5: Calcule $z = (R-m)/ s$.

Paso 6: Conclusión:

Si $z > Z_a$, entonces debería tener un comportamiento cíclico y con estacionalidad (sub mezclada).

Si $z < - Z_a$, debería tener una pendiente.

Si $z < - Z_{a/2}$, ó $z > Z_{a/2}$, rechaza la aleatoriedad.

Nota: Esta prueba es válida para los casos en los cuales n_1 y n_2 son grandes, al menos mayores que 10. Para muestras de pequeñas de tamaños, las tablas especiales deben ser utilizadas.

Por ejemplo, suponga que para una muestra dada de tamaño 50, se tienen R = 24, n_1 = 14 y n_2 = 36. Pruebe para la aleatoriedad en a = 0,05. Aplicando estos valores a las formulas anteriores se obtiene que a = 16,95, s = 2,473, y z = -2,0. De la tabla Z, tenemos Z = 1,645. Podría existir una pendiente o tendencia, que significa que la muestra no es aleatoria.

A usted podría gustarle utilizar el Javascript para la Prueba de Aleatoriedad.

Prueba de Normalidad

La prueba estándar para la normalidad es el estadístico de Lilliefors. Un histograma y un diagrama normal de la probabilidad también le ayudarán a distinguir entre una salida sistemática de la normalidad cuando este es mostrada como una curva.

La Prueba de Lilliefors para la Normalidad: Esta prueba es un caso especial de la Prueba de Bondad de Ajuste de Kolmogorov-Smirnov, desarrollada para probar la normalidad de la distribución de la población. Al aplicar la prueba de Lilliefors, una comparación es hecha entre la función de distribución acumulativa, normal estándar, y una función muestral de distribución acumulativa con variable aleatoria estandarizada. Si existe un acuerdo cercano entre las dos distribuciones acumulativas, se apoya la hipótesis de que la muestra fue dibujada de la población con una función de distribución normal. Si, sin embargo, existe una discrepancia demasiado grande entre las dos funciones de distribución acumulativas para ser atribuido un solo chance, se rechaza la hipótesis.

La diferencia entre las dos funciones de distribución acumulativas es medida por el estadístico D, el cual es la distancia vertical más grande entre las dos funciones.

A usted podría gustarle utilizar la muy bien conocida Prueba de Normalidad de Lilliefors para determinar la bondad de ajuste.

Introducción a la Estimación

Para estimar medias de valor (dar valor a). Un estimador es cualquier cantidad calculada de los datos de la muestra los cuales se utilizan para obtener información sobre una cantidad desconocida de la población. Por ejemplo, la media muestral es un estimador de la media poblacional m.

Los resultados de un estimador pueden ser expresados como un simple valor; entendido como una estimación en un punto, o un rango de valores, referido como un intervalo de confianza. Siempre que utilicemos la valoración de un punto, calculamos el margen de error asociado a la estimación de ese punto.

Los estimadores de los parámetros de la población son diferenciados a veces de los valores verdaderos mediante el uso del símbolo de "sombrero". Por ejemplo, la verdadera desviación estándar de la población s se estima de la muestra de la desviación estándar de la población.

De nuevo, el estimador usual de la media poblacional es $\bar{x} = Sx_i / n$, donde n es el tamaño n de la muestra y x_1, x_2, x_3,.....,x_n son los valores de la muestra. Si el valor del estimador en una muestra particular es 5, entonces 5 es la estimación del μ de la media de la población.

Cualidades de un buen Estimador

Un "buen" estimador, es aquel que provee una estimación con las cualidades siguientes:

Imparcialidad: Una estimación es imparcial con respecto a un parámetro cuando el valor esperado del estimador puede ser expresado como igual al parámetro que ha sido estimado. Por ejemplo, la media de una muestra es una estimación imparcial de la media de la población de la cual la muestra fue obtenida. La imparcialidad es una buena cualidad para una estimación, puesto que, usando el promedio ponderado de varias estimaciones se obtendría una mejor estimación que de cada una de ellas por separado. Por lo tanto, la imparcialidad permite que actualicemos nuestras estimaciones. Por ejemplo, si sus estimaciones de la medias poblacional μ son, digamos 10, y 11,2 con respecto a dos muestras independientes de tamaños 20, y 30 respectivamente, la mejor estimación de la media poblacional μ basada en ambas muestras es [20 (10) + 30 (11,2)] (20 + 30) = 10,75.

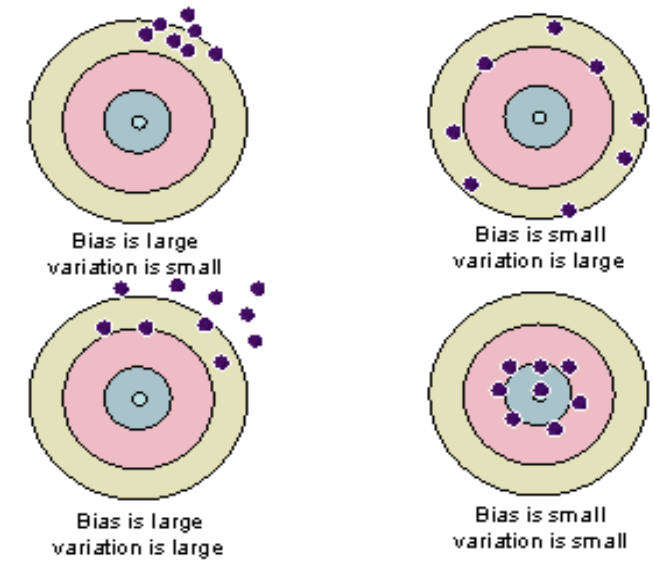
Consistencia: La desviación estándar de una estimación es llamada el error estándar de esa estimación. Mientras mas grande es el error estándar existirá más error en su estimación. La desviación estándar de una estimación es un índice comúnmente usado del error exigido al

estimar un parámetro de la población basado en la información en una muestra de tamaño n escogida al azar de la población entera.

Un estimador debe ser “consistente” si al aumentar el tamaño de la muestra se produce una estimación con un error estándar más pequeño. Por lo tanto, su estimación es “consistente” con el tamaño de la muestra. Es decir, gastando más dinero para obtener una muestra más grande produzca una mejor estimación.

Eficiencia: Una estimación eficiente es la que tiene el error estándar más pequeño entre todos los estimadores imparciales .

El “mejor” estimador es el que está más cercano al parámetro de la población que es estimado.



Accuracy versus Quality of an Estimator Using Bias and Variation as Measurable Quantities Respectively

El Concepto de Distancia para un Estimador

La figura anterior ilustra el concepto de la proximidad por medias que tienen como objetivo el centro para la imparcialidad con varianza mínima. Cada tablero de dardos tiene varias muestras:

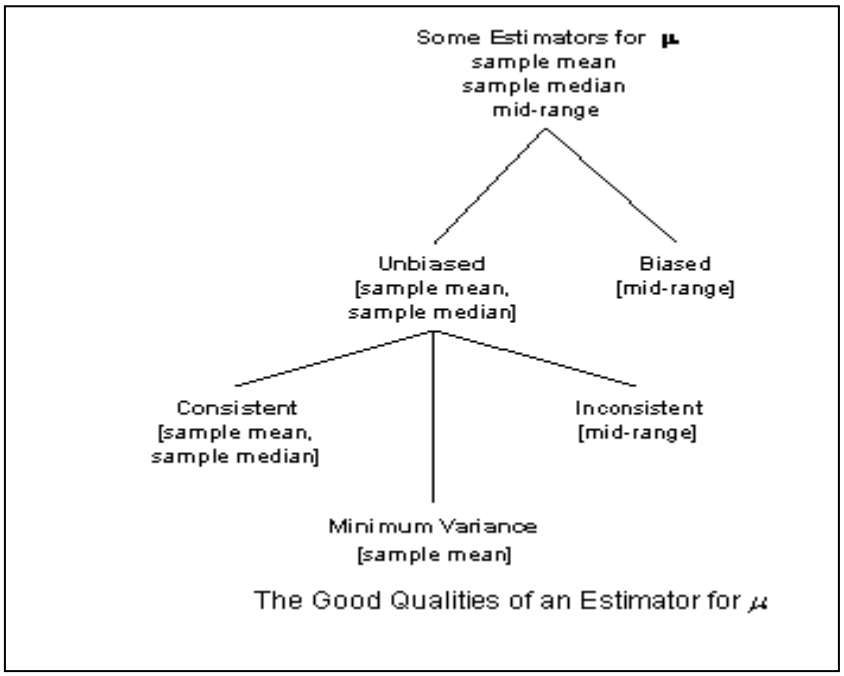
El primero tiene todos los tiros agrupados firmemente juntos, pero ningunos de ellos golpean el centro. El segundo tiene una extensión mas grande, pero alrededor del centro. El tercero es peor que los primeros dos. Solo el último tiene un grupo apretado alrededor del centro, por lo tanto tiene buena eficacia.

Si un estimador es imparcial, entonces su variabilidad determinará su confiabilidad. Si un perito es extremadamente variable, las estimaciones que produce pueden en promedio no estar tan cerca del parámetro

poblacional como lo estaría un estimador parcializado con varianza mas pequeña.

El esquema siguiente representa la calidad de algunos estimadores populares para la media poblacional μ:

El estimador mas común de la media poblacional μ es $\bar{x} = Sx_i/n$, donde n es el tamaño de la muestral x_1, x_2, x_3,....., x_n son los valores de la



muestra que tienen todas las buenas características antedichas.

Por lo tanto, es un “buen” estimador.

Si usted desea una estimación de la tendencia central como parámetro de una prueba o para comparación, los tamaños de muestra pequeños son poco probables de rendir cualquier estimación estable. La media es sensible en una distribución simétrica como medida de tendencia central; pero, por ejemplo, con diez casos, usted no podrá juzgar si usted tiene una distribución simétrica. Sin embargo, la estimación de la media es útil si usted está intentando estimar la suma de la población, o alguna otra función del valor esperado de la distribución. ¿Sería la mediana una mejor medida? En algunas distribuciones (por ejemplo, las tallas de camisas) la moda podría ser mejor. BoxPlot indicarán outliers en el conjunto de datos. Si existen outliers, la mediana es mejor que la media para la medida de tendencia central.

A usted podría gustarle usar el JavaScript de Estadística Descriptiva para obtener “buenas” estimaciones.

Estadísticos con Confianza

En la práctica, un intervalo de la confianza se utiliza para expresar la incertidumbre en una cantidad que es estimada. Hay incertidumbre porque las inferencias se basan en una muestra escogida al azar del tamaño finito de una población entera o del proceso de interés. Para juzgar el procedimiento estadístico podemos preguntar qué sucedería si repitiéramos el mismo estudio una y otra vez y que consiguiéramos repetidamente datos diferentes cada vez (y así diversos intervalos de la confianza).

En la mayoría de los estudios, los investigadores están interesados en la determinación del tamaño de las diferencias de un resultado medido entre grupos, en vez de un simple indicativo de es estadísticamente significativo. Los intervalos de la confianza presentan un rango de valores, con base en los datos de la muestra, de los cuales el valor de esta diferencia podría ser mentira.

Sepa que un intervalo de la confianza calculado a partir de una muestra será diferente de un intervalo de la calculado computado de otra muestra.

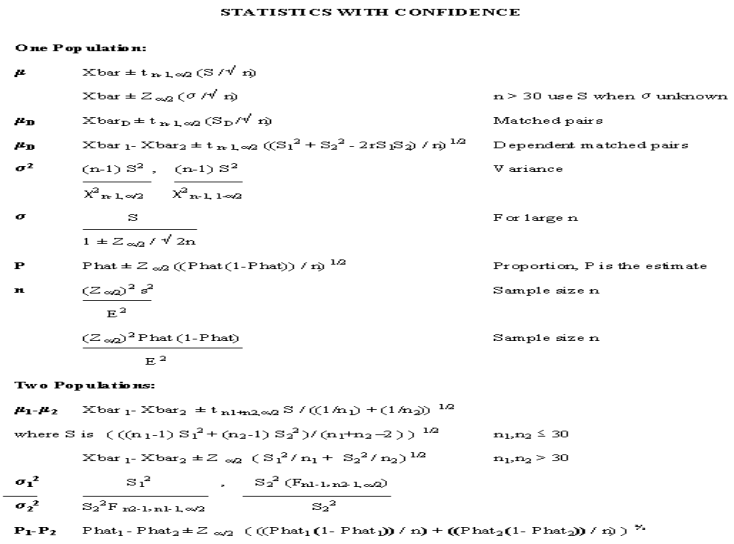
Entienda la relación entre el tamaño de muestra y la anchura del intervalo de la confianza, por otra parte, sepa que el intervalo de confianza calculado algunas veces no contiene al valor verdadero.

Digámosle que se calcula un intervalo de confianza del 95% para una media m . La manera de interpretar esto es imaginar un número infinito de muestras de la misma población, el 95% de los intervalos calculados contendrán la media m de la población, y a lo máximo el 5% no. Sin embargo, es incorrecto indicar, “tengo el 95% de confianza de la media m de la población esta dentro del intervalo.”

Una vez más la definición usual de un intervalo de confianza del 95% es un intervalo construido por un proceso tal que el intervalo contendrá el valor verdadero el 95% del tiempo. Esto significa que el “95%” es una característica del proceso, no el intervalo.

¿Es la probabilidad de ocurrencia de la media poblacional mayor en el centro del intervalo de la confianza (IC) y mas baja en los límites? ¿La probabilidad de la ocurrencia de la media poblacional en un intervalo de confianza que varía de una manera mensurable del centro a los límites? En un sentido general, se asume la condición de la normalidad, y entonces el intervalo entre los límites del IC es representado por una distribución t acampana. La expectativa (e) de otro valor es la más alta en el valor de la media calculada, y disminuye mientras los valores se acercan a los límites del IC.

Intervalo de la Tolerancia y IC: Una buena aproximación para un simple intervalo de tolerancia es veces el intervalo de confianza con respecto a la media $n^{1/2}$.



STATISTICS WITH CONFIDENCE

One Population:

μ	$Xbar \pm t_{n-1,\alpha/2} (S/\sqrt{n})$	
	$Xbar \pm Z_{\alpha/2} (\sigma/\sqrt{n})$	$n > 30$ use S when σ unknown
μ_D	$Xbar_D \pm t_{n-1,\alpha/2} (S_D/\sqrt{n})$	Matched pairs
μ_D	$Xbar_1 - Xbar_2 \pm t_{n-1,\alpha/2} ((S_1^2 + S_2^2 - 2rS_1S_2)/n)^{1/2}$	Dependent matched pairs
σ^2	$\frac{(n-1)S^2}{X^2_{n-1,\alpha/2}}, \frac{(n-1)S^2}{X^2_{n-1,1-\alpha/2}}$	Variance
σ	$\frac{S}{1 \pm Z_{\alpha/2}/\sqrt{2n}}$	For large n
P	$Phat \pm Z_{\alpha/2} ((Phat(1-Phat))/n)^{1/2}$	Proportion, P is the estimate
n	$\frac{(Z_{\alpha/2})^2 s^2}{E^2}$	Sample size n
	$\frac{(Z_{\alpha/2})^2 Phat(1-Phat)}{E^2}$	Sample size n

Two Populations:

$\mu_1-\mu_2$	$Xbar_1 - Xbar_2 \pm t_{n1+n2,\alpha/2} S/((1/n_1)+(1/n_2))^{1/2}$	
where S is	$((n_1-1)S_1^2 + (n_2-1)S_2^2)/(n_1+n_2-2))^{1/2}$	$n_1, n_2 \leq 30$
	$Xbar_1 - Xbar_2 \pm Z_{\alpha/2} (S_1^2/n_1 + S_2^2/n_2)^{1/2}$	$n_1, n_2 > 30$
$\frac{\sigma_1^2}{\sigma_2^2}$	$\frac{S_1^2}{S_2^2 F_{n2-1,n1-1,\alpha/2}}, \frac{S_2^2 (F_{n1-1,n2-1,\alpha/2})}{S_2^2}$	
P_1-P_2	$Phat_1 - Phat_2 \pm Z_{\alpha/2} ((Phat_1(1-Phat_1))/n) + ((Phat_2(1-Phat_2))/n))^{1/2}$	

Estadísticos con Confianza

Teclee en la imagen para agrandarla y LUEGO imprímala

Usted necesita utilizar el Javascript de la Determinación del Tamaño de la Muestra en el diseño de etapas en su investigación estadística en la toma de decisión con requisitos subjetivos específicos.

Una Nota sobre la Comparación Múltiple Vía Intervalos Individuales: Note que, si los intervalos de la confianza a partir de dos muestras no se superponen, existe una diferencia estadística significativa, digamos del 5%. Sin embargo, la otra manera no es verdad; dos intervalos de confianza pueden superponerse incluso cuando hay una diferencia significativa entre ellos.

Como ejemplo numérico, considere las medias de dos muestras independientes. Suponga que sus valores son 10 y 22 con error estándar igual a 4. El intervalo de confianza del 95% para los dos estadísticos (usando el valor crítico de 1,96) es: [2,2, 17,8] y [14,2, 29,8], respectivamente. Como se observa, estos exhiben una considerable superposición. Sin embargo, el estadístico z para la media de la dos poblaciones es: $|22 -10|/(16 + 16)^{1/2} = 2,12$ el cual es claramente significativo bajo las mismas condiciones aplicadas al construir los intervalos de confianza.

Se deberían examinar el intervalo de confianza para la diferencia explícita. Incluso si los intervalos de confianza están superpuestos, es difícil encontrar el nivel exacto de confianza. Sin embargo, la suma de niveles individuales de confianza pueden servir como límite superior. Esto es evidente en el hecho de que: P(A y B) £ P(A) + P(B).

¿Que es el Margen de Error?

La estimación es el proceso mediante el cual los datos de la muestra son utilizados para indicar el valor de una cantidad desconocida en una población.

Los resultados de un estimador pueden ser expresados como un simple valor; entendido como una estimación en un punto, o un rango de valores, referido como un intervalo de confianza.

Siempre que utilicemos la valoración del punto, calculamos el margen de error asociado a esa estimación del punto. Por ejemplo, para la estimación de la proporción de la población, por las medias de una muestra de proporciones (p), el margen del error se calcula ***a menudo*** como sigue:

$\pm 1{,}96\ [p(1-p)/n]^{1/2}$

En periódicos e informes de televisión sobre encuestas de la opinión pública, el margen del error aparece a menudo en ***caracteres pequeños*** en el fondo de una tabla o de una pantalla. Sin embargo, divulgar solamente la cantidad de error, no es bastante informativo por sí mismo, lo que falta es el **grado de confianza** en los resultados. El pedazo de información mas importante y que falta es el tamaño de la muestra n; Es decir, ¿Cuánta gente participó en la encuesta, 100 o 100000 Para este momento, usted sabe bien que cuanto más grande es el tamaño de muestra más exacto es el resultado, ¿no?.

El margen de error reportado es el margen de “error de muestreo”. Hay muchos errores de no-muestreo que pueden y afectar la exactitud de las encuestas. Aquí hablamos de error de muestreo. El hecho de que subgrupos pudieron tener error de muestreo más grande que el grupo de donde provienen, debería generar la inclusión de la declaración siguiente en el informe:

> “Otras fuentes del error incluyen, pero no se limitan a, individuos que rechazan participar en la entrevista, y la inhabilidad de hacer contacto con el número seleccionado. Cada esfuerzo factible fue hecho para obtener una respuesta y reducir el error, pero el lector (o el espectador) debe estar enterado de un cierto error inherente a toda investigación.”

Si usted tiene preguntas de tipo si/ no en un examen, usted probablemente desearía calcular una proporción P de los sí’s (o los no’s). En una encuesta de a una muestra aleatoria simple, la varianza de p es p(1-p)/n, no haciendo caso a la corrección de la población finita, de tamaño n, digamos mayor a 30. Ahora un intervalo de confianza del 95% es:

$p - 1{,}96\,[p(1-p)/n]^{1/2}$, $p + 1{,}96\,[p(1-p)/n]^{1/2}$.

Un intervalo conservador puede ser calculado, puesto que p(1-p) toma su valor máximo cuando p = 1/2. Substituya 1,96 por 2, ponga p = 1/2 y usted tiene un 95% de intervalo conservativo de confianza de $1/n^{1/2}$. Esta aproximación tiene un buen funcionamiento siempre y cuando p no este muy cerca de 0 o de 1. Esta aproximación útil le permite calcular intervalos aproximados de confianza de 95%.

Para las variables aleatorias continuas, tales como la estimación de la media poblacional m, el margen de error se calcula ***a menudo*** como sigue:

$\pm 1{,}96\ S/n^{1/2}$.

El margen del error se puede reducir por una o combinación de las siguientes estrategias:

13. Disminuyendo la confianza en la estimación -- una estrategia indeseable puesto que la confianza se relaciona con la oportunidad de dibujar una conclusión incorrecta (es decir, aumentos del error Tipo II).
14. Reduciendo la desviación estándar -- algo que no podemos hacer puesto que es generalmente una característica estática de la población.
15. Aumentando el tamaño de muestra -- esto proporciona más información para una mejor decisión.

A usted podría gustarle usar el JavaScript de Estadística Descriptiva para comprobar sus cálculos, y el Javascript de Determinación del Tamaño de la Muestra en la etapa del diseño de su investigación estadística en la toma de decisión con requisitos subjetivos específicos.

Técnicas de Reducción de Preferencias: Bootstrapping y Jackniffing

Algunas técnicas de estadísticas inferencial no requieren distribución de asunciones sobre la estadística implicada. Estos métodos modernos no paramétricos utilizan cantidades grandes de cálculos para explorar la variabilidad empírica de un estadístico, en vez de hacer asunciones a priori sobre esta variabilidad, como se hace en las pruebas paramétricas tradicionales t y z.

Bootstrapping: Este método se usa con correa se usa para obtener una estimación combinando los estimadores a cada uno de las muchas sub muestras de un conjunto de datos. Normalmente, M muestras aleatoriamente dibujadas de T observaciones son dibujadas de los datos originales de tamaño n con el reemplazo, donde T es menor que n.

El Estimador Jackknife: Este estimador crea una serie de estimaciones de un simple conjunto de datos, generando que el estadístico repetidamente abandone un valor de lo datos cada vez. Esto produce una estimación de la media del parámetro y una desviación estándar de las estimaciones del parámetro.

La simulación de Monte Carlo permite la evaluación del comportamiento de un estadístico cuando su análisis matemático es óptimo. Bootstrapping y Jackniffing permiten que las inferencias sean hechas de la muestra cuando la inferencia paramétrica tradicional falla. Estas técnicas son especialmente útiles para lidiar con problemas estadísticos, tales como un tamaño de muestra muy pequeño, estadísticas sin teoría distribucional bien desarrollada, y de violaciones paramétricas de la condición de la inferencia. Ambas son intensivas en el uso de computadoras. Bootstrapping significa que usted toma muestras repetidas de otras muestras y de allí saca conclusiones sobre una población. Bootstrapping y Jackknife exigen el muestreo-con-reemplazo de una muestra. Jackniffing sistemáticamente envuelve el hacer n pasos, de omitir 1 caso de una muestra a la vez, o, de manera mas general, n/ k pasos de omitir k casos; los cálculos que comparan “incluido” contra “omitidos” pueden ser utilizado (especialmente) para reducir las preferencias de los estimadores. Ambos tienen aplicaciones en reducir las preferencias en las estimaciones.

Re muestreo-- incluyendo Bootstrapping , la permutación, y otras pruebas no paramétricas -- es un método para la prueba de la hipótesis, los límites de confianza, y otros problemas aplicados en estadística y probabilidad. No implica ninguna fórmula o tabla.

Después de la primera publicación de la técnica general (y Bootstrapping) en 1969 de Julian Simon y el desarrollo subsecuente independiente por Bradley Efron, el re muestreo se ha convertido en un acercamiento alternativo para las pruebas de hipótesis.

Existen otros resultados: “Bootstrapping comenzó como buena noción de lo que el presentó, en teoría, un procedimiento estadístico elegante que estaba libre de condiciones distribucionales. La técnica del Bootstrapping no es muy eficiente en la práctica, y los intentos por modificarlo la hacen mas complicada y más confusa que los procedimientos paramétricos al cual estaba supuesto a reemplazar.”

Mientras que las técnicas de re muestreo pueden reducir las preferencias, estas alcanzan esto a expensas de aumento en la varianza. Las dos preocupaciones principales son:

16. La pérdida en la exactitud de la estimación según lo medido por la varianza puede ser muy grande.
17. La dimensión de los datos afecta drásticamente la calidad de las muestras y por lo tanto de las estimaciones.

Intervalos de Predicción

En muchos uso de la estadística de negocio, tal como pronóstico, estamos interesados en la construcción de un intervalo estadístico para la variable aleatoria, en vez de un parámetro de una distribución de la población.

- La desigualdad del Tchebysheff se utiliza a menudo para poner los límites en la probabilidad que una proporción de la variable aleatoria X estará dentro k > 1 desviación estándar con respecto a la media m para cualquier distribución de probabilidad. En otras palabras:

  P [|X - m| ³ k s] £ $1/k^2$, para cualquier k mayor a 1

  La propiedad de simetría de la desigualdad de Tchebysheff es útil; por ejemplo, construyendo límites de supervisión en el proceso de control de calidad. Sin embargo, los límites son muy conservadores debido a la carencia del conocimiento sobre la distribución subyacente.

- Los límites antedichos pueden ser mejorados (es decir, ser más apretado) si tenemos cierto conocimiento sobre la distribución de la población. Por ejemplo, si la población es homogénea; es decir, su distribución es unimodal; entonces,

  P [|X - m| ³ k s] £ $1/(2{,}25k^2)$, para cualquier k mayor a 1.

  La desigualdad anterior se conoce como la desigualdad del Campo-Meidell.

- Ahora, deje que X sea una variable aleatoria distribuida normalmente con media estimada x y desviación estándar S, entonces el intervalo de la predicción para la media muestral x con 100(1- a)% nivel de confianza confidencees:

  $x \pm t_{a/2} \cdot S \cdot (1+1/n)^{1/2}$.

  Este es el rango de una variable aleatoria x con 100(1- a)% de confianza , usando la tabla t. Descansando en la Condición de Normalidad para predicciones del intervalo de la media muestral, requiere una muestra de gran tamaño, digamos n mayor 30.

¿Que es un Error Estándar?

Para la inferencia estadística, digamos una prueba estadística y de estimación, se necesita estimar los parámetros de la población. La estimación implica la determinación, con un error posible debido al muestreo, del valor desconocido de un parámetro de la población, tal como la proporción que tiene una cualidad específica o el valor medio m de una cierta medida numérica. Para expresar la exactitud de las estimaciones de las características de la población, se debe también calcular los errores estándar de las estimaciones. Éstas son las medidas de exactitud que determinan los errores posibles que se presentan del hecho de que las estimaciones están basadas en muestras escogidas al azar de la población entera, y no en un censo completo de la población.

El error estándar es un estadístico que indica la exactitud de una estimación. Es decir, nos dice cuan diferente la estimación (como $\bar{x}$) es del parámetro de la población (como m). Por lo tanto, esta es la desviación estándar de una distribución muestral para un estimador como $\bar{x}$. Los siguientes son una colección de errores estándar para la extensamente usada estadística:

- Error Estándar para la Media $\bar{x}$is: $S/n^{½}$.

  Como cualquiera esperaría, el error estándar disminuye mientras que el tamaño de la muestra aumenta. Sin embargo la desviación estándar de la estimación disminuye por un factor del $n^{½}$ no n. Por ejemplo, si usted desea reducir el error en 50%, el tamaño de la muestra debe ser 4 veces n, lo cual es costoso. Por lo tanto, como alternativa a incrementar el tamaño de la muestra, se puede reducir el error obteniendo los datos de “calidad” el cual proporciona una estimación más exacta.

- Para una población finita de tamaño N, el error estándar de la media muestral de tamaño n, es:

  $S \acute{} [(N -n)/(nN)]^{½}$.

- El Error Estándar para la multiplicación de dos Medias independientes $\bar{x}_1 \acute{} \bar{x}_2$ es:

  $\{\bar{x}_1 S_2^2/n_2 + \bar{x}_2 S_1^2/n_1\}^{½}$.

- El Error Estándar para dos medias Dependientes $\bar{x}_1 \pm \bar{x}_2$ es:

  $\{S_1^2/n_1 + S_2^2/n_2 + 2\, r \acute{} [(S_1^2/n_1)(S_2^2/n_2)]^{½}\}^{½}$.

- El Error Estándar para la Proporción P es:

  $[P(1-P)/n]^{½}$

- El Error Estándar para $P_1 \pm P_2$, dos Proporciones dependientes es:

  $\{[P_1 + P_2 - (P_1-P_2)^2] / n\}^{½}$.

- El Error Estándar de la Proporción (P) de población finita es:

  $[P(1-P)(N -n)/(nN)]^{½}$.

  Las dos fórmulas para poblaciones finitas, normalmente se utilizan cuando se desea comparar una sub-muestra de tamaño n con una muestra más grande del tamaño N, el cual contiene la sub-muestra. En tal comparación, sería incorrecto tratar las dos muestras "como si" existieran dos muestras independientes. Por ejemplo, comparando las dos medias uno puede utilizar el estadístico t pero junto a el error de estándar:

  $S_N [(N -n)/(nN)]^{½}$

  como su denominador. Un tratamiento similar es necesario para proporciones.

- El Error Estándar de la pendiente (m) en la Regresión Lineal es:

  $S_{res} / S_{xx}^{½}$, donde S_{res} es el residuo de la desviación estándar.

- El Error Estándar de la Intercepción (b) en la Regresión Lineal es:

  $S_{res}[(S_{xx} + n \cdot \bar{x}^2) /(n \cdot S_{xx}]^{½}$.

- El Error Estándar del Valor Estimado usando la Regresión Lineal es:

  $S_y(1 - r^2)^{½}$.

  El termino $(1 - r^2)^{½}$ es llamado el coeficiente de alineación. Por lo tanto si r = 0, el error de la predicción es S_y como se esperaba.

- El Error Estándar de la Regresión Lineal es:

  $S_y (1 - r^2)^{½}$.

  Observe que si r = 0, el error estándar alcanza su valor máximo posible, que es la desviación estándar en Y.

Estabilidad de un Estimador: Un estimador es estable si, tomando dos diversas muestras del mismo tamaño, producen dos estimaciones que tienen "pequeñas" diferencia absoluta. La estabilidad de un estimador es medida por su confiabilidad:

Confiabilidad de un estimador = 1/(su error estándar) ^{2}

Cuanto más grande es el error de estándar, menos confiable es la estimación. La confiabilidad de estimadores se utiliza a menudo para seleccionar el “mejor” estimador entre todos los estimadores imparciales.

Determinación del Tamaño de la Muestra

En la etapa de planeamiento de una investigación estadística, la pregunta sobre el tamaño de la muestra (n) es crítica. Esto es una cuestión importante que NO se debe tomar a la ligera. Tomar una muestra más grande que lo necesario para alcanzar los resultados deseados es derrochar los recursos, mientras que las muestras muy pequeñas conducen a menudo a ningún uso práctico para tomar buenas decisiones. El objetivo principal es obtener tanto una exactitud deseable y un nivel deseable de la confianza con mínimos costos.

Estudiantes algunas veces me preguntan, ¿Qué fracción de la población usted necesita para una buena estimación? Yo contesto, “esto es irrelevante; la exactitud es determinada por el tamaño de la muestra.” Esta respuesta tiene que ser modificada si la muestra es una fracción importante de la población.

El nivel de la confianza de las conclusiones dibujadas de un sistema de datos depende del tamaño de los datos. Cuanto más grande es la muestra, más alta es la confianza asociada. Sin embargo, muestras más grandes también requieren más esfuerzo y recursos. De esta forma, su objetivo debe ser encontrar el tamaño de muestra más pequeño que proporcionará la confianza deseable.

Para un artículo anotado 0 o 1, para no o sí, el error estándar (EE) de la proporción estimada p, basado en sus observaciones de la muestra aleatoria, se ubica cerca de:

$$EE = [p(1-p)/n]^{1/2}$$

donde p es la proporción de obtener una cuenta de 1, y n es el tamaño de muestra. Este EE es la desviación estándar del rango de los valores posibles de la estimación.

El EE está en su máximo cuando p = 0,5, por lo tanto el peor escenario del caso ocurre cuando los 50% son sí, y los 50% son no.

Bajo esta condición extrema, el tamaño de muestra, n, se puede entonces expresar como el número entero más grande menor que o igual a:

$$n = 0{,}25/EE^2$$

Para tener cierta noción del tamaño de la muestra, por ejemplo para que el EE sea 0,01 (es decir el 1%), un tamaño de muestra de 2500 será necesario; el 2%, 625; el 3%, 278; el 4%, 156, el 5%, 100.

Nota, incidentalmente, mientras la muestra sea una fracción pequeña de la población total, el tamaño real de la población es enteramente irrelevante para los propósitos de este cálculo.

Estudios Experimentales (Pilotos): Cuando las estimaciones necesarias para el cálculo del tamaño de muestra no están disponibles en una base de datos existente, un estudio experimental es necesario para una adecuada estimación con una precisión dada. Una muestra piloto, o preliminar, debe ser dibujado de la población, y los estadísticos calculados de esta muestra se utilizan en la determinación del tamaño de muestra. Las observaciones usadas en la muestra experimental se pueden contar como parte de la muestra final, de modo que el tamaño de muestra calculada menos el tamaño de muestra experimental es el número de observaciones necesarias para satisfacer el tamaño de muestra requerido.

Tamaño de muestra con la precisión absoluta aceptable: La siguiente presenta el método mas usado para determinar el tamaño de muestra requerido para estimar la media y la proporción de la población.

Supongamos que deseamos un intervalo que extienda en d unidades en cualquier lado del estimador. Podemos escribir

d = Precisión Absoluta = (coeficiente de confiabilidad) ´ (error estándar) = $Z_{a/2}$ ´ ($S/n^{1/2}$)

Suponga, basado en una muestra experimental de tamaño n, la proporción estimada es p, el tamaño requerido de la muestra con el tamaño absoluto de error que no excede d, con 1- a de confianza:

$[t^2 \, n \, p(1-p)] / [t^2 \, p(1-p) - d^2 \, (n-1)]$,

donde $t = t_{a/2}$ siendo el valor tomado de la tabla t con parámetros gl= n = n-1, respectivamente al intervalo de confianza deseado 1- a.

Para muestras pilotos grandes (n mayor a 30), la manera mas simple de determinar el tamaño de la muestral es:

$[(Z_{a/2})^2 \, S^2] / d^2$ para la media m

$[(Z_{a/2})^2 \, p(1-p)] / d^2$ para la proporción,

donde d es el margen deseable de error (es decir, el error absoluto), que es la mitad del intervalo de confianza con 100(1- a)%.

Tamaño de Muestra con Errores Tipo I y Tipo II Aceptables: Se puede utilizar el siguiente tamaño de muestra determinado, el cual se basa en el tamaño del error tipo I y error tipo II:

$2(Z_{a/2} + Z_{b/2})^2 S^2/d^2$,

de donde a and b son los errores aceptables tipo I y tipo II, respectivamente. S^2 es la variación obtenida de la corrida piloto, y d es la diferencia entre la nula y la alternativa (m_0 -m_a).

Tamaño de Muestra con Precisión Relativamente Aceptable: Se puede utilizar el siguiente tamaño de muestra determinado para un **error relativo deseable** D en %, el cual requiere una estimación del coeficiente de variación (CV en %) de una muestra experimental con tamaño mayor a 30:

$[(Z_{a/2})^2 (C.V.)^2] / D^2$

Tamaño de Muestra Basado en la Hipótesis Nula y la Alternativa: Se puede utilizar **el poder de la prueba** para determinar el tamaño de la muestra. La relación funcional de la capacidad y del tamaño de la muestra se conoce como la **curva característica de funcionamiento**. En esta curva, cuando el tamaño de la muestra aumenta, la función de capacidad aumenta rápidamente. Dejemos que d sea:

$m_a = m_0 + d$

tal que es un alternativa para representar la salida desde la hipótesis nula. Deseamos ser razonablemente confidentes de encontrar evidencia contra la hipótesis nula, si, el hecho particular de la alternativa se mantiene. Es decir, el error tipo b, es la probabilidad de fallar por no poder encontrar evidencia de por lo menos en el nivel de a, cuando la alternativa se mantiene. Esto implica:

Tamaño de la muestra requerido= $(z_1 + z_2) S^2/ d^2$

De donde: z_1 = |media - m_0|/ EE, z_2 = |media - m_a|/ EE, la media es la estimación actual para m, y S lo es para s.

Todos los determinantes anteriores del tamaño de la muestra se podrían también utilizar para estimar la media de cualquier población unimodal, con variables aleatorias discretas o continuas, con una corrida piloto de n mayor a 30.

En la estimación del tamaño de la muestra, cuando la desviación estándar no se conoce, en vez de usar S^2 se puede utilizar 1/4 del rango del tamaño de la muestra mayor a 30 como una “buena” estimación para la desviación estándar. Esta es una buena práctica comparar los resultados con IQR/1,349.

Se podría extender la determinación del tamaño de la muestra a otra estadístico útil, tal como el coeficiente de correlación (r)basado en errores aceptables tipo I y tipo II:

$2 + [(Z_{a/2} + Z_{b/2}(1- r^2)^{½})/r]^2$

el r proporcionado no es igual a -1, 0, o 1.

El atino de aplicar cualquiera de determinantes para el tamaño de muestra anteriormente expuestos, está en mejorar sus estimaciones pilotos a costes factibles.

A usted podía gustarle utilizar el Javascript de Determinación del Tamaño de Muestra para comprobar sus cálculos.

Revisando el Valor Esperado y la Varianza

Varianzas Promediadas: ¿Cuál es la varianza media de k varianzas sin tomar en consideración el tamaño de sus muestras? La respuesta es simple:

Promedio de las Varianzas = $[SS_i^2] / k$

Sin embargo, ¿Cuál es la varianza de todos los grupos de k combinados? La respuesta debe considerar el tamaño de la muestra n_i del iesimo grupo:

Grupo de Varianzas Combinadas = $S\ n_i[S_i^2 + d_i^2]/N$,

donde d_i = $media_i$ - gran media, y $N = S\ n_i$, para todas las i = 1, 2, .., k.

Note que la fórmula anterior permite que dividamos la varianza total en sus dos componentes. Este proceso nos permite determinar el grado al cual la varianza total es afectada por la diferencia entre las medias del grupo. ¿Cuál sería la variación si todos los grupos tienen la misma media? ANOVA es una aplicación ampliamente conocida de este concepto donde la igualdad de varias medias se prueba.

Media Subjetiva y Varianza: En muchas aplicaciones, hemos visto cómo tomar decisiones basadas en datos objetivos; sin embargo, un tomador de decisiones podría tener la capacidad de combinar su interpretación subjetiva y usar las dos fuentes de información.

Una aplicación: Suponga que la siguiente información se encuentra disponible a partir de dos fuentes independientes:

Revisando el Valor Esperado y la Varianza

Fuente de Estimación	Valor Esperado	Varianza
Gerente de Ventas	$m_1 = 110$	$s_1^2 = 100$
Estudio de Mercado	$m_2 = 70$	$s_2^2 = 49$

El valor esperado combinado es:

$[m_1/s_1^2 + m_2/s_2^2] / [1/s_1^2 + 1/s_2^2]$

La varianza combinada es:

$2 / [1/s_1^2 + 1/s_2^2]$

Para nuestra aplicación, usando la información tabular anterior, la estimación combinada de las ventas es 83,15 unidades con una varianza combinada de 65,77.

A usted podía gustarle utilizar el Javascript de Revisando el Valor Esperado y la Varianza en la ejecución de ciertas experimentaciones numéricas. Usted podía aplicarla para validar el ejemplo anterior y para una comprensión más profunda del concepto de donde más de dos fuentes de información van a ser combinadas.

Evaluación Subjetiva de varias Estimaciones basadas en Relativa Precisión

En muchos casos, desearíamos comparar varias estimaciones del mismo parámetro. El acercamiento más simple es medir la mas cercana entre todas las estimaciones en un intento de determinarse que por lo menos una de las estimaciones es mayor a r veces el parámetro de distancia al otro parámetro, donde r es un número subjetivo, no negativo menor que uno.

A usted podía gustarle utilizar el Javascript de Evaluación Subjetiva de Estimaciones para aislar cualquier estimación inexacta. Repitiendo el mismo proceso usted podría eliminar las estimaciones inexactas

---

Inferencia Estadística Bayesiana: Una Introducción

La inferencia estadística describe los procedimientos mediante los cuales nosotros observamos los datos, de forma tal de *establecer conclusiones* acerca de una población de la cual los datos han sido obtenidos o con respecto al proceso mediante el cual los datos fueron generados. Nosotros asumimos que existe un proceso desconocido que genera los datos que tenemos y que este proceso puede ser descrito mediante una probabilidad de distribución, la cual, en etapas, puede ser

caracterizada por algunos parámetros desconocidos. Por lo tanto, para una distribución normal los parámetros desconocidos son m y s^2.

En términos mas generales, la inferencia estadística puede ser clasificada bajo dos encabezados: La inferencia Clásica y la Inferencia Bayesiana. La inferencia estadística clásica se basa en dos premisas:

32. La muestra de datos constituye la única información relevante.
33. La construcción y la evaluación de los diferentes procedimientos para la inferencia están basados en comportamientos a largo plazo bajo circunstancias esencialmente similares.

En la Inferencia Bayesiana se combina la información de la muestra con la información previa. Supongamos que tenemos una muestra aleatoria x1, x2,....xn de tamaño n de una población normal.

En la inferencia estadística tomamos la media muestral $\bar{x}$ como nuestra estimación de m. Su varianza es s^2 / n. La inversa de esta varianza es conocida como la precisión muestral. Por lo tanto, esta precisión es n / s^2.

En la inferencia Bayesiana tenemos información previa de m. Esto es expresado en términos de una funcionde de distribución de probabilidad conocida como *distribución previa*. Suponga que dicha distribución previa es normal con media m_0 y varianza s_0^2, esto es, precisión 1/ s_0^2 . Ahora sabemos combinar la información de la muestra para obtener lo que es conocido como la distribución posterior de μ. Esta distribución puede ser mostrada como una normal. Esto significa que es un average ponderado de la media muestral y la media anterior, ponderada por la precisión de la muestra y la precisión anterior respectivamente, por lo tanto

Media Posterior = $(W_1\bar{x} + W_2\, m_0) / (W_1 + W_2)$

Varianza Posterior = $1 / (W_1 + W_2)$

de donde

W1 = Precisión Muestral = n/S^2, y W1 = Precisión Previa = n/s_0^2

Adicionalmente, la precisión (o el inverso de la varianza) de la distribución posterior de m es $W_1 + W_2$, el cual es, la suma de la precisión muestral y la precisión previa.

La media posterior descansará entre la media muestral y la media previa. La varianza media posterior será menor que ambas, la varianza muetral y previa.

En este sitio Web no se discute la inferencia Bayesiana por que esto nos adentraría a muchos mas detalles de los que intentamos cubrir. Sin

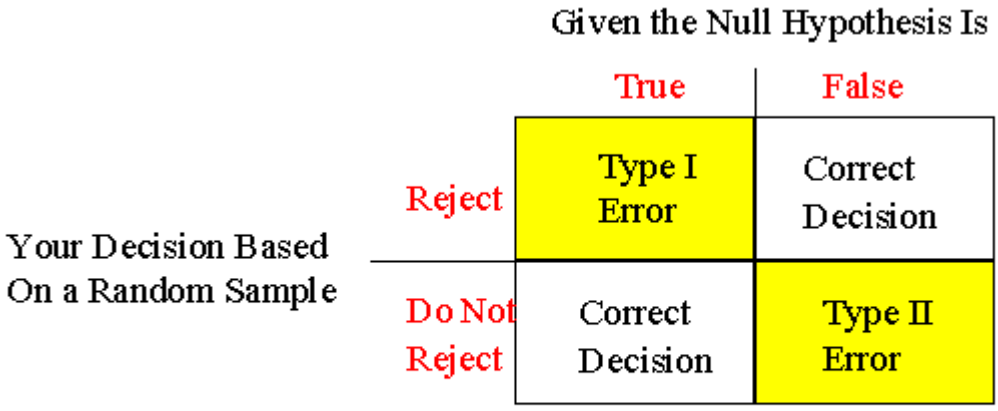
embargo, la noción básica de combinar la media muestral y la media previa en proporción inversa a sus varianzas respectivas, será interesante mientras proporcione algún uso útil.

A usted podría gustarle utilizar el JavaScript de Inferencia Estadística Bayesiana para comprobar sus cálculos y para realizar algunas experimentaciones.

Gerencia del Riesgo de los Productores y el Riesgo de los Consumidores

La lógica detrás de una prueba de hipótesis estadística es similar a la lógica siguiente. Dibuje dos líneas en un papel y determínese si están tienen diferentes longitudes. Compárelas y diga, “bueno, ciertamente no son iguales”. Por lo tanto ellas tienen que ser de longitudes diferentes. Rechazando igualdad, es decir, la hipótesis nula, usted afirma que hay una diferencia.

La potencia de la prueba estadística es mejor explicada mediante la descripción de los errores Tipo I y Tipo II. La matriz siguiente demuestra la representación básica de estos errores.



Two Types of Errors in Decision Making

Según lo indicado en la matriz anterior, un Error Tipo I ocurre cuando, basado en sus datos, usted rechaza la hipótesis nula cuando de hecho es verdad. La probabilidad de un error Tipo I es el nivel de la significancia de la prueba de la hipótesis y es denotada por a .

El error Tipo I es llamado a menudo **el riesgo del productor** de que los consumidores rechazan un buen producto o servicio indicado por la hipótesis nula. Es decir, un productor introduce un buen producto en el mercado, y de esta forma, él o ella toma el riesgo de que el consumidor lo rechazará.

Un error Tipo II ocurre cuando usted no rechaza la hipótesis nula y está es de hecho falsa. La probabilidad de un error Tipo II se denotada por b . La cantidad 1 - b se conoce como la Potencia o capacidad de una prueba. Un error Tipo II se puede evaluar para cualquier hipótesis alternativa específica, indicada mediante la forma “no igual” como la hipótesis competitiva.

El error Tipo II es a menudo llamado el **riesgo del consumidor** de no rechazar un producto o servicio posiblemente malo indicado por la hipótesis nula.

Los estudiantes a menudo fórmulas preguntas tales como ¿cuáles son los intervalos de confianza “correctos”, y porqué la mayoría de la gente utiliza el nivel de 95%? La respuesta es que los tomadores de decisiones deben considerar ambos errores Tipo I y Tipo II y obtener la mejor compensación posible. Idealmente, se desea reducir la probabilidad de hacer estos tipos de errores; sin embargo, para un tamaño de muestra fijo, no podemos reducir un tipo de error sin que al mismo tiempo estemos aumentando la probabilidad del otro tipo de error. No obstante, reducir las probabilidades de ambos tipos de errores es simultáneamente aumentar el tamaño de la muestra. Es decir, teniendo más información se toman mejores decisiones.

El siguiente ejemplo destaca este concepto. Una firma de componentes electrónicos, Big Z, fabrica y vende una pieza a un fabricante de radio, Big Y. Big Z mantiene constantemente un porcentaje de piezas defectuosas de el 10% por cada 1000 unidades producidas. Aquí Big Z es el productor y Big Y es el consumidor. Big Y, por razones del sentido práctico, probará una muestra de 10 piezas de1000 lotes. Big Y adoptará una de las dos reglas con respecto a la aceptación de una proposición:

- Regla 1: Aceptar lotes con una o menos piezas defectuosas; Por lo tanto, el lote cuanto mucho tiene 0 o 1 defectuosa.
- Regla 2: Aceptar los lotes con dos o menos piezas defectuosas; Por lo tanto, el lote cuanto mucho tiene 0, 1 o 2 defectuosas.

Con base en la distribución binomial, la P(0 o 1) es 0,7367. Esto significa que, con un índice defectuoso del 10%, Big Y aceptará el 74% de las unidades probadas y rechazará el 26% de las mismas, sin importar que los lotes sean buenos. Este 26% es el riesgo del productor o el nivel a . Este nivel a es análogo a un error Tipo I -- rechazar una hipótesis nula verdadera. O, en otras palabras, rechazando un buen lote. En este ejemplo, para propósitos de ilustración, los lotes representan una hipótesis nula. La porción rechazada va de nuevo al productor; por lo tanto, es el riesgo del productor. Si Big Y tomara la regla 2, el riesgo del productor disminuiría. La P(0, o 1, o 2) es 0,9298 por lo tanto, Big Y aceptará el 93% de todas las porciones probadas, y el 7% serán rechazadas, aunque la porción sea aceptable. La razón principal es que, aunque la probabilidad de artículos defectuosos es 10%, Big Y a través

de la regla 2 permite una tasa de aceptación de artículos defectuosos más alta. Según lo indicado anteriormente, Big Y aumenta su propio riesgo (riesgo del consumidor) como se asumio previamente.

Tomando una Buena Decisión: Dado que existe un beneficio importante (que podría ser negativo) para el resultado de su decisión, y una probabilidad previa (antes de probar) para que la hipótesis nula sea verdad, el objetivo es tomar una buena decisión. Denotemos los beneficios para cada célula en la tabla de decisión como $a, $b, $c y $d (en el orden de las columnas), respectivamente. La expectativa del beneficio es [aa + (1-a)b], y + [(1-b)c + bd], dependiendo de que si la hipótesis nula es verdadera.

Ahora teniendo una probabilidad subjetiva previa (es decir, antes de probar) de p de que la hipótesis nula es verdadera, el beneficio previsto de su decisión es:

Beneficio Neto = [aa + (1-a)b]p + [(1-b)c + bd](1-p) – Costos de Muestreo

Una buena decisión hace este beneficio tan grande como sea posible. Con este fin, debemos elegir convenientemente el tamaño de la muestra y el resto de los factores de la función de beneficio.

Observe que, puesto que estamos utilizando una probabilidad subjetiva que expresa la fuerza de la creencia de la verdad de la hipótesis nula, esta es llamada una **Aproximación Bayesiana** a la toma de decisiones estadísticas, que es un acercamiento estándar en teoría de decisiones.

A usted podía gustarle utilizar el JavaScript de Subjetividad en las Pruebas de Hipótesis en Javascript en la ejecución de una ciertas experimentación numéricas para validar los resultados anteriores y obtener una comprensión más profunda.

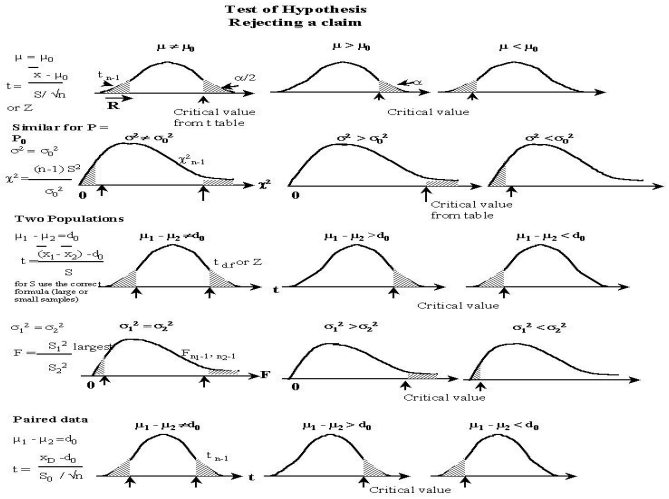
Prueba de Hipótesis: Rechazando una Proposición

Para realizar una prueba de hipótesis, se debe ser muy específico sobre la prueba que se desea realizarse. La hipótesis nula debe ser indicada claramente, y los datos se deben recoger de una manera repetible. Si existe alguna subjetividad, los resultados son técnicamente inválidos. Todos los análisis, incluyendo el tamaño de la muestra, nivel de significancia, el tiempo, y el presupuesto, se deben planear por adelantado, o bien el usuario corre el riesgo de “datos sumergidos.”

La prueba de hipótesis es una prueba matemática por contradicción. Por ejemplo, para la prueba t de student la cual compara a dos grupos, asumimos que los dos grupos vienen de la misma población (las mismas medias, desviaciones estándar, y en general las mismas distribuciones). Entonces hacemos nuestro mejor para probar que esta asunción es

falsa. Rechazar H_0 significa que, o H_0 es falso, o un acontecimiento raro ha ocurrido.

La pregunta verdadera en estadística no es saber si una hipótesis nula está correcta, es saber si está bastante cercana ser utilizada como aproximación.



Prueba de Hipótesis
Teclee en la imagen para agrandarla y LUEGO imprímala

En la mayoría de las pruebas estadísticas referentes a m, comenzamos asumiendo que s^2, y los momentos más altos, tales como oblicuidad y kurtosis, son iguales. Luego, inferimos que las a's son iguales, lo que es la hipótesis nula.

Lo “nulo” normalmente sugiere ninguna diferencia entre las medias del grupo, o ninguna relación entre las variables cuantitativas, etcétera.

Consecuentemente probamos con un valor t calculado. Por simplicidad, suponga que tenemos una prueba de dos lados. Si el t calculado está cerca de 0, decimos que “es bueno”, como esperamos. Si el t calculado está lejos de 0, decimos, “la ocasión de conseguir este valor t, dado mi asunción de que las poblaciones son estadísticamente iguales, es tan pequeña que no creería la asunción. Diremos que las poblaciones no son iguales; las medias no son específicamente iguales.”

Como ejemplo, hagamos un esquema con una distribución normal de media x̄1 - x̄2 y desviación estándar s. Si la hipótesis nula es verdadera, la media es 0. Calculamos el valor de ' t ', según la ecuación. Buscamos un valor “crítico” de t. La probabilidad de calcular un valor de t más extremo (+ o -) que esto, dado que la hipótesis nula es verdad, es igual o menor que el riesgo a que utilizamos para obtener el valor crítico de la tabla. Marque el t calculado, y el t crítico (ambos lados) en el esquema de la distribución. Ahora, si el t calculado es más extremo que el valor crítico, decimos, “el chance de conseguir este valor t, por la ocasión de descartar, cuando la hipótesis nula es verdadera, es tan pequeña que

preferiríamos decir que la hipótesis nula es falsa, y aceptar el alternativa, de que las medias no son iguales". Cuando el valor estimado es menos extremo que el valor calculado, decimos, "podría conseguir este valor de t por el chance de descarte. No puedo detectar una diferencia en las medias de los dos grupos en un nivel a de significancia."

En esta prueba, necesitamos (entre otras) la condición de que las varianzas poblacionales (es decir, el tratamiento afecta la tendencia central pero no la variabilidad) son iguales. Sin embargo, esta prueba es rígida a las violaciones de esa condición si las n's son grandes y casi del mismo tamaño. Un ejemplo contrario sería intentar una prueba t entre (11, 12, 13) y (20, 30, 40). La prueba de valores agrupados y no agrupados da un estadístico t de 3,10, pero los grados de libertad son diferentes: gl. = 4 (para los agrupados) o gl cerca de 2 (para los no agrupados). Por lo tanto la prueba reunida da p = 0,036 y la no agrupada p = 0,088. Podríamos bajar a n = 2 y todavía conseguir algo más extremo.

A usted podía gustarle utilizar Cálculo Estadístico en Línea, Probando la Media , y Probando la Varianza en la ejecución de más de estas pruebas.

Usted podría necesitar utilizar el Javascript de la Determinación del Tamaño de la Muestra en la etapa de diseño de su investigación estadística en la toma de decisión con requisitos subjetivos específicos.

Acercamiento Clásico a la Prueba de Hipótesis

En este tratamiento existen dos lados: Un lado (o una persona) propone la hipótesis nula (la proposición). El otro lado propone una hipótesis alternativa. Un nivel de significancia a y un tamaño de muestra n son convenidos por ambas partes. El paso siguiente es calcular los estadísticos relevante basados en la hipótesis nula y la muestra escogida al azar de tamaño n. Finalmente, se determina la **región del rechazo**. La conclusión basada en este acercamiento es:

Si el estadístico calculado cae dentro de la región del rechazo, entonces se rechaza la hipótesis nula; si esto no ocurre esto, No rechace la hipótesis nula (la proposición)..

Usted podría preguntarse: ¿Cómo determinar el valor crítico (por ejemplo valor z) para el intervalo del rechazo en una hipótesis de una y dos colas?. ¿Cuál es la regla?

Primero, usted tiene que elegir un nivel de significación a. Sabiendo que la hipótesis nula siempre esta en la forma de "igualdad", la hipótesis alternativa tiene una de las tres formas posibles: "mayor que", "menor que", o "no igual a". Las primeras dos formas corresponden a una

hipótesis de una cola, mientras que la tercera corresponde a una hipótesis de dos colas.

- Si su alternativa está en la forma "mayor que", entonces z es el valor que le da un área en la cola derecha de la distribución, el cual es igual a a.
- Si su alternativa está en la forma "menor que", entonces z es el valor que le da un área en la cola izquierda de la distribución, el cual es igual a a.
- Si su alternativa está en la forma de "no igual a", entonces hay dos valores de z, un positivo y otro negativo. El z positivo es el valor que le da un área de a/2 en la cola derecha de la distribución. Mientras que, el z negativo es el valor que le da un área a/2 en la cola izquierda de la distribución.

La regla anterior puede ser generalizada e implementada para determinar el valor crítico de cualquier prueba de hipótesis, usted debe primero dominar la lectura de las tablas estadísticas, porque, como usted ve, no todas las tablas en su libro de textos se presentan en el mismo formato.

Significado e Interpretación de los Valores P (¿Qué Dicen los Datos?)

El valor p depende directamente de ensayos muestrales para proporcionar una medida de fuerza de los resultados en una prueba para la hipótesis nula, en contraste con un rechazo simple o no rechazo en el acercamiento clásico a la prueba de hipótesis. Si la hipótesis nula es verdadera, y si el chance de una variación aleatoria es la única razón de las diferencias muestrales, entonces el valor p es una medida cuantitativa de sustentar como evidencia a un proceso de toma de decisión. La tabla siguiente proporciona una interpretación razonable de los valores p:

Valor P	**Interpretación**
P < 0,01	Fuerte evidencia contra H_0
0,01£ P < 0,05	Moderada evidencia contra H_0
0,05 £ P < 0,10	Evidencia sugestiva contra H_0
0,10 £ P	Poca o no evidencias reales contra H_0

Esta interpretación es ampliamente aceptada, y muchos diarios científicos publican rutinariamente investigaciones usando esta interpretación para el resultado de una prueba de la hipótesis.

Para una muestra de tamaño fijo, cuando el número de realizaciones se decide por adelantado, la distribución de p es uniforme, asumiendo que

la hipótesis nula es verdadera. Expresaríamos esto como P(p £ x) = x. Eso significa que el criterio de p £ 0,05 alcanza a a of 0,05.

Entienda que la distribución de los valores de p bajo la hipótesis nula H_0 es uniforme, y por lo tanto no depende de una forma particular de prueba estadística. En una prueba estadística de la hipótesis, el valor de P es la probabilidad de observar una prueba estadística por lo menos tan extrema como el valor realmente observado, si se asume que la hipótesis nula es verdad. El valor de p es definido con respecto a una distribución. Por lo tanto, podríamos llamarlo “hipótesis de modelo-distribución" en vez de “la hipótesis nula”.

En corto, esto simplemente significa que si la nula había sido verdadera, el valor p es la probabilidad contra la nula en ese caso. El valor p es determinado por el valor observado; sin embargo, esto hace difícil incluso para medir el inverso de p.

Finalmente, puesto que los valores p son variables aleatorias, no se puede comparar varios valores p para ninguna conclusión estadística (ni obtenerla). Esto es un error común que mucha gente comete, por lo tanto, la tabla anterior no es para tal comparación.

Usted podría necesitar utilizar Valores P para la Distribución de la Población en Javascript.

Combinando el Acercamiento Clásico y el Valor P en la Prueba de Hipótesis

Un valor p es una medida de cuánta evidencia se tiene en contra de la hipótesis nula. Note que la hipótesis nula está siempre en la forma =, y no contiene ninguna forma de desigualdad. Cuanto más pequeño es el valor p, es más la evidencia que se tiene. En este contexto, el valor p se basa en la hipótesis nula y no tiene nada hacer con una hipótesis alternativa y por lo tanto con la región del rechazo. En años recientes, algunos autores han tratado de utilizar una la mezcla del acercamiento clásico y valor p. Este se basa en el valor crítico obtenido de un a dado, del estadístico calculado y del valor p. Esta es una mezcla de dos diversas escuelas del pensamiento. En este ajuste, algunos libros de textos comparan el valor p con el nivel de significancia para tomar decisiones en una prueba de hipótesis dada. Cuanto más grande valor p es en comparación con a (en hipótesis alternativa unilateral, y a/2 para las hipótesis alternativas con dos lados), menor la evidencia que se tendrá para rechazar la hipótesis nula. En tal comparación, si el valor p es menor que un cierto umbral (generalmente 0,05, a veces un pedacito más grande como 0,1 o un pedacito más pequeño como 0,01) entonces se rechaza la hipótesis nula. El siguiente argumento esta envuelto en un acercamiento combinado.

Uso del valor P y de a: En este ajuste, debemos también considerar la hipótesis alternativa al dibujar la región de rechazamiento. Existe solamente un valor p para comparar con a (ó a/2). Sepa que, para cualquier prueba de hipótesis, existe solamente un valor p. Los siguientes, son lineamientos para calcular el valor p y el proceso de decisión envuelto en una prueba de hipótesis dada:

39. Valor P para hipótesis alternativa unilaterales: El valor p se define como el área bajo la cola derecha de la distribución, si la región de rechazamiento esta dentro en la cola derecha; si la región de rechazamiento está en la cola izquierda, entonces el valor p es el área bajo la cola izquierda (en hipótesis alternativas unilaterales).
40. Valor P para hipótesis alternativas de dos lados: Si la hipótesis alternativa es con dos lados (es decir, la región de rechazo están tanto en la cola izquierda y en la cola derecha), entonces el valor p es el área bajo la cola derecha o la cola izquierda de la distribución, dependiendo de si el estadístico calculado está más cerca de la región derecha de rechazo o la región izquierda de rechazo.

    Para densidades simétricas (tales como la densidad t), los valores p en el lado izquierdo y derecho de las colas son iguales. Sin embargo, para las densidades no simétricas (tales como Chi-cuadrado) se utiliza el más pequeño de los dos. Esto hace la prueba más conservadora. Note que, para las hipótesis alternativas de dos lados, el valor p nunca es mayor que 0,5.

41. Después de encontrar el valor p según como se ha definido aquí, compare con el valor a preestablecido para pruebas unilaterales, y con a/2 para pruebas de dos colas. Cuanto más grande es el valor p en comparación con a (en hipótesis alternativa unilateral, y a/2 para las hipótesis alternativas con dos lados), menor será la evidencia que tenemos para rechazar la hipótesis nula.

Para evitar obtener los valores p de tablas estadísticas limitadas dadas en su libro de textos, los paquetes estadísticos profesionales tales como SAS y SPSS proporcionan el valor p de dos colas. Basado en donde se encuentre la región de rechazamiento, se debe descubrir qué valor p utilizar.

Algunos libros de textos tienen muchas afirmaciones engañosas sobre el valor p y sus usos. Por ejemplo, en muchos libros de textos usted encuentra a autores que doblan el valor de p para compararlo con a cuando se trabaja con prueba de hipótesis con dos colas. Quizás se pregunte cómo se hace en el caso cuando “su” valor p excede 0,5. Note eso, mientras es correcto comparar el valor p con a para las pruebas de hipótesis unilaterales a, para la prueba de hipótesis con dos lados, se debe comparar el valor p con a/2, NO a con 2 veces el valor p, como algunos libros de textos aconsejan. Mientras que la decisión es la

misma, existe una distinción clara aquí y una diferencia importante que el lector cuidadoso observar.

¿Cómo fijar el un valor apropiado a ?a value? Usted podría preguntarse porqué el a = 0,05 es tan popular en la prueba de hipótesis. a = 0,05 es tradicional para las pruebas, pero es arbitrario en su origen como sugirió R.A. Fisher, el cual dijo que en el espíritu de 0,05 siendo el valor mas grande del valor p el cual uno pensaría quizá la hipótesis nula en un experimento estadístico debería ser considerada falsa. Esto era también una sustitución entre “error tipo I” y el “error tipo II” ; que no deseamos aceptar la hipótesis nula incorrecta, pero tampoco queremos fracasar al rechazar la hipótesis nula. Como nota final, el promedio de estos dos valores p se llama a menudo el valor p medio.

Conversiones de Probabilidades de Dos Colas a la Probabilidad Unilateral: Deje que C sea la probabilidad para un intervalo de confianza de dos lados (IC) construido para una estimación. La probabilidad (C_1) de que la estimación sea o mayor que el límite más bajo o de que sea menor que el límite mas alto puede ser calculada usando:

$C_1 = C/2 + 1/2,$ para la conversion a unilateral
Ejemplo numérico: Suponga que desea convertir un C = 90% con IC de dos lados aspectos a unilateral $C_1 = 0,90/2 + 1/2 = 95\%$.

Usted podría necesitar el Javascript Determinación del Tamaño de Muestra en la etapa del diseño de su investigación estadística en la toma de decisión con requisitos subjetivos específicos.

Método de Bonferroni para el procedimiento de Múltiples valores de P

Se podrían combinar pruebas t usando el método de Bonferroni. Este método trabaja razonablemente bien cuando hay algunas pocas pruebas, pero cuando el número de comparaciones es mayor que 8, el valor del “t”' requerido para concluir que la diferencia existe, se convierte en un valor mucho más grande que el que realmente se necesita que sea, y el método se convierte conservador en exceso.

Una forma para hacer la prueba t de Bonferroni menos conservadora es utilizar la estimación de la varianza de la población calculada entre los grupos en el análisis de la varianza.

$t = (\bar{x}1 - \bar{x}2)/ (s^2 / n1 + s^2 / n2)^{1/2},$

donde s^2 es la varianza de la población calculada entre los grupos.

Procedimiento de Múltiple Valores P de Hommel: Esta prueba puede ser resumida como sigue:

Suponga que tenemos n números de valores P: p(i), i = 1,.., n, en orden ascendente que corresponden a pruebas independientes. Deje que j sea el número entero más grande, por ejemplo:

p(n-j+k) > ka/j, para todo k=1,.., ,j.

Si no existe ningún j, rechace todas las hipótesis; si no, rechace todas las hipótesis con el p(i) £ a / j. Esto proporciona un fuerte control de la tasa de error familiar a un nivel a dado.

Existen otras mejoras en el ajuste de Bonferroni cuando las pruebas múltiples son independientes o positivamente dependientes. Sin embargo, el método del Hommel es el más poderoso comparado con otros métodos.

La Potencia de la Prueba (Test) y el Efecto Tamaño

La potencia de la prueba desempeña el mismo papel en la prueba de hipótesis que el error estándar juega en la estimación. Es una herramienta de medición para determinar la exactitud de una prueba o para comparar dos métodos de prueba en competencia.

La potencia de la prueba es la probabilidad de rechazar una hipótesis nula falsa cuando la hipótesis nula es falsa. Esta probabilidad se relaciona inversamente con la probabilidad de hacer un error Tipo II, no rechazando la hipótesis nula cuando es falsa. Recuerde que elegimos la probabilidad de hacer un error Tipo I cuando fijamos a. Si disminuimos la probabilidad de hacer un error Tipo I, entonces aumentamos la probabilidad de hacer un error Tipo II. Por lo tanto, existen básicamente dos tipos de errores posibles al conducir un análisis estadístico; error Tipo I, y error Tipo II:

- Error Tipo I - (del productor) El riesgo de rechazar la hipótesis nula cuando esta de hecho es verdadera.
- Error Tipo II- (del consumidor) El riesgo de no rechazar la hipótesis nula cuando está de hecho es falso.

Potencia y Alpha (a)(a): Así, la probabilidad de no rechazar una nula verdadera tiene la misma relación al error Tipo I que la probabilidad de rechazar correctamente una nula falsa error Tipo II. Todavía, como mencioné si disminuimos la probabilidad de hacer algún tipo de error incrementamos la probabilidad de hacer el otro tipo del error. ¿Cuál es la relación entre el error Tipo I y el Tipo II? Para un tamaño de muestra fijo, disminuir un tipo de error aumenta el tamaño del otro.

Potencia y el Efecto Tamaño: Siempre que probamos si una muestra difiere de una población, o si dos muestras vienen a partir de 2 poblaciones separadas, existe la condición de que cada una de las poblaciones que estamos comparando tenga su propia media y desviación estándar (incluso si no la sabemos). La distancia entre los dos medias de las poblaciones afectará la Potencia de nuestra prueba. Esto se conoce como el tamaño del tratamiento, también conocido como el efecto tamaño, según como se demuestra en la tabla siguiente con los tres valores mas conocidos para a:

Potencia como Función de a y del Efecto Tamaño

	a		
Efecto Tamaño	0,10	0,05	0,01
1,0	,22	,13	,03
2,0	,39	,26	,09
3,0	,59	,44	,20
4,0	,76	,64	,37
5,0	,89	,79	,57
6,0	,96	,91	,75
7,0	,99	,97	,88

Potencia y el tamaño de la varianza s^2: Cuanto mayor es la variación S^2, más bajo la potencia 1-b. Cualquier cosa que tenga efecto en el grado al cual las dos distribuciones comparten valores comunes aumentara b (la probabilidad de hacer un error Tipo II)

Potencia y el tamaño de la muestra: Cuanto más pequeños son los tamaños de muestra n, más baja es la Potencia. Una n muy pequeña produce una Potencia tan bajo que las hipótesis falsas son aceptadas.

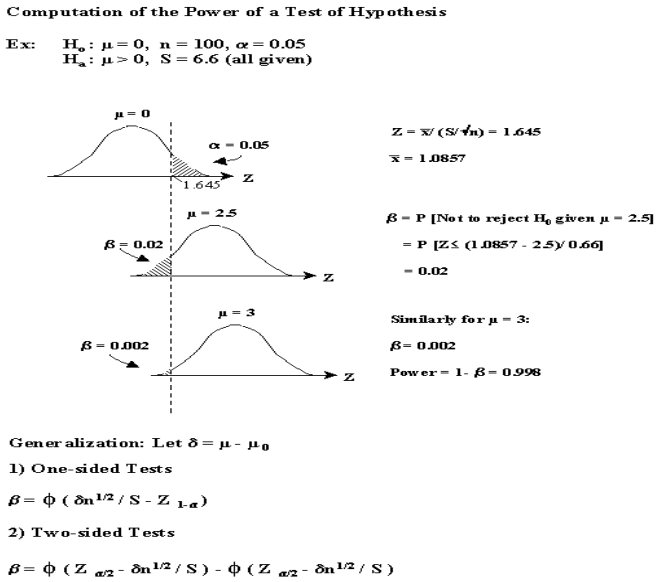
La siguiente lista detalla cuatro factores que influencian la potencia:

- Efecto tamaño (por ejemplo, la diferencia entre las medias)
- La varianza S^2
- Nivel de significancia a
- El número de observaciones, o el tamaño de la muestra n

En la práctica, los primeros tres factores normalmente son fijos. Solamente el tamaño de muestra se puede controlar por el estadístico y eso solamente dentro de la imposición del presupuesto. Existe una compensación entre el presupuesto y el logro de la exactitud deseable en cualquier análisis.

Un Ejemplo Numérico: La potencia de la pruebaes entendida más fácilmente viéndola en el contexto de una prueba compuesta. Una prueba compuesta requiere la especificación de una media poblacional como hipótesis alternativa. Por ejemplo, usando prueba Z de la hipótesis

en la figura siguiente. La Potencia es desarrollado mediante la especificación de una hipótesis alternativa como por ejemplo m = 2,5, y de m = 3. La distribución resultante bajo esta alternativa cambia 2,5 unidades a la derecha, con el área sombreada que representa la Potencia de la prueba, rechazando correctamente una información falsa.



La potencia de la Prueba

Teclee en la imagen para agrandarla

No rechazar la hipótesis nula cuando esta es falsa se define como error Tipo II, y es denotada por la región de b En la figura anterior, esta región se ubica a la izquierda del valor crítico. En la configuración mostrada en esta figura, b cae baja a la izquierda del valor crítico (y debajo de la función de densidad (o probabilidad) del estadístico bajo hipótesis alternativa H_a). El b también se define como la probabilidad de no-rechazar una hipótesis nula falsa cuando es falso, también llamada una perdida. Relaciona con el valor de b esta la potencia de una prueba. La potencia se define como la probabilidad de rechazar la hipótesis nula dado que un alternativa específica es verdadera, y se calcula como (1-b).

Una Pequeña Discusión: Considere el probar de una hipótesis nula contra una hipótesis alternativa simple. En la agrupación de Neyman-Pearson, un límite superior se fija para la probabilidad de un error Tipo I (a) dado, y luego es deseable encontrar pruebas con la probabilidad baja del error tipo II (b) dadas. La justificación general para esto es que "Estamos más preocupados sobre el error Tipo I, que fijamos un límite superior en un a que podamos tolerar." Se han visto esta clase de razonamientos en textos elementales y también en alguno ya

avanzadazos. Esto no se parece tener ningún sentido. Cuando el tamaño de muestra es grande, para la mayoría de las pruebas estándar, el cociente b/a tiende a 0. Si nos preocupamos mas sobre el error Tipo I que en el error Tipo II, ¿por qué esta incógnita se debería dispar con el aumento de tamaño de muestra?

Esto, de hecho, es una desventaja de la teoría clásica de probar hipótesis estadísticas. Una segunda desventaja es que las alternativas se encuentran entre solamente dos decisiones de la prueba: rechace la hipótesis nula o acepte la hipótesis nula. Esto es mas considerado con acercamientos que superan estas deficiencias. Esto se puede hacer, por ejemplo, por el concepto de pruebas de perfil a un "nivel" a. Ni las tasas del error Tipo I ni Tipo II son consideradas por separado, pero estas son los cocientes de una decisión correcta. Por ejemplo, aceptamos la hipótesis alternativa H_a y rechazamos la H_0, nula, si se observa un evento que es por lo menos una vez mayor debajo H_a que debajo de H_0. Inversamente, aceptamos H_0 y rechazamos H_a, si el evento observado es por lo menos una vez mayor debajo de H_0 que debajo de H_a. Este es un concepto simétrico que se formula dentro del acercamiento clásico.

Potencia de las Pruebas Paramétricas contra Pruebas no Paramétricas: Como regla general, para un tamaño de muestra dado n, las pruebas paramétricas son más poderosas que sus contrapartes las no paramétricas. La razón principal es que nos hemos acentuado pruebas paramétricas. Por otra parte, entre las pruebas paramétricas, los que utilizan la correlación son más poderosos, tales como La prueba de antes y después. Esto se conoce como Técnica de Reducción de la Varianza usada en Sistemas de Simulación para aumentar la exactitud (es decir, reducir la variación) sin incrementar el tamaño de la muestra.

Coeficiente de Correlación como Herramienta de Medición y Criterio de Decisión para el Efecto Tamaño: El coeficiente de correlación se podría obtener y utilizar como una herramienta de medida y como criterio de decisión para la fortaleza del efecto tamaño basado en el cálculo de una prueba estadística para una significativa prueba de hipótesis.

El coeficiente de correlación r se erige como un índice muy útil y accesible de la magnitud de efecto. Se acepta comúnmente que valores pequeños, medianos y grandes correspondan a valores r sobre 0,1, 0,3 y 9,5 respectivamente. Los siguientes son transformaciones necesarias de algunos estadísticos inferenciales significativos a valores r:

- Para t(gl) estadístico: $r = [t^2/(t^2 + df)]^{½}$
- Para la $F(1,gl_2)$-statistic: $r = [F/(F + df)]^{½}$
- Para la $c^2(1)$ estadístico: $r = [c^2/n]^{½}$
- Para la Normal Estándar Z: $r = (Z^2/n)^{½}$

A usted podría gustarle usar el JavaScript de Determinación del Tamaño de la Muestra en JavaScript en la etapa de diseño de su investigación

estadística para la toma de decisiones con requerimientos objetivos específicos.

Paramétrica contra no Paramétrica contra Prueba de Libre Distribución

Se debe utilizar una técnica estadística llamada no paramétrica si satisface por lo menos uno de los cinco criterios siguientes:

52. Los datos que se incorporan el análisis son enumerativos; es decir, los datos contados representan el número de observaciones en cada categoría o de categorías cruzadas.
53. Los datos se miden y /o se analizan usando una escala nominal de medida.
54. Los datos se miden y /o se analizan usando una escala ordinal de medida.
55. La inferencia no se refiere a un parámetro en la distribución de la población; por ejemplo, la hipótesis que en un sistema de tiempo ordenado de observaciones exhibe un patrón aleatorio.
56. La distribución de la probabilidad del estadístico sobre el cual se basa el análisis no es dependiente de información o condiciones específicas (es decir, asunciones) de la población(s) de la cual las muestras son dibujadas, pero solamente sobre asunciones generales, tales como una distribución de población continua y /o simétrica.

Según estos criterios, la distinción de no paramétrico es acorde ya sea por el nivel de medida usado o que sea requerido para el análisis, como en los tipos 1 a 3; el tipo de inferencia, como en el tipo 4, o la generalidad de las asunciones hechas sobre la distribución de la población, como en el tipo 5.

Por ejemplo, uno puede utilizar la Prueba de Rango de Mann- Whitney como una alternativa no paramétrica a la prueba t de Student cuando no se tienen datos distribuidos normalmente.

Mann-Whitney: Para ser utilizada con dos grupos independientes (análogos a la prueba t de grupos independientes)
Wilcoxon: Para ser utilizado con dos grupos relacionados (es decir, emparejados o repetidos) (análogos a las muestras de la prueba t relacionada)
Kruskall-Wallis: Para ser utilizado con dos o mas grupos independientes (análogos al factor simple entre objetivos ANOVA)
Friedman: Para ser utilizado con dos o mas grupos relacionados (análogos al factor simple dentro de los objetivos ANOVA)

No paramétricos contra Pruebas de Libre Distribución:

Las pruebas no paramétricas son las usadas cuando algunas condiciones específicas para las pruebas ordinarias se violan.

Las pruebas de libre distribución son las para las cuales el procedimiento es válido para toda la diversa forma de la distribución de la población.

Por ejemplo, la prueba Chi-cuadrado referente a la variación de una población dada es paramétrica puesto que esta prueba requiere que la distribución de la población sea normal. La prueba Chi-cuadrado de la independencia no asume la condición de normalidad, ó que los datos son numéricos. La Prueba de Kolmogorov-Smirnov es una prueba de libre distribución, que es aplicable para comparar a dos poblaciones con cualquier distribución de variables aleatorias continuas.

La sección siguiente es un interesante procedimiento no paramétrico con diferentes y útiles aplicaciones.

Comparación de dos Variables Aleatorias: Considere dos observaciones independientes $X = (x_1, x_2,\ldots, x_r)$ e $Y = (y_1, y_2,\ldots, y_s)$ para dos variables aleatorias X e Y respectivamente. Para estimar la función de la confiabilidad:

$R = Pr (X > Y)$
Se podría utilizar:
El estimador $RS = U/(r \acute{} s)$,
De donde U es el Número de pares (x_i, y_j) tal que $x_i > y_j$, para todo i = 1, 2, ,r, y j = 1, 2,..,s.

Este es un estimador neutral con mínima varianza para R. Es importante saber que la estimación tiene un limite superior, y un valor delta no negativo para su precisión:

$Pr\{R \geq RS - d\} \geq \max \{1- \exp(-2nd^2), 4nd^2/(1-4nd^2)\}$.
Las áreas de aplicación incluyen el problema de la ruina del seguro. Deje que la variable aleatoria Y denote las unidades de tiempo y deje que la variable aleatoria X denote los retornos en inversión (ROI) para la compañía de seguros. Finalmente, deje que z denote la el monto constante de la prima recogida; entonces la probabilidad de que la compañía de seguros sobrevivirá es:
$R = Pr [X + z > Y]$.

A usted podría gustarle usar la Prueba para Dos Poblaciones de Kolmogorov-Smirnov y Comparar dos Variables Aleatorias para comprobar el resultado de sus cálculos y realizar experimentos numéricos para una compresión mas profunda de estos conceptos.

Pruebas de Hipótesis

Recuerde que, en las pruebas t para las diferencias en las medias, existe una condición de las varianzas poblacionales iguales que deben ser examinadas. Una forma para probar las posibles diferencias en las varianzas es hacer una prueba F. Sin embargo, la prueba F es muy sensible a las violaciones de la condición de normalidad; es decir, si las poblaciones parecen no ser normales, entonces la prueba F tenderá a rechazar muy frecuentemente nulidad de no diferencias en las varianzas de la población.

A usted podría gustarle usar los siguientes JavaScripts para comprobar sus cálculos y para realizar algunos experimentos estadísticos para una compresión mas profunda de estos conceptos:

- Prueba de la Media.
- Prueba de la Varianza.
- Prueba de Dos Poblaciones.
- Prueba de las Diferencias: La prueba de Antes-y-Después .
- ANOVA.
- Para la igualdad *estadística* de dos poblaciones, a usted le podría gustar usar la Prueba de Kolmogorov-Smirnov .

Prueba t para una Población Simple

El propósito es comparar la media de la muestra con la media de la población dada. El objetivo es juzgar el valor medio demandado, basado en un sistema de observaciones aleatorias de tamaño n. Una condición necesaria para la validez del resultado es que la distribución de la población sea normal, si el tamaño de muestra n es pequeño (digamos menor a 30.)

La misión es decidir si aceptar la hipótesis nula:

$H_0 = m = m_0$

ó rechazar la hipótesis nula a favor de hipótesis alternativa:

H_a: m es significativamente diferente de m_0

El esquema de la prueba consiste en calcular un t estadístico:

$T = [(\bar{x} - m_0)\, n^{1/2}] / S$
De donde $\bar{x}$ es la media estimada y S^2 es la varianza estimada basada en n observaciones aleatorias.

El estadístico anterior se distribuye como una distribución t con parámetro de gl. = n = (n-1). Si el valor absoluto del T estadístico

calculado es “demasiado grande” comparado con el valor crítico de la tabla t, entonces se rechaza la proposición del valor para la media de la población.

Esta prueba también se podría utilizar para probar proposiciones similares para otras poblaciones unimodal incluyendo aquellos con variables aleatorias discretas, tales como proporción, con tal de que hayan suficientes observaciones (mas de 30.)

A usted podría gustarle usar la Prueba de la Media en Javascript para comprobar de sus cálculos y el Javascript de la Determinación del Tamaño de la Muestra en la etapa del diseño de su investigación estadística en la toma de decisión con requisitos subjetivos específicos.

Prueba de Dos Poblaciones Testing Two Populations.

¿Cuándo Deberíamos Agrupar las Estimaciones de las Varianzas?

Debemos reunir las estimaciones de las varianzas solamente si hay una buena razón para hacerlo, y entonces (dependiendo de esa razón) las conclusiones tienen que ser hechas explícitamente condicionales en la validez del modelo de varianzas iguales. Existen diversas buenas razones para reunirlas:

(a) para conseguir una sola estimación estable de varias muestras relativamente pequeñas, donde las fluctuaciones de las varianzas parezcan no ser sistemáticas; ó

(b) por conveniencia, cuando todas las estimaciones de las varianzas están suficientemente cerca a la igualdad; ó

(c)). cuando no hay opción diferentes a modelar varianzas (como en la regresión lineal simple sin valores replegados de X), y desviaciones del modelo de varianza constante parezcan no ser sistemáticas; ó

(d) cuando los tamaños de los grupos son grandes y casi igual, de modo que no hayan diferencias esenciales en parejas de contraste entre las estimaciones de los errores estándar reunidos y no reunidos, y los grados de libertad sean casi asintóticos.

Observe que este último racional podría caer aparte para contrastar otras parejas. En realidad, no se están reuniendo varianzas en el cado (d), en vez, se esta tomando un atajo para calcular los errores estándar en parejas de contraste.

Si se calcula la prueba sin la asunción, usted tiene que determinar los grados de libertad (gl). El fórmula funciona de manera tal que los gl serán menores si la varianza de la muestra más grande está en el grupo

con el número más pequeño de observaciones. Éste es el caso en el cual las dos pruebas diferirán considerablemente. Un estudio de la fórmula para los gl cera la mejor aclaratoria, y se deberá entender la correspondencia entre el diseño desafortunado (teniendo la mayoría de las observaciones en el grupo con poca varianza) y bajos gl y un acompañante valor t grande.

Ejemplo: Cuando se este haciendo pruebas t para las diferencias en las medias de las poblaciones (un caso clásico de muestras independiente):

63. Para diferencias en las medias que no hacen ninguna asunción sobre la igualdad de las varianzas de la población, utilice la fórmula del error estándar:

    $[S^2_1/n_1 + S^2_2/n_2]^{½}$,

    con gl= n = n_1 ó n_2 el que sea mas pequeño de los dos.

64. Con Varianzas iguales, use este estadístico:

    $$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{1/n_1 + 1/n_2}},$$

    con parámetro de gl = n = (n_1 + n_2- 2), n_1, para n_2 mas grande o igual a 1, donde la varianza agrupada es:

    $$s_p^2 = \frac{SS(x_1) + SS(x_2)}{n_1 + n_2 - 2} = \frac{s_1^2\,(n_1 - 1) + s_2^2\,(n_2 - 1)}{n_1 + n_2 - 2}$$

65. Si el N total es menor a 50 y una muestra es 1/2 el tamaño de la otra (o menos), y si la muestra mas pequeña tiene una desviación estándar de por lo menos dos veces el tamaño de la otra muestra, se debe aplicar el procedimiento no. 1, pero ajuste el parámetro de gl de la prueba t al entero mas grande o igual a:

    d.f. = n = A/(B +C),

    de donde:

    A = $[S^2_1/n_1 + S^2_2/n_2]^2$,

    B = $[S^2_1/n_1]^2$ / (n_1 -1),

    C = $[S^2_2/n_2]^2$/ (n_2 -1)

    Si esto no ocurre, no se preocupe por el problema del tener un a nivel que sea muy diferente que el qué se ha fijado.

La sección de Estadísticos de Confianza se refiere a la construcción de un intervalo de confianza donde la condición de igualdad de las varianzas es un asunto importante.

La aproximación anterior, la cual es muy general con resultados conservadores, se puede implementar usando el JavaScript de Prueba de Dos Poblaciones JavaScript.

A usted podría gustarle usar el Javascript de Prueba de Diferencias en las Medias: La Prueba de Antes y Después y laPrueba de Proporción de Paired para proporciones dependientes.

Procedimiento de Comparación Múltiple de No-paramétrico:

Prueba de Rango Múltiple de Duncan: Este es uno de los muchos procedimientos de comparación múltiple. Esta basado en el rango estadístico estandarizado mediante la comparación de todos los pares de medias mientras se controla todo el error Tipo I al nivel deseado. Mientras esto no proporcione intervalos de estimaciones de la diferencia entre cada par de medias, no indicara cuales medias son significativamente diferentes de las otras. Para determinar las diferencias significativas entre las medias de un *grupo simple de control* y las otras medias, se podría utilizar la prueba de comparaciones múltiples de Dunnett.

Introducción a las Pruebas de Igualdad Estadística de Dos o mas Poblaciones:

Dos variables aleatorias X e Y que tienen distribución $F_X(x)$ y $F_Y(y)$ respectivamente, serían equivalentes, o iguales en ley, o iguales en la distribución, si y solo si tienen la misma función de distribución. Es decir,
$F_X(z) = F_Y(z)$, para toda z,
Existen diferentes pruebas dependiendo de los usos previstos. Las pruebas ampliamente usadas para la igualdad estadística de poblaciones son las siguientes:

66. Igualdad de Dos Poblaciones Normales: Se podría utilizar la prueba Z y la prueba F para comprobar la igualdad de las medias, y la igualdad de las varianzas, respectivamente.
67. Prueba de un Cambio en Poblaciones Normales: Con frecuencia, estamos interesados en la prueba para una cambio dado en una población dada de la distribución, lo cual es que estamos probando si una variable al azar Y es igual en distribución a otra X + c para alguna c constante. Es decir, la distribución de Y es la distribución de X cambiada de puesto. En la prueba de cualquier cambio en la distribución se necesita probar para primero la

normalidad, y luego probar las diferencia en valores esperados aplicando la prueba Z con dos lados con la hipótesis nula de:

H_0: $m_Y - m_X = c$.

68. Análisis de la Varianza: La prueba de análisis de las varianzas (ANOVA) es diseñada para la prueba de igualdad simultánea de tres o más poblaciones. Las condiciones previas en la aplicación de ANOVA son la normalidad de cada distribución poblacional, y la igualdad de todas las varianzas simultáneamente (no la prueba de pares ordenados.)

    Note que ANOVA es una extensión del punto No. 1 en la prueba de igualdad de más de dos poblaciones. Se podría demostrar si se aplica ANOVA para probar la igualdad de dos poblaciones basadas en dos muestras independientes con tamaños n1 y n2 para cada población, respectivamente, los resultados de ambas pruebas son idénticos. Por otra parte, la prueba estadística obtenida por cada prueba se relaciona directamente, es decir,

    $F_{a\ ,\ (1,\ n1+\ n2\ -\ 2)} = t^2_{a/2\ ,\ (n1+\ n2\ -\ 2)}$

69. Igualdad de Proporciones en Varias Poblaciones: Esta prueba es para variables aleatorias discretas. Esta es uno de los muchos usos interesantes de las aplicaciones de la Chi-cuadrado.
70. Igualdad de Libre Distribución de Dos Poblaciones: Siempre que se este interesado en la prueba de la igualdad de dos poblaciones con una variable aleatoria continua común, sin ninguna referencia a la distribución subyacente tal como la condición de normalidad, se puede utilizar la libre distribución conocida como la prueba K-S.
71. Comparación no paramétrica de dos Variables Aleatorias: Considere dos observaciones independientes $X = (x_1, x_2,\ldots, x_r)$ e $Y = (y_1, y_2,\ldots, y_s)$ para dos poblaciones independientes con variables aleatorias X e Y, respectivamente. A menudo estamos interesados en estimar la Pr (X > Y).

Igualdad de dos Poblaciones Normales:

La Distribución Normal o Gaussiana es una distribución simétrica continua que sigue la curva acampanada familiar. Una de sus características aplicaciones interesantes es que, únicamente la media y la varianza determinan independientemente la distribución.

Por lo tanto, para probar la igualdad estadística de dos poblaciones normales independientes, se necesidad primero realiza la Prueba de Normalidad de Lilliefors para asegurar esta condición. Dado que ambas poblaciones se distribuyen normalmente, se deben realizar dos pruebas

mas, la prueba para la igualdad de las dos medias y la prueba para la igualdad de las dos varianzas. Ambas pruebas pueden ser realizadas usando el JavaScript de la Prueba de Hipótesis para Dos Poblaciones en Javascript.

Comparación de Medias Múltiples: Análisis de las Varianza (ANOVA)

Las pruebas que hemos aprendido hasta ahora, nos permiten probar hipótesis que examinan la diferencia entre dos medias solamente. El análisis de la varianza o ANOVA nos permitirá probar la diferencia entre dos o más medias examinando el cociente de la variabilidad entre dos condiciones y de la variabilidad dentro de cada condición. Por ejemplo, digamos que suministramos una droga que creamos mejorará la memoria a un grupo de personas y demos un placebo a otro grupo. Podríamos medir el funcionamiento de la memoria por el número de las palabras recordadas de una lista que pedimos a cada uno para memorizar. Una prueba t compararía la probabilidad de observar la diferencia entre los números medios de las palabras recordadas por cada grupo. Una prueba ANOVA, por otra parte, compararía la variabilidad que observamos entre las dos condiciones a la variabilidad observada dentro de cada condición. Recuerde que medimos variabilidad como la suma de la diferencia de cada valor con respecto a la media. Cuando realmente calculamos un ANOVA utilizaremos una fórmula atajo.

Por lo tanto, cuando la variabilidad que predecimos entre dos grupos es mucho mas grande que la variabilidad que no pudimos predecir dentro de cada grupo, concluiremos que nuestro tratamiento produce resultados diferentes.

Un Ejemplo Ilustrativo de ANOVA

Considere las muestras aleatorias (enteros pequeños, solo para efectos ilustrativos mientras se ahorra espacio) siguientes que corresponden a tres poblaciones diferentes.

Con hipótesis nula: H_0: $\mu 1 = \mu 2 = \mu 3$, y alternativa: H_a: al menos dos de las medias no son iguales.

A un valor de significancia de a = 0,05, el valor critico de la tabla F es: $F_{0.05, 2, 12} = 3,89$.

						Suma	Media
Muestra P1	2	3	1	3	1	10	2

Muestra P2	3	4	3	5	0	15	3
Muestra P3	5	5	5	3	2	20	4

Demostrar que, SCT = SCE + SCD.
Esto es, la suma de los cuadrados totales (SCT) igual a la suma de los cuadrados entre (SCE) los grupos mas la suma de los cuadrados dentro (SCD) de los grupos.

Cálculo de la muestra SCT: Con la media principal = 3, primero, se comienza tomando la diferencia entre cada observación y la media, y luego se eleva al cuadrado para punto de los datos.

						Suma
Muestra P1	1	0	4	0	4	9
Muestra P2	0	1	0	4	9	14
Muestra P3	4	4	4	0	1	13

Por lo tanto SCT = 36 con gl = (n-1) = 15-1 = 14.

Cálculo de la muestra SCE:

Segundo, deje que todos los datos en cada muestra tenga el mismo valor como la media principal en esa muestra. Esto remueve cualquier variación DENTRO de ella. Calcule la suma de los cuadrados de las diferencias con respecto a la media principal.

						Suma
Muestra P1	1	1	1	1	1	5
Muestra P2	0	0	0	0	0	0
Muestra P3	1	1	1	1	1	5

Por lo tanto, SCE = 10, con gl = (m-1)= 3-1 = 2 para m =3 grupos.

Cálculo de la muestra SCD:

Tercero, calcule la suma de los cuadrados de las diferencias dentro de cada muestra usando sus propias medias muestrales. Esto provee una suma de los cuadrados de las desviaciones DENTRO de todas las muestras.

						Suma
Muestra P1	0	1	1	1	1	4
Muestra P2	0	1	0	4	9	14
Muestra P3	1	1	1	1	4	8

SCD = 26 con gl = 3(5-1) = 12. Esto es, 3 grupos por (5 observaciones en cada -1)

Los resultados son: SCT = SCE + SCD, y $gl_{SCT} = gl_{SCE} + gl_{SCD}$, como se esperaba.

Ahora, construya la tabla ANOVA para este ejemplo numérico colocando los resultados de sus cálculos en esta tabla. Note que, los Cuadrados de las Medias son la Suma de los cuadrados divididos por sus Grados de Libertad. El estadístico F es el cociente de las dos Medias al Cuadrado.

Tabla ANOVA				
Origen de la Variación	Suma de Cuadrados	Grados de Libertad	Medias al Cuadrado	Estadístico F
Entre Muestras	10	2	5	2,30
Dentro de las Muestras	26	12	2.17	
Total	36	14		

Conclusión: No existe suficiente evidencia para rechazar la hipótesis nula H_0.

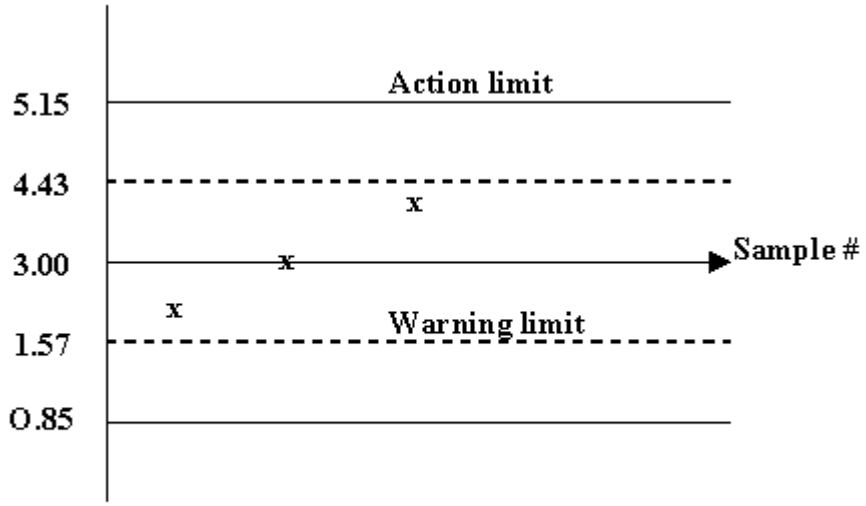
La lógica detrás de ANOVA: Primero, intentemos explicar la lógica y después ilustrarla con un ejemplo simple. En la ejecución de la prueba de ANOVA, estamos intentando determinar si un cierto número de medias poblacionales son iguales. Para hacer esto, medimos la diferencia de las medias muestrales y las comparamos con la variabilidad dentro de las observaciones de la muestra. Esta es la razón del porqué la prueba estadística es el cociente de la variación entre-muestra (VEM) y de la variación dentro-muestra (VDM). Si este cociente está cerca de 1, existe evidencia de que las medias poblacionales son iguales.

Esta es un buen uso para usted: Mucha gente cree que en el mundo de los negocios, los hombres perciben mejor salario que las mujeres, simplemente por ser del genero masculinos. Para justificar o rechazar tal proposición, se podría mirar la variación dentro de cada grupo (un grupo que es el salario percibido por las mujeres y el otro grupo el percibido por hombres) y compararlos con la variación entre las medias de las muestras aleatoriamente seleccionadas de cada población. Si la variación en los salarios de las mujeres es mucho mayor que la variación entre la media de los salarios de los hombres y de las mujeres, uno podría decir que porque la variación es muy grande dentro del grupo de las mujeres, esto podría no ser un problema relacionado al género.

Ahora, volviendo a nuestro ejemplo numérico del tratamiento de la droga para incrementar la memoria contra el placebo. Notamos que: dada la conclusión de la prueba y las condiciones de la prueba ANOVA, podemos concluir que estas tres poblaciones son de hecho, la misma

población. Por lo tanto, la técnica de ANOVA se podría utilizar como una herramienta de medición de rutina estadística para el control de calidad, según lo descrito a continuación con ejemplo numérico.

Construcción del Cuadro de Control para las Medias de la Muestra: Bajo la hipótesis nula, el ANOVA concluye que μ1 = μ2 = μ3; es decir, tenemos una "población familiar hipotética." La pregunta es, ¿cuál es su varianza? La varianza estimada (es decir, los cuadrados de las medias totales) es 36/14 = 2,57. De esta forma, la desviación estándar estimada es = 1,60 y la desviación estándar estimada para las medias es 1,6/ $5^{½}$ = 0,71. Bajo las condiciones de ANOVA, podemos construir un cuadro de control con los límites de cuidado = 3 ± 2(0,71); Los límites de acción = 3± 3(0,71). La figura siguiente representa el cuadro de control.



Control Chart for ANOVA in our example

A usted podría gustarle usarANOVA: Prueba de Igualdad de Medias para sus cálculos, y luego interpretar los resultados en términos gerenciales (no técnicos). .

Usted podría necesitar utilizar el Javascript de la Determinación del Tamaño de Muestra en la etapa de diseño de su investigación estadística en la toma de decisión con requisitos subjetivos específicos.

ANOVA para Datos Normales pero Condensados

En la prueba de la igualdad de varias medias, por lo general las informaciones en bruto no se encuentran disponibles. En tal caso, se debe realizar el análisis necesario basado en datos secundarios usando el sumario de los datos; Digamos, Preparación Triple: Los tamaños de las muestras, los medios de las muestras, y las varianzas de las muestras.

Suponga que una de las muestras es de tamaño n, que tiene media muestral $\bar{x}$, y varianza muestral S^2. Deje que:

$y_i = \bar{x} + (S^2/n)^{½}$ para todo i = 1, 2, …, n-1,
y
$y_n = n\bar{x} - (n - 1)y_1$
Entonces, la nueva variable aleatoria y_i's son datos sustitutos que tienen la misma media y varianza que el modelo original. Por lo tanto, generando los datos sustitutos para cada muestra, se puede realizar la prueba estándar de ANOVA. Los resultados son idénticos.

A usted podría gustarle usar el ANOVA para Datos Condensados para sus cálculos y experimentación.

El Javascript de la Evaluación Subjetiva de Estimaciones prueba la proposición de que por lo menos el cociente de una estimación a otra estimación sea tan grande como el valor dado de la proposición.

ANOVA para Poblaciones Dependientes

Las poblaciones pueden ser dependientes en cualquiera de las maneras siguientes:

72. Cada sujeto o individuo es probado en cada condición experimental. Esta clase de dependencia es llamada diseño repetido de medición.
73. Los sujetos bajo diversas condiciones experimentales son relacionados de ciertas maneras. Esta clase de dependencia es llamada diseño de sujetos equivalente.

Una aplicación: Suponga que estamos interesados en estudiar los efectos del alcohol en la capacidad para conducir. Diez sujetos proporcionan tres índices diferentes de alcohol, el número de errores al conducir son tabulados en la siguiente tabla:

											Media
0 oz	2	3	1	3	1	4	1	3	2	1	**2,1**
2 oz	3	2	1	4	2	3	1	5	1	2	**2,4**
4 oz	3	1	2	4	2	5	2	4	3	2	**3,1**

La hipótesis nula es:

H_0: $\mu1 = \mu2 = \mu3$,

y la alternativa:

H_a: por lo menos dos medias no son iguales.

Utilizando la ANOVA para Poblaciones Dependientes en JavaScripts, obtenemos la información necesaria para construir la tabla de ANOVA siguiente:

Tabla ANOVA				
Origen de la Variación	Suma de Cuadrados	Grados de Libertad	Medias al Cuadrado	Estadístico F
Sujetos	31,50	9	3,50	-
Entre	5,26	2	2,63	7,03
Dentro	6,70	18	0,37	
Total	43,46	29		

Conclusión: El valor p es P= 0,006, indicando una fuerte evidencia contra la hipótesis nula. Las medias poblacionales no son iguales. Aquí, se podría concluir que una persona que haya consumido mas de cierto nivel de alcohol comete mas errores cuando maneja.

Un "muestreo de diseño de bloque" implica estudiar a más de dos poblaciones dependientes. Para probar la igualdad de las medias de más de dos poblaciones basadas en el muestreo del diseño de bloque, se podría utilizar el Javascript de la Prueba de ANOVA de Dos Vías. En el caso del tener datos de diseño de bloque con las réplicas, utilice JavaScript Prueba de ANOVA de Dos Vías con reproducción para obtener la información necesaria para construir las tablas de ANOVA.

Prueba de Igualdad para Varias Proporciones de Poblaciones

La prueba del Chi-cuadrado de la homogeneidad proporciona un método alternativo para probar la hipótesis nula de que dos proporciones de la población son iguales. Por otra parte, extiende a varias poblaciones similares la prueba de ANOVA que compara varias medias.

Una aplicación: Suponga que deseamos probar la hipótesis nula:

$\mathbf{H_0}$: $P_1 = P_2 = ... = P_k$

Esto es, las tres proporciones de las poblaciones son casi idénticas. Los datos de la muestra con respecto a las tres poblaciones son dadas en la tabla siguiente:

Prueba para la Homogeneidad de Proporciones de Varias Poblaciones

Poblaciones	Si	No	Total

Muestra I	60	40	100
Muestra II	57	53	110
Muestra III	48	72	120
Total	165	165	330

El estadístico Chi-cuadrado es 8,95 con gl = (3-1)(3-1) = 4. El valor p es igual a 0,062, indicando que hay evidencia moderada contra la hipótesis nula de que las tres poblaciones son estadísticamente idénticas.

A usted podría gustarle usar la Prueba de Proporciones para realizar este experimento.

Igualdad de Libre Distribución de Dos Poblaciones

Para la igualdad *estadística* de dos poblaciones, se puede utilizar la prueba de Kolmogorov-Smirnov (prueba de K-S) para dos poblaciones. La prueba de K-S busca diferencias entre las funciones de distribución de las dos poblaciones basada en sus dos muestras independientes escogidas al azar. La prueba rechaza la hipótesis nula de ninguna diferencia entre las dos poblaciones si la diferencia entre las dos funciones de distribución empíricas es "grande."

Antes de la aplicación de la prueba de K-S es necesario arreglar cada uno de las dos observaciones de las muestras en una tabla de la frecuencia. La tabla de frecuencia debe tener una clasificación común. Por lo tanto la prueba se basa en la tabla de frecuencia, que pertenece a la familia de las pruebas de libre distribución.

El proceso de la prueba de K-S es como sigue:

74. Un cierto número de k de "clases" se selecciona, cada una típicamente cubre un rango diferente pero similar de valores.
75. Un cierto número mucho más grande de observaciones independientes (n_1, y n_2, ambos mayores que 40) son tomadas. Cada una es medida y su frecuencia es ubicada en una clase.
76. Basándose en la tabla de frecuencia, las funciones de distribución empírica acumulativa $F1_i$ y $F2_i$ para dos poblaciones de muestras son construidas, para i = 1, 2..., k.
77. El estadístico K-S es la diferencia absoluta más grande entre $F1_i$ and $F2_i$; es decir,

    Estadística de K-S = D = máximo $| F1_i - F2_i |$, para todos i = 1, 2, .., k.

Los valores críticos del estadístico K-S pueden ser encontrados en Computadoras y Estadística Computacional con Aplicaciones

Una aplicación: Las ventas diarias de dos subsidiarios de la compañía PC & Accesorios son mostrados en la tabla siguiente, con n1 = 44, y n2 = 54:

Ventas Diarias de Dos Subsidiarias en 6 Meses			
Ventas ($1000)	*Frecuencia I*	*Frecuencia II*	
0 - 2	11	1	
3 - 5	7	3	
6 - 8	8	6	
9 - 11	3	12	
12 - 14	5	12	
15 - 17	5	14	
18 - 20	5	6	
Sumas	44	54	

El gerente de la primer subsidiaria está tiene la siguiente asunción "puesto que las ventas diarias son fenómenos aleatorios, mi funcionamiento total es tan bueno como el funcionamiento del otro gerente." En otras palabras:

H_0: Las ventas diarias en los dos almacenes casi son iguales.
H_a: El funcionamiento de los gerentes es perceptiblemente diferente.

Después del proceso anterior para esta prueba, el estadístico K-S es 0,421 con valor p de 0,0009, indicando que existe fuerte evidencia en contra de la hipótesis nula. Existe suficiente evidencia que el funcionamiento del encargado de la segunda subsidiaria sea mejor.

Introducción a la Aplicaciones del Estadístico Chi-cuadrado

La varianza no es la única razón por la cual se puede utilizar la prueba Chi-cuadrado.

Las aplicaciones más comunes de la distribución Chi-cuadrado son:

La prueba Chi-cuadrado por asociación, la cual es una prueba no paramétrica; por lo tanto, puede ser utilizada también para datos nominales. Es una prueba de significancia estadística ampliamente utilizada en análisis tabular de asociación de doble variación. Típicamente, la hipótesis es si o no dos poblaciones son diferentes en cierta característica o aspecto de su comportamiento basado en dos muestras escogidas al azar. Este método de prueba también se conoce como la prueba Chi-cuadrado de Pearson.

La calidad o bondad de ajuste de la prueba Chi-cuadrado se utiliza probar si una distribución observada conforma a cualquier otra

distribución particular. El cálculo de esta calidad de ajuste es mediante la comparación de datos observados con datos esperados basados en una distribución particular.

Una de las desventajas de algunas de las pruebas Chi-cuadrado es que no permiten el cálculo de los intervalos de la confianza; por lo tanto, la determinación del tamaño de muestra no es fácilmente disponible.

Tratamiento de Casos con Muchas Categorías: Note que, aunque en la siguiente sección de tablas cruzadas se tienen solo dos categorías, existe siempre la posibilidad de convertir casos con muchas categorías tablas cruzadas similares. Por lo tanto, uno debe considerar todos los pares posibles de categorías y de sus valores numéricos mientras que se construye las “dos categorías” equivalentes de tablas cruzadas.

Prueba de Relación para Tablas Cruzadas

Tablas Cruzadas: Las tablas cruzadas se utilizan para probar relaciones entre dos tipos de datos categóricos, o la independencia de dos variables, tales como el uso del cigarrillo y uso de la droga. Si usted encuesta 1000 personas preguntando si fuman o no y si consumes drogas o no, se podrían conseguir cuatro respuestas: (no, no) (no, sí) (sí, no) (sí, sí.)

Compilando el número de personas en cada categoría, usted puede probar en última instancia si el consumo de la droga es independiente a fumar cigarrillos usando la distribución Chi-cuadrado (la cual es aproximada, pero trabaja bien). Una vez más la metodología para aplicar esto se encuentra en su libro de textos. Los grados de libertad son iguales a (número de filas-1)(número de columnas -1). Es decir, todos estos números son necesarios para completar el cuerpo entero de las tablas cruzadas, el resto será determinado usando las sumas dadas de las filas y las sumas de los valores de las columnas.

No olvide las condiciones para la validez de la prueba y Chi-cuadrado y sus valores esperados relacionados mayores a 5 en el 80% o más celdas. De otra forma, se podría usar una prueba “exacta”, usando una permutación o el acercamiento por re muestreo.

Usando la Chi-cuadrado en una tabla 2x2 requiere la corrección de Yates. Primero se resta 0,5 de la diferencia absoluta entre las frecuencias observadas y esperadas para cada uno de los tres genotipos antes de elevarlos al cuadrado, dividiéndose por la frecuencia esperada, y luego sumamos. La fórmula para el valor del Chi-cuadrado en una tabla 2x2 se puede derivar de la Teoría Normal de la comparación de dos proporciones en la tabla usando la incidencia total para producir los errores estándar. El análisis razonado de la corrección es una mejor equivalencia del área bajo la curva normal y de las

probabilidades obtenidas de las frecuencias discretas. Es decir la corrección más simple es mover el punto de corte para la distribución continua con respecto al valor observado de la distribución discreta hacia la mitad del camino entre ése punto y el valor siguiente en la dirección de la hipótesis nula esperada. Por lo tanto, la corrección esencialmente se aplica solo a las pruebas de un grado de libertad donde la “raíz cuadrada” del Chi-cuadrado se asemeja a una “prueba t normal” y donde una dirección puede ser adjuntada a la adición de 0,5.

Chi-square distribution is used as an approximation of the binomial distribution. By applying a continuity correction, we get a better approximation of the binomial distribution for the purposes of calculating tail probabilities.

Dado la siguiente tabla 2x2, se pueden calcular algunas medidas relativas al riesgo:

a	**b**
c	**d**

Las medidas más generalmente:

Tasa de diferencia: a/(a+c) - b/(b+d)
Tasa de cociente: (a/(a+c))/(b/(b+d))
Chance del cociente: ad/bc

La tasa de diferencia o la tasa de cociente son apropiadas cuando se están contrastando dos grupos, de los cuales sus tamaños (a+c y b+d) están dados. El cociente de la probabilidad es para los casos de asociación y no de diferencia.

El Cociente de riesgo (CR) es el cociente de la proporción (a/(a + b)) a la proporción (c/(c + d)):

CR = (a / (a + b)) / (c / (c + d))

El CR es por l tanto, una medida de cuánto más grande es la proporción en la primera fila cuando se compara a la segunda. Valor de CR < 1 indica una asociación “negativa” de [a/(a+b) < c/(c+d)], de 1 indica que no existe ninguna asociación y, >1 indica que la asociación es “positiva” [a/(a+b) > c/(c+d)]. Mientras mayor sea el CR que 1, mas fuerte será la asociación.

Una aplicación: Suponga que el consejero de una escuela en una ciudad pequeña está interesado si la profesión elegida por los estudiantes está relacionado con la ocupación de sus padres. Se necesitan registrar los datos según lo demostrado en la tabla siguiente de la contingencia con dos filas (r1, r2) y tres columnas (c1, c2, c3):

Relación entre la ocupación de padres y la Profesión

elegida por los estudiantes de secundarias

Profesión elegida por estudiantes

Parental Ocupación	*Prep. Universitaria*	*Vocacional*	*General*	*Totales*
Profesional	12	2	6	20
Obrero	6	6	8	20
Totales	18	8	14	

Bajo la hipótesis de que no existe relación, el valor esperado (E) de la frecuencia sería:

$E_{i,\,j} = (Sr_i)(Sc_j)/N$

Las frecuencias observadas (O) y esperadas (E) son recogidas en la siguiente tabla:

Frecuencias esperadas para los datos.

	Prep. Universitaria	*Vocacional*	*General*	*Totales*
Profesional	*O* = 12 *E* = 9	*O* = 2 *E* = 4	*O* = 6 *E* = 7	å*O* = 20 å*E* = 20
Obrero	*O* = 6 *E* = 9	*O* = 6 *E* = 4	*O* = 8 *E* = 7	å*O*= 20 å*E* = 20
Totales	å*O* = 18 å*E* = 18	å*O* = 8 å*E* = 8	å*O* = 14 å*E* = 14	

La cantidad

$c^2 = S\,[(O - E)^2 / E]$

es una medida del grado de desviación entre las frecuencias Observadas y Esperadas. Si no existe relación entre las variables de las filas y las variables de las columnas, esta medida estaría muy cerca de cero. Bajo la *hipótesis de que existe una relación* entre las filas y las columnas, esta cantidad tiene una distribución Chi-cuadrado con el parámetro igual al número de filas menos 1, multiplicado por el número de columnas menos 1.

Para este ejemplo numérico tenemos:

$c^2 = S\,[(O - E)^2 / E] = 30/7 = 4{,}3$

con gl = (2-1)(3-1) = 2, tal que tiene el valor p de 0,14, sugiriendo poca o ninguna evidencias en contra de la hipótesis nula.

La pregunta principal es cuan grande es la medida. El valor máximo de esta medida es:

$c^2_{max} = N(A-1)$,

de donde A es el número de filas o de columnas, cualquiera que sea más pequeño. Por nuestro ejemplo numérico este es: 40(2-1) = 40.

El coeficiente de determinación de el cual tiene un rango de [0, 1], proporciona la fuerza relativa de la relación, calculada tal como:

$c^2/c^2_{max} = 4,3/40 = 0,11$

Por lo tanto concluimos que el grado de la asociación es solamente de11%, el cual es bastante débil.

Alternativamente, usted podría también mirar al coeficiente de contingencia estadística f el cual es:

$f = [c^2/(N + c^2)]^{½} = 0,31$

El rango de este estadístico es entre 0 y 1 y se puede interpretar como el coeficiente de correlación. Esta medida también indica que la profesión elegida por los estudiantes esta relacionada a la ocupación de sus padres.

A usted podría gustarle utilizar la Prueba Chi-cuadrado para la Relación de Tablas Cruzadas en la ejecución de esta prueba, y el JavaScript de Valores P para Distribuciones Populares para encontrar los valores p y el estadístico Chi-cuadrado.

Prueba de Poblaciones Idénticas para Datos de Tablas Cruzadas

La prueba de homogeneidad es similar a la Prueba de Relación de Tablas Cruzadas en la medida de que ambas se ocupan de la clasificación cruzada de datos nominales; es decir, tablas de r ´ c. El método para calcular el estadístico Chi-cuadrado es igual para ambas pruebas, con los mismos gl tables.

La s dos pruebas se diferencian, sin embargo, en el siguiente aspecto. La prueba para la relación de Tablas Cruzadas es hecha mediante el dibujo de datos provenientes de una población simple (con un número total de elementos **fijo**) del cual solo se considera si un grupo de atributos es independiente con respecto otro grupo. La prueba para la homogeneidad, por otra parte, es diseñada para probar la hipótesis nula de que las muestras que dos o más muestras aleatorias sean dibujadas de la misma o de diferentes poblaciones, de acuerdo a algunos criterios aplicados a la clasificación de las muestras.

La prueba de homogeneidad se refiere a la pregunta: ¿Son las muestras obtenidas de una poblaciones homogéneas (es decir, iguales) con respecto a un cierto criterio de clasificación?

En la prueba de tablas cruzadas, ya sea la fila o la columna puede representar a las poblaciones de donde las muestras son dibujadas.

Una aplicación: Suponga a una junta directiva de una unión de trabajadores deseo encuestar la opinión de sus miembros referente a un cambio en su constitución. La tabla siguiente muestra el resultado de la encuesta enviado a tres uniones locales:

Reacción de una Muestra de tres Grupos de Miembros Locales

Union Local

Reacción	*A*	*B*	*C*
A favor	18	22	10
En contra	7	14	9
No responde	5	4	11

El problema no es determinar si los miembros de unión están en el favor del cambio o no. La pregunta es probar si existe una diferencia significativa en las proporciones de la opinión de los miembros de las tres uniones concerniente al cambio propuesto.

El estadístico del Chi-cuadrado es 9,58 con gl = (3-1)(3-1) = 4.El valor p es igual a 0,048, indicando que existe evidencia moderada contra la hipótesis nula de que los tres locales de la unión son iguales.

A usted podría gustarle utilizar la Prueba de Poblaciones Homogéneas para realizar esta prueba.

Pruebas para la Igualdad de Varias Medias Poblacionales

Generalmente, la mediana proporciona una mejor medida de localización que la media cuando hay algunas observaciones extremadamente grandes o pequeñas; es decir, cuando los datos estansesgados a la derecha o a la izquierda. Por esta razón, el ingreso mediano es utilizado como la medida de localización de la renta por hogar en los Estados Unidos.

Suponga que estamos interesados en probar las medianas de un número k de poblaciones con respecto a la misma variable aleatoria continua.

El primer paso para calcular la prueba estadística es calcular la mediana común de las muestras k combinadas. Luego, se determina para cada grupo el número de observaciones que se encuentran por arriba y por debajo de la mediana común. Las frecuencias resultantes son arregladas en una tabla cruzada de 2 por k. Si las muestras de k están, son de hecho, de poblaciones con la misma mediana, se espera que cerca de una mitad del valor en cada muestra esté sobre la mediana combinada y la otra mitad por debajo de la misma. En el caso de que algunas observaciones sean iguales a la mediana combinada, se podría desechar algunas observaciones cuando se construye la tabla cruzada 2 x k. Bajo esta condición, el estadístico Chi-cuadrado se puede calcular y comparar con el valor p de la distribución Chi-cuadrado con gl = k-1.

Una aplicación ilustrativa: ¿Existen diferencias entre los salarios de los profesores de escuelas primarias públicas y privadas? Los datos de una muestra escogida al azar son descritos en la tabla siguiente (en millares de dólares por año.)

Pública	**Privada**	**Pública**	**Privada**
35	29	25	50
26	50	27	37
27	43	45	34
21	22	46	31
27	42	33	
38	47	26	
23	42	46	
25	32	41	

La prueba de hipótesis es:

H_0: Los sueldos de los profesores de escuelas públicas y privadas son casi iguales.

La mediana de los datos (es decir, combinada) es 33,5. Ahora se determina para cada grupo el número de observaciones que caen por arriba y por debajo de 33,5. Las frecuencias resultantes se muestran en la tabla siguiente:

Tabla Cruzada para Profesores en Escuelas Públicas y Privadas			
	Públicas	Privadas	Total
Sobre la Mediana	6	8	14
Debajo de la Mediana	10	4	14
Total	16	12	28

El estadístico Chi-cuadrado basado en esta tabla es 2,33. El valor p calculado para la prueba estadística con gl = (2-1)(2-1) = 1 es 0,127, por lo tanto, no podemos rechazar la hipótesis nula.

A usted podría gustarle utilizar Prueba de las Medianas .

Prueba de Bondad de Ajuste para Funciones de Masa de Probabilidad

Hay otras pruebas que pudieron utilizar el Chi-cuadrado, por ejemplo la prueba de calidad o bondad de ajuste para variables aleatorias. discretas. Por lo tanto, el Chi-cuadrado es una prueba estadística que mide la "calidad o bondad de ajuste". En otras palabras, mide cuánto se diferencian las frecuencias observadas o reales de las frecuencias esperadas o predichas. Usar una tabla Chi-cuadrado le permitirá descubrir cuan significativa es la diferencia. Una hipótesis nula en el contexto de la prueba Chi-cuadrado es el modelo que se utiliza para calcular sus valores esperados o predichos. Si el valor que usted consigue mediante el cálculo del estadístico Chi-cuadrado es suficientemente alto (con respecto a los valores en la tabla Chi-cuadrado), significa que su hipótesis nula probablemente sea incorrecta.

Deje que Y_1, Y_2, . . ., Y_n sean un sistema de variables aleatorias **discretas** idénticamente distribuidas e independientes. Asuma que la distribución de probabilidad de la Y_i's tiene la función de masa de probabilidad $f_o(y)$. Podemos dividir el sistema de todos los valores posibles de Y_i, i = {1, 2, ..., n}, dentro de m intervalos sin superposición D_1, D_2, ..., D_m. Define los valores de probabilidad p_1, p_2 , ..., p_m como sigue;

$p_1 = P(Y_i \hat{I} D_1)$
$p_2 = P(Y_i \hat{I} D_2)$

:

$p_m = P(Y_i \hat{I} D_m)$

De donde el símbolo Î significa "un elemento de."

Por que la unión de los intervalos mutuamente excluyentes D_1, D_2, ..., D_m es el grupo de valores posibles de Y_i's, (p_1 + p_2 + ... + p_m) = 1. Se define el conjunto de variables aleatorias discretas X_1, X_2, ..., X_m, de donde

X_1= número de Y_i's del cual su valor ÎD_1
X_2= número de Y_i's del cual su valor Î D_2

:
:

X_m= número de Y_i's del cual su valor Î D_m

y (X_1+ X_2+ .... + X_m) = n. Luego el grupo de variables aleatorias discretas X_1, X_2, ..., X_m tendrán una distribución de probabilidad multinomial con parámetros *n* y grupo de probabilidades {p_1, p_2, ..., p_m}. Si los intervalos D_1, D_2, ..., D_m son escogidos tal que np_i ³ 5 para i = 1, 2, ..., m, se obtiene;

$C = S (X_i - np_i)^2 / np_i$.
La suma es sobre i = 1, 2,..., m. El resultado se distingue como c^2_{m-1}.

Para la prueba de bondad de ajuste de la muestra, se formula la hipótesis nula y alternativa como sigue

H_0 : $f_Y(y) = f_o(y)$
H_a : $f_Y(y)$ ¹ $f_o(y)$

At the a a un nivel de significancia, H_0 será rechazada a favor de H_a si

$C = S (X_i - np_i)^2 / np_i$
es mayor que c^2_m

Sin embargo, es posible que en calidad de ajuste de la prueba, uno o más de los parámetros del $f_o(y)$ sean desconocido. Entonces los valores de probabilidad p_1, p_2, ..., p_m tendrán que ser estimados asumiendo que H_0 es verdad y que calcula sus valores estimados de los datos de la muestra. Es decir, otro sistema de valores de probabilidad p'_1, p'_2, ..., p'_m necesitaran ser calculados de modo que los valores (np'_1, np'_2, ..., np'_m) sean los valores previstos estimados de la variable aleatoria multinomial (X_1, X_2, ..., X_m). En este caso, la variable aleatoria C tendrá todavía una distribución Chi-cuadrado, pero con grados de libertad reducidos. En detalle, si la función de probabilidad $f_o(y)$ tiene *r* parámetros desconocidos,

$C = S (X_i - np_i)^2 / np_i$

esta distribuida como c^2_{m-1-r}.

Para esta prueba de calidad de ajuste, formulamos las hipótesis nula y alternativa como

H_0: $f_Y(y) = f_o(y)$
H_a: $f_Y(y) \neq f_o(y)$

A un nivel de significancia a, H_0 será rechazada a favor de H_a si C es mas grande que c^2_{m-1-r}.

Una aplicación: Un dado se lanza 300 veces y las siguientes frecuencias son observadas. Pruebe la hipótesis nula de que el dado no esta influenciado a un nivel 0,05. Bajo la hipótesis nula que el dado no esta influenciado, las frecuencias previstas son todas igual a 300/6 a = 50. Ambas, la frecuencia observada (O) y la esperada (E) se registran en la tabla siguiente junto con la variable aleatoria Y, la cual representa los números en cada uno de los lados del dado:

Prueba de Bondad de Ajuste para Variables Discretas						
Y	1	2	3	4	5	6
O	57	43	59	55	63	23
E	50	50	50	50	50	50

La cantidad

$c^2 = S [(O - E)^2 / E] = 22,04$

es una medida de calidad de ajuste. Si existe un ajuste razonablemente bueno a la distribución hipotética, esta medida estará muy cerca de cero. Porque $c^2_{n-1,\ 0,95} = 11,07$, rechazamos la hipótesis nula de que el dado no esta influenciado.

A usted podría gustarle utilizar este JavaScript para realizar esta prueba.

Para la igualdad *estadística* de dos variables aleatorias caracterizando a dos poblaciones, a usted podría gustarle utilizar la Prueba de Kolmogorov-Smirnov si usted tiene dos sistemas independientes de observaciones aleatorias, una para cada población.

Comparabilidad de la Prueba de Conteos Múltiples

En algunas aplicaciones, tales como control de calidad, es necesario comprobar si el proceso está bajo control. Esto se puede hacer probando si existen diferencias significativas entre el número de "conteos", tomados sobre k períodos de tiempo iguales. Los conteos se suponen de haber sido obtenidos bajo condiciones comparables.

La hipótesis nula es:

H_0: No existe diferencia significativa entre el número de "cuentas" tomados sobre k períodos de tiempo iguales.

Bajo la hipótesis nula, el estadístico:

$S (N_i - N)^2/N$

Tiene una distribución Chi-cuadrado con gl= k-1. Donde i son los números de conteos números, N_i es sus conteos, y $N = SN_i/k$.

Se podría extender esta prueba útil hasta la duración de obtener el i^{th} conteo sea t_i. Luego la prueba estadística anterior se transforma en:

$S [(N_i - t_iN)^2/ t_iN]$

y tiene una distribución Chi-cuadrado con gl = k-1, Donde i son los conteos números, N_i es sus conteos, y $N = SN_i/St_i$.

A usted podría gustarle utilizar Comparabilidad de Conteos Múltiples en JavaScript para comprobar sus cálculos, y realizar algunas experimentaciones numéricas para una comprensión mas profunda de los conceptos.

Coediciones Necesarias para la Prueba Anterior Basada en la Chi-cuadrado

Como cualquier método de prueba estadística, la prueba basada en la Chi-cuadrado debe resolver ciertas condiciones necesarias para su aplicación; de otra forma, cualquier conclusión obtenida podía ser incorrecta o engañosa. Esto es verdad en el caso particular de usar la prueba basada en la Chi-cuadrado para los datos de tablas cruzadas.

Las condiciones necesarias para las pruebas basadas en la Chi-cuadrado para los datos de tablas cruzadas son:

78. Valores esperado mayor 5 en el 80% o más de las celdas.
79. Por otra parte, si el número de celdas es menor a 5, todos los valores esperados deben ser mayores que 5.

Un Ejemplo: Suponga que el número mensual de accidentes reportados en una fábrica en tres turnos de ocho horas es 1, 7, y 7, respectivamente. ¿Son las condiciones de trabajo y la exposición al riesgo similar para todos los turnos? Claramente, la respuesta debe ser, no, ellos no son. Sin embargo, la aplicación de la calidad de ajuste al 0,05, bajo la hipótesis nula de que no hay diferencias en el número de accidentes en los tres turnos, se esperarían 5, 5, y 5 accidentes en cada turno. El estadístico de la prueba Chi-cuadrado es:

$c^2 = S [(O - E)^2 / E] = 4.8$

sin embargo, porque $c^2_{n-1,\ 0,95} = 5,99$, no existe ninguna razón para rechazar que no existe diferencias, lo cual es una conclusión bastante extraña. ¿Que esta errado con esta aplicación?

A usted podría gustarle utilizar este JavaScript para verificar sus cálculos.

Prueba de las Varianzas: ¿Es la Calidad Tan Buena?

Suponga una población que tiene una distribución normal. El gerente debe probar una proposición específica hecha sobre la calidad de la población mediante la prueba de su varianza s^2. Entre tres escenarios posibles, el caso interesante consiste en probar la hipótesis nula siguiente basada en un sistema de n observaciones de muestra aleatoria:

H_0: La variación se encuentra alrededor del valor propuesto.
H_a: La variación es más de lo propuesto, indicando que la calidad es mucho menor que la esperada.

Sobre los cálculos de la varianza estimada S^2 basada en n observaciones, el estadístico:

$c^{½} = [(n-1) . s^2] / s^2$

tiene una distribución Chi-cuadrado con grados de libertad n = n – 1. Este estadístico se utiliza para probar la hipótesis nula anterior.

A usted podría gustarle utilizar la Prueba de las Varianzas en Javascript para comprobar sus cálculos.

Prueba de Igualdad de Varianzas Múltiples

La igualdad de varianzas a través de poblaciones se llama homogeneidad de varianzas o del homocedasticidad. Algunas pruebas estadísticas, tales como la prueba de igualdad de las medias mediante la prueba t y la ANOVA, asumen que los datos vienen de poblaciones que tienen la misma varianza, incluso si la prueba rechaza la hipótesis nula de la igualdad de las medias poblacionales. Si esta condición de la homogeneidad de varianzas no se resuelve, los resultados de la prueba estadística podrían no ser válidos. Heterocedastidad se refiere a la carencia de la homogeneidad de las varianzas.

La Prueba de Bartlett es usada para probar si k muestras tienen varianzas iguales. Compara la Media Geométrica del grupo de varianzas a la media aritmética; por lo tanto, es un estadístico Chi-cuadrado con (k-1) grados de libertad, donde k es el número de categorías en la variable independiente. La prueba es sensible a las salidas de la normalidad. Los tamaños de las muestras no tienen que ser iguales, pero cada uno debe ser por lo menos 6. Justo como la prueba t para dos poblaciones, ANOVA puede dar error cuando la igualdad de la condición de las varianzas no se resuelve.

La prueba estadística de Bartlett es diseñada para probar la igualdad de varianzas a través de grupos contra la alternativa de que las varianzas son desiguales para por lo menos dos grupos. Formalmente,

$\mathbf{H_0}$: Todas las varianzas son casi iguales.

La prueba estadística:

$B = \{S [(n_i -1)LnS^2] S [(n_i -1)LnS_i^2]\}/ C$

En la anterior, S_i^2 es la varianza del iesimo grupo, n_i es el tamaño de la muestra del iesimo grupo, k es el número de grupos, y S^2 es la varianza agrupada. La varianza agrupada es el average ponderado del grupo de varianzas y se define como:

$S^2 = \{S [(n_i -1)S_i^2]\} / S [(n_i -1)]$, sobre todos los i = 1, 2,..,k

y

$C = 1 + \{S [1/(n_i -1)] - 1/ S [1/(n_i -1)] \}/[3(k+1)]$.

A usted podría gustarle utilizar la Igualdad de Varianzas Múltiples en Javascript para comprobar sus cálculos, y realizar ciertas experimentaciones numéricas para una comprensión más profunda de los conceptos.

Regla de 2: Para 3 o más poblaciones, hay una regla práctica conocida como la “Regla de 2”. Según esta regla, se divide la varianza más alta de una muestra por la varianza más baja de la otra muestra. Dado que los tamaños de muestra son relativamente iguales, y el resultado de esta división es menor que 2, las variaciones de las poblaciones son casi iguales.

Ejemplo: Considere las tres muestras escogidas al azar siguientes a partir de tres poblaciones, P1, P2, P3:

	Muestra P1	Muestra P2	Muestra P3
	25	17	8
	25	21	10
	20	17	14
	18	25	16
	13	19	12
	6	21	14
	5	15	6
	22	16	16
	25	24	13
	10	23	6
N	10	10	10
Media	16,90	19,80	11,50
Desv. Están.	7,87	3,52	3,81
SE Media	2,49	1,11	1,20

La Tabla ANOVA

Fuente de Variación	Suma de los Cuadrados	Grados de Libertad	Media al Cuadrado	Estadístico F
Entre Muestras	79,40	2	39,70	4,38
Dentro de las Muestras	244,90	27	9,07	
Total	324,30	29		

Con un F estadístico = 4,38 y un valor p de 0,023, rechazamos la hipótesis nula a un a = 0,05. Esto no es una buena noticia, porque ANOVA, así como las otras dos pruebas t de la muestra, pueden resultar equivocadas cuando la condición de igualdad de las varianzas no se logra.

Prueba de los Coeficientes de Correlación

La Transformación Z de Fisher es una herramienta útil en las circunstancias en las cuales los dos o más coeficientes de correlación independientes deben ser comparados simultáneamente. Para realizar tal prueba se debe evaluar el estadístico Chi-cuadrado:

$c^2 = S[(n_i - 3).Z_i^2] - [S(n_i - 3).Z_i]^2 / [S(n_i - 3)]$, la suma sobre todos los i = 1, 2, .., k.

Donde la transformación Z de Fisher es

$Z_i = 0,5[Ln(1+r_i) - Ln(1-r_i)]$, provista $| r_i | ^1 1$.

Bajo la hipótesis nula:

H_0: Todos los coeficientes de correlación son casi iguales.

La prueba estadística c^2 tiene (k-1) grados de libertad, de donde k es el número de poblaciones.

Una aplicación: Considere los siguientes coeficientes de correlación obtenidos mediante muestreo aleatorio de diez poblaciones independientes.

Población P_i	Correlación r_i	Tamaño de Muestra n_i
1	0,72	67
2	0,41	93
3	0,57	73
4	0,53	98
5	0,62	82
6	0,21	39
7	0,68	91
8	0,53	27
9	0,49	75
10	0,50	49

Usando la fórmula anterior del estadístico $c^2 = 19,916$, que tiene un valor p de 0,02. Por lo tanto, existe una evidencia moderada en contra de la hipótesis nula.

En tal caso, se pueden omitir algunas outliers del grupo, luego se utiliza el JavaScript de Prueba para la Igualdad de Varios Coeficientes de Correlación. Se repite este procedimiento hasta que un subgrupo homogéneo emerja.

Usted podría necesitar el Javascript Determinación del Tamaño de la Muestra en la etapa del diseño de su investigación estadística en la toma de decisión con requisitos subjetivos específicos.

Regresión Lineal Simple: Aspectos Computacionales

El análisis de regresión tiene tres objetivos: predecir, modelar, y la caracterización. ¿Cuál debería ser el orden lógico en el cual se aborden estos tres objetivos de forma tal que uno de ellos guíe y/ o justifique los otros objetivos?. Obviamente, esto dependerá de cual es el objetivo principal. Algunas veces se necesita modelar con el objetivo de realizar mejores predicciones. Por lo tanto, el orden lógico es obvio. Algunas veces simplemente se necesita explicar los hechos, por lo tanto el modelar sería la clave, a pesar de la muestra, la predicción podría ser utilizada para probar el modelo. Con frecuencia, modelar y predecir

utilizan procesos iterativos de los cuales no existe ningún “orden lógico” en el sentido más amplio. Se podría modelar para obtener predicciones, lo cual posibilita mayor control, sin embargo, las iteraciones son fáciles de aparecer, y existen algunas aproximaciones para controlar los problemas.

Fórmulas y Notaciones:

- $\bar{x}$= Sx /n
  Esta es simplemente la media de los valores de x.
- $\bar{y}$= Sy /n
  Esta es simplemente la media de los valores de y.
- $S_{xx} = SS_{xx} = S(x(i) - \bar{x})^2 = Sx^2 - (Sx)^2 / n$
- $S_{yy} = SS_{yy} = S(y(i) - \bar{y})^2 = Sy^2 - (Sy)^2 / n$
- $S_{xy} = SS_{xy} = S(x(i) - \bar{x})(y(i) - \bar{y}) = Sx \times y - (Sx) \times (Sy) / n$
- Pendiente $m = SS_{xy} / SS_{xx}$
- Intercepto, $b = \bar{y} - m . \bar{x}$
- predicción de y = y sombrero(i) = m×x(i) + b.
- Residual(i) = Error(i) = y – y sombrero(i).
- $SSE = S_{errores} = SS_{errores} = S[y(i) - yhat(i)]^2$.
- Desviación Estándar de los residuos = $s = S_{res} = S_{errores} = [S_{res} / (n-2)]^2$.
- Error Estándar de la pendiente (m) = $S_{res} / S_{xx}^{1/2}$.
- Error Estándar del Intercepto (b) = $S_{res}[(S_{xx} + n. \bar{x}^2) /(n \times S_{xx}]^{1/2}$.

Un Ejemplo Computacional: El gerente de una línea de taxis considera que las reparaciones mensuales (Y) de los taxis se encuentran relacionadas a los años de antigüedad (X) de los mismos. Cinco taxis son seleccionados aleatoriamente, y acorde a sus record históricos obtuvimos los datos siguientes: (x, y) = {(2, 2), (3, 5), (4, 7), (5, 10), (6, 11)}. Basado en nuestro conocimiento práctico, y al diagrama de dispersión de los datos, dedujimos la hipótesis de una relación lineal entre la variable de predicción y el costo Y.

Ahora la pregunta es ¿cómo podemos usar de la mejor manera posible (es decir, mínimos cuadrados) la información muestral de manera de estimar la pendiente desconocida (m) y el intercepto (b)? El primer paso para encontrar la linea de mínimos cuadrados es construir una tabla de suma de los cuadrados para encontrar la suma de los valores de x (Sx), e y (Sy), los cuadrados de los valores de x (Sx^2), y (Sy^2), y los productos cruzados de los valores correspondiente de x e y (Sxy), como se muestra en la tabla siguiente:

x	**y**	**x^2**	**xy**	**y^2**
2	2	4	4	4
3	5	9	15	25
4	7	16	28	49
5	10	25	50	100

	6	11	36	66	121
SUMA	**20**	**35**	**90**	**163**	**299**

El segundo paso es sustituir los valores de Sx, Sy, Sx^2, Sxy, y Sy^2 dentro de las formulas siguientes:

$SS_{xy} = Sxy – (Sx)(Sy)/n = 163 - (20)(35)/5 = 163 - 140 = 23$

$SS_{xx} = Sx^2 – (Sx)^2/n = 90 - (20)^2/5 = 90- 80 = 10$

$SS_{yy} = Sy^2 – (Sy)^2/n = 299 - 245 = 54$

Utilice los primeros dos valores para calcular la pendiente estimada:

Pendiente = $m = SS_{xy} / SS_{xx} = 23 / 10 = 2,3$

Para estimar el intercepto de la línea de mínimos cuadrados, emplee el hecho de que el grafico de la línea de mínimos cuadrados siempre pasa a través del punto ($\bar{x}$, $\bar{y}$), por lo tanto,

El intercepto = $b = \bar{y} – (m)(\bar{x}) = (Sy)/ 5 – (2,3) (Sx/5) = 35/5 – (2,3)(20/5) = -2,2$

Por lo tanto la línea de mínimos cuadrados es:

predicción de y = y sombrero = $mx + b = -2,2 + 2,3x$.

Luego de estimar la pendiente y el intercepto, la pregunta es ¿cómo determinamos estadísticamente si el modelo es suficientemente bueno, digamos para predecir. El error estándar de la pendiente es:

Error estándar de la pendiente (m)= $S_m = S_{res} / S_{xx}^{1/2}$,

y su precisión relativa esta medida por los estadísticos

$t_{pendiente} = m / S_m$.

para nuestro ejemplo numérico, esto es:

$t_{pendiente} = 2,3 / [(0,6055)/ (10^{1/2})] = 12,01$

el cual es suficientemente grande, indicando que el modelo ajustado es “bueno”.

Uste se preguntara, ¿en que sentido es la línea de mínimos cuadrados la linea recta que “mejor ajusta” los 5 puntos de los datos. El criterio de mínimos cuadrados elige la línea que minimiza la suma de los

cuadrados de las desviaciones verticales, es decir, residuos = error = y – y sombrero:

SSE = S (y – y sombrero)^2 = S(error)^2 = 1,`1
El valor numérico del SSE se obtiene de la siguiente tabla computacional para nuestro ejemplo numérico.

x Factor de Predicción	**-2,2+2,3x predicción de y**	**y observada**	**error y**	**Error al cuadrado**
2	2,4	2	-0,4	0,16
3	4,7	5	0,3	0,09
4	7	7	0	0
5	9,3	10	0,7	0,49
6	11,6	11	-0,6	0,36
			Suma=0	**SumA=1,1**

Alternativamente, se podría calcular el SSE mediante:

SSE = SS_{yy} – m SS_{xy} = 54 – (2,3)(23) = 54 – 52,9 = 1,1,
como se esperaba

Note que este valor de SSE corresponde con el valor calculado directamente de la tabla anterior. El valor numérico de SSE proporciona la estimación de la variación de los errores s^2:

s^2 = SSE / (n -2) = 1,1 / (5 - 2) = 0,36667

La estimación del valor de error de la varianza, es una medida de variabilidad de los valores de y con respecto a la línea estimada. Obviamente, podríamos calcular también la desviación estándar s de los residuos mediante el cálculo de la raíz cuadrada de la varianza s^2.

Como último paso en la construcción del modelo, el análisis de la tabla de (ANOVA) es construida para lograr la bondad de ajuste general utilizando la prueba F- estadística:

Análisis de los Componentes de la Varianza

Fuente	DF	Suma de los Cuadrados	Media al Cuadrado	Valor F	Prob > F
Modelo	1	52,90000	52,90000	144,273	0,0012
Error	3	SSE = 1,1	0,36667		
Total	4	SS_{yy} = 54			

Para propósitos prácticos, el ajuste es considerado aceptable si el F-estadístico es mas de cinco veces que el valor de F de una tabla distribución F al final de su libro. Note que, el criterio de que el F-estadístico tiene que ser cinco veces mayor que el el de la tabla de distribución F, es independiente del tamaño de la muestra.

Adicionalmente note que existe una relación entre los dos estadísticos lo cual asegura la calidad de la línea de ajuste, es decir el T estadístico de la pendiente y el F estadístico en la tabla de ANOVA. La relación es:

$t^2_{pendiente} = F$

Esta relación puede ser verificada para nuestro ejemplo computacional

Predicciones Mediante la Regresión: Después de haber chequeado estadísticamente la bondad de ajuste del modelo y estar satisfecho de que el factor de predicción (X) contribuye a la predicción de (Y), nos encontramos preparados para utilizar el modelo con confianza. El intervalo de confianza proporciona una manera útil para evaluar la calidad de la predicción. Normalmente uno o mas de las siguientes construcciones son el interés en la predicción mediante la regresión:

93. Un intervalo de confianza para un valor futuro simple de Y correspondiente a un valor de X elegido.
94. Un intervalo de confianza para un simple valor sobre la línea.
95. Una región de confianza para toda la línea completa.

Estimación de Intervalo de Confianza para Valores Futuros: Un intervalo de confianza de interés puede ser utilizado para evaluar la precisión de un valor simple (futuro) de Y correspondiente a un valor seleccionado de X (digamos, X_0). Este JavaScript proporciona intervalos de confianza para un valor estimado de Y correspondiente a X_0 con un nivel de confianza deseable de 1 - a.

$Y_p \pm S_e \, . \, t_{n-2,\, a/2} \{1/n + (X_0 - \bar{x})^2 / S_x\}^{1/2}$

Estimación del Intervalo de Confianza para un Punto Sobre la Línea: Si un valor particular de la variable de predicción (por ejemplo, X_0) tiene una importancia especial, un intervalo de confianza sobre el valor del criterio de la variable (digamos, el average de Y hacia X_0) correspondiente a X_0 podría ser de interés. Este JavaScript proporciona un intervalo de confianza al valor estimado Y correspondiente a X_0 con un nivel de confianza deseable de 1 - a.

$$Y_p \pm S_e \, . \, t_{n-2,\, a/2} \{ 1 + 1/n + (X_0 - \bar{x})^2 / S_x \}^{1/2}$$

Es interesante comparar los dos diferentes intervalos de confianza anteriores. El primero tiene un intervalo de confianza mayor, el cual refleja una menor precisión proveniente de la estimación de un simple valor futuro de y en vez del valor de la media calculada para el segundo tipo de intervalo de confianza. Este último a su vez puede ser utilizado para identificar cualquier anomalía o outlier en los datos.

Región de Confianza, la Línea de Regresión como la Totalidad: Cuando estamos interesados en toda la línea, la región de confianza nos permite hacer juicios simultáneos sobre nuestras estimaciones Y para un número de valores de la variable de predicción X. Con el objetivo de que la región cubra adecuadamente el rango de interés de la variable de predicción X, el número de datos debe ser mayor a 10 pares de observaciones.

$$Y_p \pm S_e \{ (2 \, F_{2,\, n-2,\, a}) \, . \, [1/n + (X_0 - \bar{x})^2 / S_x] \}^{1/2}$$

En todos los casos, el JavaScript proporciona los resultados para los valores nominales (x). Para otros valores de X se podrían utilizar directamente otros métodos computacionales, métodos gráficos, o interpolación lineal para obtener resultados aproximados. Estas aproximaciones están en la dirección correcta, es decir, son un poco más amplio que los valores exactos.

Interpolación Lineal: Para estimar los límites inferior (superior) a un valor dado X, se podría utilizar la interpolación lineal en dos puntos vecinos conocidos a X, digamos XL y XU, como sigue:

El límite inferior aproximado a X es:

LL(XL) + [LL(XU) – LL (XL)] × [X – XL] / [XU – XL]
De igual manera para el límite superior a X es:
UL(XL) + [UL(XU) – UL (XL)] × [X – XL] / [XU – XL]
La aproximación resultante es de tipo conservadora, por lo tanto se encuentra en el lado seguro.

Modelos de Regresión y Análisis

Muchos problemas surgen cuando se describe cómo las variables están relacionadas. El más simple de todos los modelos que describe la relación entre dos variables es un modelo lineal, o de línea recta. La regresión lineal es siempre lineal en los coeficientes que son estimados, y no necesariamente lineal en las variables.

El método más simple de dibujar un modelo lineal es “calcular visualmente” una línea a través de los datos sobre un diagrama, pero un modelo mas elegante sería el método convencional de mínimos cuadrados, el cual encuentra la línea al reducir al mínimo la suma de las distancias verticales entre los puntos observados y la línea ajustada. entienda que ajustando la “mejor” línea de acuerdo a la vista es muy difícil, especialmente cuando hay mucha variabilidad residual en los datos.

Sepa que existe una conección simple entre los coeficientes numericos en la ecuación de regresión, la pendiente y la intercepción de la línea de regresión.

Sepa también que un simple sumario estadístico, como el coeficiente de correlación, no dice la historia completa. Un gráfico de dispersión es un complemento esencial para examinar la relación entre dos variables.

Una vez más, la línea de regresión es un grupo de estimaciones para la variable trazada en el eje de las y. Tiene una la forma de y = b + mx, donde m es la pendiente de la línea. La pendiente es el crecimiento sobre la corrida. Si una línea va hasta 2 por cada 1, su pendiente es 2.

La línea de regresión pasa a través de un punto con coordenadas de (media de los valores de x, media de los valores de y), conocidos como el punto media-media.

Si se introduce cada valor de x en la ecuación de regresión, se obtiene un valor estimado para y. La diferencia entre la y estimada y la y observada se llama un residual, o un término de error. Algunos errores son positivos y otros negativos. Mediante la suma de los cuadrados de los errores más la suma de los cuadrados de las estimaciones se obtiene la suma de cuadrados de Y. La línea de regresión es la línea que reduce al mínimo la varianza de los errores. El error de la media es cero; de esta forma, se reduce al mínimo la suma de los errores al cuadrado.

La razón para encontrar la línea mas apropiada es que se pueda hacer una predicción razonable de lo que sería y si x es conocida (no vise-versa.)

r^2 es la varianza de las estimaciones divididas por la varianza de Y. r es el tamaño de la pendiente de la línea de regresión, en términos de desviaciones estándar. Es decir, es la pendiente de la línea de regresión si utilizamos la X y la Y estandardizadas. Esto es cuántas desviaciones estándar de Y se movería hacia arriba, cuando se mueve una desviación estándar de X hacia la derecha.

Coeficiente de Determinación: Otra medida de la cercanía de los puntos a la línea de regresión es el Coeficiente de Determinación:

$r^2 = S_{ysombrero\ ysombrero} / S_{yy}$

el cuál es la cantidad de la desviación al cuadrado en Y, la cual es explicada por los puntos en la menor línea de regresión de los cuadrados.

Homocedasticidad y Heterocedasticidad: Homocedasticidad (homo = iguales, skedasis = dispersando) es una palabra usada para describir la distribución de los puntos de referencias alrededor de la línea del mejor ajuste. El término opuesto es Heterocedasticidad. Brevemente, la homocedasticidad significa que los puntos de referencias están distribuidos igualmente sobre la línea del mejor ajuste. Por lo tanto, el homocedasticidad significa la varianza constante sobre todos los niveles de factores. Heterocedasticidad significa que los puntos de referencias se encuentran agrupados tanto por arriba como pro debajo de la línea en un patrón no-igual.

Análisis de Regresión Estandardizada: La escala de medidas usadas para medir X e Y tiene su mayor impacto en la ecuación de la regresión y el coeficiente de correlación. Este impacto es más drástico cuando se comparan dos ecuaciones de regresión que tienen diferentes escalas de medida. Para superar estas desventajas, se deben estandardizar X e Y antes de construir la regresión e interpretar los resultados. En este modelo, la pendiente es igual al coeficiente de correlación r. Note que la derivada de la función Y con respecto a la variable dependiente X es el coeficiente de correlación. Por lo tanto, existe una semejanza en el significado de r en estadística y la derivada del cálculo, este es que su signo y su magnitud revelan el crecimiento/ decrecimiento y la tasa de variación, como lo hace la derivada de una función.

En el modelo de regresión general **la pendiente estimada y la intercepción están correlacionadas** ; Por lo tanto, cualquier error en estimar la pendiente influencia la estimación de la intercepción. Una de las ventajas principales de usar los datos estandardizados es que la intercepción es siempre igual a cero.

Regresión cuando X e Y son Aleatorias: La regresión lineal simple de los mínimos cuadrados tiene entre sus condiciones que los datos para las variables independientes (X) son conocidos sin error. De hecho, los resultados estimados son condicionados a que cualquier error que

sucediera este presente en los datos independientes. Cuando los datos de X tienen un error asociado a ellos el resultado influencia la pendiente hacia abajo. Un procedimiento conocido como la regresión de Deming puede manejar este problema perfectamente. Estimaciones de pendientes influenciadas en polarización negativa de la cuesta (debido al error en X) pueden ser evitadas usando la regresión de Deming.

Si X e Y son variables aleatorias, el coeficiente de correlación R se refiere a menudo como el **Coeficiente de Confiabilidad.**

La Relación entre la Pendiente y el Coeficiente de Correlación: con un poco de manipulación algebraica, se podría mostrar que el coeficiente de correlación está relacionado con la pendiente de las dos líneas de regresión: Y en X, y X en Y, denotada por m_{yx} y m_{xy}, respectivamente:

$R^2 = m_{yx} . m_{xy}$

Líneas de la Regresión hacia el Origen: Con frecuencia, las condiciones de un problema práctico requieren que la línea de regresión pase por el origen (x = 0, y = 0). En tal caso, la línea de regresión tiene un solo parámetro, el cual es su pendiente:

$m = S (x_i \acute{} y_i)/ Sx_i^2$

Note que, para los modelos que omiten la intercepción, es generalmente conveniente que R^2 no sea definido o si quiera considerado.

Modelos de Parábola: Las regresiones de parábola tienen tres coeficientes con forma general:

$Y = a + bX + cX^2$,

donde

$c = \{ S (x_i - xbarra)^2 \times y_i - n[S(x_i - xbarra)^2 \times Sy_i]\} / \{n S(x_i - xbarra)^4 - [S(x_i - xbarra)^2]^2\}$
$b = [S(x_i - xbarra) y_i]/[S(x_i - xbarra)^2] - 2 \times c \times xbarra$
$a = \{Sy_i - [c \times S(x_i - xbarra)^2)\}/n - (c \times xbarra \times xbarra + b \times xbarra)$,

Donde x barra es la media de x_i's.

Las aplicaciones de la regresión cuadrática incluyen el ajuste de las curvas de oferta y demanda en econometría y el ajuste de las funciones de costos de ordenes y de manutención en el control de inventario para encontrar la cantidad que orden óptima.

A usted podría gustarle utilizar el Javascript de Regresión Cuadrática para comprobar sus cálculos manuales. Para grados mayores a la

cuadrática, a usted podría gustarle utilizar el Javascript Regresiones Polinomiales.

Regresión Lineal Múltiple: Los objetivos en un problema de regresión múltiple son esencialmente iguales que para una regresión simple. Mientras que los objetivos siguen siendo iguales, mientras más predictores tenemos, los cálculos y las interpretaciones son más complicadas. Con la regresión múltiple, podemos utilizar más de un predictor. Esto siempre es mejor, sin embargo, ser parsimonioso, es decir, utilizar tan pocas variables como predictores sean necesarios para conseguir un pronóstico razonablemente exacto. La regresión múltiple es mejor modelada con el paquete estadísticos comerciales como el SAS y SPSS. El pronóstico toma la forma:

$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n,$$

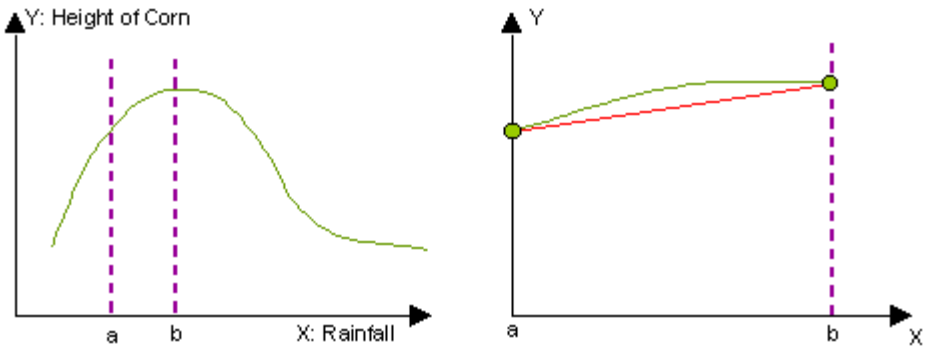
donde b_0 es la intercepción, b_1, b_2, . . . b_n son los coeficientes que representan la contribución de las variables independientes X_1, X_2,..., X_n.

Para muestras de tamaño pequeño, a usted podría gustarle utilizar el Javascript de Regresión Lineal Múltiple.

Que es la Auto-Regresión: En el análisis de series de tiempo y técnicas de pronóstico, la regresión lineal es comúnmente utilizada para combinar valores presentes y pasados de una observación para pronosticar su valor futuro. El modelo se llama un modelo auto-regresivo. Para mas detalles y para la implementación del proceso visite el Javascript Modelo Auto-regresivo JavaScript.

Que es una Regresión Logística: La regresión logística estándar es un método para modelar datos binarios (por ejemplo, ¿Esa persona fuma o no?, ¿Esa persona sobrevivirá a una enfermedad, o no?). La regresión logística *Poligámica* es un método para modelar más de dos opciones (por ejemplo, ¿Esa persona toma el autobús, conduce un coche o toma el subterráneo?, ¿En esa oficina usan WordPerfect, Word, u otro programa de oficina?).

¿Por qué la Regresión Lineal? El estudio de la altura de la cáscara del maíz (es decir, espiga de trigo) con respecto a las lluvias ha mostrado tener la curva siguiente de la regresión:



Claramente, la relación es altamente no lineal; sin embargo, si se está interesado en un rango “pequeño” (digamos, para un área geográfica específica, como la región del norte del Valparaíso) la condición de linealidad podría ser satisfactoria. Una aplicación típica se representa en la figura anterior de la cual se esta interesado en predecir la altura del maíz en un área con precipitación en el rango [a, b]. Magnificando el proceso a la escala que nos permita realizar una regresión lineal útil. Si el rango no es suficientemente corto, se podría subdividir el rango mediante el mismo proceso de ajustar algunas líneas, una para cada sub-intervalo.

Cambios Estructurales: Cuando se ha estimado un modelo de regresión usando los datos disponibles, un conjunto de datos adicionales podría llegar a estar disponible. Para probar si el modelo anterior sigue siendo válido o si los dos modelos separados son equivalentes o no, uno puede utilizar la prueba descrita en el sitio Análisis de la Covarianza.

A usted podría gustarle utilizar el Javascript Análisis de Regresión para comprobar sus cálculos y realizar experimentaciones numéricas para una comprensión mas profunda de los conceptos.

Proceso de Selección del Modelo de Regresión

Cuando se tiene más de una ecuación de regresión basada en datos, para seleccionar el “mejor modelo”, se deben comparar:

96. R-Cuadrados: Es decir, el porcentaje de variación [de hecho, la suma de los cuadrados] en Y, considerado por la variación en X capturada por el modelo.
97. Cuando se quiere comparar modelos de tamaños diferentes (diversos números de variables independientes (p) y/o diferente tamaño de muestra n), usted debe utilizar el R-Cuadrado ajustado, porque el r-cuadrado usual, tiende a crecer con el número de variables independientes.

$$r^2_a = 1 - (n - 1)(1 - r^2)/(n - p - 1)$$

98. La desviación estándar de los términos del error, es decir, el valor y observado, - el valor y predicho para cada x.
99. Tendencias en errores como función de la variable de control x. Tendencias sistemáticas no son poco frecuentes.
100. El T-estadístico de parámetros individuales.
101. Los valores de los parámetros y sus contenidos a refuerzos contenidos. El valor de
102. $F_{df1\ df2}$ para la evaluación general. Donde df1 (numerador de grados de libertad) es el número de predictores linealmente independientes en el modelo asumido menos el número de predictores linealmente independientes en el modelo restricto; es decir, el número de las restricciones linealmente independientes impuestas ante el modelo asumido, y df2 (denominador de grados de libertad) es el número de observaciones menos el número de predictores linealmente independientes en el modelo asumido.

    El estadístico F observado debe exceder no simplemente el valor crítico seleccionado de la tabla F, pero por lo menos cuatro veces el valor crítico.

Finalmente en la estadística para el negocio, existe una opinión de que con más de 4 parámetros, se puede ajustar a un elefante, por lo tanto, si se procura ajustar una función de regresión que dependa de muchos parámetros, el resultado no debe ser visto como muy confiable.

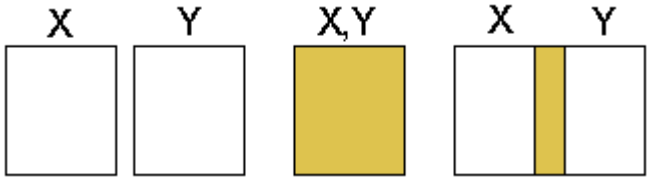
Covarianza y Correlación

Suponga que X e Y son dos variables aleatorias para el resultado de un experimento aleatorio. La covarianza de X e Y es definida por:
Cov (X, Y) = E{[X - E(X)][Y - E(Y)]}
y, dado que las varianzas son estrictamente positivas, la correlación de X e Y es definida por
r (X, Y) = Cov(X, Y) / [sd(X) . sd(Y)]
La correlación es una versión escalada de la covarianza; observe que los dos parámetros tienen siempre el mismo signo (positivo, negativo, o 0). Cuando el signo es positivo, se dice que las variables están correlacionadas positivamente; cuando el signo es negativo, se dice que las variables están correlacionadas negativamente; y cuando es 0, las variables no tienen correlación.

Note que la correlación entre dos variables aleatorias se debe a menudo solamente al hecho de que ambas variables están correlacionadas con la misma tercera variable.

Como estos términos sugieren, la covarianza y la correlación miden ciertos tipos de comportamiento en ambas variables. La correlación es muy similar a la derivada de una función que usted debe haber estudiado en secundaria.

Coeficiente de Determinación: El cuadrado del coeficiente de correlación r^2 indica la proporción de la variación en una variable que pueda asociada a la varianza de otra variable. Las tres posibilidades típicas se representan en la figura siguiente:



La proporción de varianzas compartidas por dos variables para diferentes valores de coeficiente de determinación: $r^2 = 0$, $r^2 = 1$, and $r^2 = 0{,}25$, Como es mostrado en la parte sombreada de la figura anterior.

Propiedades: Los ejercicios siguientes ofrecen algunas propiedades básicas de los valores esperados. La herramienta principal que usted necesitará es el hecho de que el valor esperado es una operación lineal.

A usted podría gustarle utilizar este Applet para realizar algunas experimentaciones numéricas para:

103. Mostrar que $E[X/Y] \neq E(X)/E(Y)$.
104. Mostrar que $E[X \cdot Y] \neq E(X) \cdot E(Y)$.
105. Mostrar que $[E(X \cdot Y)^2] \leq E(X^2) \cdot E(Y^2)$.
106. Mostrar que $[E(X/Y)^n] \geq E(X^n)/E(Y^n)$, para todo n.
107. Mostrar que $Cov(X, Y) = E(XY) - E(X)E(Y)$.
108. Mostrar que $Cov(X, Y) = Cov(Y, X)$.
109. Mostrar que $Cov(X, X) = V(X)$.
110. Mostrar que: Si X e Y son variables aleatorias independientes, entonces $Var(XY) = 2\,V(X) \cdot V(Y) + V(X)(E(Y))^2 + V(Y)(E(X))^2$.

Pearson, Spearman, y la Correlación en el Punto Biserial

Existen medidas que describen el grado en el cual dos variables se encuentran linealmente relacionadas. Para la mayoría de estas medidas, la correlación se expresa como un coeficiente que se extienda en el rango 1,00 a –1,00. Un valor de 1 indica una relación lineal perfecta, tal que sabiendo el valor de una variable permitirá la predicción perfecta del valor de la variable relacionada. Un valor de 0 indica ninguna predicción posible mediante un modelo lineal. Con valores negativos indicando que, cuando el valor de una variable es mayor al promedio, el valor de la otra es menor que el promedio (y viceversa); y valores positivos tal que, cuando el valor de una variable es alto, lo es también el valor de la otra (y viceversa).

La correlación es similar a la derivada que usted ha aprendido en cálculo (curso determinista).

La correlación del producto del Pearson es un índice de relación lineal entre dos variables.

La Correlación de Pearson es

$r = S_{xy} / (S_{xx} \acute{} S_{yy})^{0,5}$

Una relación positiva indica que si un valor individual de x está sobre la media de los x's, entonces este valor individual de x es podría tener un valor de y que esté sobre la media de las y's, y viceversa. Una relación negativa sería un valor de x sobre la media de x y un valor de y debajo de la media de y. Esta es una medida de relación entre variables y un índice de la proporción de diferencias individuales en una variable que puede estar asociada a las diferencias individuales en otra variable.

Note que, el coeficiente de correlación es la media de los valores de los productos cruzados. Por lo tanto, si se tienen tres valores para r de 0,40, 0,60 y 0,80, no se podría decir que la diferencia entre r =0,40 y r = 0,60 es igual a la diferencia entre r = 0,60 y r = 0,80, o que r = 0,80 es dos veces más grande que r = 0,40 porque la escala de valores para el coeficiente de correlación no es un intervalo o un cociente, es solamente ordinal. Por lo tanto, todo lo que se puede decir es que, por ejemplo, un coeficiente de correlación de +0,80 indica una alta relación lineal positiva y que un coeficiente de correlación de +0,40 indica una relación lineal positiva más baja.

El cuadrado del coeficiente de correlación iguala la proporción de la varianza total en Y la cual pueda ser asociada a la variación en x. Esto podría decirnos cuánto de la varianza total de una variable puede estar asociada a la varianza de otra variable.

Observe que un coeficiente de correlación proviene de la correlación lineal. Si los datos forman una parábola, entonces una correlación lineal de x e y producirá un valor r igual a cero. Por lo tanto, se debe ser cuidadoso y observar los datos note que una correlación coeficiente, hecho en correlación lineal.

Estadísticos estándar para la Prueba de Hipótesis: H_0: $r = r_0$, es la transformación normal de Fisher:

$z = 0,5[Ln(1+r) - Ln(1-r)]$, con media $m = 0,5[Ln(1+ r_0) - Ln(1-r_0)]$, y desviacón estándar $s = (n-3)^{-½}$.
Habiendo construido un intervalo de confianza deseable, digamos [a, b], basado en el estadístico Z, este tiene que ser transformado de nuevo a la escala original. Esto es, el intervalo de confianza:
$(e^{2a} -1)/ (e^{2a} +1)$, $(e^{2b} -1)/ (e^{2b} +1)$.

Provisto de que $| r_0 |$ ¹ 1, y $| r_0 |$ ¹ 1, y n es mayor a 3.

Alternativamente,

$$\{1+ r - (1-r) \exp[2z_a/(n-3)^{½}]\} / \{1+ r + (1-r) \exp[2z_a/(n-3)^{½}]\} , \quad \text{and}$$

$$\{1+ r - (1-r) \exp[-2z_a/(n-3)^{½}]\} / \{1+ r + (1-r) \exp[-2z_a/(n-3)^{½}]\}$$

A usted podría gustarle utilizar esta calculadora para sus cálculos necesarios. Usted podría realizar la Prueba del Coeficiente de Correlación de la Población.

Correlación de Spearman es usado como una versión no paramétrica del de Pearson. Se expresa como:

$r = 1 - (6 S d^2) / [n(n^2 - 1)]$,

de donde d es el rango de diferencia entre los pares X e Y.

El coeficiente de correlación Spearman se puede derivar algebraicamente de la fórmula de correlación de Pearson, haciendo uso de suma de series. Pearson contiene expresiones para S X(i), S Y(i), S $X(i)^2$, y $SY(i)^2$.

En el caso del Spearman, el X(i)'s y el Y(i)'s son rangos, y así que la suma de los rangos, y la de los rangos al cuadrado, son enteramente determinadas por el número de casos (sin ningún lazo).

$S i = (n+1)n/2$, $S i^2 = n(n+1)(2n+1)/6$.

Por lo tanto la formula de Spearman es igual a:

$[12P - 3n(n+1)^2] / [n(n^2 - 1)]$,

de donde P es la suma de los productos de cada par de filas X(i) Y(i). Esto se reduce:

$r = 1 - (6 S d^2) / [n(n^2 - 1)]$,

de donde d es la diferencia de rangos entre cada par X(i) e Y(i).

Una consecuencia importante de esto es que si incluyen filas en la fórmula de Pearson, se consigue exactamente el mismo valor numérico que es obtenido incluyendo filas en la fórmula de Spearman. Esto podría impresionar a aquellos que les gusta adoptar lemas simplistas, tales como "Pearson es para intervalo de datos, Spearman es para datos alineados". Spearman no trabaja muy bien si existen muchas filas vinculadas. Esto se debe a que la formula para calcular la suma de los rangos al cuadrado no tiene vigencia. Si se tienen muchos rangos vinculadas, utilice la formula de Pearson.

Se podría utilizar esta medida como herramienta para la toma de decisiones:

Valor de |r|	Interpretación
0,00 – 0,40	Pobre
0,41 – 0,75	Justa
0,76 – 0,85	Buena
0,86 – 1,00	Excelente

Esta interpretación es extensamente aceptada, y muchas revistas científicas publican rutinariamente trabajos que usan esta interpretación para sus resultados estimados, igualmente para la prueba de hipótesis.

Correlación del Punto-Biserial se utiliza cuando una variable aleatoria es binaria (0, 1) y la otra es una variable aleatoria continua; la fortaleza de la relación es medida por la correlación del punto-biserial:

$r = (X_1 - X_0)[pq/S^2]^{½}$

Donde X_1 y X_0 son las medias de los valores que tienen valores 0 y 1, y p y q son sus proporciones, respectivamente. S^2 es la *varianza de la muestra* de la variable aleatoria continua. Esta es una versión simplificada de la correlación de Pearson para el caso cuando una o dos variables aleatoria (0, 1) es una variable aleatoria nominal.

Observe también que r tiene la característica de puesto invariante para cualquier escala positiva. Esto es ax + cm y by + d, tienen el mismo r que x e y, para cualquier positivo a y b.

Correlación, y Niveles de Significancia

Es intuitivo que con muy pocos puntos de referencias, una alta correlación puede no ser estadísticamente significativa. Se podrían ver declaraciones como por ejemplo, “la correlación es significativa entre x e y a un nivel a = 0,05 y “la correlación es significativa a un nivel a = 0,05”. La pregunta es: ¿Cómo se determinan estos números?

Usando la simple correlación r, la fórmula para el F estadístico es:

$F= (n-2)\ r^2 / (1-r^2)$, donde n es por lo menos 2.

Como se puede ver, el estadístico F es función monotónica con respecto a ambas: r^2, y el tamaño de la muestra n.

Note que la prueba para la significancia estadística de un coeficiente de correlación requiere que las dos variables estén distribuidas como normal de doble variación.

Independencia contra Correlación

En el sentido que se utiliza en estadística; es decir, como asunción en la aplicación de una prueba estadística; una muestra aleatoria de la población entera proporciona un sistema de las variables aleatorias X1...., Xn, que se distribuyen idénticamente y que son mutuamente independiente. Mutuamente independiente es más fuerte que en pares independencia. Las variables aleatorias son mutuamente independientes si su distribución común es igual al producto de sus distribuciones marginales.

En el caso de la normalidad común, la independencia es equivalente a correlación cero, pero no en general. La independencia implicará correlación cero pero no inversamente. Note que no todas las variables aleatorias tienen un primer momento, dejan solamente un segundo momento, y podría no existir un coeficiente de correlación.

Sin embargo; si el coeficiente de correlación de dos variables aleatorias no es cero, entonces las variables aleatorias no son independientes.

Cómo Comparar Dos Coeficientes de Correlación

Dado que dos poblaciones tienen distribuciones normales, deseamos probar la hipótesis nula siguiente con respecto a la igualdad de los coeficientes de correlación:

H_o: $r_1 = r_2$,

Basado en dos coeficientes de correlación observados r_1, y r_2, obtenidos a partir de la muestra aleatoria de tamaño n_1 y n_2, respectivamente, tal que $|r_1| \neq 1$, y $|r_2| \neq 1$, y n_1, n_2 son ambos mayores a 3. Bajo condición de normalidad y de la hipótesis nula, la prueba estadística es:

$Z = (z_1 - z_2) / [\ 1/(n_1-3) + 1/(n_2-3)\ ]^{½}$
donde:
$z_1 = 0{,}5\ Ln\ [\ (1+r_1)/(1-r_1)\ ]$,
$z_2 = 0{,}5\ Ln\ [\ (1+r_2)/(1-r_2)\ ]$,

y n_1= el tamaño de muestra asociada a r_1, y n_2 = tamaño de la muestra asociada a r_2.

La distribución del Z-estadístico es la normal estándar (0,1); Por lo tanto, se puede rechazar H_0 si $|Z| > 1{,}96$ a un nivel de confianza de 95%.

Una aplicación: Suponga $r_1 = 0{,}47$, $r_2 = 0{,}63$ se obtienen a partir de dos muestras aleatorias independientes de tamaño $n_1=103$, y $n_2 = 103$,

respectivamente. Por lo tanto z_1 = 0,510, and z_2 = 0,741, with Z-statistics:

$Z = (0,510 – 0,7)/ [1/(103-3) + 1/(103-3)]^{½} = -1,63$

Este resultado no esta dentro del área de rechazamiento de los valores críticos de dos colas al a = 0,05, por lo tanto no es significativo. Consecuentemente, no hay suficiente evidencia para rechazar la hipótesis nula que los dos coeficientes de correlación son iguales.

Ciertamente, esta prueba puede ser modificada y aplicada para pruebas de hipótesis con respecto a la correlación de población r basada en un r observado obtenido de una muestra aleatoria de tamaño n:

$Z = (z_r - z_r) / [1/(n-3)]^{½}$,

dado que | r | ¹ 1, y | r | ¹ 1, y n es mayor a 3.

Prueba de la igualdad de dos correlaciones dependientes: En la prueba de hipótesis de no diferencia entre dos coeficientes de correlación de la población:

H_0: r (X, Y) = r (X, Z)

En contra de la alternativa:

H_a: r (X, Y) ¹ r (X, Z)

con una covarianza común X, se podría usar la siguiente prueba estadística:

$t = \{ (r_{xy} - r_{xz}) [(n-3)(1 + r_{yz})]^{½}] \} / \{2(1-r_{xy}^2 - r_{xz}^2 - r_{yz}^2 + 2r_{xy}r_{xz}r_{yz})\}^{½}$,

con n - 3 grados de libertad, donde n es el tamaño de la muestra triplicado-ordenado, tal que todo valor absoluto de los r's no sean iguales a 1.

Ejemplo numérico: Suponga que n = 87, r_{xy} = 0,631, r_{xz} = 0,428, y r_{yz} = 0,683, el t estadístico es igual a 3,014, con el valor p igual a 0,002, indicando una fuerte evidencia en contra de la hipótesis nula.

R^2 Ajustado: En el proceso de modelamiento basado en valores de R^2 es necesario y significativo ajustar los R^2's a sus grados de libertad. Cada **Adjustado** R^2 es calculado por:

$1 - [(n - i)(1 - R^2)] / (n - p)$,

de donde i es igual a 1 si existe una intercepción y 0 si no; n es el número de observaciones usadas para ajustar el modelo; y p es el número de parámetros en el modelo.

A usted podría gustarle utilizar el Javascript de la Prueba del Coeficiente de Correlación de la Población para ejecutar ciertas experimentaciones numéricas para validar sus cálculos y para una comprensión más profunda de los conceptos.

Condiciones y la Lista de Comprobación para Modelos lineales

Casi todos los modelos de realidad, incluyendo los modelos de regresión, tienen asunciones que deben ser verificadas para que el modelo tenga la fuerza para probar hipótesis y por lo tanto para hacer predicciones acertadas.

La siguiente lista contiene las asunciones básicas (es decir, condiciones) y las herramientas para comprobar estas condiciones necesarias.

111. Cualquier outlier desapercibido puede tener un impacto importante en el modelo de regresión. Los outliers son algunas observaciones que no se ajustan bien al "mejor" modelo disponible. En tal caso uno, se necesita primero investigar la fuente de los datos, si existe duda sobre la exactitud o la veracidad de la observación, debería ser quitada y el modelo debería ser reajustado.

    A usted podría gustarle utilizar el Javascript para la Determinación de los outliers para realizar algunas experimentaciones numéricas para validar y obtener una comprensión más profunda de los conceptos

112. La variable dependiente Y es una función lineal de la variable independiente X. Esto se puede comprobar examinando cuidadosamente todos los puntos en el diagrama de dispersión, y ver si es posible limitarlos dentro de dos líneas paralelas. Usted puede utilizar también la Prueba para Detectar la Tendencia para comprobar esta condición.
113. La distribución de la residual debe ser normal. Usted puede comprobar esta condición usando la Prueba de Lilliefors.
114. Las residuales deben tener una media igual a cero, y una desviación estándar constante (es decir, condición homocedasticidad). Usted puede comprobar esta condición dividiendo los datos de las residuales en dos o más grupos; este acercamiento se conoce como la prueba de Goldfeld-Quandt. A usted podría utilizar el Proceso de Prueba de Estacionalidad para comprobar esta condición.

115. Las residuales constituyen un sistema de variables aleatorias. Usted puede utilizar la Prueba de Aleatoriedad y la Prueba de Aleatoriedad con Fluctuaciones para comprobar esta condición.

116. El Estadístico Durbin-Watson (D-W) cuantifica la correlación serial de errores de mínimos cuadrados en su forma original. El estadístico D-W se define por:

Estadístico D-W = $S_2^n (e_j - e_{j-1})^2 / S_1^n e_j^2$,

dondee e_j es el error j-esimo. El D-W toma valores dentro [0, 4]. Para una correlación no serial, se espera un valor cerca de 2. Con una correlación serial positiva, las desviaciones adyacentes tienden a tener el mismo signo, por lo tanto la D-W se convierte en menos de 2; Mientras que, con la correlación serial negativa, alternando los signos de los errores, D-W toma los valores mayores que 2. Para un ajuste de mínimos cuadrados donde el valor de D-W es significativamente de 2, las estimaciones de las varianzas y de las covarianzas de los parámetros (es decir, coeficientes) podrían encontrarse erradas, siendo demasiado grandes o demasiado pequeñas. La correlación serial de las desviaciones se encuentran presentes también en el análisis y pronóstico de las series de tiempo. Usted puede utilizar la Medida de Exactitud en Javascript para comprobar esta condición.

La " buena" ecuación de regresión candidata es analizada mucho mas a fondo usando un diagrama de residuales contra las variables independientes. Si se observan algunos patrones en el gráfico; por ejemplo, una indicación de una variación no-constante; entonces existe la necesidad de transformar los datos. Las siguientes son las transformaciones comúnmente usadas:

- X' = 1/X, para X no-cero.
- X' = Ln (X), para X positivo.
- X' = Ln(X), Y' = Ln (Y), para X e Y positivos.
- Y' = Ln (Y), para Y positivo.
- Y' = Ln (Y) - Ln (1-Y), para Y positivo, menor que uno.
- Y' = Ln [Y/(100-Y)], conocida como la *Transformación logística*, la cual es útil para las funciones de forma S.
- Tomar la raíz cuadrada de la variable aleatoria, de Poisson, la variable transformada es mas simétrica. Esta es una transformación útil en el análisis de regresión con observaciones de Poisson. Este también estabiliza las variaciones residuales.

Transformaciones de la Caja-Cox Box-Cox Transformations: La transformación de la Caja-Cox, (abajo), se puede aplicar a un regresor, una combinación de regresores, y/o a la variable dependiente (y) en la regresión. El objetivo de hacerlo es generalmente para hacer los residuales de la regresión mas

homocedásticos (es decir, independientes y distribuidos idénticamente) y más cerca a una distribución normal:

$(y^l - 1) / l$ para un l constante, diferente de cero, y log(y) para l = 0.

A usted podría gustarle utilizar el Análisis de Regresión con Herramientas de Diagnóstico en Javascript para comprobar sus cálculos, y para realizar ciertas experimentaciones numéricas para una comprensión más profunda de los conceptos.

Análisis de Covarianza: Comparando las Pendientes

Considere las dos muestras siguientes de tratamientos independientes antes-y-después.

Valores de Covarianza X y una Variable Dependiente Y

Tratamiento-I		**Tratamiento-II**	
X	*Y*	*X*	*Y*
5	11	2	1
3	9	6	7
1	5	4	3
4	8	7	8
6	12	3	2

Deseamos probar la prueba hipótesis siguiente de dos medias de las variables dependientes Y1, y Y2:

H_0: La diferencia entre las dos medias es un valor dado M.
H_a: La diferencia entre las dos medias es absolutamente diferente al propuesto.

Puesto que nos enfrentamos a variables dependientes, es natural investigar los coeficientes de la regresión lineal de las dos muestras; digamos, las pendientes y las intercepciones.

Suponga que estamos interesados en probar la igualdad de dos pendientes. En otras palabras, deseamos determinar si dos líneas dadas son estadísticamente paralelas. Dejemos que m_1 represente el coeficiente de la regresión para la variable explicativa X_1 en la muestra 1 con el tamaño n_1. Dejemos que m_2 represente el coeficiente de la regresión para X_2 en la muestra 2 con n_2. La diferencia entre las dos pendientes estimadas tiene la variación siguiente:

$V = Var[m_1 - m_2] = \{S_{xx1} \acute{} S_{xx2}[(n_1 - 2)S_{res1}^2 + (n_2 - 2)S_{res2}^2] / [(n_1 + n_2 - 4)(S_{xx1} + S_{xx2}]$.
Por lo tanto, la cantidad:
$(m_1 - m_2) / V^{½}$

tiene una distribución t con gl = n1 + n2 - 4.

Esta prueba y su generalización en comparar más de dos pendientes son llamadas el Análisis de la Covarianza (ANOCOV). La prueba de ANOCOV es igual que la prueba de ANOVA; sin embargo, hay una variable adicional llamada covariación. ANOCOV nos permite conducir y ampliar la prueba de antes-y-después para dos poblaciones diferentes. El proceso es como:

124. Encuentre un modelo lineal para $(X_1, Y_1) = (antes_1, después_1)$, y otro para $(X_2, Y_2) = (antes_2, después_2)$ que se ajusten mejor.
125. Realice la prueba de hipótesis para $m_1 = m_2$.
126. Si el resultado de la prueba indica que las pendientes son casi iguales, entonces se calcula la pendiente común de las dos líneas paralelas de regresión:

    $Pendiente_{par} = (m_1S_{xx1} + m_2S_{xx2}) / (S_{xx1} + S_{xx2})$.

    La varianza de los residuos es:

    $S_{res}^2 = [S_{yy1} + S_{yy2} - (S_{xy1} + S_{xy2}) Pendiente_{par}] / (n_1 + n_1 -3)$.

127. Ahora, realice la prueba de diferencias entre las dos intercepciones, la cual es la diferencia vertical entre las dos líneas paralelas:

    Diferencias de las Intercepción $= \bar{y}_1 - \bar{y}_2 - (\bar{x}_1 - \bar{x}_2) Pendiente_{par}$.

    La prueba estadística es:

    (Diferencias de las Intercepción) / $\{S_{res} [1/n_1 + 1/n_2 + (\bar{x}_1 - \bar{x}_2)^2/(S_{xx1} + S_{xx2})]^{½}\}$,

    la cuál tiene una distribución t con parámetros de gl = $_1 + n_1 -3$.

    Dependiendo del resultado de la prueba anterior, se podría rechazar la hipótesis nula.

Para nuestro ejemplo numérico, usando el Javascript de Análisis de la Covarianza se obtuvieron los estadísticos siguientes:
Pendiente 1 = 1,3513514; su error estándar = 0, 2587641
Pendiente 2 = 1,4883721; su error estándar = 1,0793906

Esto indica que no hay evidencia contra la igualdad de las pendientes. Ahora, podemos probar para cualquier diferencia en las intercepciones. Suponga que deseamos probar la hipótesis nula de que la distancia vertical entre las dos líneas paralelas es cerca de 4 unidades.

Usando la segunda función en el Análisis de Covarianza en Javascript, obtuvimos los siguientes estadísticos: Pendiente Común = 1,425; Intercepción = 5,655, proporcionando una evidencia moderada contra la hipótesis nula.

Aplicación para la Valoración de Propiedades Residenciales

Estimar el valor de mercado de un de propiedades residenciales es del interés de los agentes socioeconómicos, tales como compañías de hipoteca y seguros, bancos y las agencias inmobiliarias, y compañías de propiedades de inversión, etc. Esto es tanto una ciencia como un arte. Es una ciencia, porque se basa en métodos formales, rigurosos y de prueba. Es un arte porque interactúa con agentes socioeconómicos y los métodos usados dan lugar a toda clase de compensaciones y de compromisos que los asesores y sus organizaciones deban considerar al tomar decisiones en base de su experiencia y habilidades.

La evaluación del valor de mercado de un grupo de casas seleccionadas implica realizar la evaluación por por medio de algunos evaluadores individuales para cada propiedad y luego calcular un promedio del valor proporcionado por cada evaluador.

La valoración individual se refiere al proceso de estimar el valor de intercambio de una casa basada en una comparación directa entre su perfil y los perfiles de un grupo de otras propiedades comparables vendidas en condiciones aceptables. El perfil de una propiedad consiste en todas las cualidades relevantes de cada casa, tales como la localización, el tamaño, el espacio habitable, la antigüedad, un piso, dos pisos, uno mas, el garaje, la piscina, el sótano, etc. Los datos sobre precios y características de casas individuales son disponibles; por ejemplo en la oficina de censo de los EE.UU. (para este país.)

El análisis de regresión se utiliza a menudo para determinar las características que influencian el precio de las casas. Por lo tanto, es importante corregir los elementos subjetivos en el valor de la valoración antes de realizar el análisis de la regresión. Los coeficientes que no son significativamente diferentes a cero según lo indicado por un t estadístico insignificante a un nivel del 5% son excluidos del modelo de regresión.

Existen varias preguntas prácticas que deben ser contestadas antes de que la colección de datos sea realizada.

El primer paso es utilizar técnicas estadísticas, tales como la clasificación geográfica, para definir las agrupaciones homogéneas de casas dentro de un área urbana.

¿Cuántas casas debemos observar? Idealmente, uno debe recoger la información tantas casas como el tiempo y el dinero permiten. Esta es una de esas consideraciones prácticas que hacen la estadística tan útil. Difícilmente, cualquier persona podría gastar el tiempo, dinero, y esfuerzo necesario para mirar cada casa en venta. Es poco realista obtener la información sobre cada casa de interés, o en términos estadísticos, a cada artículo de la población. De esta forma, podemos mirar solamente una muestra de casas -- un subconjunto de la población -- y esperar que esta muestra nos de la información razonablemente exacta sobre la población. Digamos que podemos mirar 16 casas.

Elegiríamos probablemente seleccionar una muestra aleatoria que, en línea general, cada casa de la población tiene igual posibilidad de ser incluida. Luego, esperamos conseguir una muestra razonablemente representativa de casas a través de un rango seleccionado del tamaño, reflejando los precios para la vecindad entera. Esta muestra debe darnos una cierta información sobre todas las casas de todos los tamaños dentro de este rango, puesto que una muestra aleatoria simple tiende a seleccionar tanto casas más grandes como casas más pequeñas, y tanto las mas costosas como las menos costosas.

Suponga que las 16 casas en nuestra muestra aleatoria tienen los tamaños, antigüedad y precios mostrados en la tabla siguiente. Si 160 casas son seleccionadas aleatoriamente, las variables Y, X1, y X2 son variables aleatorias. No tenemos ningún control sobre ellas y no podemos saber qué valores específicos serán seleccionados. Es solo el chance el que lo determina.

- Tamaño, Antiguedad, y Precio de Veinte Casas -

X1 = Tamaño	X2 = Antiguedad	Y = Precio	X1 = Tamaño	X2 = Antiguedad	Y = Precio
1,8	30	32	2,3	30	44
1,0	33	24	1,4	17	27
1,7	25	27	3,3	16	50
1,2	12	25	2,2	22	37
2,8	12	47	1,5	29	28
1,7	1	30	1,1	29	20
2,5	12	43	2,0	25	38
3,6	28	52	2,6	2	45

¿Qué podemos decir acerca de la relación entre el tamaño y el precio de nuestra muestra? Leyendo los datos de nuestra tabla anterior fila por fila, e incorporándolos en el Análisis de Regresión con Herramientas Diagnósticas en Javascript, encontramos el siguiente modelo simple de regresión:

Precio = 9,253 + 12,873(Tamaño)

Ahora considera el problema de estimar el precio (Y) de una casa sabiendo su tamaño (X1) y también su antigüedad (X2). Los tamaños y los precios serán iguales que en el problema simple de la regresión. Lo qué hemos hecho es agregar las antigüedades de las casas a los datos existentes. Observe cuidadosamente que en la vida real, primero no se saldría a recoger los datos sobre tamaños y precios y luego se analiza el problema de regresión simple. Preferiblemente, se recoge todos que pudieron ser pertinentes en las veinte casas en general. Luego el análisis realizado arrojaría los predictores que resulten no ser necesarios.

Los objetivos en un problema de regresión múltiple son esencialmente los mismos que para una regresión simple. Mientras que los objetivos siguen siendo iguales, más predictores hacen que los cálculos y las interpretaciones lleguen a ser más complicadas. Para un conjunto de datos grandes se podría utilizar el módulo de regresión múltiple de cualquier paquete estadístico tal como SAS y SPSS . Usando el Javascript de la Regresión Lineal Múltiple para nuestro ejemplo numérico con X1 = tamaño, X2 = antigüedad, e Y = precio, obtenemos el siguiente modelo estadístico:

Precio = 9,959 + 12,800(Tamaño) – 0,027(Antigüedad)
Los resultados de la regresión sugieren que, en promedio, mientras el tamaño de la casa aumente los precios también aumentan. Sin embargo, el coeficiente de la variable antigüedad es significativamente pequeño con valor negativo que indica una relación inversa. Casas más viejas tienden a costar menos que casas más nuevas. Por otra parte, la correlación entre el precio y la antigüedad es –0,236. Este resultado indica que solamente 6% de la variación en los precios se considera explicado por las diferencias en los años de antigüedad de las casas. Este resultado apoya nuestra suposición de que la antigüedad no es un predictor significativo de los precio. Por lo tanto, la regresión simple:
El precio = 9,253 + 12,873(Tamaño)
Ahora la pregunta es: ¿Es este modelo lo suficientemente bueno para satisfacer las condiciones generalmente del análisis de la regresión?.

La siguiente lista son las asunciones básicas (es decir, condiciones) y las herramientas para comprobar estas condiciones necesarias.

128. Cualquier outlier desapercibido puede tener impacto importante en el modelo de la regresión. Usando el Javascript de

Determinación de los Outliers encontraríamos que no existen outliers en el modelo anterior.

129. La variable dependiente precio es una función lineal de la variable independiente tamaño. Mediante una cuidadosa examinación del diagrama de dispersión encontraríamos que la condición de linealidad es satisfecha.

130. La distribución residual debe ser normal. Leyendo los datos de la tabla anterior fila por fila, e incorporándolos en el Análisis de Regresión con Herramientas Diagnósticas en Javascript, encontraríamos que la condición de la normalidad también es satisfecha.

131. Los residuales deben tener un media igual a cero, y una desviación estándar constante (es decir, condición homocedasticidad). Mediante el Análisis de Regresión con Herramientas Diagnósticas en Javascript, los resultados son satisfactorios.

132. Los residuales constituyen un sistema de variables aleatorias. La persistencia de no-aleatoriedad en los residuales viola la condición del mejor estimador imparcial lineal. Sin embargo, desde que los estadísticos numéricos que corresponde a los residuales obtenidos usando el Análisis de Regresión con Herramientas Diagnósticas en Javascript, no sean significativos, implica que nuestra regresión ordinaria de mínimos cuadrados es adecuado para nuestro análisis.

133. El estadístico de Durbin-Watson (D-W) cuantifica la correlación serial de los errores de mínimos cuadrados en su forma original. El estadístico de D-W para este modelo es 1,995, el cual es suficientemente bueno en rechazar cualquier correlación serial.

134. Estadísticos más útiles para el modelo: Los errores estándar para la pendiente y la intercepción son 0,881, y 1,916, respectivamente, los cuales son suficientemente pequeños. El estadístico F es 213,599, el cual es suficientemente grande indicando que el modelo es en su totalidad bueno para realizar los propósitos de la predicción.

Note que puesto que el análisis anterior se realiza en un sistema de datos específicos, como siempre, se debe tener cuidado en la generalización de los resultados.

La Introducción al Concepto de la Integración Estadístico

El razonamiento estadístico para la toma de decisiones requiere una comprensión más profunda que simplemente memorizar cada técnica aislada. La comprensión implica siempre la extensión de las redes neurales por las vías de medio de la conectividad correcta entre los conceptos. El objetivo de este capítulo es mirar de cerca algunos de los conceptos y técnicas que hemos aprendido hasta ahora en una forma unificada. Los siguientes casos de estudios, mejoran su razonamiento

estadístico para considerar la integridad y multi facetas de las herramientas estadísticas.

Como usted vera, aunque se esperaría que todas las pruebas suministren los mismos resultados, éste no es siempre el caso. Todo depende de que tan informativos sean los datos y de cuan extenso hayan sido condensados antes de ser presentados para el análisis (mientras que se convierte en un buen estadístico). Las secciones siguientes son ilustraciones que examinan cuánta información útil es proporcionada y cómo se puede dar lugar a conclusiones opuestas, si no se es suficientemente cuidado.

La Prueba de Hipótesis con Confianza

Una de las ventajas principales de construir un intervalo de confianza (IC) es proporcionar un grado de confianza para el punto de estimación para el parámetro de la población. Por otra parte, se puede utilizar el IC para propósitos de la prueba de la hipótesis. Suponga que deseamos probar la siguiente prueba de hipótesis general:

H_0: El parámetro poblacional es casi igual a un valor propuesto dado,

contra la alternativa:

H_a: El parámetro poblacional no es uniforme cerca al valor propuesto.

El proceso de ejecutar la prueba de hipótesis anterior a un nivel a de significación, usando el IC es como sigue level:

135. Ignore el valor propuesto en la hipótesis nula, mientras usa este procedimiento.
136. Construya un intervalo de confianza de 100(1- a)% basado en los datos disponibles.
137. Si el IC construido no contiene el valor propuesto, indica que no existe suficiente evidencia para rechazar la hipótesis nula; de lo contrario, no hay razón de rechazar la hipótesis nula.

A usted podría gustarle utilizar la Prueba de Hipótesis con Confianza en Javascript para realizar algunas experimentaciones numéricas, para validar las aserciones anteriores y para una comprensión más profunda de los conceptos.

El Análisis de Regresión, ANOVA, y la Prueba Chi-cuadrado

Existe una relación estrecha entre la regresión lineal, el análisis de la varianza y la prueba Chi-cuadrado. Para ilustrar la relación, considere la siguiente aplicación:

Relación entre la edad e ingreso en un vecindario dado: Una encuesta aleatoria de una muestra de 33 individuos en una vecindario reveló los siguientes pares de datos. Para cada par de edades se representa en años y el ingreso se indica en millares de pesos:

- Relación entre Edad e Ingresos (unidades en 1000 pesos) -					
Edad	Ingreso	Edad	Ingreso	Edad	Ingreso
20	15	42	19	61	13
22	13	47	17	62	14
23	17	53	13	65	9
28	19	55	18	67	7
35	15	41	21	72	7
24	21	53	39	65	22
26	26	57	28	65	24
29	27	58	22	69	27
39	31	58	29	71	22
31	16	46	27	69	9
37	19	44	35	62	21

Construyendo una regresión lineal obtenemos:

Ingreso = 22,88 – 0,05834 (Edad)

El resultado sugiere una relación negativa; mientras que la gente se hace mayor, tienen ingresos más bajos, en promedio. A pesar de que la pendiente es pequeña, no puede ser considerado como cero, puesto que el t estadístico para el es –0,70, el cual es significativo.

Ahora suponga que solo se tienen los datos secundarios, siguientes, de donde los datos originales han sido condensados:

- Relación entre Edad e Ingreso (unidades en 1000 pesos) -		
Edad (29 - 39)	Edad (40 - 59)	Edad (60 óamp; mas)
15	19	13
13	17	14
17	13	9
21	21	7
15	39	21
26	28	24
27	22	27
31	26	22
16	27	9
19	35	22
19	18	7

Se puede utilizar la ANOVA para probar que no existe relación entre la edad y el ingreso. Ejecutando el análisis proporciona un F estadística igual a 3,87, el cual es absolutamente significativo; es decir, rechazando la hipótesis de que no existe diferencia en los ingresos promedios de la población para las tres categorías de edad.

Ahora, suponga que datos secundarios condensados están proporcionados como en la tabla siguiente:

Relación entre Edad e Ingresos (unidades en 1000 pesos):			
	Edad		
Ingreso	*20-39*	*40-59*	*60 ó mas*
Hasta 20,000	7	4	6
20,000 o mas	4	7	5

Se puede utilizar la prueba Chi-cuadrado para la hipótesis nula de que la edad y el ingreso no están relacionados. El estadístico Chi-cuadrado es 1,70, lo cual no es significativo; ¡por lo tanto no hay razón para creer que el ingreso y la edad están relacionadas! Pero por supuesto, los datos están sobre condensan, porque cuando todos los datos en la muestra fueron utilizados, había una relación observable.

El Análisis de Regresión, ANOVA, la Prueba T, y el Coeficiente de Determinación

Existe una gran relación directa entre la regresión lineal, el análisis de variación, la prueba T y el coeficiente de determinación. El siguiente grupo pequeño de datos lustra las conexiones entre los procedimientos estadísticos anteriores, y por lo tanto las relaciones entre las tablas estadísticas:

X1	4	5	4	6	7	7	8	9	9	11
X2	8	6	8	10	10	11	13	14	14	16

Suponga que aplicamos la Prueba T. El T estadístico es = 3,207, con gl = 18. valor P es 0,003, el cual que indica una fuerte evidencia contra la hipótesis nula.

Ahora, introduciendo una variable simulada x con dos valores, digamos 0 y 1, representando los dos grupos de datos, respectivamente, podemos aplicar el análisis de regresión:

x	0	0	0	0	0	0	0	0	0	0
y	4	5	4	6	7	7	8	9	9	11
x	1	1	1	1	1	1	1	1	1	1
y	8	6	8	10	10	11	13	14	14	16

Entre otros estadísticos, obtenemos una pendiente larga = m = 4 ¹ 0, indicando el rechazamiento de la hipótesis nula. Note que, el T estadístico para la pendiente es: T estadístico = la Pendiente /(el error estándar de la pendiente) = 4 / 1,2472191 = 3,207, el cual es el estadístico T que obtuvimos de la prueba T. En general, el cuadrado del T estadístico de la pendiente es el estadístico F en la tabla de ANOVA; es decir; i.e.,

t_m^2 = F estadístico

Por otra parte, el coeficiente de determinación r^2 = 0,36, el cual es siempre obtenible de la prueba T, como sigue:

$r^2 = t^2 / (t^2 + d.f.)$.

Para nuestro ejemplo numérico, el r^2 es $(3,207)^2 / [(3,207)^2 + 18]$ = 0,36, según lo esperado.

Ahora, aplicando ANOVA en los dos sistemas de datos, obtenemos el estadístico F= 10,285, con $gl_1 = 1$, y $gl_2 = 18$. El estadístico F no es suficientemente grande; por lo tanto, se debe rechazar la hipótesis nula. Observe que, en el general,

$F_{a\ ,\ (1,\ n)} = t^2_{a/2\ ,\ n}$.

Para nuestro ejemplo numérico, $F = t^2 = (3,207)^2 = 10,285$, según lo esperado.

Según lo esperado, apenas mirando los datos, las tres pruebas indican fuertemente que las medias de los dos sistemas de datos son absolutamente diferentes.

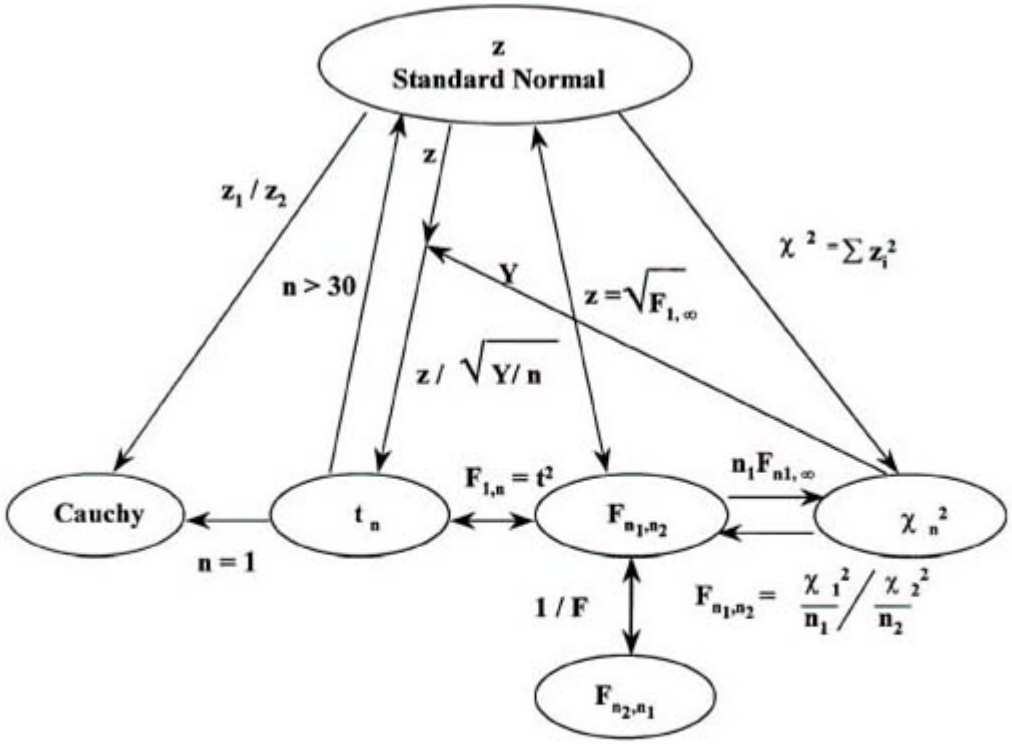
Las Relaciones entre Distribuciones y la Unificación de Tablas Estadísticas

Atención particular debe ser prestada al primer curso en estadística. Cuando primero comencé a estudiar estadística, me incomodó que había diversas tablas para diversas pruebas. Me tomó tiempo para aprender que esto no es tan casual como parecía. Las distribuciones Binomial, normal, Chi-cuadrado, T y F que usted aprenderá están estrechamente relacionadas.

Un problema con los libros de textos de estadística elemental es que no solamente no proporcionan este tipo de información que permite una comprensión útil de los principios implicados, pero generalmente tampoco proporcionan los nexos conceptuales. Si se desea entender las conexiones entre conceptos estadísticos, se debe practicar el hacer estas conexiones. El aprender haciendo estadística se presta a un aprendizaje activo en vez de pasivo. La estadística es un sistema altamente correlacionado de conceptos, y para ser acertado en ellos, se debe aprender a hacer conexiones conscientes en su mente.

Los estudiantes a menudo preguntan: ¿Por qué los valores de la tabla T con gl = 1 son mucho más grandes comparados con otros valores de grados de libertad? Algunas tablas son limitadas. ¿Qué debo hacer cuando el tamaño de la muestra es demasiado grande?, ¿Cómo me puedo familiarizar con las tablas y sus diferencias?. ¿Existe algún tipo de integración entre las tablas? ¿Hay conexión entre la prueba de hipótesis y el intervalo de la confianza bajo diversos panoramas? Por ejemplo, prueba con respecto a una, dos o más poblaciones, etcétera.

La figura siguiente muestra las relaciones útiles entre distribuciones y la unificación de tablas estadísticas:



Relationship Among Popular Distribution

Por ejemplo, los siguientes son algunas conexiones útiles entre las tablas mas importantes:

- Normal Estándar z y el F estadístico: $F = z^2$, donde F tiene ($gl_1 = 1$, y gl_2 es el mas grande disponible en la tabla F)
- Estadístico T y el F estadístico: $F = t^2$, donde F tiene ($gl_1 = 1$, y gl_2 = gl de la tabla t)
- Chi-cuadrado y el F estadístico: F = Chi-cuadrado/ gl_1, de donde F tiene (gl_1 = gl de la tabla Chi-cuadrado, y gl_2 es el valor mas grande disponible en la tabla F)
- Estadístico T y la Chi-cuadrado: $(\text{Chi-cuadrado})^{½}$ = t, de donde Chi-cuadrado tiene gl=1, y t tiene gl= ¥.
- Normal Estándar z y el estadístico T: z = t, donde t tiene gl = ¥.
- Normal Estándar z y la Chi-cuadrado: $(2\ \text{Chi-cuadrado})^{½} - (2gl-1)^{½}$ = z, de donde gl es el valor mas grande disponible en la tabla Chi-cuadrado).
- Standard normal z, Chi-square, and T- statistic: $z/[\text{Chi-aquare}/n)^{½}$ = t with d.f. = n.
- Estadístico F y su inversa: $F_a(n1, n2) = 1/F_{1-a}(n2, n1)$, por lo tanto solo se necesita tabular, las probabilidades de la cola superior.
- Coeficiente de Correlación r y el estadístico T: $t = [r(n-2)^{½}]/[1 - r^2]^{½}$.

Transformación de algunas inferencias Estadísticas a la Normal Estándar Z:

- Para el t(gl): $Z = \{gl \cdot Ln[1 + (t^2/gl)]\}^{½} \cdot \{1 - [1/(2gl)]\}^{½}$.
- Para la F(1,gl): $Z = \{gl \cdot Ln[1 + (F/gl)]\}^{½} \cdot \{1 - [1/(2gl)]\}^{½}$,

de donde Ln es el logaritmo natural where.

Visite también la Relación Entre Distribuciones Comunes .

A usted podría gustarle utilizar las tablas estadísticas que aparecen en la parte de atrás de su libro de texto y/o valores P en JavaScript para realizar algunas experimentaciones numéricas para validar las relaciones anteriores y para una comprensión mas profunda de los conceptos. Usted podría necesitar utilizar una calculadora científica, también.

Números índices con Aplicaciones

Cuando se enfrenta a una carencia en la una unidad de la medida, normalmente se utilizan indicadores como sustitutos para las medidas directas. Por ejemplo, la altura de una columna de mercurio es un indicador familiar de la temperatura. Nadie presume que la altura de la columna de mercurio constituye la temperatura en absolutamente el mismo sentido que la longitud constituye el número de centímetros de extremo a extremo. Sin embargo, la altura de una columna de mercurio es un correlativo confiable de la temperatura y sirve así como medida útil de ella. Por lo tanto, un indicador es un correlativo accesible y confiable de una dimensión de interés; este correlativo se utiliza como medida de esa dimensión porque la medida directa de la dimensión no es posible o práctica. De modo semejante los números índice sirven como sustituto para los datos reales.

El propósito primario de un número índice es proporcionar un valor útil para comparar magnitudes de los agregados de variables relacionadas, y medir los cambios en estas magnitudes en un cierto plazo. Por lo tanto, diversos números índice se han desarrollado para usos especiales. Existe un número particularmente bien conocidos, los cuales se anuncian en los medios públicos a diario. Las agencias gubernamentales a menudo reportan datos de serie de tiempo en la forma de números índice. Por ejemplo, el índice de precio al consumidor es un indicador económico importante. Por lo tanto, es útil entender cómo se construyen los números índice y cómo interpretarlos. Estos números índice se desarrollan generalmente comenzando con la base 100 que indica un cambio en magnitud concerniente a su valor en un punto especificado en tiempo.

Por ejemplo, en la determinación del coste de vida, la oficina de estadísticas de trabajo identifica primero una “cesta” de bienes y servicios en el mercado que un consumidor típico compra. Anualmente, esta oficina encuesta a consumidores para determinar que tipo de productos y servicios que compraron y el costo total de los mismos: Qué, donde, y cuánto. El índice de precios del consumidor (IPC) se utiliza para supervisar cambios en los costos de vida en un periodo determinado (de una cesta de productos seleccionados). Cuando IPC se incrementa, una familia típica tiene que gastar más Pesos para mantener el mismo estándar vivir. La meta del IPC es medir cambios en el coste de vivir. Este reporta el movimiento de los precios, pero no en términos de pesos, si no en números índices.

La Media Geométrica

The La Media Geométrica es extensivamente utilizada por la Oficina de Estadísticas de Trabajo de los Estados Unidos de Norte América, “Geomeans” como la llaman, es el cálculo del índice de precios del consumidor en este país. Las “Geomeans” también se utilizan en todos los índices de precios.

Cociente de Números Índice

La siguiente tabla proporciona el procedimiento computacional y usos para algunos números índice, incluyendo el índice del cociente, y los números índice compuestos.

Suponga que se esta interesado en la utilización de mano de obra de dos instalaciones fabriles A y B con las unidades de producción y el requerimiento de hombre/ horas, según lo demostrado en la tabla siguiente, junto con el estándar nacional de los últimos tres meses:

	Planta- A		Planta- B	
Meses	**Unidades Producto**	**Hombre/ Horas**	**Unidades Producto**	**Hombre/ Horas**
1	0283	200000	11315	680000
2	0760	300000	12470	720000
3	1195	530000	13395	750000
Estándar	**4000**	**600000**	**16000**	**800000**

La utilización de trabajo (mano de obra) en la planta A para el mes:

$L_{A,1}$ = [(200000/283)] / [(600000/4000)] = 4,69

Similarmente,

$L_{B,3}$ = 53,59/50 = 1,07.

Luego del cálculo de la utilización de trabajo de ambas plantas para cada mes, se pueden presentar los resultados mediante la representación gráfica del trabajo en un cierto período de tiempo para los estudios comparativos.

Números Índice Compuestos

Considere la fuerza laboral total, y el costo de materiales por dos años consecutivos para una planta industrial, según lo demostrado en la tabla siguiente:

		Año 2000		Año 2001	
	Unidades Necesarias	Costos por Unidad	Total	Costos por Unidad	Total
Mano de Obra	20	10	200	11	220
Aluminio	02	100	200	110	220
Electricidad	02	50	100	60	120
Total			**500**		**560**

De la información dada en la tabla anterior, los índices para dos años consecutivos son 500/500 = 1, y 560/500 = 1,12, respectivamente.

El Índice de Variación como Indicador de Calidad

Un índice comúnmente utilizado como media de variación y comparación para datos nominales y ordinales se llama el índice de dispersión:
$D = k\,(N^2 - Sf_i^2)/[N^2(k-1)]$

De donde k es el número de categorías, f_i es la escala de cada categoría, y N es el número total de escalas. D es un número entre cero y 1 dependiendo si todos las escalas caen en una categoría, o si las escalas fueran divididas igualmente entre las k categorías.

Una aplicación: Considere los datos siguientes con N = 100 participantes, k = 5 categorías, f_1 = 25, f_2 = 42 etcétera.

Categoría	Frecuencia
A	25
B	42
C	8
D	13
E	12

Por lo tanto, el índice de dispersión es: $D = 5\ (100^2 - 2766)/[100^2(4)] = 0{,}904$, indicando una buena distribución de los valores a través de las categorías.

Índice de Desempleo de la Fuerza Laboral

¿Es una ciudad dada un área económicamente deprimida? El grado de desempleo entre fuerza laboral (L) es considerado de ser un indicador apropiado de la depresión económica. Para construir el índice de desempleo, cada persona es clasificada de dos maneras con respecto a los miembros de la fuerza laboral y con respecto al grado de desempleo en valor fraccionario, extendiéndose a partir de la 0 a 1. La fracción que indica la porción de trabajo que se encuentra ocioso es:
$L = S[U_iP_i] / SP_i$, la suma de todos los valores de i = 1, 2,…, n.

De donde P_i es la proporción de una semana completa de trabajo para cada residente en el área donde se requiere el empleo y n es el número total de residentes en el área. U_i es la proporción P_i para el cual cada residente en el área esta desempleado. Por ejemplo, una persona que busca dos días de trabajo por la semana (5 días) y es empleado por solo medio día con $P_i = 2/5 = 0{,}4$, y $U_i = 1{,}5/2 = 0{,}75$. La multiplicación resultante de $U_iP_i = 0{,}3$ sería la porción de una semana completa de trabajo en la cual una persona se encuentra desempleada.

Ahora, la pregunta es que valor de L constituye un área económicamente deprimida. La respuesta la tienen los responsables de las tomas de decisiones para decidir.

El Índice Estacional y la Desestacionalización de los Datos

Los índices estacionales representan la influencia estacional para un segmento particular del año. El cálculo implica una comparación de los valores previstos de ese período a la gran media.

Se necesita conseguir una estimación del índice estacional para cada mes, u otros períodos tales como cuatrimestres, semana, etc, dependiendo de la disponibilidad de datos. La estacionalidad es un patrón que se repite para cada período. Por ejemplo el patrón estacional anual tiene un ciclo que tiene 12 períodos, si los períodos son meses, o 4 períodos si los períodos son trimestres.

Un índice estacional es la medida de cuánto un average para un periodo de tiempo determinado tiende estar por debajo (o por arriba) del average general. Por lo tanto, para conseguir una estimación exacta de el, se calcula el promedio del primer período del ciclo, y el segundo período del

ciclo, etc, y luego dividimos cada uno por el promedio total. La fórmula para calcular los factores estacionales es:

$S_i = D_i/D$,
donde:
S_i = El índice de estacionalidad para todos los periodos I,
D_i = El valor average de los periodos i,
D = Average general,
i = el iesimo periodo estacional del ciclo.

Un índice estacional de 1,00 para un mes en particular indica que el valor previsto de ese mes es 1/12 del promedio total. Un índice estacional de 1,25 indica que el valor previsto para ese mes es 25% mas grande que 1/12 del promedio total. Un índice estacional de 80 indica que el valor previsto para ese mes es 20% menos que 1/12 del promedio total.

Proceso De Desestacionalidad de los Datos: También llamado ajuste estacional es el proceso de quitar variaciones recurrentes y periódicas sobre un periodo corto de tiempo (por ejemplo, semanas, trimestres, meses). Por lo tanto, las variaciones de la estaciones están repitiendo regularmente los movimientos en los valores de la serie que se pueden atar a los acontecimientos que se repiten. La Desestacionalidad de los datos es obtenida mediante la división de cada observación de series de tiempo por el respectivo índice estacional.

Casi todas las series de tiempo publicadas por el gobierno son desestacionalizadas usando el índice estacional para desenmascarar las tendencias subyacentes en los datos, que se habrían podido causar por el factor de estacionalidad.

Una aplicación numérica: La tabla siguiente proporciona las ventas mensuales (en 000 de pesos) en una librería de la universidad.

M T	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septi.	Octub.	Novie.	Dicie.	Total
1	196	188	192	164	140	120	112	140	160	168	192	200	1972
2	200	188	192	164	140	122	132	144	176	168	196	194	2016
3	196	212	202	180	150	140	156	144	164	186	200	230	2160
4	242	240	196	220	200	192	176	184	204	228	250	260	2592
Media:	208,6	207,0	192,6	182,0	157,6	143,6	144,0	153,0	177,6	187,6	209,6	221,0	2185
Indice:	1,14	1,14	1,06	1,00	0,87	0,79	0,79	0,84	0,97	1,03	1,15	1,22	12

Las ventas demuestran un patrón estacional, con el número más grande cuando la universidad está en sesión y una disminución durante los meses del verano. Por ejemplo, para Enero el índice es:

S(Ene) = D(Ene)/D = 208,6/181,84 = 1,14,
De donde D(Ene) es la media de todos los cuatro meses de eneros, y D es la media total de todos los cuatro años de ventas.

A usted podría gustarle utilizar el Javascript de Índice Estacional para comprobar sus cálculos manuales. Como siempre, usted debería primero utilizar el Diagrama de Series de Tiempo como herramienta para el proceso inicial de la caracterización.

Para probar estacionalidad basadas en índices estacionales, a usted podría gustarle utilizar el Javascript de Prueba de Estacionalidad .

Para modelar series de tiempo que tienes componentes de estacionalidad y tendencias, visite el sitio Web Pronóstico de Negocios.

Técnicas Estadísticas y Números Índices

Se debe tener mucho cuidado cuando se aplica o generaliza cualquier técnica estadística a los números índices. Por ejemplo, la correlación de tasas genera un problema potencial. Específicamente, deje que X, Y, y Z sean tres variables independientes, de modo que en parejas las correlaciones sean cero; sin embargo, los cocientes X/ Y, y Z/ Y serán correlacionado debido al denominador común.

Deje que I = X_1/X_2 donde están las variables X_1, y X_2 son variables dependientes con la correlación r, teniendo media y coeficiente de variación m_1, c_1 y m_2, c_2, , respectivamente; por lo tanto,

Media de I = $m_1 (1-r´c_1´c_2 + c_2^2)/m_2$,

Desviación estándar de I = $_1(c_1^2 - 2\, r´c_1´c_2 + c_2^2)^{½}/m_2$

Fuente: Recopilación del mundo virtual.

Panamá, 25 de marzo de 2008.

DPS/DEI

/MRP.