Métodos Estadísticos

Salvador Pintos
Instituto de Cálculo Aplicado

julio-2000

Índice General

Capítulo 1

Revisión de conceptos básicos

1.1 Variables aleatorias

1.1.1 Estadística descriptiva en el SAS

Una descripción informativa del conjunto de datos de una muestra de una variable está dada por el histograma de frecuencias relativas. La habilidad para detectar patrones o tipos de distribución a partir de un histograma depende de la elección adecuada de las clases. El PROC GCHART es el procedimiento idóneo para hacer estos histogramas. Se sugiere hacer uso de la opción midpoints para controlar el número de clases del histograma.

Existen dos medidas de interés para cualquier conjunto de datos $\{x_1, \ldots\ldots, x_n\}$: la localización de su centro y la variabilidad.

Hay, principalmente, tres medidas de tendencia central:

La media muestral:

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

La mediana muestral: si se ordena el conjunto de datos la mediana divide la muestra en dos.

La moda muestral: es el valor más frecuente de la muestra.

Es la media, por razones teóricas, la que se usa fundamentalmente.

En cuanto a la variabilidad, la varianza muestral es la medida estadísticamente más útil.

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - m)^2}{(n - 1)}$$

A menudo se prefiere la **desviación standard**, que es su raíz cuadrada, porque se expresa en las mismas unidades físicas que la media y las observaciones. El PROC UNIVARIATE del SAS permite calcular esas medidas y otras: cuantiles, observaciones máximas y mínimas, etc.

1.1.2 Variables Aleatorias, Valor Esperado y Varianza.

Cuando se consideran variables económicas o sociales es necesario admitir que son esencialmente Variables Aleatorias y que en consecuencia tienen asociada una estructura de probabilidad que se caracteriza por la distribución de probabilidad.

Una Variable Aleatoria X es Discreta si el conjunto de valores que toma es finito o, si es infinito, puede ordenarse en una secuencia que se corresponda con los naturales.

Su distribución de probabilidad está dada por: $Prob(X = x_i) = p_i$.

Una Variable aleatoria X es Continua si el conjunto de valores que toma es uno o más intervalos de la recta real.

Su distribución de probabilidad está caracterizada por la función de densidad f.

Definición 1 f es la función de densidad de la variable aleatoria X si $f(x) \geq 0$ y para cada intervalo $[a; b]$

$$Prob(a \leq X \leq b) = \int_a^b f(x)dx$$

La Función de Distribución Acumulada es $F(x) = Prob(X <= x)$:

Con el propósito de calcular probabilidades, el SAS tabula, para las distintas distribuciones, esta Función de Distribución Acumulada.

Si X es discreta y toma valores ordenados 0,1,2 ... etc. entonces: $Prob(X = k) = F(k) - F(k - 1)$

Para una Variable Aleatoria X es necesario establecer su valor medio, (Valor Esperado), y cómo se dispersa respecto de su valor medio (Varianza).

Definición 2 Si X es continua Valor Esperado $E(X)$ es:

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x)\, dx$$

la Varianza $V(X)$: $V(X) = E(X - \mu)^2$ y la Desviación Standard: $\sigma = \sqrt{V(X)}$

1.1.3 Valores esperados y varianzas

En lo que sigue usaremos E, V y S como símbolos de Valor Esperado, Varianza y Desviación Estándar.

Proposición 3 Si X e Y son variable aleatorias y k es una constante entonces:

$E(kX) = kE(X)$ $\quad V(kX) = k^2 V(X)$ $\quad S(kX) = |k| S(X)$

Además si X e Y son independientes: $V(X+Y) = V(X-Y) = V(X) + V(Y)$

$$\sigma(X+Y) = \sqrt{V(X) + V(Y)}$$

Ejemplo: Si el peso de un níspero se distribuye normal con valor esperado 120 grs y desviación estándar 15 grs, la bolsa también normal con peso 5 grs y desviación estándar 1 gr y el precio es de Bs 100 por kilo, Hallar las siguientes probabilidades:

Que una bolsa con 3 nísperos pese más de 400 grs.

Que 10 nísperos sin bolsa pesen más de 1300 grs.

Que 1 níspero cueste más de 13 Bs.

1.2 Distribuciones discretas

1.2.1 Distribución binomial

Características:

1. Hay n ensayos independientes.
2. El resultado del ensayo es éxito(E) o fracaso(F).
3. La probabilidad p de éxito es constante en los ensayos.
4. Distribución asociada al muestreo con reemplazo.

Parámetros: n, p.

Variable aleatoria: número x de éxitos en los n ensayos.

Función de distribución acumulada en SAS: PROBBNML (p, n, x)

1.2.2 Distribución Hipergeométrica

Características:

1. De una población de tamaño NP, K elementos son del tipo A. Si se selecciona una muestra aleatoria de n elementos ¿cuál es la probabilidad de que en ésta se hallen x elementos del tipo A?

2. Muestreo sin reemplazo.

Parámetros: NP, K, n.

Variable aleatoria: número x de elementos A en la muestra.

Obsérvese que $0 \leq x <= K$ y $0 \leq n - x \leq NP - K$. Luego, los posibles valores de la variable aleatoria x están restringidos al intervalo: $M\acute{a}x(0; n + K - NP) \leq x \leq M\acute{i}n(n; K)$

Sea $p = K/NP$ la proporción de elementos del tipo A en la población:

Valor esperado y Varianza: $E(X) = np$ $\quad V(X) = np(1 - p)[1 - (n - 1)/(NP - 1)]$

Función de distribución acumulada en SAS: PROBHYPR(NP,K,n,x)
Cuando la fracción de muestreo n/NP es pequeña, la hipergeométrica se puede aproximar por la binomial con parametros p y n.

1.2.3 Distribución binomial negativa

Características:

1. El resultado del ensayo es éxito(E) o fracaso(F).

2. La probabilidad p de éxito es constante en los ensayos.

3. Se realizan ensayos independientes consecutivos hasta obtener k éxitos.

Parámetros: k, p.

Variable aleatoria: número x de fracasos antes del k éxito.

Función de distribución acumulada en SAS: PROBNEGB (p,k,x)

La distribución geométrica es un caso particular de ésta cuando $k = 1$. es decir la variable aleatoria de la distribución geométrica representa el número de fracasos antes de obtener el primer éxito.

1.2.4 Distribución de Poisson

Características:

1. Eventos que ocurren con velocidad constante en el tiempo (número de carros que llegan a un autobanco), o en el espacio (número de fallas por metro en un rollo de tela, etc).

2. Se utiliza como aproximación al modelo binomial cuando n es grande y p pequeño (ley de los sucesos raros). $p < 0.1$ $np < 5$, tomando como parámetro $\lambda = np$.

Parámetros: λ promedio de ocurrencias del suceso por unidad de tiempo o espacio.

Función de distribución acumulada en SAS: POISSON (λ, x)

1.3 Distribuciones continuas

1.3.1 La Distribución Normal

La distribución Normal es indudablemente la distribución continua fundamental, tanto por sus aplicaciones como por el rol que juega dentro de la Teoría Estadística. Es la piedra angular de la Inferencia ya que muchas estadísticas muestrales tienden hacia la distribución Normal cuando el tamaño de la muestra crece.

Se afirma que una variable aleatoria X es Normal $N(\mu, \sigma^2)$ si su función de densidad está dada por:

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$
$$-\infty < \mu < \infty \qquad \sigma > 0$$

Los parámetros μ y σ son, además, el Valor Esperado y la Desviación Standard de X. Si X es $N(\mu, \sigma^2)$ entonces, $Z = (X - \mu)/\sigma$, es $N(0;1)$ y se le designa como Normal Standard. Esta propiedad permite relacionar la función de distribución acumulada de X con la de Z, ya que: Luego, es suficiente con tabular las probabilidades de la distribución Normal Standard. La distribución acumulada de Z en el SAS es PROBNORM (z) y su inversa PROBIT (p).

Ejemplo: $prob(Z \leq 2)$ = PROBNORM (2)

Ejemplo: ¿Cuál es el z tal que la $Prob(Z < z) = 0.8$?. z = PROBIT (0.8).

Ejemplo: Si el salario de un obrero de la Industria Petrolera se distribuye Normal con μ = Bs 17.000 y σ = Bs 3.500, ¿cuál es la probabilidad de que un obrero gane más de Bs 20.500? p= PROBNORM ((20500-17000)/3500) es la probabilidad de que gane menos de 20500 y 1-p, su complemento, es la probabilidad de que gane más de 20500.

Ejercicio. Hacer un programa en SAS para calcular esta probabilidad.

La distribución Normal tiene además la propiedad fundamental siguiente:

Si una variable aleatoria es la combinación lineal de variables aleatorias Normales independientes, entonces es también una variable aleatoria Normal.

1.3.2 Distribución Gamma

La distribución Gamma es una distribución continua con dos parámetros: θ= Parámetro de escala, α parámetro de forma.
Su función de densidad es:

$$\Gamma(\alpha;\theta;x) = \frac{x^{\alpha-1}\, e^{-(\frac{x}{\theta})}}{\Gamma(\alpha)\theta^{\alpha}}$$

La forma depende de α. Si $\alpha > 1$ la densidad crece y luego decrece con la moda en $\frac{\alpha-1}{\theta}$

Cuando $\alpha > 1$ a la distribución se le llama exponencial. Esta distribución juega un rol esencial en la teoría de colas ya que representa el tiempo entre dos llegadas consecutivas, cuando las llegadas son independientes.

La Distribución Gamma en el SAS.

El SAS tabula solamente la distribución Gamma estándar($\theta = 1$). La función de distribución acumulada es: PROBGAM (x,α) y su inversa: GAMINV (prob,α).

¿Cómo obtener entonces la función de distribución acumulada de una Gamma de parámetros α,θ ? Se logra a partir de una Gamma standard con $\theta = 1$. Si se define como PROBGAM2(x,α,θ) a la función con los dos parámetros se cumple que:

PROBGAM2(x, α, θ) = PROBGAM ($\frac{x}{\theta}$, α)
GAMINV2(prob, α, θ) = θ * GAMINV (prob, α)

1.3.3 Distribución Exponencial

La distribución Exponencial de parámetro θ es una Gamma generalizada con $\alpha = 1$ entonces:

PROBEXP (x;θ) = PROBGAM2 (x;1;θ) = PROBGAM ($\frac{x}{\theta}$;1)
EXPINV (prob;θ) = θ * GAMINV (prob;1)

1.3.4 Distribución Beta

La distribución Beta es una distribución continua con dos parámetros: α y β, ambos parámetros de forma.
Se la utiliza para representar variables aleatorias cuyos valores se encuentran

restringidos a intervalos de longitud finita. Ejemplos: PERT, evaluación de programas, límites estadísticos de tolerancia, estadística del orden, etc.

Su función de densidad es 0 en todos lados salvo en el intervalo [0,1] donde está definida por:

$$f(\alpha;\beta;x) = \frac{\Gamma(\alpha+\beta)x^{\alpha-1}\ x^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)}$$

$$0 < x < 1 \quad \alpha > 0 \quad \beta > 0$$

La forma depende de la posición de los parámetros α y β respecto de 1.
Ejercicio: ejecute el programa beta.plt y observe la forma de la curva en función de los parámetros.

Si α y β son mayores que 1 la moda es: $(\alpha-1)/(\alpha-1+\beta-1)$

La distribución Beta en el SAS

PROBBETA (x, α, β) es la función de distribución acumulada definida para todo x en $(0;1)$: BETAINV (p, α, β) su función inversa.

1.3.5 La Distribución Uniforme

La distribución uniforme o rectangular tiene densidad constante en un intervalo $[a;b]$ y vale 0 fuera del mismo. En consecuencia está dada por:

$$f(x) = \tfrac{1}{b-a} \text{ en } [a;b]$$

$E(X) = \frac{a+b}{2}$ $\qquad VAR(X) = \frac{(b-a)^2}{12}$

Esta distribución se aplica en teoría de errores y en la generación de observaciones simuladas que sigan una distribución dada.

Aproximación Normal de la Binomial

Cuando en el modelo Binomial n es grande, y tanto np como $n(1-p)$ son mayores que 5, se puede aproximar la binomial con una Normal de igual esperanza y varianza. Esto es con una $N[np; np(1-p)]$.

Sin embargo se tendrá en cuenta la siguiente:

Prob(X = n) se aproxima por Prob(n-0.5 < X < n+ 0.5)

Prob(n1 <= X <= n2) se aproxima por:

Prob(n1 - 0.5 <= X <= n2 + 0.5)

Ejemplo: Si una coneja tiene 50 hijos ¿cuál es la probabilidad de tener exactamente 25 hembras?. ¿Cuál la de tener entre 20 y 30 machos?.

Si p es pequeño, $(p < 0.1)$, y $np < 5$, la binomial se aproxima por la distribución de Poisson con valor $\lambda = np$.

1.4 Simulación de variables aleatorias con el SAS

El SAS dispone de varias funciones para generar variables aleatorias que sigan una distribución dada. Dos muy útiles son:

Z= RANNOR(semilla), genera valores de una Normal STANDARD, N(0,1)

U= RANUNI(semilla), genera valores de una Uniforme en el intervalo [0,1]

Puesto que en esencia son seudo-aleatorias, todas dependen de una semilla inicial que genera la secuencia aleatoria. Si se cambia la semilla se genera otra secuencia de números independientes de la secuencia anterior. Elija para la semilla cualquier entero positivo para garantizar aleatoriedad, si elige 0 el SAS la genera a partir del tiempo del computador. Por ejemplo, son válidas las siguientes semillas: SE=3719241; SE1=6258583.

tabla de generación aleatoria con el sas

Distribución	parámetros	generación aleatoria
Uniforme	a, b	u= a+ (b-a)* ranuni(se)
Normal	mu, sigma	nor= mu + sigma* rannor(se)
Gamma	theta, alfa	gam= theta* rangam(se2, alfa)
Poisson	lambda	poisson= ranpoi(se, lambda)
Binomial	n, p	binomi= ranbin(se,n,p)
Geometrica	p	g= floor(-ranexp(se)/log(1-p))

Para una variable aleatoria continua X cuya Función de distribución F(x) es conocida, y donde su inversa es INV F(u), es posible generar valores aleatorios de X mediante el siguiente procedimiento:

1. genere un valor distribuido uniformemente en [0,1]
2. halle X = INV F(u)

Los valores de RANUNI, básicos en la generación seudo aleatoria, son generados por el método de congruencias donde el módulo es 2^{31} -1 y el multiplicador 397204094. La semilla debe ser menor que el módulo.

Capítulo 2

Inferencia

2.1 Estimación puntual

2.1.1 Muestreo aleatorio simple

Definición 4 Dada una Variable aleatoria X, un conjunto $\{X_1; X_2; ...; X_n\}$ de variables es una **Muestra Aleatoria Simple** (MAS), de X si:
1. Cada X_i tiene la misma distribución que la X.
2. Las variables $X_1; X_2; ...; X_n$ son estadísticamente independientes entre sí.

En consecuencia, una muestra aleatoria simple (MAS), es un caso particular de distribución multivariada donde las variables $\{X_1; X_2; ...; X_n\}$ son independientes y todas tienen por distribución común la distribución de X. A menudo se confunde este concepto ya que también se le llama muestra al resultado de observar el valor que toman las n variables $\{X_1; X_2; ...; X_n\}$. Sin embargo, y para que no queden dudas, las propiedades estadísticas que aquí se estudian se refieren a la MAS como variable aleatoria multivariada y no como al conjunto de números resultantes de la observación de la misma.

Definición 5 Un estadístico T es una función de la muestra $\{X_1; X_2; ...; X_n\}$. Es una **nueva variable aleatoria** uni-dimensional construida a partir de una muestra aleatoria simple.

Todo estadístico, por ser una variable aleatoria de una dimensión, tiene su distribución de probabilidad y suele ser más sencillo operar con él que con la muestra como tal. Por ejemplo, el estadístico más importante es la media muestral: $m = \frac{\sum_{i=1}^{n} x_i}{n}$ y cuando se hace referencia a las propiedades de la media muestral se sobreentiende que es sobre el estadístico como variable

aleatoria.
Otros estadísticos son: $\max\{X_1, X_2, \ldots, X_n\}$; $\min\{X_1, X_2, \ldots, X_n\}$; rango = $\max - \min$.

El propósito del muestreo estadístico es hacer inferencias acerca de la población. Más específicamente: estimar los parámetros desconocidos de la variable aleatoria X que representa a la población.

Una vez identificados los parámetros, se determina totalmente la variable aleatoria en el sentido de que se determina totalmente su distribución de probabilidad. Para ello se usan estadísticos adecuados, útiles, cuyas distribuciones se conocen, como la media muestral, la varianza muestral, el máximo de la muestra, etc.

Definición 6 *Si se emplea un estadístico T para estimar un parametro θ, entonces se dice que T es un* **estimador** *de θ, y al resultado de T, al calcularlo sobre los valores específicos de la muestra, se le llama* **estimación** *de θ.*

2.1.2 Propiedades de los estimadores puntuales

Una estimación obtenida a partir del valor de un estimador, se la define como estimación puntual, por oposición a la estimación por intervalos que se verá más adelante. Alguna de las propiedades deseables más importantes de los estimadores puntuales se incluyen a continuación.

Definición 7 *T es un* **estimador insesgado** *de θ si $E(T) = \theta$ para todos los posibles valores de θ. Se entiende por sesgo del estimador a la diferencia: $E(T) - \theta$. Si T es insesgado su distribución se encuentra centrada en torno a θ.*

Es razonable esperar que un buen estimador T_n, (donde se ha agregado la n en la notación para hacer énfasis en la dependencia del tamaño muestral n), se concentre en torno al parámetro θ a medida que la muestra crece (o sea cuando la información aumenta).

Definición 8 *T_n es un* **estimador consistente** *de θ si:*

$$\lim_{n\to\infty} Prob(|T_n - \theta| \leq \epsilon) = 1 \tag{2.1}$$

para todos los valores de θ y donde ϵ es un número positivo arbitrario.

El requisito establecido por el límite anterior constituye lo que se denomina convergencia en probabilidad. Luego, T_n es un estimador consistente de µ si converge en probabilidad a µ. La consistencia significa la concentración del estimador en torno µ a medida que la muestra crece.
Es esta propiedad de consistencia de un estimador lo que permite conformarse con el valor observado de un estimador y asumirlo como representativo del verdadero valor del parámetro µ. Ya que, el error que se comete al sustituir el verdadero valor del parámetro por su estimación tiene una alta probabilidad de ser menor que ².

La varianza de un estimador insesgado es la cantidad más importante para decidir qué tan bueno es el estimador.

Definición 9 Dado dos estimadores insesgados Ta y Tb de µ, se dice que Ta es más eficiente que Tb si:

$$Var(Ta) <= Var(Tb) \tag{2.2}$$

Definición 10 T es un estimador insesgado de varianza mínima de µ si es más eficiente que cualquier otro estimador insesgado para todos los valores posibles de µ.

Es decir, si es el que tiene mínima varianza entre todos los estimadores insesgados de µ.

2.1.3 Ley de los Grandes Números y el Teorema Central del Límite

La media muestral es el estadístico más usado en los procesos de inferencia:

$$m = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2.3}$$

Este estadístico tiene propiedades resaltantes que justifican su relevancia en el muestreo. Estas se resumen en dos grandes propiedades:

- Ley de los Grandes Números
- Teorema Central del Límite

Ambas válidas cuando la variable aleatoria X tiene valor esperado y varianza finitos.

Considérese una MAS de tamaño n, de una variable aleatoria X de valor esperado μ y desviación estándar σ, entonces por las propiedades ya establecidas sobre la suma de variables aleatorias independientes de la ecuación 2.3 se deduce:

$$E(m) = \mu \qquad Var(m) = \frac{\sigma^2}{n} \tag{2.4}$$

Luego, m es un estimador insesgado de μ.

La media muestral tiene el mismo valor esperado μ pero está más concentrada en torno a μ, ya que la desviación estándar disminuye con la raíz cuadrada de n y $\frac{\sigma^2}{n} \to 0$ cuando $n \to \infty$. Esta concentración se refleja en la siguiente propiedad que es el fundamento teórico para la estimación de la media poblacional por la media muestral:

Theorem 11 Ley de los Grandes Números:
La media muestral m es un estimador consistente de μ.

La ley de los Grandes Números indica que m se concentra entorno a μ, pero no indica cómo es la distribución de la media muestral. El siguiente teorema tiene ese propósito:

Theorem 12 Teorema central del límite
Cuando $n \to \infty$ la distribución de la media muestral, m, se aproxima a la de una Normal; más precisamente el estadístico de la media estandarizada:

$$Z = \frac{m - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{2.5}$$

tiene una distribución que se aproxima a la Normal estándar cuando $n \to \infty$.

Este último teorema es el centro de la estadística y el que le otorga a la Normal su rol preponderante, ya que para muestras aleatorias simples de tamaño grande, de cualquier variable aleatoria de distribución no especificada, se puede trabajar con la hipótesis de que m es Normal, independientemente de la distribución de origen.

Proposición 13 La varianza muestral

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - m)^2}{(n-1)} \tag{2.6}$$

es un estimador insesgado de la varianza de la población, ($E(S^2) = V(X)$), y además, consistente

Pero no existe para ésta un teorema de convergencia como el teorema central del límite, tal como existe para la media muestral.

2.1.4 Distribuciones asociadas a la Normal en el muestreo

En lo que sigue se considerará una muestra aleatoria simple $\{X_1; X_2; ...; X_n\}$ de una variable aleatoria Normal X, $N(\mu; \sigma^2)$. Luego, su media muestral m se distribuye también Normal pero $N(\mu; \sigma^2/n)$ cualquiera sea el tamaño de la muestra.

La distribución Chi-Cuadrado (χ^2_p)

$$\chi^2_p = \sum_{k=1}^{p} Z_k^2 \tag{2.7}$$

donde $Z_1; ...; Z_p$ es una muestra aleatoria simple de Normales estandar ($N(0;1)$).

Proposición 14 La distribución de la siguiente función de la varianza muestral S^2:

$$\frac{(n-1)S^2}{\sigma^2} \tag{2.8}$$

es una χ^2 (Chi-Cuadrado) con n-1 grados de libertad.

Esta distribución que depende de un solo parámetro gl (grados de libertad), es un caso particular de la distribución GAMMA donde el parámetro de forma, $\alpha = gl/2$, y el parámetro de escala, $\theta = 2$.

Su distribución acumulada en SAS es PROBCHI(x,gl), y su inversa CINV(p,gl), que es la más útil en el muestreo; donde gl significa grados de libertad.

Si X es una CHI-SQUARE entonces: $E(X) = gl$ y $V(X) = 2\,gl$

De la ecuación 2.8 se deduce que $S^2 = \frac{\sigma^2}{n-1}\chi^2(n-1)$, luego: $E(S^2) = \sigma^2$ y $V(S^2) = 2\sigma^4/(n-1)$

Se prueba asi que S^2 es un estimador insesgado y consistente de la varianza poblacional σ^2, ya que la varianza tiende a 0 cuando el tamaño muestral crece.

La distribución t de Student

Definición 15 La distribución t_p es el cociente entre las variables aleatorias independientes siguientes:

$$\frac{N(0;1)}{\sqrt{\frac{\chi_p^2}{p}}} \tag{2.9}$$

Esta distribución depende solamente de un parámetro p, (grados de libertad). $E(t_p) = 0$ y $Var(t_p) = \frac{p}{p-2}$ si:p > 2 Cuando los grados de libertad crecen, ($p \to \infty$) la distribución t_p tiende a la normal estandard.

La distribución F de Fischer

Definición 16 Si X e Y son variables aleatorias con distribuciones Chi-cuadrado de n_x y n_y grados respectivamente, se dice que la variable:

$$F = \frac{\frac{X}{nx}}{\frac{Y}{ny}} \tag{2.10}$$

tiene una distribución F con n_x y n_y grados de libertad.

Esta distribución depende de dos parámetros, (gln y gld), grados de libertad del numerador y denominador.
Su distribución acumulada en SAS es PROBF(x,gln,gld) y su inversa (que es la más útil en el muestreo) FINV (p,gln,gld). El archivo dist_f.sas permite graficar la distribución para distintos valores de los parámetros. Obsérvese la asimetría positiva para cualquier valor de los parámetros.

Su importancia reside en que:

Proposición 17 si $\{X_1;\ldots;X_{nx}\}$ es una muestra de una variable $X \sim N(\mu_x, \sigma_x^2)$ y $\{Y_1;\ldots;Y_{ny}\}$ es una muestra de una variable $Y \sim N(\mu_y, \sigma_y^2)$, entonces:

$$Fm = \frac{\frac{s_x^2}{\sigma_x^2}}{\frac{s_y^2}{\sigma_y^2}} \tag{2.11}$$

tiene por distribución una F con n_x - 1 y n_y - 1 grados de libertad.

Además su valor esperado está dado por: $E(F_{nx-1,ny-1}) = \frac{ny}{ny-2}$ ny > 2

Si las variables X e Y tienen igual varianza (aunque sea desconocida), el valor muestral F_m depende sólo de las varianzas muestrales, luego, F_m puede calcularse, y compararse con el valor teórico de la distribución F.

Ejemplo En un laboratorio dos máquinas A y B llenan con el producto, XYZ, sendos envases. Se afirma que ambas son igualmente precisas en el volumen que llenan. Para probarlo se toman muestras de tamaño 28 de A y 21 de B obteniendo varianzas muestrales de valor 100 y 40 respectivamente. Si las varianzas teóricas son iguales Fm= 2.5 y PROB F(2.5,27,20)= 0.98. Luego, el valor Fm= 2.5 está entonces ubicado en un extremo de la distribución ($p = 0.98$) y es poco probable, por lo que se rechaza la hipótesis.

2.2 Estimación por intervalos

Un estimador puntual proporciona un valor aproximado de un parámetro θ de una población, pero hay situaciones donde esa información es insuficiente y se desea responder a la pregunta: ¿cuán próxima es la estimación al parámetro? La estimación por intervalos de confianza da respuesta a la interrogante anterior.
Sea el conjunto $\{X_1, X_2, \ldots, X_n\}$ una Muestra Aleatoria Simple (MAS), de una Variable aleatoria X, y el parámetro θ el que se desea estimar. Para dar una respuesta en términos probabilísticos se construye un intervalo de confianza, [inf ; sup] que con un nivel de confianza previamente especificado contenga al parámetro θ. Más específicamente:

Definición 18 Para establecer un intervalo de confianza, [inf, sup], de un parámetro θ, donde 1-α es el nivel de confianza deseado, se hallarán dos estadísticos inf y sup tal que la probabilidad de que el intervalo cubra a θ es $1 - \alpha$.

$$Prob(\inf\{X_1, \ldots, X_n\} <= \theta <= \sup\{X_1, \ldots, X_n\}) = 1 - \alpha \tag{2.12}$$

¿Cómo interpretar este concepto en términos frecuenciales?
Si se toman 100 muestras aleatorias de la población, y se construye un intervalo de confianza [inf,sup] del 95% para cada una de ellas, se espera que en 95 de ellas θ se encuentre entre inf y sup, o sea: que el intervalo [inf,sup] contenga al parámetro. Obsérvese con detenimiento las siguientes consideraciones:

- que lo aleatorio es el intervalo [inf,sup] que se modifica para cada muestra mientras que el parámetro θ es un desconocido pero fijo

- que es el intervalo aleatorio a priori, creado con los estadísticos inf y sup, el que tiene el nivel de confianza, pero, una vez observados los valores de la muestra y estimados inf y sup a partir de los valores muestrales, ya este intervalo concreto no puede interpretarse en términos de probabilidades, y contendrá o no al parámetro.

2.2.1 Intervalos de la media y de la varianza

Estimación de la media μ cuando se conoce σ

Se determinará el intervalo de confianza de probabilidad $(1-\alpha)$ del valor esperado μ de la variable aleatoria X, $N(\mu;\sigma^2)$, cuando se conoce σ.

Si X es $N(\mu;\sigma^2)$ y la muestra de tamaño N, su media muestral m se distribuye también Normal pero $N(\mu;\frac{\sigma^2}{N})$. Por lo tanto:

$$PROBIT\left(\frac{\alpha}{2}\right) \leq \frac{m-\mu}{\frac{\sigma}{\sqrt{N}}} \leq PROBIT\left(1-\frac{\alpha}{2}\right) \tag{2.13}$$

Luego, el intervalo de confianza esta dado por:

$$m - PROBIT\left(1-\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{N}} \leq \mu \leq m + PROBIT\left(1-\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{N}} \tag{2.14}$$

Una de las aplicaciones principales de la distribución χ^2 es la estimación de la varianza σ^2 de la población.

Estimación de la varianza σ^2

Un intervalo de confianza de probabilidad $(1-\alpha)$ de la varianza σ^2 es:

$$\left[\frac{(n-1)S^2}{CINV(1-\frac{\alpha}{2};n-1)};\frac{(n-1)S^2}{CINV(\frac{\alpha}{2};n-1)}\right]$$

Ejercicio Se toma una muestra del número de empleados de Maraven transportados en taxi diariamente. Los resultados son:
189, 231, 178, 220, 220, 213, 195, 170, 205, 223, 189, 198, 210
199, 211, 198, 210, 222, 203, 198, 194, 205, 203, 199, 205, 211

Halle un intervalo de un 88% de confianza de σ la desviación estándar.

Estimación de la media cuando la varianza es desconocida.

Cuando se desea estimar la media μ a partir de una muestra de tamaño n, y se desconoce σ, se debe recurrir a la distribución t de Student con $n-1$ grados de libertad. Si m y s son la media y la desviación estándar muestral respectivamente de una muestra $\{X_1,\ldots,X_n\}$ de una variable $X \sim N(\mu,\sigma^2)$ entonces de las ecuaciones 2.8, 2.9 se deduce que:

Proposición 19 $t = \frac{m-\mu}{\frac{s}{\sqrt{n}}}$ *se distribuye como una t de Student con n-1 grados de libertad.*

La distribución acumulada en SAS de la distribución t_{gl} es PROBT(x,gl) y su inversa TINV(p, gl).

Su aplicación inmediata es que:

$$m \pm \frac{s}{\sqrt{n}} TINV(1-\frac{\alpha}{2}; n-1) \tag{2.15}$$

es un intervalo de confianza de probabilidad (1-α) del valor esperado, μ.

Ejercicio halle un intervalo de confianza del 90% del valor esperado del número de pasajeros diarios del ejercicio anterior.

2.2.2 Tamaño muestral

Cuando se estima μ y σ es conocido, el intervalo de confianza tiene una longitud de:

$$L = 2\frac{\sigma}{\sqrt{N}} PROBIT(1-\frac{\alpha}{2})$$

estableciéndose una relación simple entre la confianza, la precisión L del resultado, (longitud del intervalo), y el tamaño de la muestra. También se considera en algunos textos el error $E = L/2$. Obsérvese que no se puede fijar arbitrariamente las 3 cantidades. Si deseo una cierta precisión con un grado de confianza $1-\alpha$, entonces el tamaño de la muestra queda determinado por la fórmula anterior. Mas precisamente:

$$N = CEIL[(\frac{2\sigma PROBIT(1-\frac{\alpha}{2})}{L})^2] \tag{2.16}$$

obsérvese que la función ceil es para obtener el entero más próximo por exceso.

Si σ es desconocido, en la ecuación 2.16 debería cambiarse σ por S, y PROBIT por TINV, pero ésta depende de los grados de libertad $N-1$, además, como la varianza muestral S^2 se obtiene del muestreo no se puede conocer a priori. Es por esta razón que para este caso se necesita:
a) una estimación anterior de σ con su intervalo de confianza adecuado.
b) un N razonablemente grande como para sustituir la distribución T por la normal.

Si no se dispone de una estimación anterior de σ, es necesario hacer un muestreo previo pequeño para estimarla (muestreo preliminar):

- Calcule un intervalo de confianza de σ, (con confianza $= 0.8$, por ejemplo), para usar su extremo superior como σ si desea ser cauteloso aunque el N obtenido sea grande.

- Úsese directamente la desviación estándar muestral estimada si quiere ser pragmático, y confía en la estimación preliminar, corriendo el riesgo de que el N sea subestimado y el intervalo final supere la precisión deseada.

2.2.3 Intervalo de confianza para proporciones

La información de que se dispone para estimar una proporción, (p), es el número x de veces que el evento considerado ocurre en n ensayos (tamaño muestral). Como es una distribución binomial el valor esperado de x es np, el estadístico $\frac{x}{n}$, (frecuencia relativa), es un estimador insesgado de p ($E(\frac{x}{n}) = p$). La varianza de este estimador $Var(\frac{x}{n}) = \frac{p(1-p)}{n} \to 0$ si $n \to \infty$; luego, $\frac{x}{n}$ es además consistente.

Usando la distribución binomial, para cada n es posible construir intervalos exactos de confianza de cada proporción p, pero esto es algo complicado. Es por ello que en general se intenta construir intervalos aproximados por un método más simple usando la distribución Normal.
Si se cumple la regla práctica para aproximar una binomial con una normal: $np > 5$ $n(1-p) > 5$ podemos construir intervalos aproximados de una proporción asimilándolos a los de una normal con igual media y varianza:

El intervalo de confianza aproximado es:

$$\frac{x}{n} \pm \sqrt{\frac{p(1-p)}{n}}\, PROBIT(1 - \frac{\alpha}{2}) \qquad (2.17)$$

Como p es desconocido se tienen dos opciones:

- Sustituir $p(1-p)$ por $\frac{1}{4}$ en el intervalo anterior de modo de introducir el valor máximo que el producto $p(1-p)$ puede tener, y así obtener un intervalo más grande pero seguro.

- Usar como valor de p su estimador $\frac{x}{n}$, criterio este más pragmático para definir el intervalo pero menos confiable.

  Cuando p es próximo a $1/2$ los intervalos son semejantes. En cuanto al tamaño de la muestra para obtener una precisión L con confianza $(1-\alpha)$, de la ecuación 2.17:

$$N = 4p_e(1-p_e)\left(\frac{PROBIT(1-\frac{\alpha}{2})}{L}\right)^2 \qquad (2.18)$$

donde p_e es una estimación previa de p. Si se quiere una estimación conservadora pero segura, o en caso de no disponer de p_e, sustituir p_e por $1/2$ en la ecuación 2.18 para obtener una estimación por exceso pero segura del tamaño muestral.

$$N = \left(\frac{PROBIT(1-\frac{\alpha}{2})}{L}\right)^2 \qquad (2.19)$$

2.3 Prueba de hipótesis

Una hipótesis es una afirmación acerca de una característica desconocida de una población. Por ejemplo, $\mu = 123$; si existe suficiente evidencia experimental que apoye la hipótesis.

A pesar de que el significado real de la hipótesis se refiera a características significativas de la población, (al valor esperado, a la desviación estándar, al máximo, etc), toda hipótesis en un test se estable en términos de parámetros de una variable aleatoria. Por fortuna, para la mayoría de las distribuciones estos parámetros coinciden con el valor esperado, la varianza, etc. Así cuando se decide entre dos hipótesis lo que realmente se hace es decidir entre posibles valores de los parámetros. Más precisamente, se opta entre valores de los parámetros para un mismo tipo de distribución. Por ejemplo, si una población es Normal $N(20;\sigma)$ o Normal $N(23;\sigma)$:

Ilustrémoslo con el siguiente ejemplo: disponemos de una empaquetadora XX que llena 120 bolsas por minuto, con una desviación estándar $\sigma = 5$. Se nos ofrece comprar otra ZZ de la cual se afirma que su media es superior, y que someteremos a un test. No se invertirá dinero en comprar ZZ salvo que exista clara evidencia experimental que indique que ZZ tiene una media mayor. Se supondrá que ambas máquinas tienen la misma desviación estándar conocida. Se nos permite hacer una prueba sobre ZZ de modo de tomar una muestra de varios minutos de producción $X_1, \ldots, X_n$.

La hipótesis nula, la hipótesis que no deseamos abandonar salvo que exista suficiente evidencia en su contra, es que la media de ZZ es igual a la de XX. La hipótesis alternativa es que la media es mayor.

2.3.1 Conceptos básicos

Se formulará este problema en general:

$Ho -> \mu = \mu_0 = 120$ Prueba unilateral
$Ha -> \mu = \mu_a > \mu_0$ σ conocida

Cuando la hipótesis alternativa es que la media es diferente pero no importa si es mayor o menor se formula un test bilateral:

$Ho -> \mu = \mu_0 = 120$ Prueba bilateral
$Ha -> \mu = \mu_a \neq \mu_0$ σ conocida

Se toma la muestra aleatoria $X_1, \ldots, X_n$, del número de bolsas producido por minuto. Se calcula la media muestral m y se compara con un número L preestablecido, que más adelante se determinará:

CRITERIO DE DECISIÓN

Región de aceptación de Ho | Región de rechazo de Ho
L

OPTAR POR Ho $m < L$ OPTAR POR Ha $m > L$

Entre la realidad y la decisión tomada, pueden establecerse 4 situaciones diferentes, 2 aciertos y 2 errores que se reflejan en el siguiente cuadro:

	Decido Ho	Decido Ha
Realidad Ho		error de primer tipo α
Realidad Ha	error de segundo tipo β	Potencia $= 1 - \beta$

Lo deseable sería reducir al mínimo la probabilidad de los errores pero para una muestra de tamaño fijo esto no es posible. De los dos errores indicados, al más importante, comprar la nueva máquina siendo su media igual a la que tengo, se le designa como **error de primer tipo** o nivel de significación de la prueba α: Es la probabilidad de optar por Ha siendo verdadera la Ho α = prob(Ha/Ho)

Es el error que no deseo cometer y lo controlo asignándole una probabilidad baja α (nivel de significación). Fijado el nivel de significación, es deseable que la potencia del test sea alta o lo que es equivalente que el error de segundo tipo sea bajo (Potencia = $1-\beta$).

Error de tipo II β = prob(Ho/Ha):

Potencia del test= prob(Ha/Ha).

Para establecer el criterio de decisión se determina L de manera que: Prob(m > L/Ho) = α.

Fijado L, el error de segundo es tipo: Prob(m < L/Ha) = β

En términos jurídicos, el error de primer tipo es condenar al inocente, y el de segundo, absolver al culpable.

Retomando el ejemplo, como la desviación estándar es conocida, por las propiedades de la media muestral, L está dado por:

$$L = \mu_o + \frac{\sigma}{\sqrt{N}} PROBIT(1-\alpha) \tag{2.20}$$

y el error de segundo tipo:

$$\beta = PROBNORM\left(\frac{L-\mu_a}{\frac{\sigma}{\sqrt{}}}\right) \tag{2.21}$$

Tamaño muestral en función de α y β

En la página anterior se fijaron el error de primer tipo y el tamaño de la muestra; en consecuencia, el error de segundo tipo quedó determinado por α y n. Si, por el contrario, se desea controlar a priori ambos errores y se fijan α y β, el tamaño adecuado de la muestra n se puede calcular a partir de α y β.

Se incluye a continuación los resultados para los 3 posibles test, se observará que la alternativa, μ_a, debe fijarse en un valor preestablecido.

Caso: Hipótesis simple contra alternativa simple con desviación estándar conocida.

I)

Ho $\rightarrow \mu = \mu_0$ Prueba unilateral, σ conocida

Ha $\rightarrow \mu > \mu_0$ o $\mu < \mu_0$

$$N = \frac{\sigma^2 \left[PROBIT(1-\alpha) - PROBIT(\beta)\right]^2}{(\mu_a - \mu_o)^2}$$

II) Si la alternativa es una prueba bilateral, $\mu_a \neq \mu_o$, el paréntesis del numerador debe sustituirse por: $PROBIT(1-\frac{\alpha}{2}) + PROBIT(1-\beta)$, pero el valor de N es aproximado.

2.3.2 Como programar las pruebas

En la formulación anterior se tomó la decisión comparando la media muestral m con el valor L. Sin embargo, lo habitual es estandarizar el resultado como se indica en los test siguientes:

Prueba Sobre la media (con desviación estándar σ conocida).

Estadístico de prueba:

$$Est = \left(\frac{m - \mu_o}{\frac{\sigma}{\sqrt{N}}}\right)$$

Hipótesis nula $\mu = \mu_o$

Prueba sobre la media σ conocida

Hip.alternativa	Acepta Ho	valor p
$\mu_a > \mu_o$	Est < PROBIT(1-α)	1-PROBNORM(Est)
$\mu_a < \mu_o$	Est > PROBIT(α)	PROBNORM(Est)
$\mu_a \neq \mu_o$	PROBIT($\frac{\alpha}{2}$)< Est < PROBIT(1-$\frac{\alpha}{2}$)	2(1-PROBNORM($\lvert Est \rvert$))

Prueba sobre la media (con desviación. estándar desconocida).

Estadístico de prueba:

$$Est = \left(\frac{m - \mu_o}{\frac{s}{\sqrt{N}}}\right)$$

Hipótesis nula $\mu = \mu_o$

Prueba sobre la media σ desconocida

Hip.alternativa	Acepta Ho	valor p
$\mu_a > \mu_o$	Est < TINV(1-α;N $-$ 1)	1-PROBT(Est,N $-$ 1)
$\mu_a < \mu_o$	Est > TINV(α;N $-$ 1)	PROBT(Est,N $-$ 1)
$\mu_a \neq \mu_o$	PROBT($\frac{\alpha}{2}$,N $-$ 1)< Est < PROBT(1-$\frac{\alpha}{2}$,N $-$ 1)	2(1-PROBT($\lvert Est \rvert$;N $-$ 1))

Prueba sobre la varianza

Estadístico de prueba:

$$Est = (- 1)\frac{S^2}{\sigma^2}$$

Hipótesis nula $\sigma = \sigma_o$

prueba acerca de la varianza

Hip.alternativa	Acepta Ho	valor p
$\sigma_a > \sigma_o$	Est < CINV(1-α,N-1)	1-PROBCHI(Est,N-1)
$\sigma_a < \sigma_o$	Est > CINV(α,N-1)	PROBCHI(Est,N-1)
$\sigma_a \neq \sigma_o$	CINV($\frac{\alpha}{2}$,N-1)< Est < CINV(1-$\frac{\alpha}{2}$,N-1)	2(1-PROBCHI(|Est|,N-1))

Valor p

Es posible que el valor del estadístico (Est) sea próximo al límite de decisión y se quiere expresar esta proximidad en términos probabilísticos. Si se desea saber cuan cerca se está de la probabilidad α, se determina el valor de probabilidad, valor p, que es:

El mínimo nivel de significación para el cual los datos observados indican que se tendría que rechazar la hipótesis nula.

Esta probabilidad se calcula de manera idéntica a la forma en que se asignó la región de rechazo, de modo que si la media muestral coincide con el o los límites de la región el valor p es α.

En el caso anterior, p= 1-PROBNORM (Est) y se le compara con el error de primer tipo. Por ejemplo, si se diseño un test con $\alpha = 0.05$ y el valor p es 0.06, Ud acepta la hipótesis nula pero le queda la duda ya que con otra muestra similar podría haber dado un valor p= 0.047 y optar por la alternativa. Sin embargo, si p= 0.30 Ud. aceptaría la hipótesis nula con mayor convicción.

2.3.3 Pruebas para dos poblaciones

Prueba sobre la diferencia de medias

Sean $\{X_1, \ldots, X_{nx}\}$ una muestra de una variable $X \sim N(\mu_x, \sigma_x^2)$ y $\{Y_1, \ldots, Y_{ny}\}$ una muestra de una variable $Y \sim N(\mu_y, \sigma_y^2)$.

CASO: Varianzas de ambas poblaciones conocidas.

Estadístico de prueba:

$$Est = \left(\frac{m_y - m_x}{\sqrt{\frac{\sigma_x^2}{nx} + \frac{\sigma_y^2}{ny}}} \right) \tag{2.22}$$

Hipótesis nula $\mu_x = \mu_y$

Prueba de igualdad de medias con varianzas conocidas

Hip.alternativa	Acepta Ho	valor p
$\mu_y > \mu_x$	Est < PROBIT (1-α)	1-PROBNORM (Est)
$\mu_y \neq \mu_x$	PROBIT ($\frac{\alpha}{2}$)< Est< PROBIT (1-$\frac{\alpha}{2}$)	2(1-PROBNORM (|Est|))

CASO: Varianzas de ambas poblaciones iguales pero desconocidas.

Supóngase que las varianzas son iguales pero que no se conoce su valor: $\sigma_x = \sigma_y = \sigma$ (desconocida).

Estadístico de prueba:

$$Est = \left(\frac{m_y - m_x}{S\sqrt{\frac{1}{nx} + \frac{1}{ny}}} \right) \tag{2.23}$$

Donde S^2 es un promedio de las varianzas muestrales ponderado por el tamaño muestral:

$$S^2 = \frac{(n_x - 1)S_x^2 + (n_y - 1)S_y^2}{n_x + n_y - 2}$$

Sea $nz = n_x + n_y - 2$ la suma de los grados de libertad asociados a cada muestra.

Hipótesis nula $\mu_x = \mu_y$

Prueba de igualdad de medias con varianzas desconocidas iguales

Hip.alternativa	Acepta Ho	valor p
$\mu_y > \mu_x$	Est < TINV (1 $-\alpha$; nz)	1-PROBT (Est; nz)
$\mu_y \neq \mu_x$	TINV ($\frac{\alpha}{2}$,Nz)< Est< TINV (1-$\frac{\alpha}{2}$,Nz)	2(1-PROBT (|Est|,Nz))

Pruebas sobre observaciones apareadas

Cuando las observaciones de las dos variables se haccn sobre las mismas unidades experimentales
y en consecuencia los tamaños muestrales son iguales, debe considerarse una nueva variable:
$Z = Y - X$, y reducir el análisis de la diferencias de medias al estudio de si la media de la variable Z es 0. Es decir, en el caso de observaciones apareadas, las pruebas sobre dos poblaciones se reducen a las pruebas de una población si se considera como nueva variable la diferencia de las variables que se desea comparar.

Prueba sobre la igualdad de varianzas

Sean $\{X_1, \ldots, X_{nx}\}$ una muestra de una variable X $N(\mu_x, \sigma_x^2)$ y $\{Y_1, \ldots, Y_{ny}\}$ una muestra de una variable Y $N(\mu_y, \sigma_y^2)$.

Estadístico de prueba:

$$Est = \frac{S_y^2}{S_x^2} \qquad (2.24)$$

Hipótesis nula $\sigma_x = \sigma_y = \sigma$ (desconocida)

Prueba de igualdad de varianzas

Hip.alternativa	Acepta Ho	valor p
$\sigma_y > \sigma_x$	$Est < FINV(1-\alpha, N_y-1, N_x-1)$	$1-PROBF(Est, N_y-1, N_x-1)$
$\sigma_y \neq \sigma_x$	$FINV(\frac{\alpha}{2}, N_y-1, N_x-1) < Est < FINV(1-\frac{\alpha}{2}, N_y-1, N_x-1)$	$2(1-PROBF(\lvert Est \rvert, N_y-1, N_x-1)$

Pruebas clásicos para poblaciones Normales con el SAS

Prueba sobre Medias.

Una población.

$H_o \to \mu = \mu_0$ Prueba unilateral, σ desconocida

$H_a \to \mu_a \neq \mu_0$

Probar que el valor esperado de X es μ_o es lo mismo que probar que el de la variable $Z = X - \mu_o$ es 0. Para hacer ese test se toma una muestra $\{Z_1; Z_2; \ldots; Z_n\}$ de ella y se aplica el PROC UNIVARIATE que determina la probabilidad de que la media muestral se desvíe de 0 un valor igual al observado. El SAS incluye el valor p.

- $H_o \to \mu = \mu_0$ Prueba unilateral, σ desconocida
- $H_a \to \mu_a > \mu_0$ o $\mu_a < \mu_0$

Idéntico al anterior pero observando que el valor p debe ser la mitad de lo que indica el Univariate y además, la media muestral debe tener el signo adecuado al test.

Dos poblaciones

El Procedimiento PROC TTEST permite comparar medias de dos variables X e Y, a partir de sendas muestras de tamaño arbitrario, tanto en el caso de que sus varianzas sean iguales o distintas.

Prueba:

$H_o \rightarrow \mu_x = \mu_y$

$H_a \rightarrow \mu_x \neq \mu_y$

Para probar si dos variables X e Y tienen el mismo valor esperado se toman muestras $\{X_1, \ldots, X_{nx}\}$ de X e $\{Y_1, \ldots, Y_{ny}\}$ de Y. El PROC TTEST del SAS permite realizar la prueba de igualdad de medias.

Ejemplo incluya 'ttest.sas' del directorio de ejercicios.

Prueba:

$H_o \rightarrow \mu_x = \mu_y$

$H_a \rightarrow \mu_x < \mu_y$ o $\mu_x > \mu_y$

Idéntico al anterior pero observando que el valor p debe ser la mitad de lo que indica el TTEST y además la media muestral de Y mayor que la de X si $\mu_x < \mu_y$.

Prueba:

$H_o \rightarrow \mu_x + d = \mu_y$

$H_a \rightarrow \mu_x + d < \mu_y$

Para probar si las variables X e Y cumplen que el valor esperado de Y es d unidades mayor que el de X, se toman muestras $\{X_1, \ldots, X_{nx}\}$ de X e $\{Y_1, \ldots, Y_{ny}\}$ de Y. Se crea una nueva variable $Z = X + d$ que tendrá por valor esperado el de $X + d$, y se reduce el problema al Test anterior usando el PROC TTEST del SAS con Z e Y.

Prueba sobre varianzas

Para probar si dos variables X e Y tienen varianzas iguales se toman muestras $\{X_1, \ldots, X_{nx}\}$ de X e $\{Y_1, \ldots, Y_{ny}\}$ de Y. El PROC TTEST del SAS permite realizar el test de igualdad de varianzas.

Prueba bilateral:

$H_o \rightarrow \sigma_x^2 = \sigma_y^2$

$H_a \rightarrow \sigma_x^2 \neq \sigma_y^2$

Incluye el valor del estadístico F y el valor p asociado.

o-o-o-o-o

Prueba unilateral:

$H_o \rightarrow \sigma_x^2 = \sigma_y^2$

$H_a \rightarrow \sigma_x^2 > \sigma_y^2$

Si el test es unilateral, debe tomarse como valor p la mitad, y observar si el test tiene el signo coherente con la desigualdad de la Hipótesis alternativa, o sea, la varianza muestral de Y debe ser mayor que la de X.

2.3.4 Pruebas de ajuste de distribuciones

Prueba de Kolmogoroff para el ajuste de distribuciones muestrales

Alcance: distribuciones continuas, donde los parámetros estén totalmente identificados.

La siguiente función de Kolmogoroff juega un rol fundamental en los test de ajuste de distribuciones empiricas.

$$K(z) = \sum_{k=-\infty}^{\infty} (-1)^k \exp(-2k^2 z^2)$$

Algunos valores significativos de esta función se incluyen a continuación:

K(z)	.75	.80	.90	.95	.975	.99
z	1.02	1.07	1.22	1.36	1.48	1.64

Dadas dos distribuciones empíricas $F_n(x)$ y $G_m(x)$ y una teórica $F(x)$, sean:

$$D_n = \text{máximo} \, |F_{n(x)} - F(x)|$$
$$D_{nm} = \text{máximo} \, |F_{n(x)} - G_{m(x)}|$$

donde desde un punto de vista práctico los máximos se obtienen considerando solamente los valores discretos de las distribuciones muestrales empíricas.

Prueba de ajuste de la distribución empírica $F_n(x)$ a la teórica $F(x)$ totalmente especificada.

Hipótesis nula: Los valores muestrales $X_1; \ldots; X_n$, provienen de la distribución teórica $F(x)$.

Soporte estadístico: para n "razonablemente" grande se cumple:

$$Prob[\sqrt{n} \, D_n < z] = K(z)$$

Diseño de la Prueba.

Establecido un nivel de significación α, este valor debe ser igual a $1 - K(z)$.
Se halla, entonces, el z asociado a K(z).
Cómo decidir?
Dada una muestra de tamaño n, se debe hallar D_n y verificar la ecuación anterior. Si

$$\sqrt{(N)}\, D_N \le z$$

se acepta la hipótesis nula.
Valor p: si se considera el valor de z que satisface la igualdad anterior, entonces el valor p del test está dado por $1 - K(z)$:

Prueba de ajuste sobre si ambas distribuciones empíricas $F_n(x)$ y $G_m(x)$ provienen de la misma distribución teórica $F(x)$ desconocida.

Para n y m razonablemente grandes se cumple:

$$Prob\left[\sqrt{\frac{nm}{n+m}}\, D_{nm} < z\right] = K(z)$$

Luego: Dadas sendas muestras de tamaños m y n, se debe hallar D_{nm} y aplicar la ecuación anterior. El valor p del test está dado por $1 - K(z)$.

2.3.5 Prueba de independencia

X e Y son independientes cuando la estructura de probabilidad de X no cambia por los valores que toma Y.

Dadas dos Variables Aleatorias X e Y con valores esperados μ_x y μ_y y desviaciones standard σ_x y σ_y respectivamente, se define covarianza de X e Y:

$$COV(X;Y) = E[(X - \mu_x)(Y - \mu_y)]$$

y correlación entre X e Y:

$$\rho = CORR(X;Y) = \frac{COV(X;Y)}{\sigma_x \sigma_y}$$

La correlación expresa el grado de asociación lineal entre dos variables y siempre es un número entre -1 y 1. Si las variables son independientes su correlación es 0.

El PROC CORR del SAS calcula las correlaciones muestrales entre dos o más variables y al mismo tiempo realiza una prueba de hipótesis acerca de si las variables son independientes.