Análisis y Visualización de Datos

Clase 3

Conceptos de Estadística Inferencial

Docentes : Soledad Palacios(UNLP)

Milagro Teruel (UNC)

- Definición de parámetros poblacionales y muestrales
- Esperanza y Varianza de una variable aleatoria
- Estimadores: características
- EMV
- Teorema central del límite
- Intervalos de confianza
- Test de hipótesis

Introducción

La inferencia estadística es el conjunto de métodos y técnicas que permiten inducir, a partir de la información empírica proporcionada por una muestra, cual es el comportamiento de una determinada población con un riesgo de error medible en términos de probabilidad.

Introducción

El campo de la inferencia estadística está formado por los métodos utilizados para tomar decisiones o para obtener conclusiones sobre el o los parámetros de una población. Estos métodos utilizan la información contenida en una muestra de la población para obtener conclusiones.

La inferencia estadística puede dividirse en dos grandes áreas:

- estimación de parámetros
- pruebas de hipótesis

Introducción

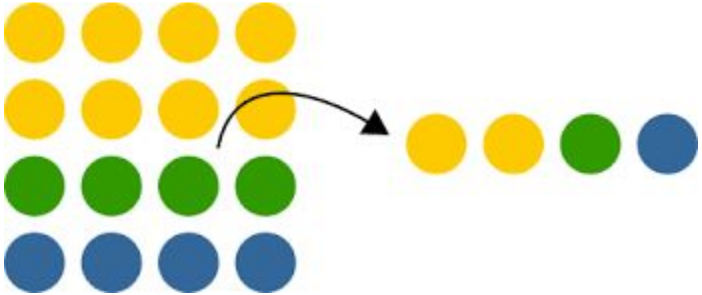
Los métodos de la inferencia nos permiten proponer el valor de una cantidad desconocida (estimación) o decidir entre dos teorías contrapuestas cuál de ellas explica mejor los datos observados (test de hipótesis).

El fin último de cualquier estudio es aprender sobre las poblaciones. Pero es usualmente necesario, y más práctico, estudiar solo una muestra de cada una de las poblaciones.

Muestreo aleatorio

Definamos algunos términos

Una ***población*** está formada por la totalidad de las observaciones en las cuales se tiene cierto interés



Una ***muestra*** es un subconjunto de observaciones seleccionada de una población

Características de la muestra

Para que las inferencias sean válidas, la muestra debe ser representativa de la población. Se selecciona una muestra aleatoria como el resultado de un mecanismo aleatorio. En consecuencia, la selección de una muestra es un experimento aleatorio, y cada observación de la muestra es el valor observado de una variable aleatoria. Las observaciones en la población determinan la distribución de probabilidad de la variable aleatoria.

Estadísticos

Un estadístico es cualquier **función** de la muestra aleatoria

Es una medida cuantitativa, derivada de un conjunto de datos de una muestra, con el objetivo de estimar o inferir características de una población o modelo estadístico.

Usamos una muestra para conocer o estimar características de la población, denominamos:

PARÁMETRO ⇒ una medida resumen calculada sobre la población
ESTADÍSTICO ⇒ una medida resumen calculada sobre la muestra

Parámetros y Estadísticos

La característica de interés de la población se conoce como parámetro y la característica correspondiente de la muestra es el estadístico de la muestra o la estimación de parámetro. Debido a que el estadístico es un resumen de información acerca de un parámetro obtenido a partir de la muestra, el valor de un estadístico depende de la muestra específica que fue extraída de la población.

Sus valores cambian aleatoriamente de una muestra aleatoria a la siguiente, por lo que un estadístico es una cantidad aleatoria (variable).

Acerca de los parámetros

Los parámetros son medidas descriptivas de una población completa que se pueden utilizar como las entradas para que una función de distribución de probabilidad (PDF, por sus siglas en inglés) genere curvas de distribución. Los parámetros generalmente se representan con letras griegas para distinguirlos de los estadísticos de muestra. Por ejemplo, la media de la población se representa con la letra griega mu (μ) y la desviación estándar de la población, con la letra griega sigma (σ). Los parámetros son constantes fijas, es decir, no varían como las variables. Sin embargo, sus valores por lo general se desconocen, porque es poco factible medir una población entera.

Esperanza de una variable aleatoria

Sea X una v.a. discreta con rango R_X . La esperanza , valor medio o valor esperado de X , lo anotamos E (X) , y se define como

$$E(x) = \begin{cases} \int_{-\infty}^{\infty} x f(x)\,dx & \text{caso continuo} \\ \sum_{\forall x} x_i P(x_i) & \textit{caso discreto} \end{cases}$$

Ejemplo discreto

Sea la variable aleatoria
X='número de crías de una camada'

X toma los valores x = 0,1,2,3, con probabilidades

$P(X=0) = 0.2;$
$P(X=1) = 0.3;$
$P(X=2) = 0.3;$
$P(X=3) = 0.2;$

$$E[X]= 0*0.2+1*0.3+2*0.3+3*0.2= 1.5$$

Ejemplo continuo

La altura de un cierto árbol sigue una v.a. con función de densidad, $f(x) = x/12$, con $1<x<5$.
Calcular la Esperanza de X

$$E[X] = \int_{-\infty}^{\infty} x f(x)\,dx = \int_{1}^{5} \frac{x^2}{12}\,dx = \frac{1}{36}\left[x^3\right]_1^5 = \frac{31}{9}$$

Observaciones de la esperanza

1- La esperanza de una v.a. no tiene que coincidir necesariamente con algún valor del rango de la variable

2- Se puede interpretar a la esperanza de una v.a. como un promedio “pesado” o “ponderado” de los valores del rango de la variable , donde el “peso” de cada x_i es la probabilidad $P(X = x_i)$

3- $E(aX+b)= aE(X)+ b$ siendo $a, b \in \mathbb{R}$

4-La esperanza de una v.a. mide dónde está centrada la distribución de probabilidad

Varianza de una variable aleatoria

Sea X una v.a. discreta con rango R_X, función de distribución de probabilidad p(x) y esperanza E (X) = μ , Entonces la varianza de X , que anotamos **V (X)**, $\boldsymbol{\sigma^2}$ o $\boldsymbol{\sigma_X^2}$ es

$$\mathbf{V(X) =}$$

$$\mathbf{E(X-\mu)^2 =}$$

$$\mathbf{E[X^2]-(E[X])^2}$$

Valor esperado y varianza de una variable aleatoria continua

X: Variable Aleatoria Continua, Ejemplo

Valor esperado:

$$E[X] = \mu_x = \int_{-\infty}^{+\infty} x f(X)dx$$

Varianza:

$$V[X] = \sigma_x^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx$$

$$V[X] = \sigma_x^2 = E[x^2] - (E[X])^2$$

Ejemplo caso discreto

Se desea realizar un estudio sobre el número de crías en una camada.

X:"Número de crías en una camada"

X toma los valores x= 0,1,2,3 con probabilidades

$P\{X= 0\} = 0.2;$

$P\{X= 1\} = 0.3;$

$P\{X=2\} = 0.3;$

$P\{X=3\} = 0.2$

Calcular la varianza de dicha variable aleatoria.

Solución

$$E\left[X^2\right] = 0^2 \times 0.2 + 1^2 \times 0.3 + 2^2 \times 0.3 + 3^2 \times 0.2 = 3.3$$

La esperanza de X ya fue calculada : $E[X] = 1.5$

Por lo tanto:

$$Var[X] = E\left[X^2\right] - E[X]^2 = 3.3 - 1.5^2 = 1.05$$

Ejemplo caso continuo

La altura de un cierto árbol sigue una v.a. con función de densidad,

$$f(x) = x/12 \quad \text{con } 1<x<5$$

Calcular la Varianza de X

Solución

$$E\left[X^2\right] = \int_{-\infty}^{\infty} x^2 f(x)\,dx = \frac{1}{12}\int_1^5 x^3\,dx = \frac{1}{48}\left[x^4\right]_1^5 = 13$$

La esperanza de X ya fue calculada y es: $E[X] = 31/9$.

Por lo tanto:

$$Var[X] = E\left[X^2\right] - E[X]^2 = 13 - \left(\frac{31}{9}\right)^2 = 1.1358$$

Observaciones de la varianza

1- La varianza de una v.a. nunca es negativa

2- La cantidad $h (X) = (X - \mu)^2$ es el cuadrado de la desviación de X desde su media, y la varianza de X es la esperanza de la desviación al cuadrado. Si la mayor parte de la distribución de probabilidad está cerca de μ , entonces σ^2 será relativamente pequeña. Si hay valores de la variable alejados de μ que tengan alta probabilidad, entonces σ^2 será grande.

3- σ^2 está expresado en las unidades de medida de X al cuadrado, mientras que σ está expresada en las mismas unidades de medida que X .

4- -> $V (aX) = a 2 V (X)$
-> $V (X + b) = V (X)$

Teorema Central del límite

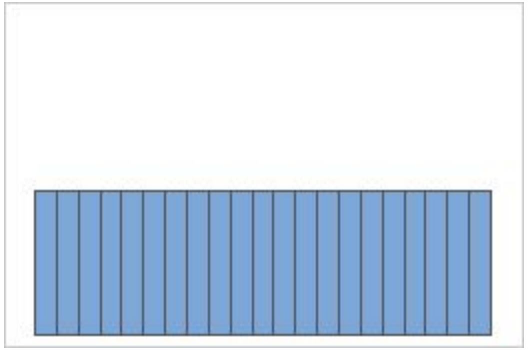
Cuando el tamaño de la muestra es lo suficientemente grande, la distribución de las medias sigue aproximadamente una distribución normal. El teorema se aplica independientemente de la forma de la distribución de la población. Muchos procedimientos estadísticos comunes requieren que los datos sean aproximadamente normales. El teorema de límite central le permite aplicar estos procedimientos útiles a poblaciones que son considerablemente no normales. El tamaño que debe tener la muestra depende de la forma de la distribución original. Si la distribución de la población es simétrica, un tamaño de muestra de 5 podría producir una aproximación adecuada. Si la distribución de la población es considerablemente asimétrica, es necesario un tamaño de muestra más grande

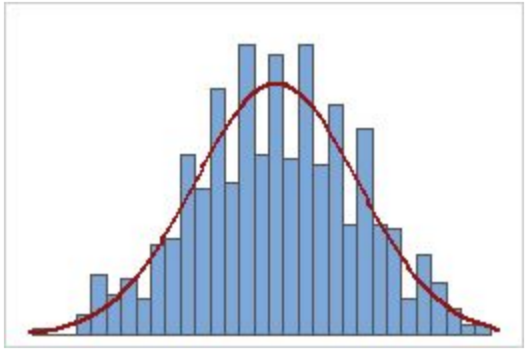
Propiedades del TCL

- El teorema del límite central garantiza una distribución aproximadamente normal cuando *n* es suficientemente grande.
- Existen diferentes versiones del teorema, en función de las condiciones utilizadas para asegurar la convergencia. Una de las más simples establece que es suficiente que las variables que se suman sean independientes, idénticamente distribuidas, con valor esperado y varianza finitas.
- La aproximación entre las dos distribuciones es, en general, mayor en el centro de las mismas que en sus extremos o colas, motivo por el cual se prefiere el nombre "teorema del límite central" ("central" califica al límite, más que al teorema).
- Este teorema, perteneciente a la teoría de la probabilidad, encuentra aplicación en muchos campos relacionados, tales como la inferencia estadística

Teorema Central del Límite

Distribución uniforme



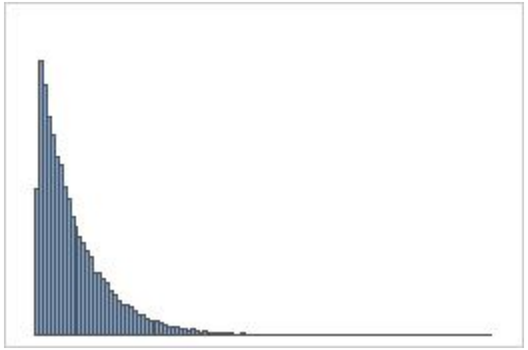


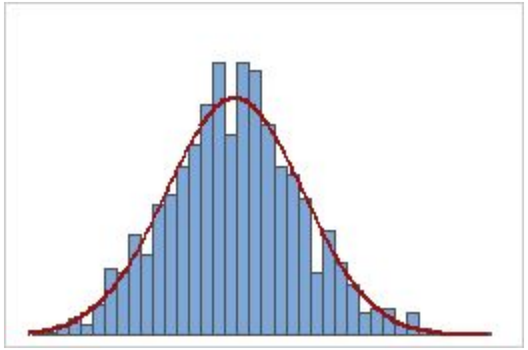
En la medición de pocos casos vemos que es marcadamente simétrica pero no se acerca a la gráfica de una normal

Histograma que conseguimos luego de medir 1000 veces una característica de los individuos de esa población

Teorema Central del límite

Función Exponencial





La gráfica de pocos casos nos mostrará un dibujo asimétrico y no normal

Con gran cantidad de casos nos acercamos a una campana de Gauss

Estadísticos usuales

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una variable aleatoria X donde $E(X) = \mu$ y $V(X) = \sigma^2$

Si desconocemos μ un estadístico que se utiliza para estimar este parámetro es la media o promedio

Análogamente si se desconoce σ^2 un estadístico usado para tener alguna información sobre ese parámetro es la varianza muestral

También se usa frecuentemente la desviación estándar muestral

Estimadores puntuales

Cuando un estadístico se utiliza para estimar un parámetro desconocido se lo llama estimador puntual. Es habitual simbolizar en forma genérica a un parámetro con la letra θ y al estadístico que se utiliza como estimador puntual de θ , simbolizarlo con $\hat{\Theta}$.

Lo que se desea de un estimador puntual es que tome valores "próximos" al verdadero parámetro.

Puede ocurrir que se tenga más de un estimador para un parámetro, por ejemplo para estimar la media muestral se pueden considerar el promedio muestral, o también la semisuma entre X_1 y X_n

Estimadores puntuales

La media de la muestra (por ejemplo el valor que tomará nuestro estadístico) generalmente no es exactamente la misma que la media de la población. Esta diferencia puede ser causada por muchos factores, incluido un diseño de encuesta deficiente, métodos de muestreo sesgados y la aleatoriedad inherente a extraer una muestra de una población

Criterios para evaluar estimadores puntuales

$$\bar{X} = \sum_{i=0}^{n} \frac{x_i}{n}$$

vs

$$semisuma = \frac{x_1 + x_n}{2}$$

Sesgo de un estimador puntual

Se dice que el estimador puntual $\hat{\Theta}$ es un estimador insesgado del parámetro θ si $E(\hat{\Theta}) = \theta$ cualquiera sea el valor verdadero de θ

$$sesgo = E(\hat{\theta}) - \theta$$

Ejemplo

$$sesgo = E(\hat{\theta}) - \theta$$

$$sesgo = E(\sum_{i=0}^{n} \frac{x_i}{n}) - \mu$$

$$sesgo = E(\frac{1}{n} \sum_{i=0}^{n} x_i) - \mu$$

$$sesgo = \frac{1}{n} E(\sum_{i=0}^{n} x_i) - \mu$$

Pero, tratándose de las componentes de una muestra aleatoria es:

$$E(X_i) = E(X) = \mu$$

$$sesgo = \frac{1}{n}.n.\mu - \mu$$

$$sesgo = 0$$

Supongamos que tomamos otro estimador para μ, lo anotamos $\hat{\mu}_2 = \dfrac{X_1 + X_n}{2}$

Entonces como

$$E(\hat{\mu}_2) = E\left(\frac{X_1 + X_n}{2}\right) = \frac{1}{2}\left(E(X_1) + E(X_2)\right) = \frac{1}{2}(\mu + \mu) = \frac{1}{2}2\mu = \mu \ ,$$

$\hat{\mu}_2 = \dfrac{X_1 + X_n}{2}$ es también un estimador insesgado de μ

¿Cuál de los dos estimadores es mejor?

Calculamos la varianza de cada uno utilizando las propiedades de la varianza.

Eficiencia

Diremos que un estimador es más *eficiente* o más *preciso* que otro estimador, si la varianza del primero es menor que la del segundo. Por ejemplo, si θ_1 < y θ_2 son ambos estimadores de θ y

$$\mathrm{Var}(\hat{\theta}_1) < \mathrm{Var}(\hat{\theta}_2)$$

diremos que θ_1 es más **eficiente** que θ_2. Un estimador es más eficiente (más preciso), por tanto, cuanto menor es su varianza.

Eficiencia

$$V(\overline{X}) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} V\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i),$$

donde en la última igualdad hemos tenido en cuenta que, por tratarse de una muestra aleatoria, las X i con i=1,2,...,n son variables aleatorias independientes y, en consecuencia, la varianza de la suma de ellas es la suma de las varianzas. Si tenemos en cuenta que además todas tienen la misma distribución que X y por lo tanto la misma varianza:

Eficiencia

$V(X_i) = V(X) = \sigma^2 \quad \forall i = 1,2,\ldots,n$, tenemos

$$V(\overline{X}) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}.$$

Análogamente calculamos la varianza de $\hat{\mu}_2 = \dfrac{X_1 + X_n}{2}$:

$$V(\hat{\mu}_2) = V\left(\frac{X_1 + X_n}{2}\right) = \frac{1}{4}\left(V(X_1) + V(X_2)\right) = \frac{1}{4}\left(\sigma^2 + \sigma^2\right) = \frac{\sigma^2}{2}$$

Vemos que si $n > 2$ entonces $V(\hat{\mu}_1) < V(\hat{\mu}_2)$. Por lo tanto si $n > 2$ es mejor estimador $\hat{\mu}_1$

Error cuadrático medio

Supongamos ahora que $\hat{\Theta}_1$ y $\hat{\Theta}$, 2 son dos estimadores de un parámetro θ y alguno de ellos no es insesgado.
A veces es necesario utilizar un estimador sesgado. En esos casos puede ser importante el error cuadrático medio del estimador.

El ***error cuadrático medio*** de un estimador $\hat{\Theta}$ de un parámetro θ está definido como

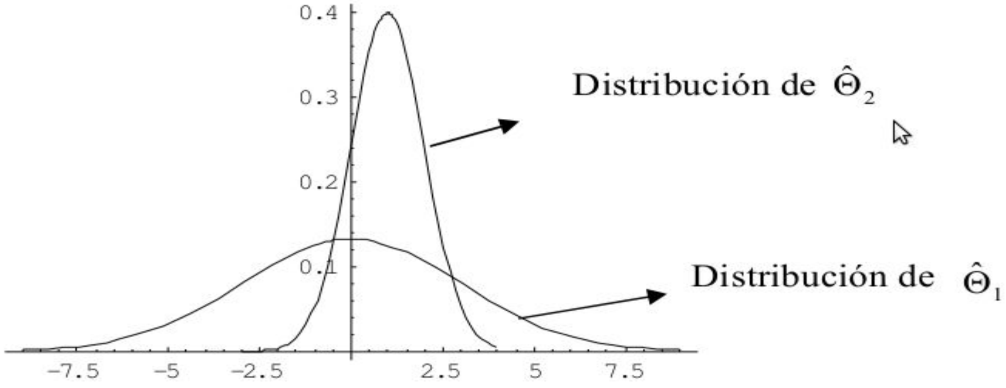
$$ECM\left(\hat{\Theta}\right)=E\left[\left(\hat{\Theta}-\theta\right)^2\right]$$

El error cuadrático medio puede escribirse de la siguiente forma:

$$ECM\left(\hat{\Theta}\right)=V\left(\hat{\Theta}\right)+\left(b\left(\hat{\Theta}\right)\right)^2$$

Cómo elegir entre 2 estimadores

A veces es preferible utilizar estimadores sesgados que estimadores insesgados, si es que tienen un error cuadrático medio menor.
En el error cuadrático medio se consideran tanto la varianza como el sesgo del estimador.
Si $\hat{\Theta}_1$ y $\hat{\Theta}_2$ son dos estimadores de un parámetro θ, tales que $E(\hat{\Theta}_1) = \theta$; $E(\hat{\Theta}_2) \neq \theta$ y $V(\hat{\Theta}_2) < V(\hat{\Theta}_1)$, habría que calcular el error cuadrático medio de cada uno, y tomar el que tenga menor error cuadrático medio. Pues puede ocurrir que $\hat{\Theta}_2$, aunque sea sesgado, al tener menor varianza tome valores mas cercanos al verdadero parámetro que $\hat{\Theta}_1$



Ejemplo

Supóngase que $\hat{\Theta}_1$, $\hat{\Theta}_2$ y $\hat{\Theta}_3$ son dos estimadores de un parámetro θ, y que $E(\hat{\Theta}_1) = E(\hat{\Theta}_2) = \theta$; $E(\hat{\Theta}_3) \neq \theta$, $V(\hat{\theta}_1) = 10$, $V(\hat{\Theta}_2) = 6$ y $E\left[(\hat{\Theta}_3 - \theta)^2\right] = 4$. Haga una comparación de estos estimadores. ¿Cuál prefiere y por qué?

<u>Solución</u>: Calculamos el error cuadrático medio de cada estimador

$ECM(\hat{\Theta}_1) = V(\hat{\Theta}_1) = 10$ pues $\hat{\Theta}_1$ es insesgado

$ECM(\hat{\Theta}_2) = V(\hat{\Theta}_2) = 6$ pues $\hat{\Theta}_2$ es insesgado

$ECM(\hat{\Theta}_3) = E\left[(\hat{\Theta}_3 - \theta)^2\right] = 4$ es dato

En consecuencia $\hat{\Theta}_3$ es el mejor estimador de los tres dados porque tiene menor error cuadrático medio.

Métodos de estimación puntual

Los criterios anteriores establecen propiedades que es deseable que sean verificadas por los estimadores. Entre dos estimadores posibles para un dado parámetro poblacional es razonable elegir aquél que cumple la mayor cantidad de criterios o alguno en particular que se considera importante para el problema que se esté analizando. Sin embargo estos criterios no nos enseñan por sí mismos a construir los estimadores. Existen una serie de métodos para construir estimadores los cuales en general se basan en principios básicos de razonabilidad. Entre éstos podemos mencionar:

- **Método de los momentos**
- **Método de máxima verosimilitud**

EMV

Supongamos que X es una v.a. discreta con función de distribución de probabilidad $p(x, \theta)$, donde θ es un parámetro desconocido. Sean $x_1, x_2, \ldots, x_n$ los valores observados de una muestra aleatoria de tamaño n.

Se define la función de verosimilitud como la función de distribución conjunta de las observaciones:

$$L(x_1, x_2, \ldots, x_n, \theta) =$$

$$P(X_1 = x_1)\, P(X_2 = x_2) \ldots P(X_n = x_n) =$$

$$p(x_1, \theta) . p(x_2, \theta) \ldots . p(x_n, \theta)$$

Notar que la función de verosimilitud es una función de θ . El estimador de máxima verosimilitud de θ es aquel valor de θ que maximiza la función de verosimilitud

EMV

La interpretación del método sería: el estimador de máxima verosimilitud es aquel valor del parámetro que maximiza la probabilidad de ocurrencia de los valores muestrales

EMV

El tiempo de fallar T de una componente tiene una distribución exponencial con parámetro λ :
T ~ Exp (λ) , es decir la fdp es

$$f(t;\lambda)=\begin{cases}\lambda e^{-\lambda t} & 0\le t<\infty\\ 0 & demás \quad valores\end{cases}$$

Donde E (T) = 1/λ y V (T) = 1 / λ^2 , respectivamente.

Se desea calcular el estimador de máxima verosimilitud del parámetro λ para una muestra de tamaño n.

Solución:

La función de probabilidad es:

$$L(t_1, t_2, \ldots, t_n; \lambda) = f(t_1; \lambda) f(t_2; \lambda) \ldots f(t_n; \lambda) = \left[\lambda e^{-\lambda t_1}\right] \times \left[\lambda e^{-\lambda t_2}\right] \times \ldots \times \left[\lambda e^{-\lambda t_n}\right],$$

que puede escribirse:

$$L(t_1, t_2, \ldots, t_n; \lambda) = (\lambda)^n e^{-\lambda \sum_{i=1}^{n} t_i}$$

Nuevamente tomamos logaritmo natural

$$\ln L(t_1, t_2, \ldots, t_n; \sigma) = n \ln \lambda - \lambda \sum_{i=1}^{n} t_i$$

$$\frac{\partial \ln L(t_1, t_2, \ldots, t_n; \lambda)}{\partial \lambda} = n \frac{1}{\lambda} - \sum_{i=1}^{n} T_i = 0$$

de donde podemos despejar λ :

$\lambda = \dfrac{n}{\sum_{i=1}^{n} t_i} = \bar{t}$, entonces el estimador de λ es $\hat{\lambda} = \dfrac{n}{\sum_{i=1}^{n} T_i}$

Para maximizar la función de verosimilitud y facilitar los cálculos tomamos el logaritmo natural de L
Pues maximizar L es equivalente a maximizar ln(L) y al tomar logaritmos transformamos productos en sumas.

El EMV de mu para la normal

Ejemplo:

La variable aleatoria X tiene distribución $N(\mu, \sigma^2)$ con μ y σ^2 ambos parámetros desconocidos para los cuales se desea encontrar los estimadores máxima verosimilitud. La *fdp* es

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

La función de verosimilitud para una muestra aleatoria de tamaño n es

$$L(x_1, x_2, \ldots, x_n; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x_1-\mu}{\sigma}\right)^2} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x_2-\mu}{\sigma}\right)^2} \ldots \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x_n-\mu}{\sigma}\right)^2} =$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2}$$

$$\ln L\left(x_1, x_2, \ldots, x_n; \mu, \sigma^2\right) = -\frac{n}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2$$

y el sistema de ecuaciones de verosimilitud queda:

$$\begin{cases} \dfrac{\partial \ln L\left(x_1, x_2, \ldots, x_n; \mu, \sigma^2\right)}{\partial \mu} = \displaystyle\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right) = 0 \\ \dfrac{\partial \ln L\left(x_1, x_2, \ldots, x_n; \mu, \sigma^2\right)}{\partial \sigma^2} = -\dfrac{n}{2\sigma^2} + \dfrac{1}{2}\displaystyle\sum_{i=1}^{n}\frac{(x_i - \mu)^2}{\sigma^4} = 0 \end{cases}$$

Resolvemos con respecto a μ y σ^2 :

$$\begin{cases} \mu = \dfrac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \\ \sigma^2 = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \end{cases}$$

Entonces los estimadores máxima verosimilitud de μ y σ^2 son

$$\begin{cases} \hat{\mu} = \dfrac{1}{n}\sum_{i=1}^{n} X_i = \bar{X} \\ \hat{\sigma}^2 = \dfrac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \end{cases}$$

Propiedades de los EMV

Propiedades de los estimadores máxima verosimilitud

1- Los EMV pueden ser ***sesgados***, pero en general si $\hat{\Theta}$ es el EMV de un parámetro θ basado en una muestra de tamaño n, entonces $\lim_{n\to\infty} E(\hat{\Theta}) = \theta$, es decir son ***asintóticamente insesgados***

2- Bajo condiciones bastantes generales se puede probar que los EMV son ***asintóticamente consistentes***

3- Bajo condiciones bastantes generales se puede probar que los EMV ***asintóticamente tienen varianza mínima***

4-Los EMV cumplen la ***propiedad de invarianza*** es decir:
si $\hat{\Theta}$ es un EMV de un parámetro θ, el EMV de $g(\theta)$ es $g(\hat{\Theta})$, si $g(x)$ es una función inyectiva.

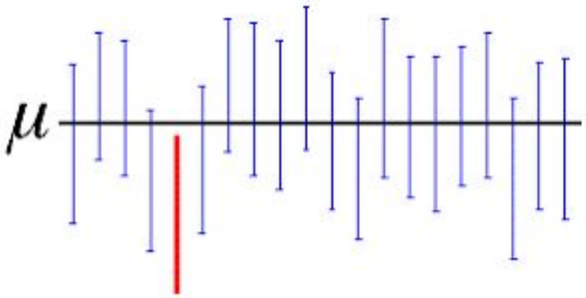
Intervalos de confianza

A veces resulta más conveniente dar un intervalo de valores posibles del parámetro desconocido, de manera tal que dicho intervalo contenga al verdadero parámetro con determinada probabilidad.

Un intervalo de confianza es un rango de valores, derivado de los estadísticos de la muestra, que posiblemente incluya el valor de un parámetro de población desconocido. Debido a su naturaleza aleatoria, es poco probable que dos muestras de una población en particular produzcan intervalos de confianza idénticos

Intervalos de confianza

En este caso, la línea negra horizontal representa el valor fijo de la media desconocida de la población, μ. Los intervalos de confianza azules verticales que se sobreponen a la línea horizontal contienen el valor de la media de la población. El intervalo de confianza rojo que está completamente por debajo de la línea horizontal no lo contiene. Un intervalo de confianza de 95% indica que 19 de 20 muestras (95%) de la misma población producirán intervalos de confianza que contendrán el parámetro de población.



Intervalos de confianza

Específicamente, a partir de una muestra aleatoria se construye un intervalo $(\hat{\theta}_1, \hat{\theta}_2)$ donde los extremos son dos estadísticos tal que

$$P(\theta \in (\hat{\theta}_1, \hat{\theta}_2)) = 1 - \alpha$$

Parámetro desconocido a estimar

es un valor real entre cero y uno dado de antemano

Ejemplo

Por ejemplo si pedimos un **α**=0.05 esto implica que

$$P(\hat{\theta}_1 \leq \theta \leq \hat{\theta}_2) = 0.95$$

Es decir, una probabilidad del 95% que el verdadero parámetro se encuentre en el intervalo propuesto

Al valor 1 - α se lo llama nivel de confianza del intervalo.

Método del pivote

Supongamos el siguiente caso particular, sea $X_1, X_2, ..., X_n$ una muestra aleatoria de tamaño n de una variable aleatoria X donde $X \sim N(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$, $\boldsymbol{\sigma}^2$ conocido, se quiere construir un intervalo de confianza para de nivel $(1 - \boldsymbol{\alpha})$. Supongamos $\boldsymbol{\alpha}=0.05$

1- tomamos un estimador puntual de $\boldsymbol{\mu}$, sabemos que es un estimador con buenas propiedades.

2- a partir de $\bar{X} - \mu$ construimos el estadístico

$$Z = \frac{\bar{X} - \mu}{\sigma - \sqrt{n}}$$

Notar que $Z \sim N(0,1)$

3- como conocemos la distribución de Z, podemos plantear: hallar un número z tal que

$$P(-z \leq Z \leq z) = 0.95$$

Recordemos la estandarización

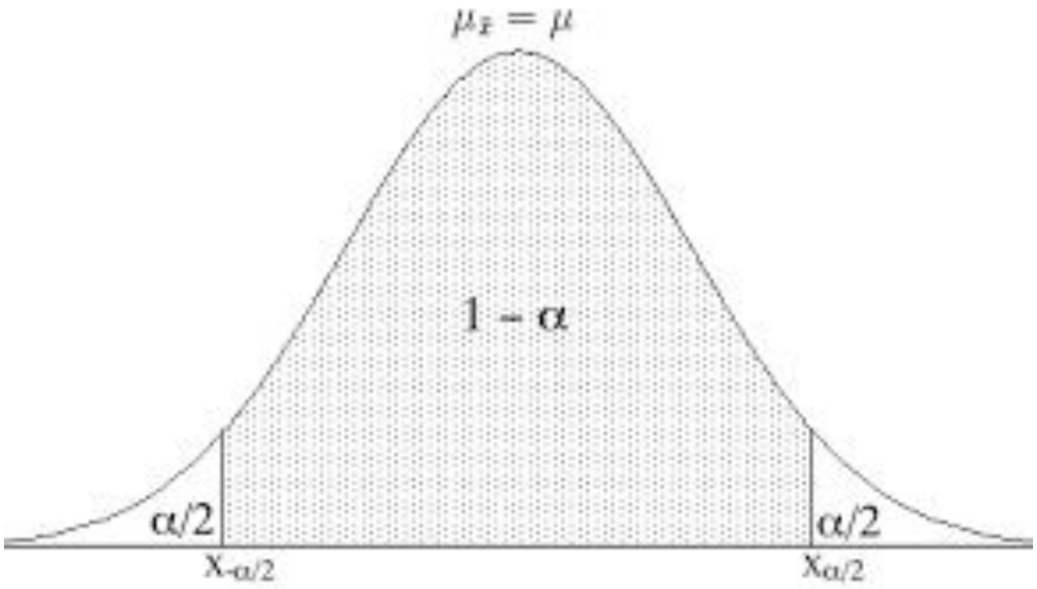
Si las variables originales (antes de aplicar el proceso de estandarización) tenían una distribución normal, el proceso de estandarización no brinda una nueva escala (común) y esta distribución sigue siendo una Normal conocida como distribución Normal estándar.

¿Por qué estandarizamos?

Para poder comparar, por ejemplo, puntuaciones de dos sujetos en distintas distribuciones.

La distribución normal estándar está tabulada (habitualmente en la forma de el valor de la función de distribución Φ) y las otras distribuciones normales pueden obtenerse como transformaciones simples.

Gráficamente



La longitud del intervalo de confianza

Recordando que $\Phi(-z)=1-\Phi(z)$, esta ecuación queda:

$\Phi(z)-\Phi(-z)=2\Phi(z)-1=1\text{-}\alpha$, o bien (ver figura anterior),

$\Phi(z)=1-\dfrac{\alpha}{2}$ o de otra forma $P(Z>z)=\dfrac{\alpha}{2}$.

Al valor de z que verifica esta ecuación se lo suele indicar $z_{\frac{\alpha}{2}}$. En consecuencia, el intervalo de confianza bilateral al nivel de significación $1\text{-}\alpha$ queda:

$$\left[\hat{\Theta}_1,\hat{\Theta}_2\right]=\left[\overline{X}-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},\overline{X}+z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$$

En consecuencia:

> Si $(X_1,X_2,...,X_n)$ una muestra aleatoria de tamaño n de una v.a. X donde $X\sim N(\mu,\sigma^2)$, σ^2 conocido, un intervalo de confianza para μ de nivel $1-\alpha$ es
>
> $$\left[\overline{X}-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},\overline{X}+z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right] \qquad (8.1)$$

Ejemplo

Un ingeniero civil analiza la resistencia a la compresión del concreto. La resistencia está distribuida aproximadamente de manera normal, con varianza 1000 (psi) 2 . Al tomar una muestra aleatoria de 12 especímenes, se tiene que x 3250 psi.

a) Construya un intervalo de confianza del 95% para la resistencia a la compresión promedio.

b) Construya un intervalo de confianza del 99% para la resistencia a la compresión promedio. Compare el ancho de este intervalo de confianza con el ancho encontrado en el inciso a).

Solución

Solución:

La v. a. de interés es X_i: "resistencia a la compresión del concreto en un espécimen i"

Tenemos una muestra de $n = 12$ especímenes.

Asumimos que $X_i \sim N(\mu, \sigma^2)$ para $i = 1,2,3,\ldots,12$ con $\sigma^2 = 1000$

a) Queremos un intervalo de confianza para μ de nivel 95%. Por lo tanto $\alpha = 0.05$

El intervalo a utilizar es $\left[\overline{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right]$.

Buscamos en la tabla de la normal estándar el valor de $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$

Reemplazando:

$$\left[3250 - 1.96 \times \frac{\sqrt{1000}}{\sqrt{12}}, 3250 + 1.96 \times \frac{\sqrt{1000}}{\sqrt{12}}\right] = \left[3232.10773, \quad 3267.89227\right]$$

b) repetimos lo anterior pero ahora $\alpha = 0.01$

La longitud del intervalo

Notar que la seguridad de que el verdadero parámetro se encuentre en el intervalo hallado es mayor en el intervalo b) que en el a), pero la longitud del intervalo b) es mayor que la del intervalo a).

Al aumentar el nivel de confianza se perdió precisión en la estimación, ya que a menor longitud hay mayor precisión en la estimación.

En general la longitud del intervalo es

$$L = 2z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Propiedades de la longitud

Notar que:

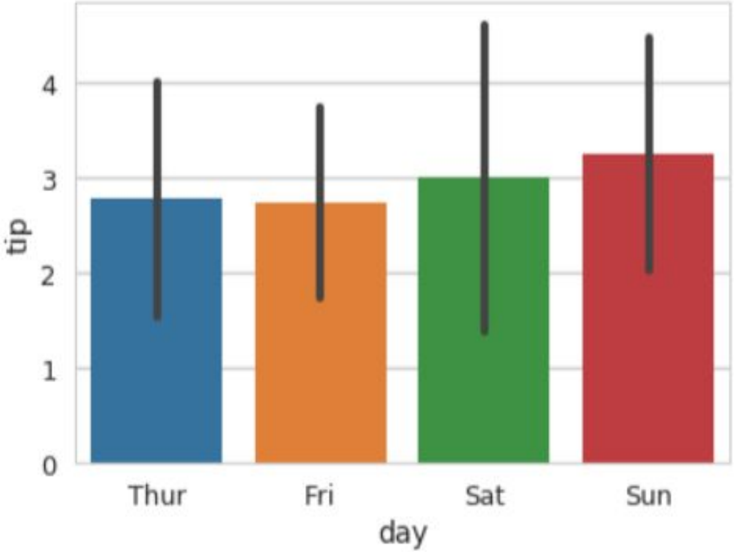
a) si n y σ están fijos, a medida que disminuye tenemos que $z_{\alpha/2}$ aumenta, por lo tanto L aumenta.

b) si α y σ están fijos, entonces a medida que n aumenta tenemos que L disminuye.

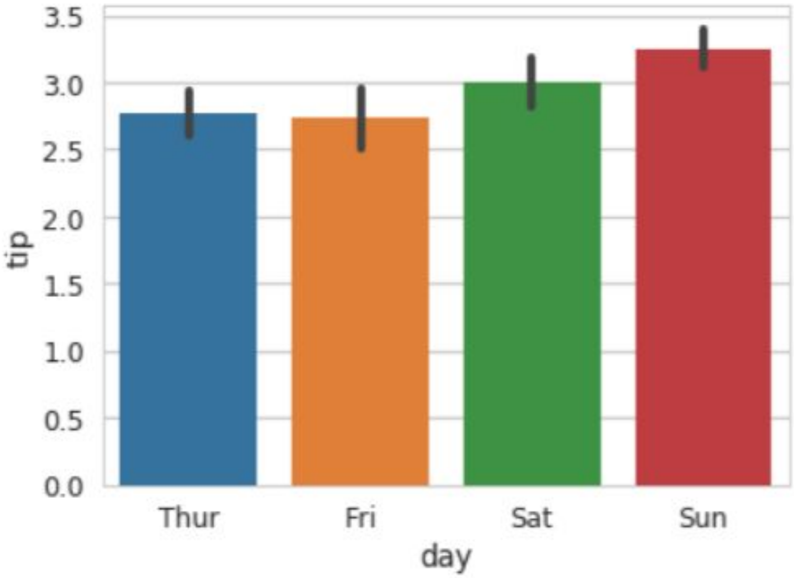
Podemos plantearnos la siguiente pregunta relacionada con el ejemplo anterior: ¿qué tamaño n demuestra se necesita para que el intervalo tenga nivel de confianza 95% y longitud la mitad de la longitud del intervalo hallado en a)?

La longitud del intervalo

```
ax = sns.barplot(x="day", y="tip", data=tips, ci="sd")
```




```
ax = sns.barplot(x="day", y="tip", data=tips, ci=68)
```



Cuántas muestras tomo?

En general, si queremos hallar n tal que $L = 2z_{\frac{\alpha}{2}} \dfrac{\sigma}{\sqrt{n}} \leq l$, donde l es un valor dado, entonces

despejando n

$$n \geq \left(\frac{2z_{\frac{\alpha}{2}}\sigma}{l} \right)^2$$

Intervalo de confianza para la media de una distribución normal, varianza desconocida

Nuevamente vamos a estimar el parámetro μ con el estimador de esperanza muestral que utilizamos anteriormente. Pero ahora no podemos usar el pivote Z, porque nos pedían como condición conocer la distribución y el parámetro σ.

Es entonces que vamos a sustituir a σ por un estimador conocido

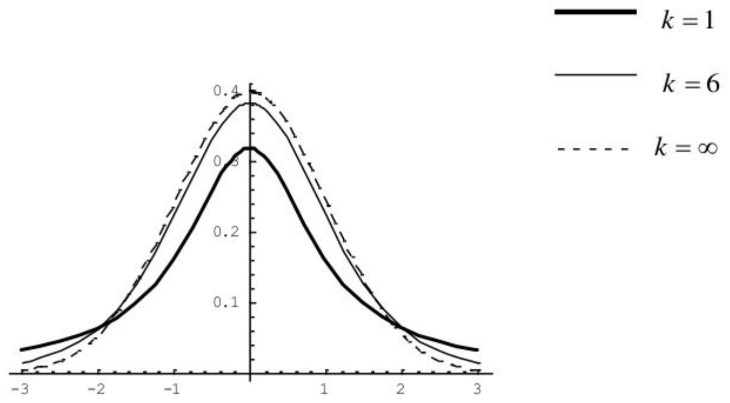
$$S^2 = \frac{1}{n-1}\sum_{1=1}^{n}\left(X_i - \bar{X}\right)^2$$

y proponemos como pivote a la variable aleatoria

$$T = \frac{X - \mu}{S/\sqrt{n}}.$$

La t de Student

Pero para poder usar a T como pivote debemos conocer su distribución.
Se puede probar que la distribución de T es una distribución llamada Student con parámetro n-1.
La gráfica de la f.d.p. de la distribución Student tiene forma de campana como la normal, pero tiende a cero más lentamente. Se puede probar que cuando k   la fdp de la Student tiende a la fdp de la N (0 , 1) .



Los grados de libertad

El parámetro ***n-1*** se conoce como los grados de libertad.

“Se definen como el número de valor es que podemos escoger 5 libremente” . (Levin 1996, p388)

“Los grados de libertad de una prueba estadística son el número de datos que son libres de variar cuando se calcula tal prueba

(Pagano 2009,p321)

Ejemplo de grados de libertad

Un ejemplo ilustrativo podría ser el popular juego japonés Sudoku, en el cual no se puede repetir el mismo número en la misma columna o fila. Si tomamos una columna, y se nos indica que ubiquemos los números del 1 al 9, entonces podemos elegir libremente los primeros 8 números, dado que el noveno número no puede ser elegido por nosotros, quedando restringido por los 8 números anteriores. El primer número puede ser cualquiera de los nueve, el segundo debe ser elegido de entre los ocho restantes, el tercero debe ser escogido de entre los siete que quedan, y así hasta que solo nos quede elegir entre dos números, tras lo cual, solo nos quedará un último y único número, el cual ya no puede ser elegido libremente. Por lo tanto, tenemos 8 grados de libertad

5	3			7				
6			1	9	5			
	9	8					6	
8				6				3
4			8		3			1
7				2				6
	6					2	8	
			4	1	9			5
				8			7	9

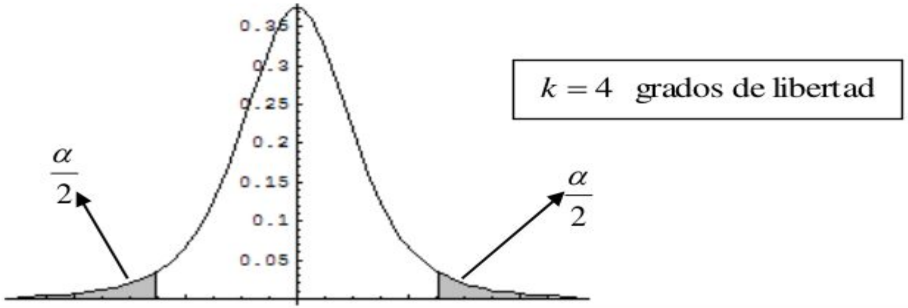
Construimos el intervalo

$$P\left(\overline{X}-t\frac{S}{\sqrt{n}}\le\mu\le\overline{X}+t\frac{S}{\sqrt{n}}\right)=1\text{-}\alpha$$

Evidentemente, si definimos

$\begin{cases}\hat{\Theta}_1=\overline{X}-t\dfrac{S}{\sqrt{n}}\\ \hat{\Theta}_2=\overline{X}+t\dfrac{S}{\sqrt{n}}\end{cases}$, hemos construido dos estadísticos $\hat{\Theta}_1$ y $\hat{\Theta}_2$ tales que $P\left(\hat{\Theta}_1\le\mu\le\hat{\Theta}_2\right)=1\text{-}\alpha$,

veamos quien es el número *t* que verifica la ecuación, es decir (ver figura):



Construcción del intervalo de confianza

Si $(X_1, X_2, ..., X_n)$ una muestra aleatoria de tamaño n de una v.a. X donde $X \sim N(\mu, \sigma^2)$, σ^2 desconocido, un intervalo de confianza para μ de nivel $1-\alpha$ es

$$\left[\overline{X} - t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \overline{X} + t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right] \qquad (8.2)$$

Ejemplo de un caso

Ejemplo:

Se hicieron 10 mediciones sobre la resistencia de cierto tipo de alambre que dieron valores $x_1, x_2, \ldots, x_{10}$ tales que $\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = 10.48$ ohms y $S = \sqrt{\frac{1}{9}\sum_{i=!}^{10}(x_i - \bar{x})^2} = 1.36$ ohms. Supóngase que $X \sim N(\mu, \sigma^2)$.

Se desea obtener un intervalo de confianza para la esperanza poblacional μ al 90 %.

Tenemos que $1-\alpha = 0.90 \rightarrow \alpha = 0.1 \rightarrow \alpha/2 = 0.05$

De la Tabla de la t de Student tenemos que $t_{0.05,9} = 1.8331$. Entonces el intervalo de confianza buscado es:

$$\left[\bar{X} - t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}, \bar{X} + t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}\right] = \left[10.48 - 1.8331\frac{1.36}{\sqrt{10}}, 10.48 + 1.8331\frac{1.36}{\sqrt{10}}\right]$$

Esto es: $[9.69,\ 11.27]$.

Intervalo de confianza para la varianza de una distribución normal

Supongamos que se quiere hallar un intervalo de confianza para la varianza 2 de una distribución normal.
Sea X 1 , X 2 ,..., X n una muestra aleatoria de una v.a. X, donde X~N(μ, σ^2).

Tomamos como estimador puntual de σ^2 a

$$S^2 = \frac{1}{n-1}\sum_{1=1}^{n}\left(X_i - \bar{X}\right)^2$$

Nuestro estadístico será

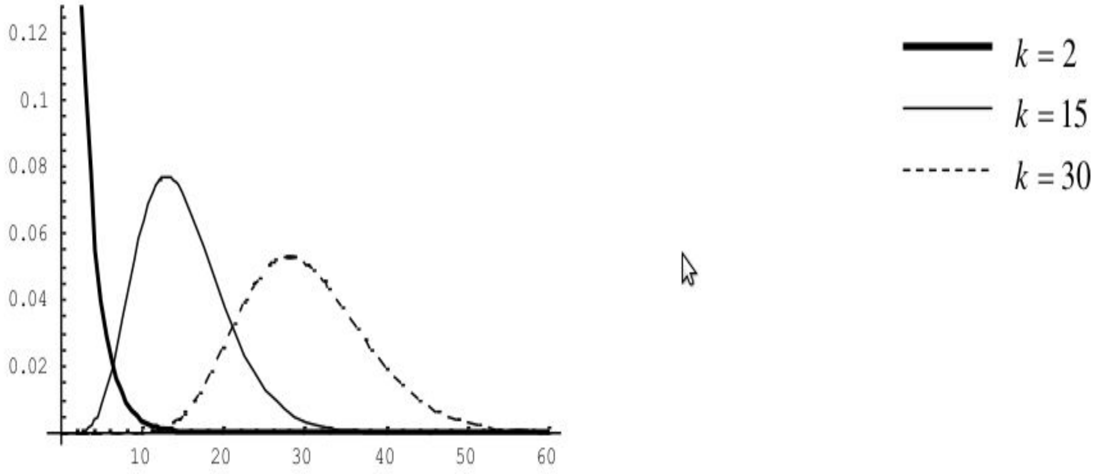
$$X = \frac{(n-1)S^2}{\sigma^2}$$

La ji-cuadrada

Este estadístico contiene al parámetro desconocido σ^2 a estimar y tiene una distribución conocida, se puede probar que X tiene una distribución llamada **ji-cuadrado** con n-1 grados de libertad.

Notación: $X \sim \chi_k^2$

La distribución ji-cuadrdo es asimétrica. En la figura siguiente se grafica la densidad para diferentes valores de k



Propiedades de la ji cuadrada

Propiedades de la distribución Chi-cuadrado

1. La χ^2 no toma valores negativos, es cero o positiva.
2. La χ^2 no es simétrica; está sesgada hacía la derecha, con la condición de que a mayor tamaño de muestra (mayor número de grados de libertad), se hará menos sesgada y tiende a la normalidad.
3. La media de la distribución χ^2 está dada por los grados de libertad ($\mu = \upsilon$) así como la varianza está determinada por el doble de los grados de libertad ($\sigma^2 = 2\upsilon$.

Ejemplo

Un fabricante de detergente líquido está interesado en la uniformidad de la máquina utilizada para llenar las botellas. De manera específica, es deseable que la desviación estándar del proceso de llenado sea menor que 0.15 onzas de líquido; de otro modo, existe un porcentaje mayor del deseable de botellas con un contenido menor de detergente. Supongamos que la distribución del volumen de llenado es aproximadamente normal. Al tomar una muestra aleatoria de 20 botellas, se obtiene una varianza muestral S^2 0 . 0153 . Hallar un intervalo de confianza de nivel 0.95 para la verdadera varianza del volumen de llenado.

Solución

Solución:

La v.a. de interés es X: " volumen de llenado de una botella"

Se asume que $X \sim N(\mu, \sigma^2)$ con σ desconocido.

Estamos en las condiciones para aplicar (8.8)

Tenemos que $1-\alpha = 0.95 \rightarrow \alpha = 0.05 \rightarrow \chi^2_{1-\frac{\alpha}{2}, n-1} = \chi^2_{0.975,19} = 8.91$ y $\chi^2_{\frac{\alpha}{2}, n-1} = \chi^2_{0.025,19} = 32.85$

Además $S^2 = 0.0153$

Por lo tanto el intervalo es

$$\left(\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}, n-1}}; \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}, n-1}} \right) = \left(\frac{(20-1)\times 0.0153}{32.85}; \frac{(20-1)\times 0.0153}{8.91} \right) = (0.00884;\ 0.0326)$$

Y un intervalo para σ es $\left(\sqrt{0.00884};\ \sqrt{0.0326}\right) = (0.09;\ 0.1805)$

Por lo tanto con un nivel de 0.95 los datos ***no apoyan la afirmación que*** $\sigma < 0.15$

Conclusión intervalo de confianza

La estimación de los parámetros de la población a través del muestreo es una forma simple pero poderosa de inferencia. Las estimaciones de parámetros combinadas con los márgenes de error nos permiten crear intervalos de confianza que capturan el parámetro de población real con alta probabilidad.

Test o prueba de hipótesis

En muchos problemas se requiere tomar una decisión entre aceptar o rechazar una proposición sobre algún parámetro. Esta proposición recibe el nombre de hipótesis estadística, y el procedimiento de toma de decisión sobre la hipótesis se conoce como prueba o test de hipótesis.

Como se emplean distribuciones de probabilidad para representar poblaciones, también podemos decir que una hipótesis estadística es una proposición sobre la distribución de probabilidad de una variable aleatoria, donde la hipótesis involucra a uno más parámetros de esta distribución.

¿Qué son los test?

Una hipótesis estadística es una afirmación respecto a alguno de los parámetros desconocidos poblacionales.

Ejemplos

- El legislador X asegura que el sueldo promedio es de $18000
- La proporción de ciudadanos que cree que la inflación va a ser del 15 % es del 65%

Vamos a usar información muestral para validar estas evidencias empíricas

Conceptos básicos

Quienes son H_0 y H_1? Son quienes van a formular la afirmación de mi hipótesis en un escenario dicotómico. Es decir el caso en el que ocurra o no la hipótesis que se está planteando.

H_0 se conoce como hipótesis nula
H_1 se conoce como hipótesis alternativa

A partir de la muestra vamos a construir un estadístico (recordemos que es una función en base a la información muestral) para afirmar o rechazar con determinado nivel de significancia la H_0. Se supone que la evidencia empírica me dará las suficientes razones para aceptar o rechazar H_0.

Tipos de errores

	Decisión respecto a la H_0	
	Rechazo H_0	No rechazo H_0
V	α Error de tipo I Falso positivo	$(1-\alpha)$ Decisión correcta
F	$(1-\beta)$ Potencia del test Decisión correcta	β Error de tipo II Falso negativo

Ambos errores no pueden darse simultáneamente, porque habremos de rechazar o no H_0.

Para ciertos valores del estadístico la decisión será rechazar H_0. Estos valores se conocen como los valores críticos y determinan una región crítica o de rechazo y una zona de aceptación. El valor crítico será aquel que expongamos como límite en el análisis. En los ejemplos vistos los $18000 pesos de salario promedio o el 65% de aceptación del estimado de inflación anual.

Estructura de un test de hipótesis

1. Formular la H_0 a contrastar
2. Establecer el nivel de significancia del test
3. Cálculo del estadístico de contraste
4. Regla de decisión
5. Conclusión

Construyendo reglas de decisión

En la mayoría de las situaciones la hipótesis nula dice que el efecto que indica la muestra es atribuible solamente a la variación aleatoria del estadístico de prueba. La hipótesis alternativa establece que el efecto que indica la muestra es verdadero.

Por ejemplo

$H_0 : \mu \geq 18000$ contra $H_1 : \mu < 18000$ (hipótesis alternativa unilateral)

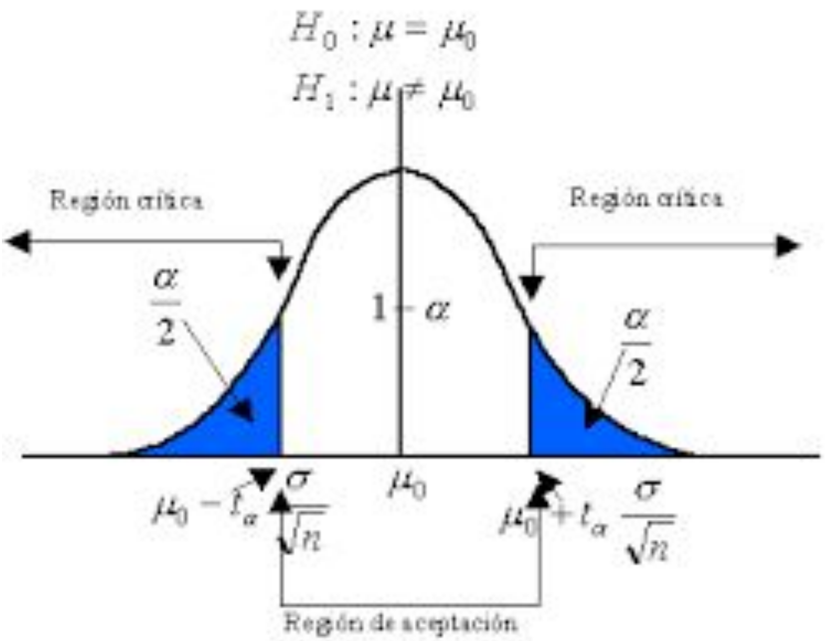
ó

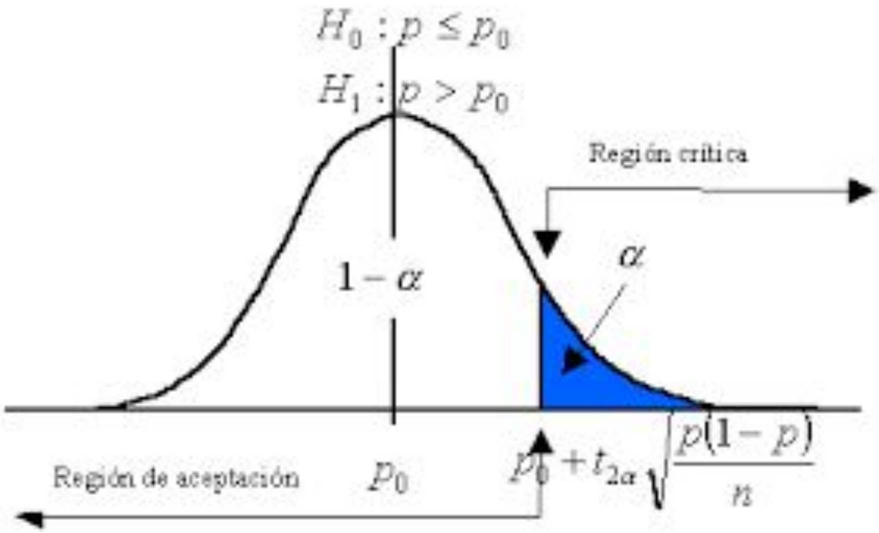
$H\ 0 : \mu \leq 18000$ contra $H\ 1 : \mu > 18000$ (hipótesis alternativa unilateral)

ó

$H\ 0 : \mu = 18000$ contra $H\ 1 : \mu \neq 18000$ (hipótesis alternativa bilateral)

Test de 1 cola y de 2 colas





Como podemos manejar los errores?

Debido a que la decisión se basa en variables aleatorias es posible asociar probabilidades a los errores de tipo I y II, específicamente anotamos

$$\alpha = P\ (\text{ error de tipo I })$$

$$\beta = P\ (\text{ error de tipo II })$$

A $\alpha = P$ (error de tipo I) se lo conoce como nivel de significancia del test .
Para calcular estas probabilidades debemos conocer la distribución del estadístico de prueba en el caso de ser H_0 verdadera, es decir debemos conocer la distribución del estadístico de prueba “bajo H_0”

Revisemos con un ejemplo

ejemplo

Por ejemplo, supongamos que cierto tipo de motor de automóvil emite una media de 100 mg de óxidos de nitrógeno (NO x) por segundo con 100 caballos de fuerza. Se ha propuesto una modificación al diseño del motor para reducir las emisiones de NO x . El nuevo diseño se producirá si se demuestra que la media de su tasa de emisiones es menor de 100 mg/s. Se construye y se prueba una muestra de 50 motores modificados. La media muestral de emisiones de NO x es de 92 mg/s, y la desviación estándar muestral es de 21 mg/s.
La variable aleatoria de interés en este caso es
X: “tasa de emisión de un motor modificado tomado al azar”.

Ejemplo

La preocupación de los fabricantes consiste en que los motores modificados no puedan reducir todas las emisiones; es decir que la media poblacional pudiera ser 100 o mayor que 100.

Entonces, la pregunta es: ¿es factible que esta muestra pueda provenir de una v.a. con media 100 o mayor?

Solución

Se ha observado una muestra con media X = 92 .
Hay dos interpretaciones posibles de esta observación:
1- La media poblacional es realmente mayor o igual que 100, y la media muestral es menor que 100 debido a la variabilidad propia de la variable aleatoria X $\rightarrow$ H_0

2- La media poblacional es en realidad menor que 100, y la media muestral refleja este hecho. $\rightarrow$ H_1

Que nos dicen H_0 y H_1

En la mayoría de las situaciones la hipótesis nula dice que el efecto que indica la muestra es atribuible solamente a la variación aleatoria del estadístico de prueba.
La hipótesis alternativa establece que el efecto que indica la muestra es verdadero.

Analicemos lo errores

Para calcular estas probabilidades debemos conocer la distribución del estadístico de prueba en el caso de ser H_0 verdadera, es decir debemos conocer la distribución del estadístico de prueba “bajo H_0 ”.
Como la muestra es grande, ya sabemos que por T.C.L. el estadístico

$$Z = \frac{\overline{X} - 100}{s/\sqrt{n}} \approx N(0,1) \quad \textbf{\textit{si}} \quad H_0 \ \textbf{\textit{es verdadera,}} \ \text{o sea} \quad Z = \frac{\overline{X} - 100}{21/\sqrt{50}} \approx N(0,1)$$

Regla de decisión

Y vamos a construir la regla de decisión. Por ahora mostraremos esta y más adelante hablaremos de construir una nueva

Por ahora tengamos la siguiente:

- se rechaza H_0 si $X < 95$
- se acepta H_0 si $X \geq 95$

El intervalo $95 , \infty$ es la zona de aceptación.
La región $-\infty ; 95$ es la zona de rechazo o región crítica.
Mientras que 95 es el punto crítico .

Cálculo del error de tipo I

Entonces para calcular α planteamos:

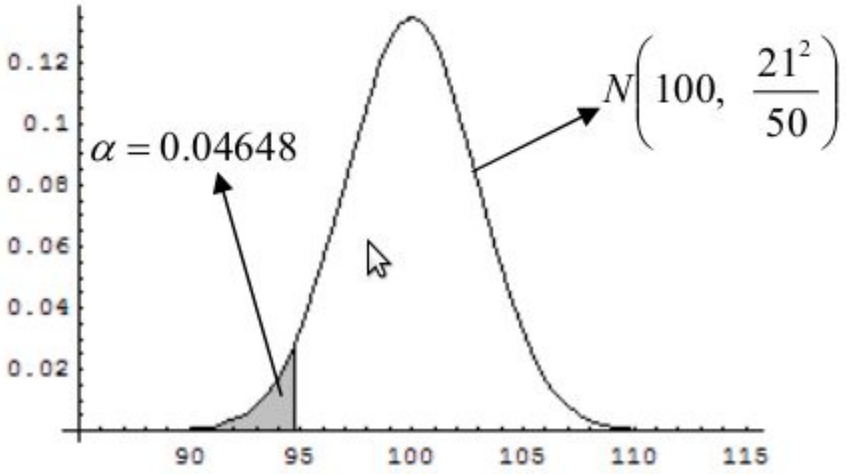
$$\alpha = P(error \quad de \quad tipo \quad I) = P\left(rechazar \quad H_0 / H_0 \quad es \quad V \right) = P\left(\bar{X} < 95 / \mu = 100\right) =$$

$$= P\left(\frac{\bar{X} - 100}{21/\sqrt{50}} < \frac{95 - 100}{21/\sqrt{50}} \right) \approx \Phi\left(\frac{95 - 100}{21/\sqrt{50}} \right) = \Phi(-1.6835) = 1 - 0.95352 = 0.04648$$

Esto significa que el 4.64% de las muestras aleatorias conducirán al rechazo de la hipótesis H_0 : μ ≥ 100 cuando el verdadero μ sea mayor o igual que 100.

Graficamente

En este caso el gráfico de la zona de rechazo es



Del gráfico vemos que podemos reducir α al aumentar la zona de aceptación. Por ejemplo supongamos que ahora la regla de decisión es

se rechaza H 0 si X < 93
se acepta H 0 si X ≥ 93

Recalculamos α

$$\text{Entonces} \quad \alpha = P(error \quad de \quad tipo \quad I) = P\left(rechazar \quad H_0 / H_0 \quad es \quad V\right) = P\left(\overline{X} < 93 / \mu = 100\right) =$$

$$= P\left(\frac{\overline{X} - 100}{21 / \sqrt{50}} < \frac{93 - 100}{21 / \sqrt{50}}\right) \approx \Phi\left(\frac{93 - 100}{21 / \sqrt{50}}\right) = \Phi(-2.357) = 1 - 0.99061 = 0.00939$$

También se puede reducir α aumentando el tamaño de la muestra.
Supongamos que n = 85 , entonces

$$= P\left(\frac{\overline{X} - 100}{21 / \sqrt{85}} < \frac{95 - 100}{21 / \sqrt{85}}\right) \approx \Phi\left(\frac{95 - 100}{21 / \sqrt{85}}\right) = \Phi(-2.195) = 1 - 0.98574 = 0.01426$$

Pasemos a β

$\beta = P$ (error de tipo II) $= P$ (aceptar H_0 / H_0 es falsa)

Pero en este caso para llegar a un valor numérico necesitamos tener una alternativa específica pues en nuestro ejemplo:
$\beta = P$ (error de tipo II) $= P$ (aceptar H_0 / H_0 es falsa) $= P\,(\, X \geq 95 \,/\, \mu \neq 100 \,) =$

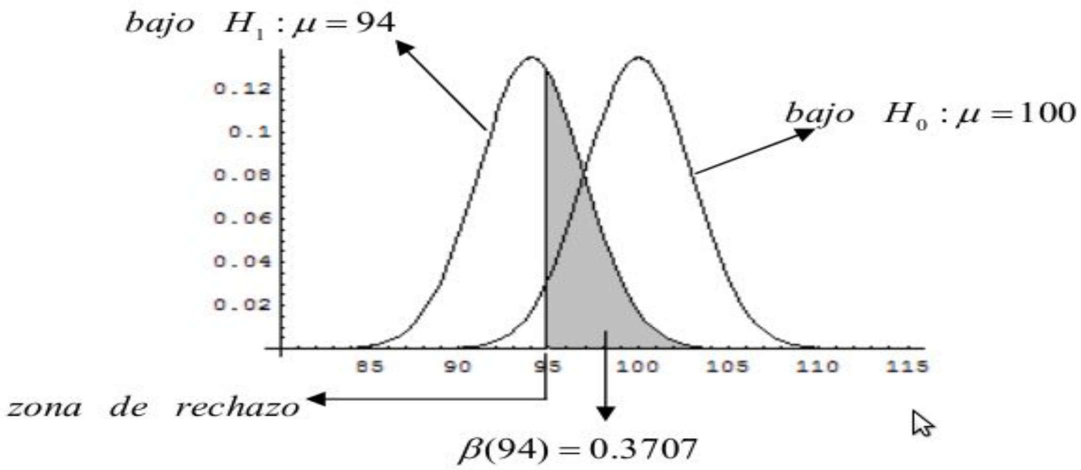
$$= P\left(\frac{\overline{X} - \mu}{21 \big/ \sqrt{50}} \geq \frac{95 - \mu}{21 \big/ \sqrt{50}} \right) = 1 - \Phi\left(\frac{95 - \mu}{21 \big/ \sqrt{50}} \right) = \beta(\mu)$$

Donde anotamos con μ a la verdadera media poblacional desconocida .
Podemos entonces calcular β para un valor particular de μ , por ejemplo nos puede interesar como se comporta el test cuando la verdadera media es $\mu = 94$

Cálculo de β

$$\beta(94) = P\left(\frac{\overline{X}-94}{21/\sqrt{50}} \geq \frac{95-94}{21/\sqrt{50}}\right) = 1-\Phi\left(\frac{95-94}{21/\sqrt{50}}\right) = 1-\Phi(0.3367) = 1-0.62930 = 0.3707$$

Gráficamente:



La probabilidad β de cometer error de tipo II crece a medida que el valor verdadero de μ se acerca al valor hipotético

Conclusiones sobre los errores

- El tamaño de la región crítica, y en consecuencia la probabilidad α de cometer error de tipo I, siempre pueden reducirse mediante una selección apropiada de los valores críticos.
- Los errores tipo I y II están relacionados. Una disminución en la probabilidad en un tipo de error siempre da como resultado un aumento en la probabilidad del otro, siempre que el tamaño de la muestra no cambie.
- En general, un aumento en el tamaño de la muestra reduce tanto a α como a β, siempre que los valores críticos se mantengan constantes.
- Cuando la hipótesis nula es falsa, β aumenta a medida que el valor verdadero del parámetro tiende al valor hipotético propuesto por la hipótesis nula. El valor de β disminuye a medida que aumenta la diferencia entre el verdadero valor medio y el propuesto.

Conclusiones sobre los errores

En general el investigador controla la probabilidad α del error de tipo I cuando selecciona los valores críticos. Por lo tanto el rechazo de la hipótesis nula de manera errónea se puede fijar de antemano. Eso hace que rechazar la hipótesis nula sea una ***conclusión fuerte*** . La probabilidad β de error de tipo II no es constante, sino que depende del valor verdadero del parámetro. También depende β del tamaño de la muestra que se haya seleccionado. Como β está en función del tamaño de la muestra y del valor verdadero del parámetro, la decisión de aceptar la hipótesis nula se la considera una ***conclusión débil*** , a menos que se sepa que β es aceptablemente pequeño. ***Por lo tanto cuando se acepta H_0 en realidad se es incapaz de rechazar H_0 . No se puede rechazar H_0 pues no hay evidencia en contra H_0 .***

Prueba de hipótesis sobre la media, varianza conocida

Veamos ahora cómo construir una regla de decisión sobre la media de una población.
Supongamos que la variable aleatoria de interés X tiene una media μ y una varianza σ^2 conocida.
Asumimos que X tiene distribución normal, es decir $X \sim N(\mu, \sigma 2)$.

Supongamos que tenemos las hipótesis

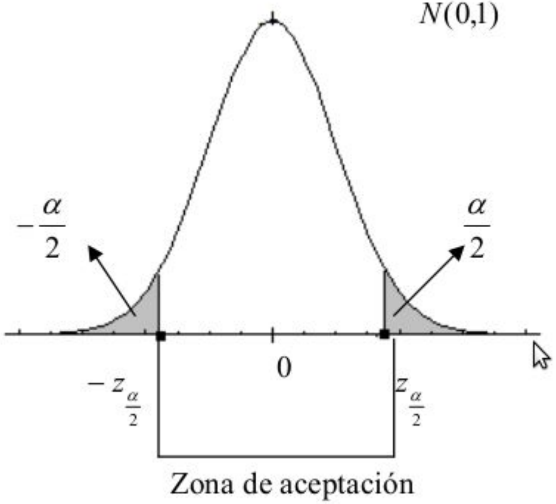
$$H_0 : \mu = \mu_0$$

contra

$$H_1 : \mu \neq \mu_0$$

Prueba de hipótesis sobre la media, varianza conocida

Ya vimos que este tipo de regla de decisión se denomina **bilateral** o de **2 colas** y podemos visualizarla de la siguiente manera



Con los datos que nos brinda la situación tomaremos como estadístico de prueba el pivote Z (porque conocemos la varianza poblacional y ademas se corresponde con una distribución normal)

Prueba de hipótesis sobre la media, varianza conocida

Es evidente que una muestra que produce un valor del estadístico de prueba que cae en las colas de la distribución de Z será inusual si $H_0 : \mu = \mu_0$ es verdadera, por lo tanto esto es un indicador que H_0 es falsa.
Entonces la regla de decisión es:

$$\begin{cases} rechazar \quad H_0 \quad si \quad Z > z_{\frac{\alpha}{2}} \\ aceptar \quad H_0 \quad si \quad |Z| \leq z_{\frac{\alpha}{2}} \end{cases}$$

Ejemplo

El porcentaje deseado de SiO_2 en cierto tipo de cemento aluminoso es 5.5. Para probar si el verdadero promedio de porcentaje es 5.5 para una planta de producción en particular, se analizaron 16 muestras obtenidas de manera independiente. Supongamos que el porcentaje de SiO_2 en una muestra está normalmente distribuido con $\sigma = 3$, y que $\bar{x} = 5.25$.
¿Indica esto de manera concluyente que el verdadero promedio de porcentaje ***difiere*** de 5.5?. Utilice $\alpha = 0.01$

Solución

La v.a. de interés es X: “porcentaje de SiO_2 en cierto tipo de cemento aluminoso”

Asumimos que $X \sim N(\mu, 3^2)$

Podemos plantear las hipótesis

$$H_0: \mu = 5.5$$

contra

$$H_1: \mu \neq 5.5$$

Tenemos una muestra de tamaño n = 16 que dio un promedio muestral x = 5.25

Como $\alpha = 0.01$ entonces $z_{\alpha/2} = z_{0.005} = 2.575$

Revisemos quien es $z_{0,05}$

Fda de la normal

Normal Deviate z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-4.0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.9	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.8	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.7	.0001	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
-3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048

Solución analítica

Por lo tanto la regla de decisión es
$$\begin{cases} \text{rechazar } H_0 \text{ si } \left|\dfrac{\bar{X}-5.5}{3/\sqrt{16}}\right| > 2.575 \\[2ex] \text{aceptar } H_0 \text{ si } \left|\dfrac{\bar{X}-5.5}{3/\sqrt{16}}\right| \leq 2.575 \end{cases}$$

El estadístico $\left|\dfrac{\bar{X}-5.5}{3/\sqrt{16}}\right|$ toma el valor $z_0 = \left|\dfrac{5.25-5.5}{3/\sqrt{16}}\right| = 0.333333$

Como $z_0 = 0.333333 < 2.575 = z_{\frac{0.01}{2}}$ se acepta H_0

p-valor

Hasta ahora se dieron los resultados de una prueba de hipótesis estableciendo si la hipótesis nula fue o no rechazada con un valor especificado de α o nivel de significancia. A menudo este planteamiento resulta inadecuado, ya que no proporciona ninguna idea sobre si el valor calculado del estadístico está apenas en la región de rechazo o bien ubicado dentro de ella.

Para evitar estas dificultades, se adopta el enfoque del p-valor . El valor p o p-valor es la probabilidad que el estadístico de prueba tome un valor que sea al menos tan extremo como el valor observado del estadístico de prueba cuando la hipótesis nula es verdadera. Es así como el p-valor da mucha información sobre el peso de la evidencia contra H_0 , de modo que el investigador pueda llegar a una conclusión para cualquier nivel de significancia especificado.

≤

El valor p es el nivel de significancia más pequeño que conduce al rechazo de la hipótesis nula H_0

A menor p-value > evidencia contra H_0

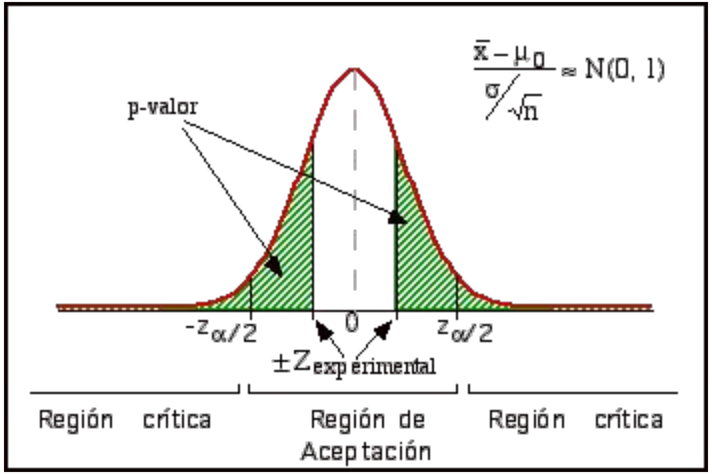
P-valor para unilaterales

Para las pruebas de distribuciones normales presentadas hasta el momento, es sencillo calcular el p-valor.
Si z_0 es el valor calculado del estadístico de prueba Z , entonces el p-valor es

a) si las hipótesis son $H_0 : \mu = \mu_0$ contra $H\,1 : \mu \neq \mu_0$

$$p - \text{valor} = P\,(\,Z > z_0) = 1 - P\,(\,Z < z_0\,) =$$
$$1 - [\,\Phi\,(\,z_0\,) - \Phi\,(\,-\,z_0\,)\,] =$$
$$1 - [\,2\,\Phi\,(\,z_0\,) - 1\,] =$$
$$2\,[\,1 - \Phi\,(\,z_0\,)\,]$$

La unilateral



Este gráfico nos muestra que el p-valor es más grande que el nivel de significancia, lo que nos lleva a decidir que no podemos rechazar la $H_{0,}$ por ser el p-valor > α.
En este caso si no pudiésemos llegar a mejorar nuestra muestra la decisión que podemos tomar es tener un mayor nivel de tolerancia recalculando nuestro nivel de significancia con el valor de p-value

Las bilaterales

b) si las hipótesis son

H_0: $\mu = \mu_0$
Contra
$H\,1 : \mu > \mu_0$

$p - \text{valor} = P\,(\,Z > z_0\,) =$
$1 - P\,(\,Z \leq z_0\,) =$
$1 - \Phi\,(\,z_0\,)$

c)si las hipótesis son

H_0: $\mu = \mu_0$
Contra
$H\,1 : \mu < \mu_0$
$p - \text{valor} =$
$P\,(\,Z < z_{\,0}\,) =$
$\Phi\,(\,z_{\,0}\,)$

Relación entre test de hipótesis e intervalos de confianza

Existe una estrecha relación entre la prueba de hipótesis bilateral sobre un parámetro μ y el intervalo de confianza de nivel $1 - \alpha$ para μ . Específicamente supongamos que tenemos las hipótesis

$$H_0 : \mu = \mu_0 \text{ contra } H1 : \mu \neq \mu_0$$

La regla de decisión es

$$\begin{cases} \text{rechazar } H_0 \text{ si } \left| \dfrac{\overline{X} - \mu_0}{\sigma / \sqrt{n}} \right| > z_{\frac{\alpha}{2}} \\ \text{aceptar } H_0 \text{ si } \left| \dfrac{\overline{X} - \mu_0}{\sigma / \sqrt{n}} \right| \leq z_{\frac{\alpha}{2}} \end{cases}$$

Relación entre test de hipótesis e intervalos de confianza

Aceptar H_0 si $\left|\dfrac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}\right| \le z_{\frac{\alpha}{2}}$ es equivalente a: aceptar H_0 si $-z_{\frac{\alpha}{2}} \le \dfrac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \le z_{\frac{\alpha}{2}}$; y esto es a su vez equivalente, despejando μ_0, a:

aceptar H_0 si $\overline{X} - z_{\alpha/2}\,\sigma/\sqrt{n} \le \mu_0 \le \overline{X} + z_{\alpha/2}\,\sigma/\sqrt{n}$; es decir si

$$\mu_0 \in \left[\overline{X} - z_{\alpha/2}\,\sigma/\sqrt{n} ; \overline{X} + z_{\alpha/2}\,\sigma/\sqrt{n}\right]$$

Pero resulta que $\left[\overline{X} - z_{\alpha/2}\,\sigma/\sqrt{n} ; \overline{X} + z_{\alpha/2}\,\sigma/\sqrt{n}\right]$ es el intervalo de confianza que se construiría para el verdadero parámetro μ de nivel $1-\alpha$.

Ejemplo de la relación

Ejemplo:
En el ejemplo referido al porcentaje deseado de SiO_2 en cierto tipo de cemento aluminoso las hipótesis eran:

$$H_0 : \mu = 5.5 \text{ contra } H_1 : \mu \neq 5.5$$

y teníamos $n = 16$; $\sigma = 3$; un promedio muestral $x = 5.25$
Como $\alpha = 0.01$ entonces $z_{\alpha/2} = z_{0.005} = 2.575$

Construimos un intervalo de confianza de nivel $1 - \alpha = 1 - 0.01 = 0.99$

Solución

$$\left[\bar{X} - z_{\alpha/2}\,\sigma/\sqrt{n}\,;\bar{X} + z_{\alpha/2}\,\sigma/\sqrt{n}\right] = \left[5.25 - 2.575\frac{3}{\sqrt{16}};5.25 + 2.575\frac{3}{\sqrt{16}}\right] = [3.31875;\ 7.18125]$$

Entonces la regla de decisión es:

$$\begin{cases} rechazar \quad H_0 \quad si \quad 5.5 \notin [3.31875;\ 7.18125] \\ aceptar \quad H_0 \quad si \quad 5.5 \in [3.31875;\ 7.18125] \end{cases}$$

Como $5.5 \in [3.31875;\ 7.18125]$, entonces se acepta H_0.

Como realizamos test en otras condiciones?

Qué otras situaciones se nos pueden presentar?

- Que tengamos más de 30 datos en nuestra muestra, conozcamos la varianza poblacional y no sepamos su distribución
- Que tengamos mas de 30 datos en nuestra muestra, desconozcamos la varianza poblacional y que tiene una distribución normal
- Tests de hipótesis sobre la varianza

También podemos crear test de hipótesis de la diferencias de 2 medias poblacionales o de 2 varianza (comparaciones)

Otras condiciones del test

Para resolver estas situaciones vamos a proceder de manera similar a los intervalos de confianza, dependiendo qué situación queramos analizar qué estadístico de prueba tomar y qué regla de decisión vamos a contrastar. El resto se reduce a pura álgebra.
Es importante tener en cuenta las condiciones en las que realizamos el test

- Qué parámetro poblacional vamos a contrastar
- Qué distribución tiene la población
- El tamaño de la muestra
- En base a estos últimos que estadístico nos ayudará a generar la decisión

La prueba de bondad de la chi-cuadrado

La prueba de bondad de ajuste chi-cuadrado es un análogo de la prueba t unidireccional para variables categóricas: prueba si la distribución de datos categóricos de muestra coincide con una distribución esperada. Por ejemplo, podría usar una prueba de bondad de ajuste chi-cuadrado para verificar si los datos demográficos de la cantidad de médicos, abogados y maestros de toda la población de Argentina. O si las preferencias de sus amigos coinciden con las de los usos de Internet.

Al trabajar con datos categóricos, los valores de las observaciones en sí mismas no son de mucha utilidad para las pruebas estadísticas porque las categorías como "masculino", "femenino" y "otro" no tienen ningún significado matemático. Las pruebas que tratan con variables categóricas se basan en recuentos de variables en lugar del valor real de las variables mismas.

La prueba de chi cuadrado para independencia de variables categóricas

Supongamos que tenemos los datos de 4 tratamientos médicos y con una descripción de cómo lo han recibido los pacientes. Los plasmamos en una **tabla de contingencia**

i x j = 4*3	Peor	igual	mejor	
Tratamiento1	7	28	115	
Tratamiento2	15	20	85	
Tratamiento3	10	30	90	
Tratamiento4	5	40	115	
				560

La prueba de chi cuadrado para independencia de variables categóricas

En primer lugar vamos a plantear nuestra hipótesis

H_0="las dos variables en estudio son independientes"

H_1="las dos variables en estudio están relacionadas"

Como en los juicios, en estadística, las variables son independientes hasta que demostremos lo contrario.

La prueba de chi cuadrado para independencia de variables categóricas

El tratamiento 1 y el 4 fueron mejores?

Si examinamos la tabla

i x j = 4*3	Peor	igual	mejor	
Tratamiento1	7	28	115	
Tratamiento2	15	20	85	
Tratamiento3	10	30	90	
Tratamiento4	5	40	115	
				560

La prueba de chi cuadrado para independencia de variables categóricas

Para tener un análisis certero debemos calcular las frecuencias marginales ya que no todos los tratamientos tuvieron igual cantidad de pacientes.

i x j = 4*3	Peor	igual	mejor	
Tratamiento1	7	28	115	150
Tratamiento2	15	20	85	120
Tratamiento3	10	30	90	130
Tratamiento4	5	40	115	160
	37	118	405	560

Ya no es lo mismo 115/160 que 15/150

La prueba de chi cuadrado para independencia de variables categóricas

Entonces como datos disponibles tenemos las frecuencias observadas, las frecuencias marginales y el número ‘gran total’ de la muestra

i x j = 4*3	Peor	igual	mejor	
Tratamiento1	7	28	115	150
Tratamiento2	15	20	85	120
Tratamiento3	10	30	90	130
Tratamiento4	5	40	115	160
	37	118	405	560

La prueba de chi cuadrado para independencia de variables categóricas

Ahora pierde sentido hacer operaciones como medianas, promedio etc con este tipo de datos que pierden sentido matemáticamente. No podemos sacar un promedio de los tratamientos y decir que un tratamiento 3,4 por ejemplo sería el correcto.

Las pruebas que tratan con variables categóricas se basan en recuentos de variables en lugar del valor real de las variables mismas.

La frecuencia esperada

¿cómo se contrasta la H_0?

$$fe_{ij} = \frac{(\text{total fila i-ésima})*(\text{total columna j-ésima})}{\text{gran total}}$$

Se calculan las frecuencias que cabría esperar si las 2 variables fueran independientes

Armamos la tabla fe

Rearmamos la tabla con las frecuencias esperadas. Para eso rellenaremos las casillas con el cociente de la slide anterior, que en definitiva es nuestro estadistico

i x j = 4*3	Peor	igual	mejor	
Tratamiento1	7 / 9.91	28 / 31.61	115 / 108.48	150
Tratamiento2	15 / 7.93	20 / 25.28	85 / 86.79	120
Tratamiento3	10 / 8.59	30 / 27.39	90 / 94.02	130
Tratamiento4	5 / 10.57	40 / 33.72	115 / 115.71	160
	37	118	405	560

Como medimos las discrepancias?

Se calcula la diferencia entre las frecuencias observadas y las frecuencias esperadas **(fo_{ij} - fe_{ij})** para todas y cada una de las casillas de la tabla

Armamos el estadístico de prueba

$$\chi^2 = \sum \left[\frac{(f_0 - f_e)^2}{f_e} \right]$$

Matemáticamente se ha demostrado que sila hipótesis nula es cierta el estadístico sigue una distribución o modelo chi cuadrado con (i-1)(j-1) grados de libertad que tienen que ver con las dimensiones de la tabla. Como se ajusta a un modelo podemos establecer un nivel de riesgo y compara el experimental con el valor crítico del modelo

Tabla chi

Aplicado al ejemplo

$$\chi^2_{experimental} = (7-9.91)^2/9.91 \ + \ \ldots\ldots \ + \ (115-115.71)^2/115.71 = 13.87$$

	0.9950	0.9750	0.950	0.900	0.200	0.10	0.050	0.025	0.010	0.001
1	0.0000393	0.000982	0.00393	0.0158	1.642	2.706	3.841	5.024	6.635	10.828
2	0.010	0.0506	0.103	0.211	3.219	4.605	5.991	7.378	9.510	13.816
3	0.0717	0.216	0.352	0.584	4.642	6.251	7.851	9.348	11.345	16.266
4	0.207	0.484	0.711	1.064	5.989	7.779	9.488	1.143	13.277	18.467
5	0.412	0.831	1.145	1.610	7.289	9.236	11.070	12.833	15.086	20.515
6	0.676	1.237	1.635	2.204	8.558	10.645	12.592	14.449	16.812	22.458
7	0.989	1.690	2.167	2.833	9.803	12.017	14.067	16.013	18.475	24.322
8	1.344	2.180	2.733	3.490	11.030	13.362	15.507	17.535	20.090	26.124
9	1.735	2.700	3.325	4.168	12.242	14.684	16.919	19.023	21.666	27.877
10	2.156	3.247	3.940	4.865	13.442	15.987	18.307	20.483	23.209	29.588
11	2.603	3.816	4.575	5.578	14.631	17.275	19.675	21.920	24.725	31.264
12	3.074	4.404	5.226	6.304	15.812	18.549	21.026	13.337	26.217	32.909
13	3.565	5.009	5.892	7.042	16.985	19.812	22.362	24.736	27.688	34.528
14	4.075	5.629	6.571	7.790	18.151	21.064	23.685	26.119	29.141	36.123

Redondeando

Observamos que para 6 grados de libertad (los correspondientes para nuestra tabla de contingencia => (4-1)*(3-1)=6) el modelo nos da un resultado teórico de 12.592 con un nivel de significancia de 0.05.

$$\chi^2_{6\ ,\ 0.05} = 12.592$$

$$13.87 > 12.592$$

Por lo que la conclusión es que la respuesta depende del tratamiento ya que mi regla es si el $\chi^{2\ \text{muestral}} > \chi^{2\ \text{experimental}}$ rechazar la H_0

La moraleja

Las hipótesis deben ser previas a los resultados del estudio

Schwartz-Woloshin / Ventura

¿Dudas?