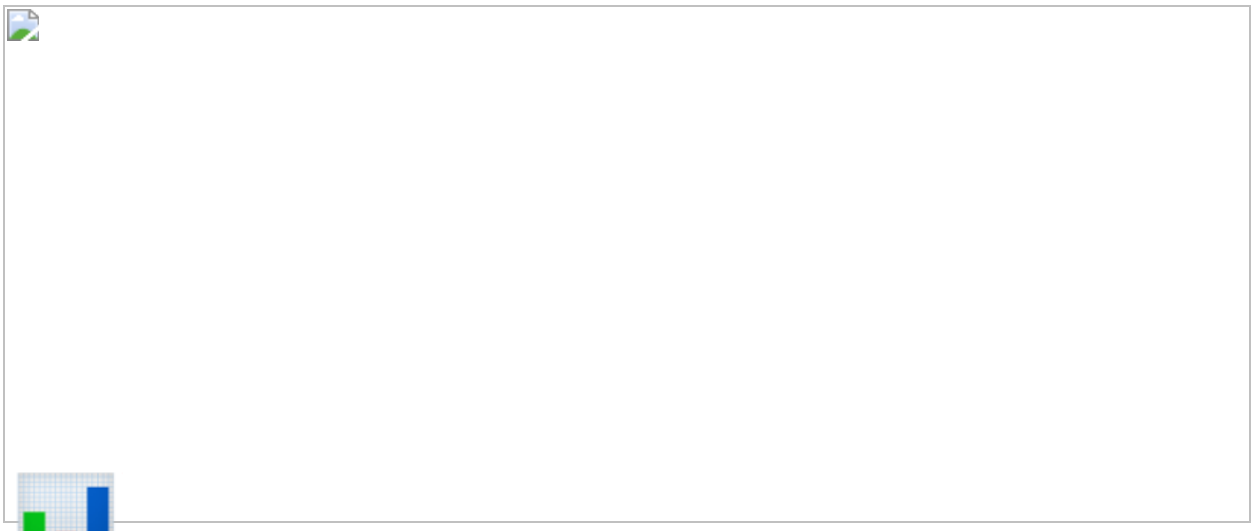


Estadistica

Age	2023

Estadistica Descriptiva

Tabla de frecuencias

- Elegimos un intervalo (a, b) que contenga todos los datos. Tomar aproximadamente n subintervalos. Los subintervalos se llaman intervalos de clase o simplemente clases. Resulta satisfactorio utilizar no menos de 5 clases ni más de 20. La longitud de cada una sería (b-a)/r. Mayor frecuencia = mayor area

Medidas descriptivas

- Media

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Mediana

$$\tilde{x} = \begin{cases} x_{((n+1)/2)} & n \text{ impar} \\ \dfrac{x_{(n/2)} + x_{((n/2)+1)}}{2} & n \text{ par} \end{cases}$$

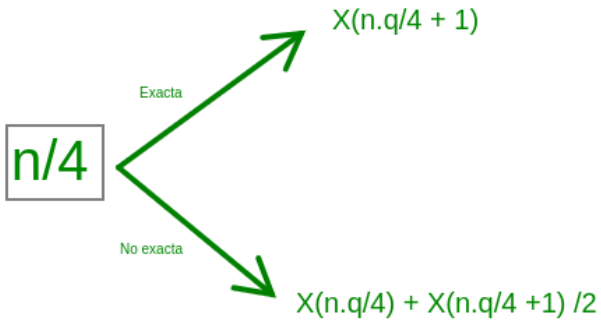
- Moda: observacion con mayor frecuencia, puede haber mas de una moda por muestra.
- Percentil

Una regla práctica para calcular los percentiles de un conjunto de n datos es la siguiente:
para calcular p_k hacemos el producto nk

si nk es un número entero i entonces $p_k = \dfrac{x_{(i)} + x_{(i+1)}}{2}$

si nk no es un número entero entonces tomamos la parte entera de nk : $\lfloor nk \rfloor = i$ y entonces $p_k = x_{(i+1)}$

- Cuartiles: son los puntos de division que separan a la muestra en 4 partes iguales.
  - El primer cuartil o cuartil inferior, q1 , es un valor que tiene aproximadamente la cuarta parte (25%) de las observaciones por debajo de él, y el 75% restante, por encima de él. El segundo cuartil , q2 , tiene aproximadamente la mitad (50%) de las observaciones por debajo de él. El segundo cuartil coincide con la mediana. El tercer cuartil o cuartil superior, q3 , tiene aproximadamente las tres cuartas partes (75%) de las observaciones por debajo de él.



- Rangos
  - Muestral: max-min

- Intercuartilico: q3-q1
- A mayor rango, mayor variabilidad de datos

Varianza muestral y desviacion estandar muestral

Si $x_1, x_2, \ldots, x_n$ es una muestra de n observaciones, entonces la ***varianza muestral*** es

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

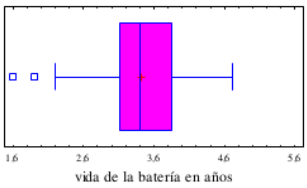
La ***desviación estándar muestral***, s, es la raíz cuadrada positiva de la varianza muestral

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

No es una medida robusta, ya que un dato atipico distorsiona.

Diagrama de caja

tamaño de muestra = 40
Media = 3.4125
Mediana = 3.4
Mínimo = 1.6
Máximo = 4.7
Rango = 3.1
1º cuartil = 3.1
3º cuartil = 3.85
rango untercuartílico (*RIC*) = 0.75
1.5 *RIC* = 1.125
3 *RIC* = 2.25



LI: q1 - 1.5RIC. LS: q3 + 1.5RIC.

Estimacion puntual

Las variables aleatorias (X 1 , X 2 ,..., X n) constituyen una muestra aleatoria de tamaño n de una v.a. X si X 1 , X 2 ,..., X n son independientes idénticamente distribuidas.

Un estadistico es cualquier funcion de una muestra aleatoria. Una de sus aplicaciones es obtener estimaciones puntuales de los parametros desconocidos de una distribucion. Si un estadistico se usa para estimar un parametro desconocido se lo llama **estimador puntual.**

Estadisticos usuales

Sea $X_1, X_2, ..., X_n$ una muestra aleatoria de una v.a. X donde $E(X) = \mu$ y $V(X) = \sigma^2$
Si desconocemos μ un estadístico que se utiliza para estimar ese parámetro es la ***media o promedio muestral*** $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Análogamente si se desconoce σ^2 un estadístico usado para tener alguna información sobre ese parámetro es la ***varianza muestral*** que se define como $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$

Otro estadístico es la ***desviación estándar muestral*** $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$

Estimador = simbolo. Estimacion = valor.

En este caso el estimador puntual de p sería $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ donde

$$X_i = \begin{cases} 1 \text{ si la } i\text{-ésima observación tiene la característica de interés} \\ 0 \quad \text{caso contrario} \end{cases} \qquad i = 1,2,...,n$$

Por lo tanto $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ es la ***proporción de objetos en la muestra*** cumplen la característica de interés

Criterios para evaluar un estimador

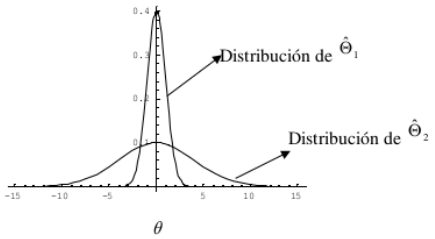
Se dice que el estimador puntual $\hat{\Theta}$ es un ***estimador insesgado*** del parámetro θ si $E(\hat{\Theta}) = \theta$ cualquiera sea el valor verdadero de θ

Podemos exigir que el estimador $\hat{\Theta}$ tenga una distribución cuya media sea θ.

Si un estimador es insesgado, su sesgo es cero

Varianza y error cuadratico medio de un estimador puntual

Supongamos que $\hat{\Theta}_1$ y $\hat{\Theta}_2$ son dos estimadores insegados de un parámetro θ. Esto indica que la distribución de cada estimador está centrada en el verdadero parámetro θ. Sin embargo las varianzas de estas distribuciones pueden ser diferentes. La figura siguiente ilustra este hecho.



Como $\hat{\Theta}_1$ tiene menor varianza que $\hat{\Theta}_2$, entonces es más probable que el estimador $\hat{\Theta}_1$ produzca una estimación más cercana al verdadero valor de θ. Por lo tanto si tenemos dos estimadores insesgados se seleccionará aquel que tenga menor varianza.

Si $\hat{\Theta}_1$ y $\hat{\Theta}_2$ son dos estimadores de un parámetro θ.

La eficiencia relativa de $\hat{\Theta}_2$ con respecto a $\hat{\Theta}_1$ se define como $\frac{ECM(\hat{\Theta}_1)}{ECM(\hat{\Theta}_2)}$

Si la eficiencia relativa es menor que 1 entonces $\hat{\Theta}_1$ tiene menor error cuadrático medio que $\hat{\Theta}_2$

Por lo tanto $\hat{\Theta}_1$ ***es más eficiente que*** $\hat{\Theta}_2$

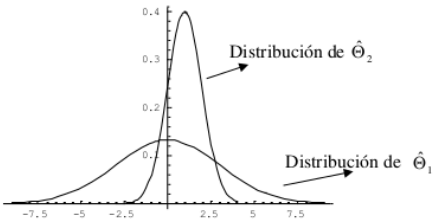
Observaciones:

1- Si $\hat{\Theta}$ es un estimador insesgado de θ, entonces $ECM(\hat{\Theta}) = V(\hat{\Theta})$

2- 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 más cercanos al verdadero parámetro que $\hat{\Theta}_1$



Consistencia de estimadores puntuales

Teorema. Sea $\hat{\Theta}_n$ un estimador del parámetro θ basado en una muestra aleatoria $(X_1, X_2, ..., X_n)$. Si $\lim_{n\to\infty} E(\hat{\Theta}_n) = \theta$ y $\lim_{n\to\infty} V(\hat{\Theta}_n) = 0$, entonces $\hat{\Theta}_n$ es un estimador consistente de θ.

Metodo de estimacion puntual

- Metodo de los momentos:

Definimos los ***momentos de orden k de una variable aleatoria*** como:

$$\mu_k = E(X^k) = \sum_{x_i \in R_X} x_i^k \, p(x_i) \qquad (k = 0,1,2,...) \qquad \text{Si } X \text{ es discreta}$$

$$\mu_k = E(X^k) = \int_{-\infty}^{+\infty} x^k f(x) dx \qquad (k = 0,1,2,...) \qquad \text{Si } X \text{ es continua,}$$

y definimos los correspondientes momentos ***muestrales de orden k*** como:

$$M_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \qquad (k = 0,1,2,...),$$

$$\begin{cases} \mu_1 = M_1 \\ \mu_2 = M_2 \\ \vdots \quad \vdots \\ \mu_r = M_r \end{cases}$$

En general es válido que v(x) = e(x²) - u²

- Metodo de maxima verosimilitud:

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

$$L(x_1, x_2, \ldots, x_n, \theta) = P(X_1 = x_1)P(X_2 = x_2)..P(X_n = x_n) = p(x_1, \theta).p(x_2, \theta)....p(x_n, \theta)$$

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

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

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

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.
Entonces

$$\ln(L(x_1, x_2, \ldots, x_n; p)) = \left(\sum_{i=1}^{n} x_i\right) \ln p + \left(n - \sum_{i=1}^{n} x_i\right) \ln(1-p)$$

Y ahora podemos maximizar la función derivando e igualando a cero

$$\frac{\partial \ln L(x_1, x_2, \ldots, x_n; p)}{\partial p} = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p} = 0$$

de donde despejando p

$p = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$ la proporción de defectuosos en la muestra

Por lo tanto se toma como estimador a $\hat{p} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

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 ***consistentes***

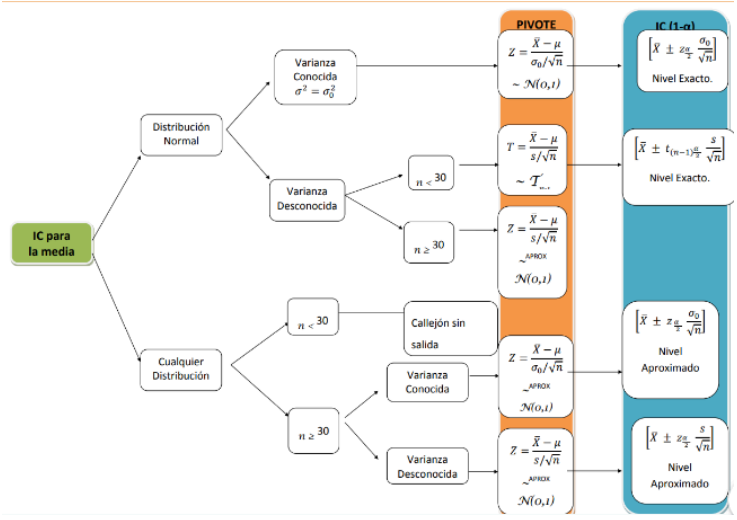
3- Bajo condiciones bastantes generales se puede probar que los EMV ***asintóticamente tienen varianza mínima***

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

$\alpha = 0.05$, se quiere construir un intervalo $(\hat{\Theta}_1, \hat{\Theta}_2)$ tal que $P(\theta \in (\hat{\Theta}_1, \hat{\Theta}_2)) = 0.95$, o escrito de otra forma $P(\hat{\Theta}_1 \leq \theta \leq \hat{\Theta}_2) = 0.95$

Nivel de confianza: 1-α.



Intervalos de confianza junto a pivotes y distribuciones.

IC para varianza conocida

- Longitud del intervalo:

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

Notar que:

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

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

- Precision del estimador:

Si estimamos puntualmente al parámetro μ con $\bar{X}$ estamos cometiendo un error en la estimación menor o igual a $\frac{L}{2} = z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$, que se conoce como ***precisión del estimador***

Ejemplo: Se estima que el tiempo de reacción a un estímulo de cierto dispositivo electrónico está distribuido normalmente con desviación estándar de 0.05 segundos. ¿Cuál es el número de mediciones temporales que deberá hacerse para que la confianza de que el error de la estimación de la esperanza no exceda de 0.01 sea del 95%?

Nos piden calcular n tal que $\frac{L}{2} = z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < 0.01$ con $\alpha = 0.05$.

Por lo tanto $n \geq \left(z_{0.025}\frac{0.05}{0.01}\right)^2$.

Además $z_{0.025} = 1.96$. Entonces $n \geq \left(z_{0.975}\frac{0.05}{0.01}\right)^2 = (1.96 \times 5)^2 = 96.04$.

O sea hay que tomar por lo menos 97 mediciones temporales.

IC para una proporcion

$$p(x)=\begin{cases} p(1)=P(X_i=1)=p \\ p(0)=P(X_i=0)=1-p, \end{cases}$$

Tiene distribucion X ~ B(1,p)

Supongamos que consideramos una muestra aleatoria $(X_1, X_2..., X_n)$ de tamaño n. Si formamos el estadístico $X = X_1 + X_2 + ... + X_n$, es evidente que esta v.a. mide el número de individuos de la muestra de tamaño n que verifican la propiedad A. Por lo tanto por su significado X es una v.a. cuya distribución es binomial con parámetros n y p: $X \sim B(n,p)$. De acuerdo con esto, la variable aleatoria $\hat{P}$ definida: $\hat{P} = \frac{X}{n}$ representa la proporción de individuos de la muestra que verifican la propiedad A.

$$Z = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \underset{n \text{ grande}}{\sim} N(0,1),$$

$$\left[\hat{P} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}},\ \hat{P} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \right]$$

Importante! N>30

Observaciones:

1- Este procedimiento depende de la aproximación normal a la distribución binomial. Por lo tanto el intervalo (8.10) se puede utilizar si $n\hat{P} > 10$ y $n(1-\hat{P}) > 10$, es decir, ***la muestra debe contener un mínimo de diez éxitos y diez fracasos.***

2- La longitud del intervalo es $L = 2z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$, pero esta expresión está en función de $\hat{P}$

Si nos interesa hallar un valor de n de manera tal que la longitud L sea menor que un valor determinado, podemos hacer dos cosas:

a) tomar una muestra preliminar, con ella estimar p con $\hat{P}$ y de la expresión anterior despejar n, lo que lleva a

$$L = 2z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \le l \quad \Rightarrow \quad n \ge \left(\frac{2z_{\frac{\alpha}{2}}}{l} \right)^2 \hat{P}(1-\hat{P})$$

b) si no tomamos una muestra preliminar, entonces acotamos $\hat{P}(1-\hat{P}) \le 0.5 \times (1-0.5)$, entonces

$$L = 2z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \le 2z_{\frac{\alpha}{2}} \sqrt{\frac{0.5(1-0.5)}{n}} \le l \quad \Rightarrow \quad n \ge \left(\frac{z_{\frac{\alpha}{2}}}{l} \right)^2$$

Test de hipotesis

Consiste en construir un estadistico a traves de una muestra y segun el valor que tome el mismo, aceptar o rechazar una hipotesis. Para realizar una prueba de hipotesis se pone en juicio la hipotesis nula, asumiendo que Ho es verdadera.

H_0	Aceptamos	Rechazamos
Verdadera	✓ *Decisión correcta*	*Error tipo I* α
Falsa	*Error tipo II* β	✓ *Decisión correcta*

Existen 2 tipos de errores, el error de tipo alpha es el mas grave, y se le llama nivel de significancia.

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.
Un concepto importante es el siguiente:
La ***potencia*** de un test es la probabilidad de rechazar la hipótesis nula. La simbolizamos $\pi(\mu)$.
Para los ***valores de*** μ ***tal que la alternativa es verdadera*** se tiene

$$\pi(\mu)=P\left(rechazar \quad H_0 / H_0 \quad es \quad falsa\right)=1-\beta(\mu)$$

Varianza conocida

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*** $X \sim N(\mu,\sigma^2)$.
Nuevamente, como en el ejemplo introductorio, es razonable tomar como estadístico de prueba al promedio muestral $\overline{X}$. Bajo las suposiciones hechas tenemos que $\overline{X} \sim N\left(\mu,\frac{\sigma^2}{n}\right)$.

Supongamos que tenemos las hipótesis

$$H_0:\mu=\mu_0 \qquad \text{contra} \qquad H_1:\mu\neq\mu_0$$

Donde μ_0 es una constante específica. Se toma una muestra aleatoria $X_1,X_2,...,X_n$ de la población.

Si $H_0:\mu=\mu_0$ es verdadera, entonces $\overline{X} \sim N\left(\mu_0,\frac{\sigma^2}{n}\right)$, por lo tanto el estadístico

$Z=\dfrac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$ tiene distribución $N(0,1)$ si $H_0:\mu=\mu_0$ es verdadera

Tomamos a Z como ***estadístico de prueba***

Si $H_0:\mu=\mu_0$ es verdadera entonces $P\left(-z_{\frac{\alpha}{2}}\leq Z\leq z_{\frac{\alpha}{2}}\right)=1-\alpha$

Importante! El pivote se plantea si Ho es verdadera.

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

$$H_0 : \mu = \mu_0 \quad \text{contra} \quad H_1 : \mu > \mu_0$$

En este caso la región crítica debe colocarse en la cola superior de la distribución normal estándar y el rechazo de H_0 se hará cuando el valor calculado de z_0 sea muy grande, esto es la regla de decisión será

$$\begin{cases} rechazar \quad H_0 \quad si \quad \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} > z_\alpha \\ aceptar \quad H_0 \quad si \quad \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \leq z_\alpha \end{cases}$$

De manera similar para las hipótesis

$$H_0 : \mu = \mu_0 \quad \text{contra} \quad H_1 : \mu < \mu_0$$

se calcula el valor del estadístico de prueba z_0 y se rechaza H_0 si el valor de z_0 es muy pequeño, es decir la regla de decisión será

$$\begin{cases} rechazar \quad H_0 \quad si \quad \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < -z_\alpha \\ aceptar \quad H_0 \quad si \quad \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \geq -z_\alpha \end{cases}$$

P - Valor (Distribucion normal)

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

a) si las hipótesis son $H_0 : \mu = \mu_0$ contra $H_1 : \mu \neq \mu_0$

$$p - valor = P(|Z| > |z_0|) = 1 - P(|Z| < |z_0|) = 1 - [\Phi(|z_0|) - \Phi(-|z_0|)] = 1 - [2\Phi(|z_0|) - 1] = 2[1 - \Phi(|z_0|)]$$

b) si las hipótesis son $H_0 : \mu = \mu_0$ contra $H_1 : \mu > \mu_0$

$$p - valor = P(Z > z_0) = 1 - P(Z \leq z_0) = 1 - \Phi(z_0)$$

c) si las hipótesis son $H_0 : \mu = \mu_0$ contra $H_1 : \mu < \mu_0$

$$p - valor = P(Z < z_0) = \Phi(z_0)$$

Un p-valor muy chico significa mucha evidencia en contra de H_0; un p-valor alto significa que no hay evidencia en contra H_0

Notar que:

Si $\alpha < p - valor$ entonces se acepta H_0 con nivel de significancia α

Si $\alpha > p - valor$ entonces se rechaza H_0 con nivel de significancia α

Varianza desconocida

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

El cual, ***si la hipótesis nula es verdadera***, tiene distribución ***Student con n-1 grados de libertad***. Entonces, para un nivel α prefijado, la regla de decisión es

$$\begin{cases} rechazar \quad H_0 \quad si \quad |T| > t_{\frac{\alpha}{2}n-1} \\ aceptar \quad H_0 \quad si \quad |T| \leq t_{\frac{\alpha}{2}n-1} \end{cases} \quad \text{es decir} \quad \begin{cases} rechazar \quad H_0 \quad si \quad \left|\dfrac{\overline{X} - \mu_0}{S/\sqrt{n}}\right| > t_{\frac{\alpha}{2}n-1} \\ aceptar \quad H_0 \quad si \quad \left|\dfrac{\overline{X} - \mu_0}{S/\sqrt{n}}\right| \leq t_{\frac{\alpha}{2}n-1} \end{cases}$$

Si la alternativa es $H_1 : \mu > \mu_0$ entonces la regla de decisión es $\begin{cases} rechazar \quad H_0 \quad si \quad T > t_{\alpha,n-1} \\ aceptar \quad H_0 \quad si \quad T \le t_{\alpha,n-1} \end{cases}$

Si la alternativa es $H_1 : \mu < \mu_0$ entonces la regla de decisión es $\begin{cases} rechazar \quad H_0 \quad si \quad T < -t_{\alpha,n-1} \\ aceptar \quad H_0 \quad si \quad T \ge -t_{\alpha,n-1} \end{cases}$

P - Valor (Distribucion T-Student)

a) las hipótesis son $H_0 : \mu = \mu_0$ contra $H_1 : \mu \neq \mu_0$

$p - valor = P(|T| > |t_0|) = 1 - P(|T| \le |t_0|) = 2(1 - P(T \le t_0))$

b) las hipótesis son $H_0 : \mu = \mu_0$ contra $H_1 : \mu > \mu_0$

$p - valor = P(T > t_0) = 1 - P(T \le t_0)$

c) las hipótesis son $H_0 : \mu = \mu_0$ contra $H_1 : \mu < \mu_0$

$p - valor = P(T \le t_0)$

Varianza desconocida para muestras grandes

$$Z = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \approx N(0,1) \quad \text{aproximadamente, si } n \ge 30 \text{ si } H_0 : \mu = \mu_0$$

Además, si no podemos decir que la muestra aleatoria proviene de una población normal, sea σ^2 conocida o no, por T.C.L. los estadísticos

$$Z = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \approx N(0,1) \quad \text{aproximadamente, si } n \ge 30 \text{ si } H_0 : \mu = \mu_0$$

Y

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \approx N(0,1) \quad \text{aproximadamente, si } n \ge 30 \text{ si } H_0 : \mu = \mu_0$$

Las pruebas de hipótesis tendrán entonces un nivel de significancia ***aproximadamente de*** α

Test para una proporcion

Si $H_0 : p = p_0$ es verdadera entonces $Z = \dfrac{\hat{P} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} \approx N(0,1)$ aproximadamente por T.C.L.

Por lo tanto la regla de decisión es

$\begin{cases} rechazar \quad H_0 \quad si \quad |Z| > z_{\frac{\alpha}{2}} \\ aceptar \quad H_0 \quad si \quad |Z| \le z_{\frac{\alpha}{2}} \end{cases}$ donde $Z = \dfrac{\hat{P} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$

Si $H_1 : p > p_0$ entonces la regla de decisión es $\begin{cases} rechazar \quad H_0 \quad si \quad Z > z_\alpha \\ aceptar \quad H_0 \quad si \quad Z \le z_\alpha \end{cases}$

Si $H_1 : p < p_0$ entonces la regla de decisión es $\begin{cases} rechazar \quad H_0 \quad si \quad Z < -z_\alpha \\ aceptar \quad H_0 \quad si \quad Z \ge -z_\alpha \end{cases}$

Observaciones:
1- La prueba descrita anteriormente requiere que la proporción muestral esté normalmente distribuida. Esta suposición estará justificada siempre que $np_0 > 10$ y $n(1-p_0) > 10$, donde p_0 es la proporción poblacional que se especificó en la hipótesis nula.

Regresion lineal simple

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

donde x es, por ahora, una variable no aleatoria, ε ***es la v.a. del error*** y ***asumimos que***

$$E(\varepsilon) = 0 \qquad \text{y} \qquad V(\varepsilon) = \sigma^2$$

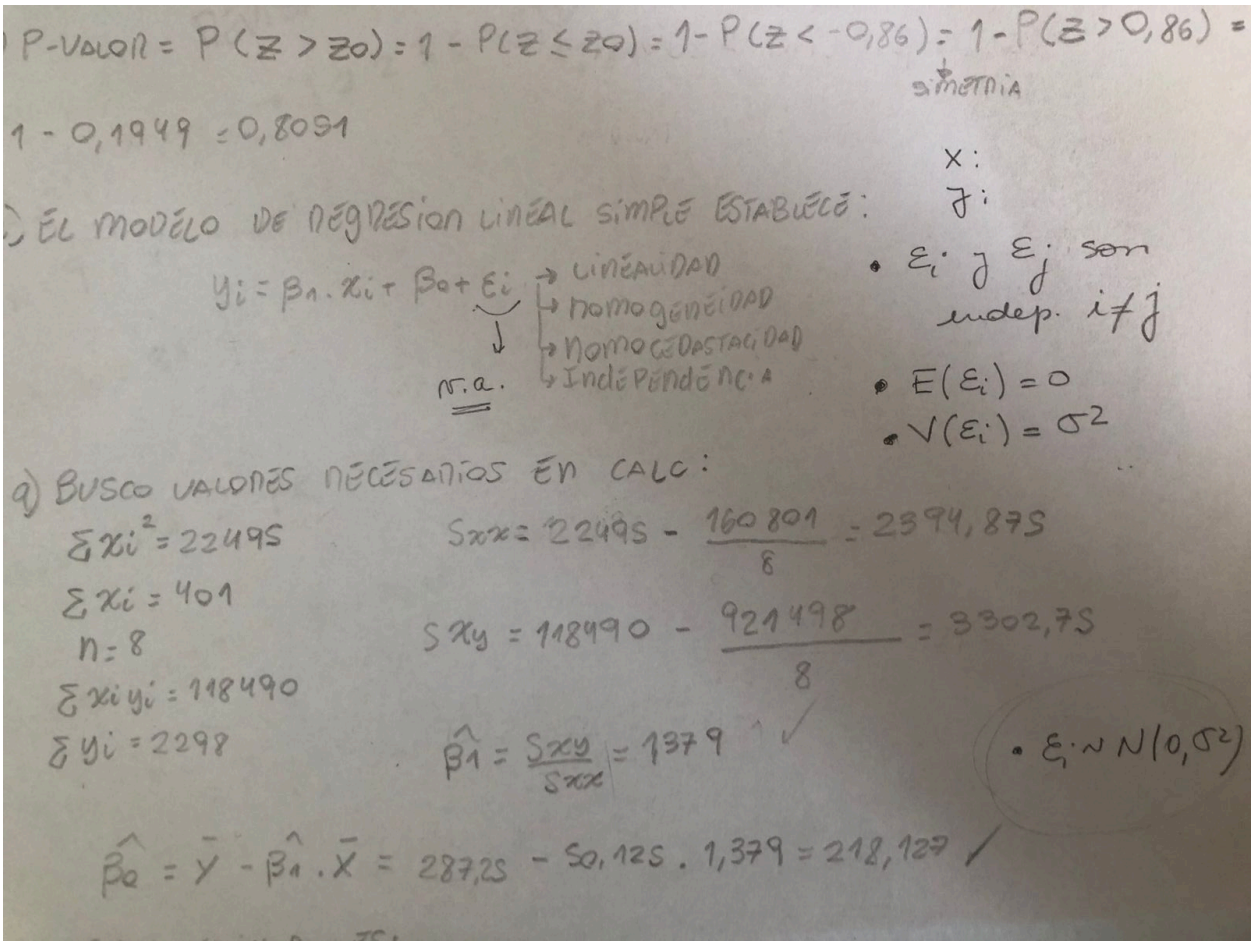
Entonces Y es una variable aleatoria tal que

$$E(Y/x) = E(\beta_0 + \beta_1 x + \varepsilon) = \beta_0 + \beta_1 x + E(\varepsilon) = \beta_0 + \beta_1 x$$

$$V(Y/x) = V(\beta_0 + \beta_1 x + \varepsilon) = V(\varepsilon) = \sigma^2$$

Se utilizo una distribucion normal para Ɛ. Por lo que Y tendra distribucion normal.

Condiciones del modelo



Estimacion de parametros

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\cdot\ \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

$$S_{xy} = \sum_{i=1}^{n} y_i (x_i - \bar{x}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$$

Esta en tabla de formulas

Observaciones:

1- Las estimaciones de mínimos cuadrados $\hat{\beta}_1$ y $\hat{\beta}_0$ son valores de variables aleatorias y dicho valor varía con las muestras. Los coeficientes de regresión β_0 y β_1 son constantes desconocidas que estimamos con $\hat{\beta}_1$ y $\hat{\beta}_0$.

2- Los residuos e_i no son lo mismo que los errores ε_i. Cada residuo es la diferencia $e_i = y_i - \hat{y}_i$ entre el valor observado y el valor ajustado, y se pueden calcular a partir de los datos. Los errores ε_i representan la diferencia entre los valores medidos y_i y los valores $\beta_0 + \beta_1 x_i$. Como los valores verdaderos de β_0 y β_1 no se conocen entonces, los errores no se pueden calcular.

Propiedades de los parametros y Pivotes

$$E(\hat{\beta}_1) = \beta_1 \quad \text{y} \quad V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

$$E(\hat{\beta}_0) = \beta_0 \quad \text{y} \quad V(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$$

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right) \quad \text{y} \quad \hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

En resumen $SS_R = S_{yy} - \hat{\beta}_1 S_{xy}$ ó $SS_R = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim N(0,1) \quad \text{y} \quad \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0,1)$$

Por lo tanto $\hat{\sigma}^2 = \frac{S_{yy} - \frac{S_{xy}^2}{S_{xx}}}{n-2}$

Tambien sigma² = (Syy - B1Sxy) / n-2

Interferencias estadisticas sobre los parametros de regresion

Para B1:

Tests de hipótesis sobre β_1

Se desea probar la hipótesis de que la pendiente β_1 es igual a una constante, por ejemplo β_{10}. Supongamos las hipótesis

$$H_0 : \beta_1 = \beta_{10} \qquad \text{contra} \qquad H_0 : \beta_1 \neq \beta_{10}$$

El estadístico de prueba es $T = \dfrac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\dfrac{\hat{\sigma}^2}{S_{xx}}}}$ que bajo H_0 tiene distribución Student con *n-2* grados de libertad.

Por lo tanto la regla de decisión es $\begin{cases} rechazar\ \ H_0 \ \ si\ \ |T| > t_{\frac{\alpha}{2},n-2} \\ aceptar\ \ H_0 \ \ si\ \ |T| \leq t_{\frac{\alpha}{2}n-2} \end{cases}$

Si $H_1 : \beta_1 > \beta_{10}$ se rechaza $H_0 : \beta_1 = \beta_{10}$ si $T > t_{\alpha,n-2}$

Si $H_1 : \beta_1 < \beta_{10}$ se rechaza $H_0 : \beta_1 = \beta_{10}$ si $T < -t_{\alpha,n-2}$

Un caso especial importante es cuando $H_0 : \beta_1 = 0$ contra $H_0 : \beta_1 \neq 0$

Estas hipótesis están relacionadas con la ***significancia de la regresión***.

Aceptar $H_0 : \beta_1 = 0$ es equivalente a concluir que no hay ninguna relación lineal entre *x* e *Y*.

Si $H_0 : \beta_1 = 0$ se rechaza implica que *x* tiene importancia al explicar la variabilidad en *Y*. También puede significar que el modelo lineal es adecuado, o que aunque existe efecto lineal pueden obtenerse mejores resultados agregando términos polinomiales de mayor grado en *x*.

Intervalos de confianza para β_1

Podemos construir intervalos de confianza para β_1 de nivel $1-\alpha$ utilizando el hecho que el estadístico $\dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{\dfrac{\hat{\sigma}^2}{S_{xx}}}} \sim t_{n-2}$. El intervalo sería

$$\left[\hat{\beta}_1 - t_{\frac{\alpha}{2},n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}};\ \hat{\beta}_1 + t_{\frac{\alpha}{2}n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}\right] \qquad (19)$$

Para Bo:

$$H_0 : \beta_0 = \beta_{00} \qquad \text{contra} \qquad H_0 : \beta_0 \neq \beta_{00}$$

El estadístico de prueba es $T\dfrac{\hat{\beta}_0 - \beta_{00}}{\sqrt{\hat{\sigma}^2\left(\dfrac{1}{n} + \dfrac{\bar{x}^2}{S_{xx}}\right)}}$ y bajo $H_0 : \beta_0 = \beta_{00}$ tenemos que $T \sim t_{n-2}$

Por lo tanto la regla de decisión es $\begin{cases} rechazar\ \ H_0 \ \ si\ \ |T| > t_{\frac{\alpha}{2},n-2} \\ aceptar\ \ H_0 \ \ si\ \ |T| \leq t_{\frac{\alpha}{2}n-2} \end{cases}$

Si $H_1 : \beta_0 > \beta_{00}$ se rechaza $H_0 : \beta_0 = \beta_{00}$ si $T > t_{\alpha,n-2}$

Si $H_1 : \beta_0 < \beta_{00}$ se rechaza $H_0 : \beta_0 = \beta_{00}$ si $T < -t_{\alpha,n-2}$

Intervalos de confianza de nivel $1-\alpha$ se deducen de manera análoga a lo visto anteriormente, donde usamos el hecho que el estadístico $T\dfrac{\hat{\beta}_0 - \beta_0}{\sqrt{\hat{\sigma}^2\left(\dfrac{1}{n} + \dfrac{\bar{x}^2}{S_{xx}}\right)}} \sim t_{n-2}$

El intervalo es $\left[\hat{\beta}_0 - t_{\frac{\alpha}{2},n-2}\sqrt{\hat{\sigma}^2\left(\dfrac{1}{n} + \dfrac{\bar{x}^2}{S_{xx}}\right)};\ \hat{\beta}_0 + t_{\frac{\alpha}{2}n-2}\sqrt{\hat{\sigma}^2\left(\dfrac{1}{n} + \dfrac{\bar{x}^2}{S_{xx}}\right)}\right]$ (20)

IC para la respuesta media

Por lo tanto

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \sim N\left(\beta_0 + \beta_1 x_0;\ \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]\right) \qquad (21)$$

Como σ^2 es desconocido lo reemplazamos por $\hat{\sigma}^2 = \frac{SS_R}{n-2}$, y puede probarse que

$\frac{\hat{\beta}_0 + \hat{\beta}_1 x_0 - (\beta_0 + \beta_1 x_0)}{\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}}$ tiene distribución Student con $n-2$ grados de libertad

Razonando como en casos anteriores, el intervalo de confianza para $\beta_0 + \beta_1 x_0$ de nivel $1-\alpha$ es

$$\left[\hat{\beta}_0 + \hat{\beta}_1 x_0 - t_{\frac{\alpha}{2}, n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]};\ \hat{\beta}_0 + \hat{\beta}_1 x_0 + t_{\frac{\alpha}{2}, n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}\right] \qquad (22)$$

Coeficiente de determinacion

$$R^2 = 1 - \frac{SS_R}{S_{YY}} \qquad (27)$$

y es llamado ***coeficiente de determinación***. Vemos que R^2 será cero si $\beta_1 = 0$ y será uno si $SS_R = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 0$, lo que significa ajuste lineal perfecto.

En general $0 \le R^2 \le 1$. El valor de R^2 se interpreta como la proporción de variación de la respuesta Y que es explicada por el modelo. La cantidad $\sqrt{R^2}$ es llamada ***índice de ajuste***, y es a menudo usada como un indicador de qué tan bien el modelo de regresión ajusta los datos. Pero un valor alto de R no significa necesariamente que el modelo de regresión sea correcto.

El indice de ajuste R es a menudo llamado **coeficiente de correlacion muestral p**.