Estadística aplicada
Temas del segundo parcial

Ultima modificación:1 de diciembre de 2004

Comparación de dos medias $\mu_1 \quad vs \quad \mu_2$

IMPORTANTE: debe tomarse como población #1 la que tenga S^2 mayor
Todo esto sirve cuando las poblaciones estudiadas son **INDEPENDIENTES**

Saber cuanto vale $\mu_1 \;\; y \;\; \mu_2$ es irrelevante para la comparación.

α Ahora es totalmente tendencioso, si es chico, marca solo las grandes diferencias, si es grande, rechazamos casi siempre.

Utilizo el parámetro $\delta=\mu_1-\mu_2$; si es $\begin{array}{l}\delta>0\rightarrow\mu_1>\mu_2\\ \delta<0\rightarrow\mu_1<\mu_2\\ \delta=0\rightarrow\mu_1=\mu_2\end{array}$

$d=\bar{x}_1-\bar{x}_2$ estimador de δ de comportamiento normal.

Primero vemos las **comparaciones** de ν

Tenemos $\nu_1^2 \;\; y \;\; \nu_2^2$; $\varphi^2=\frac{\nu_1^2}{\nu_2^2}$ si es $\begin{array}{l}\varphi^2=1\rightarrow\nu_1^2=\nu_2^2\\ \varphi^2<1\rightarrow\nu_1^2<\nu_2^2\\ \varphi^2>1\rightarrow\nu_1^2>\nu_2^2\end{array}$

Tengo también $q^2=\frac{S_1^2}{S_2^2}$; {demostración} ; luego $\frac{q^2}{\varphi^2}=F(1-\alpha, N_n=n_1-1 \;\; ; \;\; N_d=n_2-1)$

Entonces, estimamos φ^2 así: $P(A\leq\varphi^2\leq B)=1-\alpha$

$$A=\frac{\frac{S_1^2}{S_2^2}}{F(1-\frac{\alpha}{2};n_1-1;n_2-1)} \quad \text{y} \quad B=\frac{\frac{S_1^2}{S_2^2}}{F(\frac{\alpha}{2};n_1-1;n_2-1)}$$

Si quiero estimar φ ; entonces: $P(A'\leq\varphi\leq B')=1-\alpha$

$A'=\sqrt{A}$ y $B'=\sqrt{B}$

Planteo de hipótesis – Único caso que se plantea

$$\begin{array}{l}\mathrm{H0})\,\varphi^2\leq\varphi_0^2\\ \mathrm{H1})\,\varphi^2>\varphi_0^2\end{array} \quad \text{con} \quad \alpha, n_1, n_2$$

$$q_c^2=\varphi_0^2 F(1-\alpha, n_1-1, n_2-1)$$

C.R : si $q^2>q_c^2 \rightarrow$ rechazo H0

Comparación de ν_n

$H0)\,\frac{\nu_1}{\nu_2}\leq 1$; $q_c^2=F(1-\alpha, n_1-1, n_2-1)$ (ver ejemplo marcado en la carpeta ejercicio 5-6...)

$d=\bar{x}_1-\bar{x}_2$ V.A Normal } Atención, aparece luego en las formulas que siguen

$\delta=\mu_1-\mu_2 \quad = \quad \mu_d \quad = \quad \mu_{\bar{x}_1}-\mu_{\bar{x}_2}$ } Atención, aparece en las formulas que siguen

Varianzas: $\nu_d^2 \;=\; \nu_{\bar{x}_1}^2+\nu_{\bar{x}_2}^2 \;=\; \frac{\nu_1^2}{n_1}+\frac{\nu_2^2}{n_2}$

Desvíos : $\nu_d=\sqrt{\frac{\nu_1^2}{n_1}+\frac{\nu_2^2}{n_2}}$

$Z=\frac{d-\delta}{\sqrt{\frac{\nu_1^2}{n_1}+\frac{\nu_2^2}{n_2}}} \;=\; \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\nu_1^2}{n_1}+\frac{\nu_2^2}{n_2}}}$ } comportamiento de distribución normal estandarizada Z

Estimación de δ conociendo ν_1 y ν_2 (Desvíos poblacionales)

$P(A\leq\delta\leq B)=1-\alpha$; $B;A=d\pm Z(1-\frac{\alpha}{2})\sqrt{\frac{\nu_1^2}{n_1}+\frac{\nu_2^2}{n_2}}$ (la parte luego del +/- es el error muestral.)

Para poder despejar "n", hago la misma muestra en ambas $n_1=n_2=n$; y tomo factor común y despejo n.
Tengo dos poblaciones, $n_{1,}n_2$, con estimadores $\bar{x}_{1,}\bar{x}_{2,}S_{1,}S_2$

Atención: con una Estimación nunca puedo tomar una decisión, probar algo o similar. El único método es hacer un ensayo de hipótesis.

Ensayos de hipótesis con δ

La hipótesis se plantea en función de lo que quiero detectar, **no** hay criterio optimista/pesimista.
Poner alfa chico para asegurar que son distintos, grande para asegurar que son parecidos.

Caso 1:

$H_0)\,\delta\leq\delta_0\}\,\alpha\,,n_{1,}n_2$ (para demostrar que es mayor)

C.R : si $\frac{(\bar{x}_1-\bar{x}_2)-\delta_0}{\sqrt{\frac{\nu_1^2}{n_1}+\frac{\nu_2^2}{n_2}}} \;>\; Z(1-\alpha)$ entonces rechazo H0

Otra manera mas fácil, con d_c critico:
$d_c(Critico)=\delta_0+Z(1-\alpha)\sqrt{\frac{\nu_1^2}{n_1}+\frac{\nu_2^2}{n_2}}$; si d > d_c => rechazo H_0

Caso 2:

$H_0)\,\delta\geq\delta_0\}\,\alpha\,,n_{1,}n_2$ (para demostrar que es menor)

C.R : si $\frac{(\bar{x}_1-\bar{x}_2)-\delta_0}{\sqrt{\frac{\nu_1^2}{n_1}+\frac{\nu_2^2}{n_2}}} \;<\; Z(1-\alpha)$ entonces rechazo H0

Otra manera mas fácil, con d_c critico:
$d_c(Critico)=\delta_0-Z(1-\alpha)\sqrt{\frac{\nu_1^2}{n_1}+\frac{\nu_2^2}{n_2}}$; si d < d_c => rechazo H_0

Caso 3

$H_0)\ \delta = \delta_0\}\ \alpha, n_1, n_2$ (para demostrar que es distinto, no tiene sentido en la practica...)

Se hacen 2 d_criticos:

$d_c(Critico)=\delta_0 \pm Z(1-\frac{\alpha}{2})\sqrt{\frac{\nu_1^2}{n_1}+\frac{\nu_2^2}{n_2}}$; si d < d_c1(-) o d > dc2(+) => rechazo H_0

Explicación

Decir $\delta \leq \delta_0$ es lo mismo que decir $\mu_1-\mu_2 \leq \delta_0$, lo que hago es decir que la diferencia de u de las dos poblaciones es < o > que un valor δ_0 . Poniendo $\delta_0=0$, me doy cuenta si es <, > o =, sin saber en cuanto.

Atención: si quiero probar que el u1 supera en un cierto porcentaje o cantidad a u2, agrego una constante k, por ejemplo: k = 1,1 es decir 110% ; entonces : $\mu_1 > 1{,}1 \cdot \mu_2$

Generalizo: $H_0)\ \mu_1 - k \cdot \mu_2 \leq 0$

Parámetro: $\delta = \mu_1 - k\,\mu_2$

Estimador: $d = \bar{x}_1 - k\,\bar{x}_2$

$$\mu_d = \mu_{\bar{x}_1} - k\,\mu_{\bar{x}_2} = \delta$$

$\nu_d = \sqrt{\frac{\nu_1^2}{n_1} + k^2 \cdot \frac{\nu_2^2}{n_2}}$ } la variable k que se agrego aca, **debe** agregarse en todas las formulas de hipótesis antes vistas para poder usar esto.

Nota: en estos casos (generales, con o sin k), no puedo plantear un ensayo apriori de la experiencia, en estos casos, no se genera un procedimiento periódico, el valor critico y la comparación sirven solo para esta comparación. El criterio **no** es reusable!
Hay que calcularlo cada vez que se toman nuevas muestras.

Nota: a continuación, para saber si planteo $\nu_1=\nu_2$ o $\nu_1 \neq \nu_2$, uso Comparación con ensayo de hipótesis:

$H_0)\ \nu_1=\nu_2$

$H_1)\ \nu_1 \neq \nu_2$

C.R : Si $F_{calc}=\frac{S_1^2}{S_2^2} > F(1-\alpha, n_1-1, n_2-1) \rightarrow$ R.H0

y con esto determino cual usar.

NO LO HAGO SI SE DE ANTEMANO CON CERTEZA SI SON IGUALES O DISTINTOS!

Estimación de δ con ν_1 y ν_2 desconocidos y $\nu_1=\nu_2$

Nota: si conozco alguno de los dos v_n, lo descarto y hago de cuenta que no conozco ninguno.

Preciso estimar ν_1 y ν_2 ; en este primer caso, hago una suposición : v_1 = v_2
entonces, amalgamo la muestra, y obtengo un "S amalgamado"

$$S_a=\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$$

$t_n=\frac{d-\delta}{S_a\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ } t de student, grados de libertad (n_1+n_2-2)

Como estimar δ con ν_1, ν_2 desconocidos, pero suponiendo $\nu_1=\nu_2$

$$P(A \leq \delta \leq B)=1-\alpha \qquad B\,;A=(\bar{x}_1-\bar{x}_2)\pm\left[t(1-\frac{\alpha}{2}, n_1+n_2-2)S_a\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})}\right]$$

Ensayos de hipótesis

Caso 1:

$H_0) \delta \leq \delta_0 \} \alpha, n_1, n_2$ (para demostrar que es mayor)

C.R : si $\frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{S_a \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > t(1-\alpha, n_1 + n_2 - 2)$ entonces rechazo H0

Nota: Si quiero saber si es <, >,= sin importar el valor, pongo $\delta_0 = 0$

Con valor critico:

$d_c(Critico) = \delta_0 + t(1-\alpha, n_1 + n_2 - 2) S_a \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$; si d > d_c => rechazo H_0

Y así para todos los casos anteriores ya vistos, se cambia Z por T Student, y se reemplaza por Sa[...]

Si tiene k, es $S_d = S_a \sqrt{\frac{1}{n_1} + \frac{k^2}{n_2}}$; Sa se calcula igual con k o sin k.

Estimación de δ con ν_1 y ν_2 desconocidos y $\nu_1 \neq \nu_2$

$$P(A \leq \delta \leq B) = 1 - \alpha$$

$$B; A = d \pm \left[t(1 - \frac{\alpha}{2}, NAW) \sqrt{(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2})} \right]$$

Calcular **NAW** : $NAW = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{1}{(n_1 - 1)}\left(\frac{S_1^2}{n_1}\right)^2 + \frac{1}{(n_2 - 1)}\left(\frac{S_2^2}{n_2}\right)^2}$ <- SE REDONDEA HACIA **ABAJO**! (14,9 = 14)

Lo que cambia en todas las expresiones de planteo de hipótesis antes vistas:

Caso 1:

$H_0) \delta \leq \delta_0 \} \alpha, n_1, n_2$ (para demostrar que es mayor)

C.R : si $\frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} > t(1-\alpha, NAW)$ entonces rechazo H0

Con valor critico:

$d_c(Critico) = \delta_0 + t(1-\alpha, NAW) \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$; si d > d_c => rechazo H_0

y así en los demás casos.

Con k se hace : $Sd = \sqrt{\frac{S_1^2}{n_1} + k^2 \frac{S_2^2}{n_2}}$

Todo lo antes visto, sirve cuando las poblaciones estudiadas son **INDEPENDIENTES**

A CONTINUACION

TEORIA PARA CUANDO LAS POBLACIONES NO SON INDEPENDIENTES

$\delta = \mu_1 - \mu_2$; si es = 0, u1= u2 ; <0 => u1 < u2 ; >0 => u1 > u2

Estoy haciendo pruebas en la misma unidad experimental, en el mismo individuo.

La muestra es n , no $n_1 + n_2$; tamaño de muestra = n ; 2 pruebas sobre el mismo individuo.

Tengo 2 datos de un mismo individuo, se llevan ambos datos a 1 solo mediante un apareamiento de las muestras

diferencias apreciadas: $d_{ai} = (x_1 i - x_2 i)$

	t_1	t_2	d_{ai}
1	x_{11}	x_{21}	$(x_{11} - x_{21})$
2	x_{12}	x_{22}	$(x_{12} - x_{22})$
...	...	...	...
n	$x_1 n$	$x_2 n$	$(x_{1n} - x_{2n})$

Promedio: bar $\bar{d}_a = \frac{\sum d_{ai}}{n}$; Desvío = $S_{da} = \sqrt{\frac{\sum (d_{ai} - \bar{d}_a)^2}{n-1}}$

T. Student: $T_\eta = \frac{\bar{d}_a - \delta}{S \frac{d_a}{\sqrt{n}}}$; $\eta = n - 1$

Si hago experimentos sobre la misma unidad experimental, hay que aparear! Para pode aparear, se **precisan** los datos individuales. (Ver guia 6, problemas 14,13,12,etc)

Planteo hipótesis:

Para $\mu_1 > \mu_2$

$H_0) \mu_1 - \mu_2 < \delta_0 = 0$; alfa normal: 0,05 ; usualmente $\delta_0 = 0$; salvo que quiera comparar incrementos numericos determinados (ej: $\delta_0 = 0{,}1$, es decir, $\mu_1 > \mu_2$ en 0,1 unidades)

$\bar{d}_{ac} = \delta_0 + t(1-\alpha, n-1) \cdot \frac{S_{da}}{\sqrt{n}}$

C.R : si $\bar{d}_a > \bar{d}_{ac}$ => R. H0

<u>Contrastes chi-cuadrado</u> X^2

Comparo una situación/una realidad contra un patrón/base/parámetro de referencia.

Dos contrastes X2:

- **Pruebas de independencia**

  Probar o tratar de probar si una situación compleja es independiente de otra.
  La hipótesis en todos los casos es la misma:

  H0) La situación A es independiente de la situación B

  Jamas puedo probar la independencia, solo puedo probar que son dependientes. Se usan alfa medios o chicos, nunca grandes.

- **Pruebas de bondad de ajuste**
  Ver si la situación observada o experimentación realizada se ajusta a un modelo o ley.
  H0) La observación sigue la {ley o modelo}
  Si no rechazo, no puedo asegurar nada.
  Se usan alfa medios o chicos.

<u>Pruebas de independencia</u>

Busca datos que muestren relación entre A y B (para probar que son dependientes).

H0) La situación A es independiente de la situación B

Hacer una tabla de valores observados (la realidad).

B (filas)		j				Totales
A (columnas)		1	2	...	C	
i	1	O_11	O_12	...	O_1C	t_i1
	2	O_21	O_22	...	O_2C	t_i2
	...	...	...	...	...	...
	R	O_R1	O_R2	...	O_RC	t_iR
Totales		t_j1	t_j2	t_j...	t_jC	**t_T** (gran total)

El t_T tiene que dar lo mismo la fila que la columna ; sirven de **<u>control</u>** para ver si colocamos correctamente los datos.

En la fila de B se ponen las clasificaciones de B ; en la columna de A las clasificaciones de A ; y en o_ij se van poniendo los valores **observados**.

Luego, se hace una tabla **exactamente igual**, pero se colocan los **valores teóricos** (los que se deberían haber observado si A es independiente de B).
Esta tabla se llama "<u>tabla de valores esperados</u>" ; cada valor se llama E_ij.

Estos valores se pueden razonar, o se puede utilizar una formula apropiada.

La **formula** de los valores esperados es : $\frac{total_i \cdot total_j}{gran_{total}} = \frac{t_i \cdot t_j}{t_T}$

Luego, se contrastan los valores en ambas tablas, y se ve si son muy diferentes.
Si son muy diferentes, son dependientes.
Si no son tan diferentes, no puedo afirmar que son independientes.

Para observar si la diferencias son significativas, hago :

$$\sum_{i=1}^{R} \sum_{j=1}^{C} \left[\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \right] = X^2_{observado}$$

Condición de rechazo: si $X^2_{observado} > X^2(1-\alpha, (R-1)\cdot(C-1))$ rechazo H0.

Ejemplo:

	zona a	zona b	zona c	totales
mi empresa	...	...	...	...
competencia	...	...	...	...
totales	...	...	...	...

Se hacen dos tablas así, una con los valores observados, y otra con los esperados.

Pruebas de bondad de ajuste

H0) La observación sigue la ley o modelo

	F_Oi (observado)		F_Ei (esperado)
	i : V_Oi		V_Ei
	1 : V_O1		V_E1
Clases	2 : V_O2	Valores	V_E2
	...		...
	k : V_Ok		V_Ek
	sumatoria		sumatoria

La sumatoria es el n_{obs}
Nota: cada V_Ei es un valor esperado que lo da la distribución que se supone que ajusta; ej: normal, entonces prob de normal para ese intervalo*sumatoria = V_Ei. Ver ejemplo 7-9 de clase.

$$X_{obs}^2 = \sum_{i=1}^{k} \left[\frac{(V_{Oi} - V_{Ei})^2}{V_{Ei}} \right]$$

C.R: si $X_{obs}^2 > X^2(1-\alpha, k-1-p)$ entonces Rechazo H0
En la formula anterior, **k** es la cantidad de clases, y **p** es la cantidad de parámetros desconocidos a estimar ; puede ser 0.

Importante:
NO SE PUEDE aplicar bondad de ajuste con menos de 60 observaciones!
NO puede habar valores esperados menores a **5** (en realidad < a 4,...)
Si algún valor da menor a 5, se **amalgama** (suma V_o1 + V_o2 ; V_e1 + V_e2) ; si todavía no da mayor a 5, se sigue amalgamando con el cuadrito siguiente (o anterior).
Si amalgamo, la cantidad de clases cambia a la cantidad amalgamada.

Nota: si rechazo H0, puedo afirmar que **no** ajusta el modelo ; si no rechazo, estoy en problemas ; conviene para saber que modelo es el mejor, es ver que modelo no rechazado se ajusta mejor ; se ajusta mejor el que de el X_{obs}^2 menor.

Alfa estandard para usar en caso que no lo den: 0,05

Nota: si hay intervalos (ej: 20-30, etc) en la calculadora se toma el punto medio (ejemplo: 20-30, se toma 25).

Teoría de la regresión y análisis de correlación

Regresión: estudiar la variable que nos interesa en función de variables que la influyen.

$y = f(x_{1,} x_{2,} x_{3,} \ldots, x_n)$; y es la variable que nos interesa, x_n son los factores que la influyen.

El análisis de correlación permite ver si el modelo encontrado es valido.

Modelos (hay lineales, y no lineales, los no lineales no se ven en este curso):

$y = f(x_{1,} x_{2,} x_{3,} \ldots, x_n)$ } lineal múltiple

$y = f(x)$ } lineal simple

Cualquier modelo no lineal se puede estudiar como lineal utilizando una transformación apropiada.

Modelo lineal simple

$Y=\beta_0+\beta_1 x$ Modelo de regresión simple lineal poblacional promedio

(significa que da el comportamiento de todo "y" para cualquier valor de "x", dando el comportamiento promedio de "y")

β_1 es el coeficiente de regresión.

$y=\beta_0+\beta_1 x+\epsilon$ Modelo de regresión simple lineal poblacional puntual (valor puntual de y)

épsilon no pertenece al modelo; es la perturbación del entorno, algo que ocurre que hace que el modelo no se cumpla (un imprevisto no es épsilon), son cosas que no se espera que ocurran dentro del modelo.

Elementos de $Y=\beta_0+\beta_1 x$ (regresión simple lineal poblacional promedio)
Y = variable a explicar (dependiente); es aleatoria
x = (independiente) ; variable explicativa, explica el comportamiento de Y, puede o no ser aleatoria, exogena (la manipula el ambiente, aleatoria), o endógena (no aleatoria)

β_0 = (ordenada al origen) ; el valor que toma Y cuando x tiene una influencia nula.

β_1 = (tangente) ; incremento de Y para un incremento unitario de x ; llamado en economía "contribución marginal". Se llama coeficiente de regresión ; si es igual a 0, no hay coeficiente de regresión.

$\beta_0\, y\, \beta_1$ son parámetros, son desconocidos y hay que estimarlos estadísticamente.

Recta de **mínimos cuadrados:** $\hat{y}=b_0+b_1 x$ es variable aleatoria, es estimador de $Y=\beta_0+\beta_1 x$
Si es valido, entonces $\mu\cdot\hat{y}=y=\beta_0+\beta_1 x$

$S=\sqrt{\frac{Q}{N-2}}$; N es la cantidad de datos muestreados, precisa mas de 2.

$\eta=N-2$ } numero de grados de libertad

Para hallar el modelo lineal promedio, hay que hallar el baricentro de la nube de puntos $(\bar{x},\bar{y})$

$Q=\sum\left[(y_i-b_0-b_1 x_i)^2\right]$; entonces; $S=\sqrt{\frac{\sum y^2-b_0\sum y-b_1\sum y\cdot x}{N-2}}$

$$\sum y=b_0 N+b_1\sum x \quad \sim\sim \quad \sum yx=b_0\sum x+b_1\sum x^2$$

$$b_0 = \frac{\sum y-b_1\sum x}{N} = \bar{y}-b_1\bar{x} \quad \sim\sim \quad b_1 = \frac{\sum yx-\frac{\sum y\sum x}{N}}{\sum x^2-\frac{(\sum x)^2}{N}} = \frac{\sum yx-N\,\bar{y}\,\bar{x}}{\sum x^2-N\,\bar{x}^2}$$

$$\bar{x}=\frac{\sum x}{N} \quad \sim\sim \quad \bar{y}=\frac{\sum y}{N}$$

Con los datos, hay que armar una tabla así : [IMPORTANTE: identificar correctamente y,x ; "y" es la que esta en función de "x", ej: y=costo del viaje, x = distancia a viajar]

	x	y	x^2	y^2	yx
1 .. N	..	..	..	..	..
	$\sum x$	$\sum y$	$\sum x^2$	$\sum y^2$	$\sum y x$

[continua en la próxima hoja]

Una vez que tenemos $b_0 \ y \ b_1$, tenemos el modelo, pero **hay que validarlo antes de usarlo!**
Si el modelo es valido, $b_0 \ y \ b_1$ estiman bien $\beta_0 \ y \ \beta_1$; $\hat{y}$ estima bien Y e y.

Validar el modelo
Se realiza el coeficiente de correlación poblacional ρ (rho); mide la relación entre variables, oscila entre 1 y -1
0 = ausencia de relación
1 = relación total directa
-1 = relación total indirecta
A ρ no lo tenemos, lo estimamos con **r**

$$r=\frac{\sum yx-N\bar{y}\bar{x}}{\sqrt{(\sum y^2-N\bar{y}^2)(\sum x^2-N\bar{x}^2)}}$$; **-1 <= r <= 1**

si **|r| >= 0,7** y si hay relación, se puede considerar que esa relación es OK. (notar que es una condición necesaria, pero NO suficiente para validar el modelo).

Luego, se realiza el coeficiente de determinación (ρ^2)
Mide el % de influencia de la variable x en la explicación de y ; lo estimamos con $r^2=(r)^2$; va entre 0 y 1 (0% al 100% de influencia) , mas abajo se explica como estimarlo.

Probamos si hay relación entre variables (planteo hipótesis)

H0) p = 0
H1) p <> 0

Usar un alfa grande, tipo 0,10
Tengo que saber como se comporta r ; r es una V.A
Si p = 0, se comporta como T de Student ; si p <> 0 se comporta como log normal.

Como **H0) p=0, asumo T. Student**

$$t_{obs}=\frac{r}{\sqrt{\frac{1-r^2}{N-2}}}$$ => C.R : si $|t_{obs}|>t(1-\alpha, N-2)$ rechazo H0, entonces existe relación lineal entre variables, luego verificamos que |r| >= 0,7, y recién hay aceptamos el modelo como valido.

Notar que se precisan estas dos condiciones a la vez para aceptar el modelo: rechazar H0, y |r|>=0,7
Una vez validado el modelo, puedo:

Para **estimar** ρ (rho) ; **fuerza** de la relación entre las variables

$P(A\leq\rho\leq B)=1-\alpha$ (Log Normal, p'q' p<>0)

$$Z_r=\frac{1}{2}\ln(\frac{1+r}{1-r}) \sim\sim Z_{r\,\text{max\_B;min\_A}}=Z_r\pm Z(1-\frac{\alpha}{2})\sqrt{\frac{1}{N-3}}$$

y con estos datos, calculo A y B

$$A=\frac{e^{2zrmin}-1}{e^{2zrmin}+1} \quad B=\frac{e^{2zrmax}-1}{e^{2zrmax}+1}$$

Si quiero ρ^2 (**influencia porcentual** %), elevo A y B al cuadrado.

Estimar β_1 ; contribución marginal o coeficiente de regresión

$$P(A\leq\beta_1\leq B)=1-\alpha$$

$$B;A=b_1\pm t(1-\frac{\alpha}{2}, N-2)Sb_1$$

Ensayo de hipótesis para el coeficiente de regresión β_1

(es lo mismo que hacer el de H0) $\rho=0$)

Caso **1**:
H0) $\beta_1=0$

$t_{obs}=\frac{b_1}{Sb_1}$; C.R : si tobs > $t(1-\alpha, N-2)$ rechazo H_0

b_1 = v.a = t. student

$$\mu b_1=\beta_1 \quad \sim\sim \quad Sb_1=\frac{S}{\sqrt{\sum x^2-N\bar{x}^2}}$$

(el S es el $S=\sqrt{\frac{Q}{N-2}}$ de paginas anteriores)

$$t_n=\frac{b_1-\beta_1}{Sb_1} \qquad \eta=N-2$$

Caso **2**:

$\text{H0)}\,\beta_1\leq\beta_{1_0}$

$\beta_{1_c}=\beta_{1_0}+t(1-\alpha, N-2)Sb_1$

C.R: si $b_1>b_{1_c}$ rechazo H0

Caso **3**:

$\text{H0)}\,\beta_1\geq\beta_{1_0}$

$\beta_{1_c}=\beta_{1_0}-t(1-\alpha, N-2)Sb_1$

C.R: si $b_1<b_{1_c}$ rechazo H0

Estimando valor de $Y=\beta_0+\beta_1 x$ (media/promedio)

$x=x_0$; $\hat{y}_0=b_0+b_1 x_0$

Y para X=x_0 esta entre A y B con un nivel de confianza $1-\alpha$

$P(A\leq Y(x=x_0)\leq B)=1-\alpha$

(aquí el S es también el S = raíz de Q blah blah)

$$B;A=\hat{y}_0\pm t(1-\frac{\alpha}{2};N-2)\cdot S\sqrt{\frac{1}{N}+\frac{(x_0-\bar{x})^2}{\sum x^2-N\bar{x}^2}}$$ intervalo de estimación de la ley promedio

Nota: el modelo es valido dentro del entorno de observación, o sea entre el x mínimo y máximo de la muestra ; para valores fuera, es impredecible ; por eso se agrega el termino que mide la lejanía al entorno de observación.

Estimar el valor puntual de $y=\beta_0+\beta_1 x+\epsilon$

al de la ley promedio, le agrego el error E

$P(A\leq y(x=x_0)\leq B)=1-\alpha$ intervalo de predicción/pronostico

$$B;A=\hat{y}_0\pm t(1-\frac{\alpha}{2};N-2)\cdot S\sqrt{\frac{1}{N}+\frac{(x_0-\bar{x})^2}{\sum x^2-N\bar{x}^2}+1}$$ (idem anterior, pero con + 1)

Cuando usarlos: si tiene la palabra "medio/promedio/etc", uso el primero ; si no tiene, uso el segundo.

Para **estimar** β_0 ; uso las ecuaciones anteriores con $x_0=0$, y estimo b_0 y β_0

Análisis de sensibilidad

[debug : no agregado : ver apéndice de la guia que dio el profesor, pagina 53 también]

Lo que se hace : Análisis de sensibilidad : elegir de un listado de modelos candidatos, el mejor. Luego se hace el test global, y los test individuales sobre el mejor modelo elegido objetivamente.

Usualmente, se usa una tabla similar a esta:
(ejemplo con modelo de 3 variables explicativas : $y=\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\epsilon$

modelos candidatos y sus parámetros, analizo sensiblemente:

<table>
<tr><th>y=f(...)</th><th>R^2</th><th>S^2</th><th>PRESS</th><th>DET</th></tr>
<tr><td>x_1</td><td>0,78</td><td>0,22</td><td rowspan="7">busco el mas chico</td><td rowspan="7">busco uno mayor a 0,1, lo mas cercano a 1 posible.</td></tr>
<tr><td>x_2</td><td>0,76</td><td>0,23</td></tr>
<tr><td>x_3</td><td>0,69</td><td>0,25</td></tr>
<tr><td>x_1 ; x_2</td><td>aumenta</td><td>aumenta o disminuye</td></tr>
<tr><td>x_1 ; x_3</td><td>aumenta</td><td>aumenta o disminuye</td></tr>
<tr><td>x_2 ; x_3</td><td>aumenta</td><td>aumenta o disminuye</td></tr>
<tr><td>x_1 ; x_2 ; x_3</td><td>aumenta</td><td>aumenta o disminuye</td></tr>
</table>

R^2 siempre aumenta
S^2 : si aumenta, significa que la incorporación de esa variable distorsiona el modelo (aumentando la dispersión) ; por lo tanto, descartamos la variable.

PRESS: suma de las predicciones cuadradas [calcularlo : ver carpeta] ; mide la capacidad del modelo para predecir , cuanto mas chico, mejor.

DET : matriz con las variables explicativas, si el DET es 1, es el mejor, preferimos el mas cercano a 1 posible, (siempre que sea > 0,1)

Si tengo varios modelos candidatos, elijo el mas sencillo, el que tenga menos variables.
Luego de seleccionar el modelo, hago el test global y eso.

Una vez elegido el modelo:
Coeficiente de determinación parcial: indica la influencia porcentual (%) de la variable x en la explicación de la variable y cuando están funcionando otras variables. [para calcularlo, ver la carpeta]

Análisis de varianza ANOVA unifactorial

Hay un factor que hace cambiar el comportamiento de lo que estamos estudiando. Un solo factor influye sobre el sistema.
Se toman muestras por cada nivel del factor evaluado.

j -> (fila, datos)	1	...	n_j
i (columna)			
1			
2			
...			
k			
grupos (columna)			

<- datos agrupados por fila

-> de cada uno sale $\bar{x}$ y S

k = cantidad de grupos

Por cada fila se calcula

$n_1 ; \bar{x}_1 ; S_1$

$n_2 ; \bar{x}_2 ; S_2$

$n_3 ; \bar{x}_3 ; S_3$

Suma de todos los datos: $N = \sum_{j=1}^{k} n_j$

Saco el promedio de los promedios (x doble raya): $\bar{\bar{x}} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n} x_{ij}}{N} = \frac{\sum_{j}^{k} n_j \bar{x}_j}{N}$

Planteo de hipótesis ANOVA

H0) $\mu_1 = \mu_2 = \mu_3 = ... = \mu_k$

$$F_{obs} = \frac{S_b^2}{S_w^2} = \frac{\text{varianza entre grupos}}{\text{varianza dentro de los grupos}}$$

C.R: si $F_{obs} > F(1-\alpha, k-1, N-k)$ => Rechazo H0

$$S_b^2 = \frac{\sum \left[n_j (\bar{x}_j - \bar{\bar{x}})^2 \right]}{k-1}$$

$$S_w^2 = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_j} \left[x_{ij} - \bar{x}_j \right]^2}{N-k} = \frac{\sum_{j=1}^{k} \left[(n_j - 1) S_j^2 \right]}{N-k}$$

Con esto, ya esta, si rechazo la hipótesis, significa que el factor influye, si no rechazo, no afirmo nada.

(Ver ejemplo en la **ultima** hoja de la guia)