Ejercicios Resueltos y Propuestos
Curso EYP 2113
Tomo II

Primera Edición

Trabajo de Recopilación , Organización y Elaboración
Patricia Jiménez P. & Ricardo Olea O.
Dpto. de Estadística - Facultad de Matemáticas
Pontificia Universidad Católica de Chile

Santiago, Diciembre 2004

Prefacio

Con la intención de apoyar la labor docente que desarrolla el Departamento de Estadística de la Facultad de Matemáticas de la Pontificia Universidad Católica de Chile, se ha realizado un trabajo de recopilación y elaboración de ejercicios resueltos y propuestos para el curso EYP2113, algunos de los cuales fueron desarrollados en ayudantías y han sido parte de interrogaciones en semestres anteriores.

Queremos agradecer muy en especial a FONDEDOC, por haber confiado en este proyecto y habernos entregado todo su apoyo para poder ver realizada esta necesidad tanto para el Departamento de Estadística, como para todos los alumnos y alumnas que son beneficiados de los cursos de servicio que ofrece el mismo.

Este trabajo ha sido fruto de la labor que desarrollaron docentes y ayudantes que dictaron el curso durante los años 2002 y 2003.

Específicamente deseamos agradecer a los profesores

- Ricardo Aravena
- Alejandro Jara
- Ignacio Vidal

Además quisiéramos agradecer el aporte de Jorge González, Mario Tagle y Joaquín Rojas, tanto por el material donado, como por la revisión de este libro.

Atentamente.

Dirección
Departamento de Estadística
Facultad de Matemáticas

Santiago, Diciembre 2004

Índice general

4. Test de Hipótesis - Dos Muestras **103**
4.1. Ejercicios Resueltos 103
4.1.1. Test de Hipótesis 103
4.1.2. Test de Homogeneidad e Independencia 127
4.2. Ejercicios Propuestos 142

5. Regresión Clásica **149**
5.1. Ejercicios Resueltos 149
5.2. Ejercicios Propuestos 201

6. Ejercicios Resueltos de Interrogaciones **211**
6.1. Interrogaciones III 211
6.2. Soluciones 216

A. Formulario de Distribuciones **I**

B. Formulario de Análisis de Regresión Simple **III**

C. Tablas de distribución **VII**
C.1. Distribución t de Student VII
C.2. Distribución χ^2 VIII
C.3. Distribución F ($\alpha = 0{,}05$) IX
C.4. Distribución Normal XI

Capítulo 4

Test de Hipótesis - Dos Muestras

4.1. Ejercicios Resueltos

4.1.1. Test de Hipótesis

EJERCICIO 37

Dos centrales telefónicas pertenecientes a una prestigiosa tienda comercial, reciben a diario cierta cantidad de llamadas, referentes a quejas de carácter comercial en una y en la otra quejas de carácter técnico. De una muestra de tamaño 10 días, se obtuvieron las siguientes medias

Tipo de Llamada	Media
Comercial	15
Técnico	20

Asumiendo que el número de quejas en ambas centrales es Poisson con parámetros λ_c y λ_t respectivamente, obtenga el TRV para docimar la hipótesis $H_0 : \lambda_c = \lambda_t$ versus $H_1 : \lambda_c \neq \lambda_t$. Utilice $\alpha = 0{,}05$.

SOLUCIÓN

Cuando se trabaja con dos poblaciones y asumiendo que las dos poblaciones son independientes, el TRV es de la forma:

$$\Lambda = \frac{L_{\hat{\lambda}}(\boldsymbol{x}, \boldsymbol{y})}{L_{\hat{\lambda_1}, \hat{\lambda_2}}(\boldsymbol{x}, \boldsymbol{y})}$$

En donde se tiene que

$$\hat{\boldsymbol{\lambda}} = \frac{\partial}{\partial \lambda} \ln(L_\lambda(\boldsymbol{x}) \cdot L_\lambda(\boldsymbol{y}))$$

$$\hat{\lambda}_i = \left[\frac{\partial}{\partial \lambda_i} \ln(L_{\lambda_1}(\boldsymbol{x}) \cdot L_{\lambda_2}(\boldsymbol{y})) = 0 \right]$$

Luego en este caso de la Poisson, queda:

$$\hat{\lambda} = \left[\frac{\partial}{\partial \lambda} \ln \left(\prod_{i=1}^{n} \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \prod_{i=1}^{n} \frac{e^{-\lambda} \lambda^{y_i}}{y_i!} \right) = 0 \right]$$

$$\longrightarrow \frac{\partial}{\partial \lambda} \ln \left(\frac{e^{-n\lambda} \lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!} \frac{e^{-n\lambda} \lambda^{\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} y_i!} \right)$$

$$\longrightarrow \frac{\partial}{\partial \lambda} \ln \left(\frac{e^{-2n\lambda} \lambda^{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} x_i! \prod_{i=1}^{n} y_i!} \right)$$

$$\longrightarrow \frac{\partial}{\partial \lambda} \left(-2n\lambda + \left(\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i \right) \ln(\lambda) - \sum_{i=1}^{n} \ln(x_i!) - \sum_{i=1}^{n} \ln(y_i!) \right)$$

$$\longrightarrow -2n + \frac{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}{\lambda} = 0$$

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}{2n}$$

Ahora los EMV bajo la hipótesis H_1, se obtienen de la siguiente forma:

$$\hat{\lambda_c} = \left[\frac{\partial}{\partial \lambda_c} \ln \left(\prod_{i=1}^{n} \frac{e^{-\lambda_c} \lambda_c^{x_i}}{x_i!} \prod_{i=1}^{n} \frac{e^{-\lambda_t} \lambda_t^{y_i}}{y_i!} \right) = 0 \right]$$

$$\longrightarrow \frac{\partial}{\partial \lambda_c} \ln \left(\frac{e^{-n\lambda_c} \lambda_c^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!} \frac{e^{-n\lambda_t} \lambda_t^{\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} y_i!} \right)$$

$$\longrightarrow \frac{\partial}{\partial \lambda_c} \left(-n\lambda_c + \sum_{i=1}^{n} x_i \ln(\lambda_c) - n\lambda_t + \sum_{i=1}^{n} y_i \ln(\lambda_t) - \sum_{i=1}^{n} \ln(x_i!) - \sum_{i=1}^{n} \ln(y_i!) \right)$$

$$\longrightarrow -n + \frac{\sum_{i=1}^{n} x_i}{\lambda_c} = 0$$

$$\hat{\lambda}_c = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

Y por simetría de la verosimilitud, se tiene que $\hat{\lambda}_t = \frac{\sum_{i=1}^{n} y_i}{n} = \bar{y}$.

Luego el TRV queda de la siguiente forma, al reemplazar los EMV encontrados.

$$\Lambda = \frac{\exp\left\{-2n\dfrac{\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}y_i}{2n}\right\}\left(\dfrac{\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}y_i}{2n}\right)^{\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}y_i}}{\exp\left\{-n\dfrac{\sum_{i=1}^{n}x_i}{n}\right\}\left(\dfrac{\sum_{i=1}^{n}x_i}{n}\right)^{\sum_{i=1}^{n}x_i}\exp\left\{-n\dfrac{\sum_{i=1}^{n}y_i}{n}\right\}\left(\dfrac{\sum_{i=1}^{n}y_i}{n}\right)^{\sum_{i=1}^{n}y_i}}$$

$$\Lambda = \frac{\left(\dfrac{\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}y_i}{2n}\right)^{\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}y_i}}{\left(\dfrac{\sum_{i=1}^{n}x_i}{n}\right)^{\sum_{i=1}^{n}x_i}\left(\dfrac{\sum_{i=1}^{n}y_i}{n}\right)^{\sum_{i=1}^{n}y_i}}$$

Luego considerando que $-2\ln(\Lambda) \sim \chi^2_{(m)}$ y reemplazando los datos entregados, tenemos que:

$$-2\ln(\Lambda) = -2\left[\left(\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}y_i\right)\ln\left(\frac{\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}y_i}{2n}\right)-\sum_{i=1}^{n}x_i\ln\left(\frac{\sum_{i=1}^{n}x_i}{n}\right)-\sum_{i=1}^{n}y_i\ln\left(\frac{\sum_{i=1}^{n}y_i}{n}\right)\right]$$

$$= -2[350\ln(17{,}5) - 150\ln(15) - 200\ln(20)]$$

$$= 7{,}167$$

Luego como $-2\ln(\Lambda) = 7{,}167 > 3{,}843 = \chi^2_{0{,}05}(1)$ se rechaza H_0.

EJERCICIO 38

Se investiga el diámetro de las varillas de acero fabricadas por dos máquinas diferentes de extrusion. Para ello se toman dos muestras aleatorias de tamaño $n_1 = 15$ y $n_2 = 18$; las medias y las varianzas muestrales son $\bar{x} = 8{,}73$, $S_1^2 = 0{,}35$, $\bar{x}_2 = 8{,}68$ y $S_2^2 = 0{,}4$, respectivamente.

(a) Suponga que $\sigma_1^2 = \sigma_2^2$. Construya un intervalo de confianza bilateral del 95 % para la diferencia en el diámetro promedio de la varilla.

(b) Construya un intervalo de confianza bilateral del 95 % para el cuociente de las varianzas poblacionales $\frac{\sigma_1^2}{\sigma_2^2}$. Parece razonable concluir que las varianzas son iguales?

(c) Pruebe la hipótesis $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$. Utilice $\alpha = 0{,}05$ y obtenga conclusiones.

(d) Calcule el valor p aproximado para esta prueba.

SOLUCIÓN

(a) Para el caso dado, tenemos el siguiente pivote:

$$\frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

donde

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

Por lo tanto reemplazando los datos entregados, en el pivote y en S_p^2 correctamente, se tiene que $S_p^2 = 0{,}38$, $t_{15+18-2,1-\frac{\alpha}{2}} = 2{,}042$ y

$$IC(\mu_1 - \mu_2) = 8{,}73 - 8{,}68 \pm \sqrt{0{,}38}\sqrt{\frac{1}{15} + \frac{1}{18}} \cdot 2{,}042$$

Luego se obtiene que con un 95 % de confianza la diferencia de los diámetros promedios de las varillas se encuentra en

$$\mu_1 - \mu_2 \in [-0{,}39, 0{,}49]$$

Note que el $0 \in$ al intervalo, luego esto quiere decir que las medias se pueden considerar iguales con un 95 % de confianza.

(b) Para el caso dado, tenemos el siguiente pivote:

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}$$

Luego el intervalo queda de la forma

$$IC\left(\frac{\sigma_2^2}{\sigma_1^2}\right) = \left[\frac{S_2^2}{S_1^2}F_{n_1-1,n_2-1,\alpha/2}, \frac{S_2^2}{S_1^2}F_{n_1-1,n_2-1,1-\alpha/2}\right]$$

Luego reemplazando se tiene

$$\frac{\sigma_2^2}{\sigma_1^2} \in [0{,}393, 3{,}109]$$

Note que el 1 pertenece al intervalo, luego con un 95 % de confianza, se puede decir que las varianzas son iguales.

(c) Para tal test de hipótesis y considerando los resultados de las letra (b), el estadístico de prueba es

$$T = \frac{\bar{x} - \bar{y}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

El cual rechaza la hipótesis nula si $T > t_{n_1+n_2-2,1-\alpha/2}$ o bien $T < -t_{n_1+n_2-2,1-\alpha/2}$.

Luego reemplazando y evaluando se tiene que

$$T = \frac{\bar{x} - \bar{y}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$= \frac{8{,}73 - 8{,}63}{\sqrt{0{,}38}\sqrt{\frac{1}{15} + \frac{1}{18}}}$$

$$= 0{,}23$$

Luego como $T \not> 2{,}039$ y $T \not< -2{,}039$, no existe evidencia presente en los datos para rechazar H_0.

(d) El $Valor-p = P(Z > 0{,}23) = 0{,}492$ y como éste es mayor que 0.05 (α), no se rechaza H_0.

EJERCICIO 39

Los siguientes datos fueron recabados en un experimento diseñado para verificar si existe diferencia sistemática en los pesos obtenidos con dos balanzas diferentes.

	Peso en Gramos	
Roca	Balanza 1	Balanza 2
1	11.23	11.27
2	14.36	14.41
3	8.33	8.35
4	10.50	10.52
5	23.42	23.41
6	9.15	9.17
7	13.47	13.52
8	6.47	6.46
9	12.40	12.45
10	19.38	19.35

Pruebe si la diferencia de las medias de los pesos obtenidos con las balanzas es significativa.

SOLUCIÓN

En este caso lo que tenemos son muestras pareadas y lo que se pide es testear las siguientes hipótesis

$$H_0 : \mu_x = \mu_y \quad vs \quad H_1 : \mu_x \neq \mu_y$$

Cuyo estadístico de prueba es

$$T = \frac{\overline{X} - \overline{Y}}{S_D/\sqrt{n}} \sim t_{n-1}$$

donde $S_D^2 = \dfrac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1}$, $d_i = X_i - Y_i$ y $\bar{d} = \dfrac{\sum_{i=1}^{n} d_i}{n}$.

el cual rechaza H_0 si $T > t_{n-1,1-\alpha/2}$ o bien si $T < -t_{n-1,1-\alpha/2}$

Luego tenemos que

i	1	2	3	4	5	6	7	8	9	10
$X_i - Y_i$	−0,04	−0,05	−0,02	−0,02	0,01	−0,02	−0,05	0,01	−0,05	0,03

de donde se obtiene que $\bar{d} = -0{,}02$ y $S_D^2 = 0{,}02867^2$
Luego el estadístico queda

$$T = \frac{\overline{X} - \overline{Y}}{S_D/\sqrt{n}}$$

$$= \frac{12{,}871 - 12{,}891}{0{,}0286/\sqrt{10}}$$

$$= -2{,}2114$$

Por lo tanto, como $T \not> 2{,}26 = t_{9,0,975}$ y $T \not< -2{,}26 = -t_{9,0,975}$, no existe evidencia en los datos para rechazar H_0.

EJERCICIO 40

Una compañía de taxis está tratando de decidir si compra la marca A o la marca B de neumáticos para su flota de automóviles. Para estimar la diferencia entre las dos marcas, se lleva a cabo un experimento con 12 neumáticos de cada marca. Los números se utilizan hasta que se gastan. Los resultados son:

Marca	Media (Km)	Desv. Stand. (Km)
A	36.300	5.000
B	38.100	6.100

(a) Calcule un intervalo de confianza para $\mu_1 - \mu_2$, suponiendo que las poblaciones tienen distribución normal con varianzas iguales.

(b) Encuentre un intervalo de confianza para $\mu_1 - \mu_2$, si se asigna un neumático de cada compañía en forma aleatoria a las ruedas traseras de ocho taxis y se registran, en kilómetros las siguientes distancias:

Taxi	Marca A	Marca B
1	34.400	36.700
2	45.500	46.800
3	36.700	37.700
4	32.000	31.100
5	48.400	47.800
6	32.800	36.400
7	38.100	38.900
8	30.100	31.500

Asuma que las diferencias de las distancias están distribuidas aproximadamente en forma normal.

(c) Determine un intervalo de confianza de 90 % para σ_2^2/σ_1^2. ¿Qué puede concluir?

SOLUCIÓN

Tenemos que $n_A = n_B = 12$.

(a) I.C para $\mu_1 - \mu_2$ suponiendo varianzas iguales y desconocidas es el siguiente:

$$(\overline{X}-\overline{Y})-S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\,t_{(n_1+n_2-2,1-\frac{\alpha}{2})} \leq \mu_1-\mu_2 \leq (\overline{X}-\overline{Y})+S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\,t_{(n_1+n_2-2,1-\frac{\alpha}{2})}$$

Luego necesitamos

$$\begin{aligned} S_p^2 &= \frac{(n_1-1)S_2^2+(n_2-1)S_1^2}{n_1+n-2-2} \\ &= \frac{(12-1)5000^2+(12-1)6100^2}{12+12-2} \\ &= \frac{684310000}{22} \\ &= 31105000 \end{aligned}$$

De aquí obtenemos $S_p = \sqrt{31105000} = 5577{,}18$.

$\therefore$ el I.C. queda

$$(36300-38100)-5577{,}18\sqrt{\frac{1}{12}+\frac{1}{12}}\,t_{(22;0,975)} \leq \mu_1-\mu_2 \leq (36300-38100)+5577{,}18\sqrt{\frac{1}{12}+\frac{1}{12}}\,t_{(22;0,975)}$$

Reemplazando $t_{(22;0,975)} = 2{,}0738$, el I.C. al 95 % de Confianza para $\mu_1 - \mu_2$ es:

$$-6521{,}94 \leq \mu_1 - \mu_2 \leq 2921{,}94$$

(b) Lo que nos piden es un I.C. para datos pareados.

Primero que todo tenemos que calcular las diferencias para cada par de datos como sigue:

Taxi	Marca A	Marca B	d_i
1	34.400	36.700	-2300
2	45.500	46.800	-1300
3	36.700	37.700	-1000
4	32.000	31.100	900
5	48.400	47.800	600
6	32.800	36.400	-3600
7	38.100	38.900	-800
8	30.100	31.500	-1400

Un I.C. para $\mu_1 - \mu_2$ esta definido como:

$$\overline{d} - \frac{S_D}{\sqrt{n}} t_{(n-1;1-\frac{\alpha}{2})} \leq \mu_1 - \mu_2 \leq \overline{d} + \frac{S_D}{\sqrt{n}} t_{(n-1;1-\frac{\alpha}{2})}$$

donde,

$$\overline{d} = \overline{X}_A - \overline{X}_B = 37250 - 38362{,}5 = -1112{,}5$$

$$S_D^2 = \frac{\sum_{i=1}^{n}(d_i - \overline{d})^2}{n-1}$$

luego

$$S_D^2 = \frac{14808750}{7} = 2115535{,}71428$$

$$\Rightarrow \qquad S_D = 1454{,}488127$$

Ocupando un $\alpha = 0{,}05$ el valor para $t_{(n-1;1-\frac{\alpha}{2})}$ es $t_{(7;0{,}975)} = 2{,}3646$.
Reemplazando el I.C. para $\mu_1 - \mu_2$ queda:

$$-111{,}5 - \frac{1454{,}4881}{\sqrt{8}} \times 2{,}3646 \leq \mu_1 - \mu_2 \leq -111{,}5 + \frac{1454{,}4881}{\sqrt{8}} \times 2{,}3646$$

es decir, un I.C. al 95 % para $\mu_1 - \mu_2$ es

$$-2328{,}47 \leq \mu_1 - \mu_2 \leq 103{,}47$$

(c) Un I.C. para σ_2^2/σ_1^2 al 90 % esta determinado por:

$$\begin{aligned}
\tfrac{S_2^2}{S_1^2} F_{n_1-1;n_2-1;\frac{\alpha}{2}} &\leq \tfrac{\sigma_2^2}{\sigma_1^2} \leq \tfrac{S_2^2}{S_1^2} F_{n_1-1;n_2-1;1-\frac{\alpha}{2}} \\
\tfrac{6100^2}{5000^2} F_{11;11;0{,}05} &\leq \tfrac{\sigma_2^2}{\sigma_1^2} \leq \tfrac{6100^2}{5000^2} F_{11;11;0{,}95} \\
\tfrac{6100^2}{5000^2} \times 0{,}3548 &\leq \tfrac{\sigma_2^2}{\sigma_1^2} \leq \tfrac{6100^2}{5000^2} \times 2{,}82 \\
0{,}528 &\leq \tfrac{\sigma_2^2}{\sigma_1^2} \leq 4{,}1973
\end{aligned}$$

Como el 1 se encuentra en el I.C. no se rechaza que $\sigma_1^2 = \sigma_2^2$.

EJERCICIO 41

Dos tipos diferentes de aleación, A y B, se han utilizado para fabricar especímenes experimentales de un pequeño eslabón de tensión, empleado en cierta aplicación de ingeniería. Se determinó la resistencia máxima (en ksi) de cada espécimen y los resultados se resumen en la siguiente tabla de distribución de frecuencia.

	A	B
26-30	6	4
30-34	12	9
34-38	15	19
38-42	7	10
	40	42

Calcule un intervalo de confianza de 95 % para la diferencia entre las proporciones reales de todos los especímenes de aleaciones A y B que tengan una resistencia máxima de por lo menos 34 ksi.

SOLUCIÓN

Un I.C para las diferencias de proporciones esta definido por:

$$(\widehat{p}_A-\widehat{p}_B)-\sqrt{\frac{\widehat{p}_A(1-\widehat{p}_A)}{n_1}+\frac{\widehat{p}_B(1-\widehat{p}_B)}{n_2}}\times Z_{1-\frac{\alpha}{2}} \leq p_A-p_B \leq (\widehat{p}_A-\widehat{p}_B)+\sqrt{\frac{\widehat{p}_A(1-\widehat{p}_A)}{n_1}+\frac{\widehat{p}_B(1-\widehat{p}_B)}{n_2}}\times Z_{1-\frac{\alpha}{2}}$$

Mirando en la tabla los rangos, sumamos las frecuencias de los rangos que cumplen tener una resistencia mayor de 34 ksi, luego reemplazando $\widehat{p}_A = \frac{22}{40}$, $\widehat{p}_B = \frac{29}{42}$ y $Z_{0,975} = 1{,}96$.
$\therefore$ el I.C. al 95 % para $p_A - p_B$ es:

$$-0{,}348 \leq p_A - p_B \leq 0{,}067$$

Como el $0 \in$ al Intervalo, se puede decir con un 95 % de confianza que $p_A = p_B$.

EJERCICIO 42

Una firma decide estudiar una muestra aleatoria de 20 proyectos que envió para ser evaluados, tanto a consultores externos, como a su propio departamento de proyectos. Las variables medidas fueron

X: nº de días que demoro la evaluación.

Y: nº de variables consideradas en la evaluación.

Z: Consultor al que se le envió el proyecto

$$Z = \begin{cases} -1 & ; \text{Depto. de Evaluación} \\ 0 & ; \text{Robani Consultores} \\ 1 & ; \text{Tanaka Ltda.} \end{cases}$$

W: Costo de la evaluación (en U.F.)

Los resultados de este muestreo son:

N°	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
X	4	2	8	10	1	3	8	3	2	2	4	4	5	6	7	2	1	3	4	9
Y	3	1	6	8	3	2	6	2	1	1	4	4	4	7	10	3	2	4	5	10
Z	-1	-1	0	0	0	0	1	0	0	1	-1	-1	0	1	1	-1	-1	0	1	-1
W	40	30.5	80.3	68.5	24.7	40.5	90.6	38.5	50.4	50.2	60.1	60.8	70.9	80	90	30	27	40	50	40

Explicitando los supuestos necesarios:

(a) Estime con un 90 % de confianza el costo medio de los proyectos.

(b) Estime con un 90 % de confianza la proporción de proyectos cuyo costo fue inferior a 50 U.F. dado que no involucraron más de 6 variables y que fueron resueltos en un tiempo superior a 2 días.

(c) El Depto. de control afirma que el costo medio de enviar los proyectos a asesores externos es significativamente mayor que el de evaluarlos allí mismo. ¿Qué concluye usando $\alpha = 0{,}05$?

(d) Tanaka Ltda. Afirma que la proporción de proyectos que ellos evalúan, que toman más tiempo de más de 4 días, no es superior a la proporción de proyectos que evalúa Robani Consultores, que toman un tiempo más de 4 días, no es superior a la proporción de proyectos que evalúa Robani Consultores, que toman un tiempo más de 4 días. Concluya si la afirmación de Tanaka Ltda. es correcta. (Use $\alpha = 0{,}01$)

SOLUCIÓN

(a) $\text{IC}(\mu_W) = \overline{W} \mp t_{n-1;1-\frac{\alpha}{2}} \frac{S_W}{\sqrt{n}}$, donde $\alpha = 5\,\% = 0{,}05 \rightarrow t_{20-1;1-\frac{0{,}05}{2}} = t_{19;0{,}975} = 2{,}093$

Luego con

$\overline{W}$	S_W	S_W^2	n
53,15	20,948	438,828	20

tenemos que el $\text{IC}(\mu_W)$ es:

$$\text{IC}(\mu_W) = 53{,}15 \mp 2{,}093 \cdot \frac{20{,}978}{\sqrt{20}}$$

$$\Rightarrow \mu_W \in (43{,}346; 62{,}953)$$

(b) $IC(p) = \hat{p} \mp Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}}$, donde $\alpha = 10\,\% = 0{,}1 \rightarrow Z_{1-\frac{0{,}1}{2}} = Z_{0{,}95} = 1{,}645$

luego con

$\hat{p}$	$\hat{q}$	n
$\frac{1}{2}$	$\frac{1}{2}$	8

tenemos que el $\text{IC}(p)$ es:

$$IC(p) = \frac{1}{2} \mp 1{,}645 \cdot \frac{0{,}5 \cdot 0{,}5}{8}$$

$$\Rightarrow p \in (0{,}209; 0{,}790)$$

(c) Definamos primero:

E: Asesores externos
L: Asesores locales (internos)

Luego tenemos la siguiente tabla resumen

	n	$\overline{X}$	S	S^2
E	13	59,584	21,629	467,838
L	7	41,2	14,058	197,636

Sea

μ_E:costo medio asesores externos
μ_L:costo medio asesores locales (internos)

Las hipótesis son

$$\mathrm{H}_0 : \mu_E = \mu_L \qquad \text{vs} \qquad \mathrm{H}_1 : \mu_E > \mu_L$$

primero haremos un test de varianzas donde la hipótesis es:

$$\mathrm{H}_0 : \frac{\sigma_E^2}{\sigma_L^2} = 1 \qquad \text{vs} \qquad \mathrm{H}_1 : \frac{\sigma_E^2}{\sigma_L^2} \neq 1$$

donde el estadístico F es:

$$F = \frac{S_E^2}{S_L^2} = 2{,}367$$

Se rechaza H_0 si:

$$F > F_1 \vee F < F_2$$

donde F_1 y F_2 considerando $\alpha = 0{,}05$ son:

$$F_1 = F_{n_E-1;n_L-1;1-\frac{\alpha}{2}} = F_{12;6;0,975} = 5{,}37$$

$$F_2 = F_{n_E-1;n_L-1;\frac{\alpha}{2}} = F_{12;6;0,025} = \frac{1}{F_{6;12;0,975}} = \frac{1}{3{,}73} = 0{,}268$$

Como F no es mayor que F_1 ni menor que F_2, no existe suficiente evidencia bajo un 95 % de confianza para rechazar H_0, es decir, se pueden considerar las varianzas desconocidas pero iguales.

Ahora hacemos un test de diferencias de medias, donde el estadístico T_c es:

$$T_c = \frac{\overline{X}_E - \overline{X}_L}{S_p\sqrt{\frac{1}{n_E} + \frac{1}{n_L}}}$$

donde

$$S_p^2 = \frac{(n_E - 1)S_E^2 + (n_L - 1)S_L^2}{n_E + n_L - 2}$$

$$= \frac{12 \cdot 467{,}838^2 + 6 \cdot 197{,}636^2}{18}$$

$$= 158934{,}925$$

$$\Rightarrow S_p = 398{,}666$$

Reemplazando el estadístico queda:

$$T_c = \frac{59{,}584 - 41{,}2}{398{,}666 \cdot \sqrt{\frac{1}{13} + \frac{1}{7}}} = 0{,}098$$

Luego se rechaza H_0 si $T_c > t_{\nu,1-\alpha}$.

donde $t_{\nu,1-\alpha}$ es:

$$t_{\nu,1-\alpha} = t_{n_E+n_L-2,1-\alpha} = t_{18;0,95} = 1{,}734$$

como T_c no es mayor que $t_{\nu,1-\alpha}$, no existe suficiente evidencia para rechazar H_0, luego la opinión del Depto. no es correcta.

(d) la hipótesis para este caso es la siguiente:

$$\mathrm{H}_0 : p_T \leq p_R \qquad \text{vs} \qquad \mathrm{H}_1 : p_T > p_R$$

necesitaremos la siguiente información

$\hat{p}_T$	$\hat{p}_R$	n_T	n_R	$\hat{q}_T$	$\hat{q}_R$
0,8	0,25	5	8	0,2	0,75

El estadístico Z_c es

$$Z_c = \frac{\hat{p}_T - \hat{p}_R}{\sqrt{\frac{\hat{p}_T \cdot \hat{q}_T}{n_T} + \frac{\hat{p}_R \cdot \hat{q}_T}{n_R}}} = \frac{0{,}8 - 0{,}25}{\sqrt{\frac{0{,}8 \cdot 0{,}2}{5} + \frac{0{,}25 \cdot 0{,}75}{8}}} = 2{,}335$$

Se rechaza H_0 si $Z_c > Z_{1-\alpha}$.

$$Z_{1-\alpha} = Z_{1-0,001} = Z_{0,99} = 2{,}325$$

Como $Z_c > Z_{0,99}$, se rechaza H_0, por lo tanto Tanaka Ltda tiene razón.

EJERCICIO 43

En un estudio sobre hábitos de alimentación en pelícanos, se marcan 25 hembras y 11 machos, y se les rastrea por radio. La variable de interés es la distancia (en mts.) que recorren volando en una pasada, en busca de alimento. Se obtuvieron estos resultados:

Hembras:	Distancia Media	205 mts.
	Desv. Estándar	100 mts.
Machos:	Distancia Media	135 mts.
	Desv. Estándar	90 mts.

¿Puede afirmarse que el comportamiento es diferente, respecto a la distancia media recorrida?

SOLUCIÓN

Resumiendo tenemos:

$n_h = 25$	$\bar{x}_h = 205$	$s_h = 100$
$n_m = 11$	$\bar{x}_m = 135$	$s_m = 90$

Para contestar la pregunta, debemos hacer un test de hipótesis para la diferencia de medias, es decir:

$$H_0 : \mu_h - \mu_m = 0 \quad H_1 : \mu_h - \mu_m \neq 0$$

Pero para esto necesitamos saber el comportamiento de las varianzas en ambas poblaciones. Luego debemos probar si son iguales o no.

$$H_0 : \frac{\sigma_h^2}{\sigma_m^2} = 1 \quad H_1 : \frac{\sigma_h^2}{\sigma_m^2} \neq 1$$

Esta hipótesis, la rechazamos si:

$$\frac{s_h^2}{s_m^2} > F_{n_h-1,n_m-1,1-\frac{\alpha}{2}} \quad o \quad \frac{s_h^2}{s_m^2} < F_{n_h-1,n_m-1,\frac{\alpha}{2}}$$

con $\alpha = 0{,}05$, luego reemplazando nos preguntamos:

$$\frac{100^2}{90^2} > F_{24,10,0,975}? \quad o \quad \frac{100^2}{90^2} < F_{24,10,0,025}?$$

donde $F_{24,10,0,975} = 3{,}36$ y $F_{24,10,0,025} = \frac{1}{F_{10,24,0,975}} = 0{,}3788$, luego comparando, observamos que las desigualdades no se dan, por lo tanto no existe evidencia en los datos para rechazar

que las varianzas de ambas poblaciones son iguales.

Luego ahora docimamos nuestra hipótesis original, ya sabiendo que las varianzas poblaciones son iguales pero desconocidas, con el estadístico :

$$T = \frac{\bar{x}_h - \bar{x}_m}{S_p\sqrt{\frac{1}{n_h} + \frac{1}{n_m}}}$$

donde

$$S_p = \sqrt{\frac{(n_h - 1)s_h^2 + (n_m - 1)s_m^2}{n_h + n_m - 2}} = 97{,}165$$

Luego reemplazando en el estadístico, tenemos que:

$$T = \frac{205 - 135}{97{,}165\sqrt{\frac{1}{25} + \frac{1}{11}}} = 1{,}99$$

Ahora, H_0 la rechazamos si:

$$\frac{\bar{x}_h - \bar{x}_m}{S_p\sqrt{\frac{1}{n_h} + \frac{1}{n_m}}} > t_{\nu,1-\frac{\alpha}{2}} \quad o \quad \frac{\bar{x}_h - \bar{x}_m}{S_p\sqrt{\frac{1}{n_h} + \frac{1}{n_m}}} < -t_{\nu,1-\frac{\alpha}{2}}$$

con $\nu = n_h + n_m - 2$, luego $t_{\nu,1-\frac{\alpha}{2}} = t_{34,0,975} = 2{,}03$. Luego haciendo las comparaciones respectivas, concluimos que no existe evidencia en los datos para rechazar H_0, es decir, las distancias medias recorridas para ambos sexos, en los pelicanos no difieren significativamente.

EJERCICIO 44

De dos procesos de producción de plástico se seleccionaron de cada 10 especímenes en forma independiente. Las mediciones de resistencia fueron:

Plástico A	3.03	5.53	5.6	9.3	9.92	12.51	12.95	15.21	16.04	16.84
Plástico B	3.19	4.26	4.47	4.53	4.67	4.69	12.87	6.79	9.37	12.75

(a) Utilice la teoría normal para testear la hipótesis que no existe diferencia entre los procesos de producción.

(b) Realice un test no paramétrico.

(c) ¿Cual de los dos método prefiere?

SOLUCIÓN

(a) Bajo la teoría de Normalidad tenemos:

$$H_0 : \mu_A = \mu_B \quad vs \quad H_1 : \mu_A \neq \mu_B$$

lo cual se testea con el estadístico

$$T = \frac{\overline{X}_A - \overline{X}_B}{S_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} \sim t_{n_A+n_B-2}$$

el cual rechaza H_0 si $T > t_{n_A+n_B-2,1-\alpha/2}$ o bien si $T < -t_{n_A+n_B-2,1-\alpha/2}$.

Luego reemplazando tenemos que $S_p = \sqrt{18{,}096} = 4{,}25$ y por lo tanto

$$T = \frac{\overline{X}_A - \overline{X}_B}{S_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}$$

$$= \frac{10{,}693 - 6{,}75}{4{,}25\sqrt{2/10}}$$

$$= 2{,}075$$

Finalmente como $T = 2{,}075 \not> 2{,}10 = t_{0{,}975,18}$ y $T = 2{,}075 \not< -2{,}10 = -t_{0{,}975,18}$, no existe evidencia presente en los datos para rechazar H_0, es decir, con un 95 % de confianza, no existe diferencia entre los procesos.

(b) Test para muestras independientes

1. Ordenar las observaciones de menor a mayor, igualando el grupo al que pertenecen

3,03	3,19	4,26	4,47	4,53	4,67	4,69	5,53	5,6	6,79
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
9,3	9,37	9,92	12,51	12,75	12,78	12,95	15,21	16,04	16,84
(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)

2. Sumar los rangos por separado de cada grupo

$$T_X = (1) + (8) + (9) + (11) + (13) + (14) + (17) + (18) + (19) + (20) = 130$$

$$T_Y = (2) + (3) + (4) + (5) + (6) + (7) + (10) + (12) + (15) + (16) = 80$$

Ahora $E(T_Y) = \frac{n_2(n_2+n_1+1)}{2}$ y $Var(T_Y) = \frac{n_1 n_2(n_1+n_2+1)}{12}$, donde n_2 es el tamaño de la menor muestra.

Y las hipótesis en este caso son:

$$H_0 : \text{Las distribuciones de los datos son idénticos}$$
$$vs$$
$$H_1 : \text{Las distribuciones de los datos son distintas}$$

las cuales se dociman con el estadístico

$$Z_0 = \frac{T_Y - E(T_Y)}{\sqrt{Var(T_Y)}}$$
$$= \frac{80 - 105}{\sqrt{175}}$$
$$= -\frac{25}{13{,}23}$$
$$= -1{,}89$$

donde H_0 se rechaza si $|Z_0| > Z_{1-\alpha/2} = Z_{0{,}975} = 1{,}96$

En este caso $|-1{,}89| \not> 1{,}96$ luego no existe evidencia en los datos para rechazar H_0 con un 95 % de confianza.

(c) Se prefiere el test no paramétrico ya que hay muy pocas observaciones (n pequeño) y al graficar los datos en un histograma se ve que no hay normalidad, como se muestra a continuación.

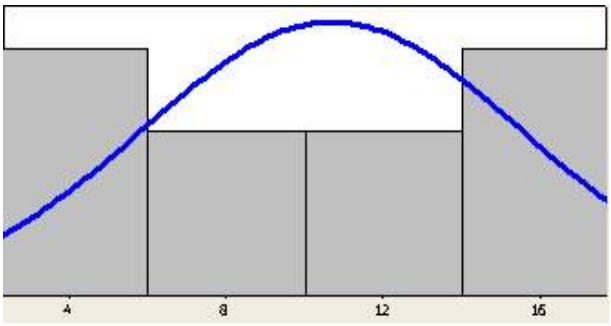

Figura 4.1: Plástico A, con curva Normal

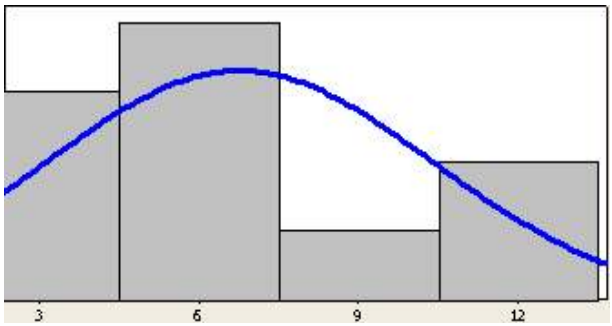

Figura 4.2: Plástico B, con curva Normal

EJERCICIO 45

Los siguientes datos se refieren a los efectos de un fármaco en la presión sanguínea de pacientes hipertensos. Los valores corresponden a la presión sistólica de los pacientes después del período placebo y después del tratamiento con la droga (se realizó una prueba cruzada, actuando cada paciente como su propio control).

Placebo	211	210	210	191	196	190	191	177	173	170	156
Droga	181	172	196	203	167	161	178	160	149	119	163

(a) ¿Sugieren estos datos que la nueva droga reduce significativamente el sistólico de la presión sanguínea?. Use $\alpha = 0{,}05$.

(b) Utilice un método no paramétrico.

SOLUCIÓN

(a) Bajo la teoría de normalidad tenemos que las hipótesis a testear son:

$$H_0 : \mu_P \leq \mu_D \quad vs \quad H_1 : \mu_P > \mu_D$$

Considerando que tenemos observaciones pareadas, es decir, dos observaciones a cada individuo (antes y después). Por lo tanto el estadístico a utilizar es:

$$T = \frac{\bar{x}_P - \bar{x}_D}{S_D/\sqrt{n}} \sim t_{n-1}$$

donde $\overline{d} = \overline{X}_A - \overline{X}_B$ y $S_D^2 = \frac{\sum_{i=1}^{n}(d_i - \overline{d})^2}{n-1} = 340{,}77$, el cual rechaza H_0, si $T > t_{n-1,1-\alpha}$.

$$T = \frac{\bar{x}_P - \bar{x}_D}{S_D/\sqrt{n}}$$

$$= \frac{188{,}6 - 168{,}1}{18{,}46/\sqrt{11}}$$

$$= 3{,}68$$

Por lo tanto como $T = 3{,}68 > 1{,}812 = t_{10,0,95}$, se rechaza H_0, es decir, con un 95 % de confianza la nueva droga reduce la presión.

(b) Para el test no paramétrico, los pasos a seguir son los siguientes:

1. Obtener las diferencias $D_i = X_i - Y_i$.
2. Ordenar por rango (de menor a mayor)las diferencias D_i.
3. Señalar los rangos, es decir, dar signos + o − de acuerdo a la diferencia original.
4. Calcular W_+, la suma de aquellos rangos con signo positivo.

Por lo tanto, obtenemos:

Placebo	211	210	210	191	196	190	191	177	173	170	156
Droga	181	172	196	203	167	161	178	160	149	119	163
D_i	30	38	14	-12	29	29	13	17	24	51	-7

Ahora, ordenamos de menor a mayor las diferencias y asignamos número con signo.

Orden	-7	-12	13	14	17	24	29	29	30	38	51
Signo	-1	-2	3	4	5	6	7	8	9	10	11

Con esto obtenemos que

$$W_+ = 3 + 4 + 5 + 6 + 7 + 8 + 9 + 10 + 11 = 63$$

$$W_- = 1 + 2 = 3$$

y las hipótesis para el caso no paramétrico, se traducen a:

H_0 : Las distribuciones de probabilidad para presión sistólica para los tratamientos son idéntica

$$vs$$

H_1 : Las distribuciones de probabilidad para presión sistólica para los tratamientos son distintas las cuales se dociman con el estadístico

$$Z_0 = \frac{W_+ - E(W_+)}{\sqrt{Var(W_+)}}$$

donde $E(W_+) = \frac{n(n+1)}{4} = 33$ y $Var(W_+) = \frac{n(n+1)(2n-1)}{24} = 126{,}5$.

Luego tenemos:

$$Z_0 = \frac{W_+ - E(W_+)}{\sqrt{Var(W_+)}}$$

$$= \frac{63 - 33}{\sqrt{126{,}5}}$$

$$= 2{,}67$$

donde H_0 se rechaza si $|Z_0| > Z_{1-\alpha/2} = Z_{0{,}975} = 1{,}96$

En este caso $|2{,}67| > 1{,}96$ luego se rechaza H_0 con un 95 % de confianza.

EJERCICIO 46

Un instructor de perros está entrenando a 27 animales para que obedezcan cierto mandato. El instructor utiliza dos técnicas de entrenamiento diferentes, una en la que recompensa y alimenta (I), y otra en la que no se da recompensa alguna (II). La tabla siguiente muestra el número de sesiones de obediencia que fueron necesarias antes de que un can obedeciera el mandato. ¿Tiene el instructor la evidencia suficiente para aseverar que el método de la recompensa requerirá, en promedio, menos tiempo de entrenamiento?. Plantee las hipótesis, llegue a conclusiones utilizando un nivel de significancia de $\alpha = 0{,}05$.

Entrenamiento I	29	27	32	25	27	28	23	31	37	28	22	24	28	31	34
Entrenamiento II	40	44	33	26	31	29	34	31	38	33	42	35			

(a) Asumiendo normalidad de los datos.

(b) Si no puede asumir normalidad de los datos.

SOLUCIÓN

(a) Tenemos que son dos muestras independientes, y por simplicidad asumiremos que las varianzas poblacionales de cada una de las muestras son iguales. Luego las hipótesis quedan:

$$H_0 : \mu_I \geq \mu_{II} \quad vs \quad H_1 : \mu_I < \mu_{II}$$

Las cuales se testean con el estadístico

$$T = \frac{\overline{X}_I - \overline{X}_{II}}{S_p\sqrt{\frac{1}{n_I} + \frac{1}{n_{II}}}} \sim t_{n_I+n_{II}-2}$$

donde $S_p = \sqrt{\frac{(n_I-1)S_I^2+(n_{II}-1)S_{II}^2}{n_I+n_{II}-2}} = 5$, el cual rechaza H_0 cuando $T < -t_{n_I+n_{II}-2,1-\alpha}$. Luego en este caso tenemos

$$T = \frac{\overline{X}_I - \overline{X}_{II}}{S_p\sqrt{\frac{1}{n_I} + \frac{1}{n_{II}}}}$$

$$= \frac{28{,}4 - 34{,}7}{5\sqrt{\frac{1}{15} + \frac{1}{12}}}$$

$$= -3{,}25$$

Por lo tanto como $T = -3{,}25 < -1{,}7 = t_{25,0,05}$ se rechaza H_0, es decir, el instructor tiene evidencia para aseverar que el método de la recompensa, requiere menos sesiones de entrenamiento.

(b) Debemos comparar las hipótesis $H_0 : \mu_I \geq \mu_{II}$ contra $H_1 : \mu_I < \mu_{II}$. Si no suponemos normalidad de las observaciones debemos aplicar un test de Mann-Whitney y para esto buscaremos los rangos de todos los datos.

Obs.	22	23	24	25	26	27	27	28	28
Rango	1	2	3	4	5	6.5	6.5	9	9
Obs.	28	29	29	31	31	31	31	32	33
Rango	9	11.5	11.5	14.5	14.5	14.5	14.5	17	18.5
Obs.	33	34	34	35	37	38	40	42	44
Rango	18.5	20.5	20.5	22	23	24	25	26	27

De aquí se tiene que

$$T_Y = 25 + 27 + 18{,}5 + 5 + 14{,}5 + 11{,}5 + 20{,}5 + 14{,}5 + 24 + 18{,}5 + 26 + 22 = 227{,}0$$

$$E(T_Y) = \frac{m(m+n+1)}{2} = \frac{12(12+15+1)}{2} = 168$$

$$Var(T_Y) = \frac{nm(n+m+1)}{2} = \frac{15 \cdot 12(15+12+1)}{2} = 420$$

Luego para docimar las hipótesis tenemos que

$$Z = \frac{T_Y - E(T_Y)}{\sqrt{Var(T_Y)}}$$

$$= \frac{227 - 168}{\sqrt{420}}$$

$$= 2{,}87$$

El cual rechaza H_0 si $Z = 2{,}87 > 1{,}645 = Z_{1-0{,}05}$. Por lo tanto en este caso se rechaza y se concluye que el método de recompensa tiene en promedio menos tiempo de entrenamiento.

EJERCICIO 47[1]

La siguiente tabla muestra los resultados obtenidos al realizar pruebas en referencia al tiempo que se requiere para llevar a cabo un procedimiento de ensamblaje con dos tipos de capacitación.

	Método de Enseñanza	
Estadística	Presencial	Virtual
n	16	9
Promedio	34	27
Desv. Estándar	5	3

(a) ¿Hay evidencia que indique que el método virtual es más eficiente, con respecto al presencial, en el sentido que reduce el tiempo medio de ensamblaje en más de 4 minutos?. Justifique.

[1] I2 II Sem. 03

(b) ¿Para qué niveles de α el test rechaza? ¿Cuál sería la forma de la curva de potencia?. Bosquéjela.

Asuma normalidad de los tiempos de ensamblaje.

SOLUCIÓN

(a) Las hipótesis apropiadas son

$$H_0 : \mu_p \leq \mu_v + 4 \quad vs \quad H_1 : \mu_p > \mu_v + 4$$

El estadístico de prueba para testear estas hipótesis es

$$T = \frac{\bar{x}_p - \bar{x}_v - 4}{S_p\sqrt{\frac{1}{n_p} + \frac{1}{n_v}}}$$

donde $S_p^2 = \frac{15 \cdot 5^2 + 8 \cdot 3^2}{16+9-2} = 19{,}435$ y rechaza H_0 cuando $T > t_{n_p+n_v-2,1-\alpha}$.

Por lo tanto en este caso tenemos

$$T = \frac{\bar{x}_p - \bar{x}_v - 4}{S_p\sqrt{\frac{1}{n_p} + \frac{1}{n_v}}}$$

$$= \frac{34 - 27 - 4}{4{,}41\sqrt{\frac{1}{16} + \frac{1}{9}}}$$

$$= 1{,}63$$

Luego como $T = 1{,}63 \not> 1{,}714 = t_{23,0,95}$ no existe evidencia presente en los datos para rechazar H_0, es decir, con un 95 % de confianza el método virtual no reduce el tiempo medio de ensamblaje en más de 4 minutos.

(b) Tenemos que el $Valor - p = P(t_{23} > 1{,}63)$ buscando en la tabla se observa que 1.63 $\in [t_{0,90} = 1{,}319; t_{0,95} = 1{,}714]$, luego el $0{,}05 < Valor - p < 0{,}10$.

Por lo tanto $\forall \alpha > valor - p$ se rechaza H_0.

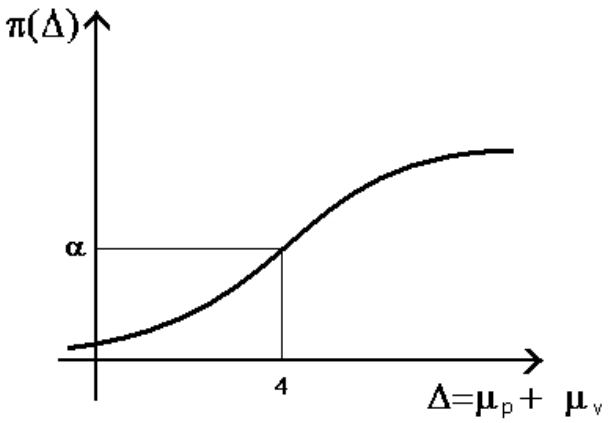

Figura 4.3: Potencia

4.1.2. Test de Homogeneidad e Independencia

EJERCICIO 48

Una empresa quiere contratar a cierta cantidad de personas y de los postulantes que se presentan se hace una preselección de 24 hombres y 24 mujeres que el jefe de personal decida quien será contratado y quien no. Después de que el jefe de personal hizo la selección de los contratados los resultados fueron los siguientes,

	Hombre	Mujer
Contratado	21	14
No contratado	3	10

Alguien acusa al empleador de tener un sesgo de selección a favor de los hombres ya que 21 de 24 hombres fueron contratados y sólo 14 de 24 mujeres también lo fueron. ¿Existirá discriminación por parte del jefe de personal?. Plantee las hipótesis con palabras y paramétricamente, llegue a conclusiones utilizando un nivel de significancia de $\alpha = 0{,}05$.

SOLUCIÓN

Hipótesis:

$$H_0 : \text{No existe discriminación (Homogeneidad)}$$

$$vs$$

$$H_1 : \text{Existe discriminación (No Homogeneidad)}$$

Equivalentemente

$$H_0 : p_{1j} = p_{2j} \;\; j = 1, 2 \quad vs \quad H_1 : p_{1j} \neq p_{2j} \text{ para algún } j$$

Para testear tales hipótesis, se ocupa el estadístico

$$\chi^2 = \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{(o_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}$$

donde $\hat{e}_{ij} = \frac{n_{i\cdot}n_{\cdot j}}{n_{\cdot\cdot}}$, el cual rechaza H_0 cuando $\chi^2 \geq \chi^2_{1-\alpha,(I-1)(J-1)}$.

Luego la tabla de valores esperados es:

	$Hombre$	$Mujer$	$Total\ n_{i\cdot}$
$Contratado$	17,5	17,5	35
$No\ contratado$	6,5	6,5	13
$Total\ n_{\cdot j}$	24	24	48

Por lo tanto el estadístico de prueba queda

$$\chi^2 = \frac{(21-17{,}5)^2}{17{,}5} + \frac{(14-17{,}5)^2}{17{,}5} + \frac{(3-6{,}5)^2}{6{,}5} + \frac{(10-6{,}5)^2}{6{,}5} = 5{,}1692$$

Como $\chi^2 = 5{,}1692 > 3{,}84 = \chi^2_{0{,}95,1}$, se rechaza H_0, es decir, con un 95 % de confianza existe discriminación hacia la mujer por parte del jefe de personal.

EJERCICIO 49

De cada una de tres comunidades se sacó una muestra de jóvenes casados. A cada pareja se le pidió que especificara la cantidad mínima de educación que esperaba que sus hijos recibieran. La siguiente tabla muestra los resultados que se observaron en la muestra:

	Comunidad			
Nivel Mínimo	A	B	C	Total
Colegio	30	28	24	82
Educ. comercial	30	19	46	95
Universitario	90	78	130	298
Total	150	125	200	475

¿Qué se puede concluir respecto a la homogeneidad de las aspiraciones en la educación de los hijos?

SOLUCIÓN

Las hipótesis son:

H_0 : Las 3 poblaciones son homogéneas respecto de las aspiraciones de educación para sus hijos. ($p_{11} = p_{12} = p_{13}$).

H_1 : Las 3 poblaciones no son homogéneas (Por lo menos 2 proporciones de una misma fila no son iguales entre si.)
Para testear tales hipótesis, se ocupa el estadístico

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(o_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}$$

donde $\hat{e}_{ij} = \frac{n_{i\cdot}n_{\cdot j}}{n_{\cdot\cdot}}$, el cual rechaza H_0 cuando $\chi^2 \geq \chi^2_{1-\alpha,(I-1)(J-1)}$.

Luego la tabla de valores esperados es:

	Comunidad			
Nivel Mínimo	A	B	C	Total
Colegio	25.89	21.58	34.53	82
Educ. comercial	30.00	25.00	40.00	95
Universitario	94.11	78.42	125.5	298
Total	150	125	200	475

Por lo tanto el estadístico de prueba queda

$$\chi^2 = \frac{(30-25{,}89)^2}{25{,}89} + \frac{(28-21{,}58)^2}{21{,}58} + \frac{(24-34{,}53)^2}{34{,}53} + \frac{(30-30)^2}{30} + \frac{(19-25)^2}{25}$$

$$+ \frac{(46-40)^2}{40} + \frac{(90-94{,}11)^2}{94{,}11} + \frac{(78-78{,}42)^2}{78{,}42} + \frac{(130-125{,}5)^2}{125{,}5}$$

$$= 8{,}455$$

Como $\chi^2 = 8{,}455 < 9{,}488 = \chi^2_{0{,}95,4}$, no existe evidencia en los datos para rechazar H_0, es decir, con un 95 % de confianza existe homogeneidad entre las comunidades.

EJERCICIO 50

Se seleccionó una muestra al azar de 275 alumnos de último año de colegio de cada uno de los siguientes tres grupos de rendimiento atlético: alto, medio y bajo. Los muchachos se clasificaron de acuerdo con la inteligencia tal como aparece en la tabla. ¿Indican estos datos una diferencia en la distribución de la inteligencia entre los tres grupos?

	Rendimiento			
Inteligencia	Alto	Medio	Bajo	Total
Alta	45	60	68	173
Media	10	15	25	50
Baja	5	15	32	52
Total	60	90	125	100

SOLUCIÓN

Las hipótesis son:

H_0 : Los 3 niveles de inteligencia son homogéneos respecto del rendimiento. $(p_{1j} = p_{2j} = p_{3j})$.

H_1 : Los 3 niveles de inteligencia no son homogéneos respecto del rendimiento (Por lo menos 2 proporciones de una misma columna no son iguales entre si.)

Para testear tales hipótesis, se ocupa el estadístico

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(o_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}$$

donde $\hat{e}_{ij} = \frac{n_{i\cdot}.n_{\cdot j}}{n_{\cdot\cdot}}$, el cual rechaza H_0 cuando $\chi^2 \geq \chi^2_{1-\alpha,(I-1)(J-1)}$.

Luego la tabla de valores esperados es:

	Rendimiento			
Inteligencia	Alto	Medio	Bajo	Total
Alta	37.77	56.62	78.64	173
Media	10.91	16.36	36.36	50
Baja	11.35	17.02	23.64	52
Total	60	90	125	100

Por lo tanto el estadístico de prueba queda

$$\chi^2 = \frac{(45 - 37{,}77)^2}{37{,}77} + \frac{(60 - 56{,}62)^2}{56{,}62} + \frac{(68 - 78{,}64)^2}{78{,}64} + \frac{(10 - 10{,}91)^2}{10{,}91} + \frac{(15 - 16{,}36)^2}{16{,}36}$$

$$+ \frac{(25 - 36{,}36)^2}{36{,}36} + \frac{(5 - 11{,}35)^2}{11{,}5} + \frac{(15 - 17{,}02)^2}{17{,}02} + \frac{(32 - 23{,}64)^2}{23{,}64}$$

$$= 10{,}199$$

Como $\chi^2 = 10{,}199 > 9{,}488 = \chi^2_{0,95,4}$, se rechaza H_0, es decir, con un 95 % de confianza no existe homogeneidad entre los niveles intelectuales.

EJERCICIO 51

Una empresa empaca determinado producto de latas de tres tamaños distintos, cada uno en distinta línea de producción. La mayor parte de las latas se apegan a las especificaciones, pero un ingeniero de control de calidad ha identificado los siguientes defectos:

- Mancha en la lata.
- Grieta en la lata.
- Ubicación incorrecta del anillo de apertura.
- Falta del anillo de apertura.
- Otras.

Se selecciona una muestra de unidades defectuosas de cada una de las tres líneas, y cada unidad se clasifica según el defecto, la siguiente tabla de contingencia incluye esos datos:

		Defecto					
		Mancha	Grieta	Ubicación	Falta	Otras	Tamaño de la muestra
Línea	1	34	65	17	21	13	150
de	2	23	52	25	19	6	125
Producción	3	32	28	16	14	10	100
	Total	89	145	58	54	29	375

¿Los datos sugieren desigualdad en las proporciones que caen en las distintas categorías de las tres líneas?

SOLUCIÓN

Los parámetros de interés son las diversas proporciones y las hipótesis relevantes son:

H_0: Las líneas de producción son homogéneas con respecto a las 5 categorías que no cumplen las especificaciones.

H_1: Las líneas de producción no son homogéneas con respecto a las 5 categorías que no cumplen las especificaciones.

Ahora se presenta una tabla resumen con los valores esperados y el valor de $(Obs. - Esp.)^2/Esp.$

	C1	C2	C3	C4	C5	Total
1	34	65	17	21	13	150
	35,60	58,00	23,20	21,60	11,60	
	0,072	0,845	1,657	0,017	0,169	
2	23	52	25	19	6	125
	29,67	48,33	19,33	18,00	9,67	
	1,498	0,278	1,661	0,056	1,391	
3	32	28	16	14	10	100
	23,73	38,67	15,47	14,40	7,73	
	2,879	2,943	0,018	0,011	0,664	
Total	89	145	58	54	29	375

luego, bajo un 95 % de confianza

$$\chi^2 = 14{,}159 \not> 15{,}50731 = \chi^2_{0{,}95;(3-1)\cdot(5-1)}$$

lo que indica que no existe suficiente evidencia para rechazar H_0, es decir las líneas de producción serían homogéneas con respecto a las 5 categorías que no cumplen las especificaciones.

Si disminuimos la confianza a un 90 % tenemos que

$$\chi^2 = 14{,}159 > 13{,}36157 = \chi^2_{0{,}90;(3-1)\cdot(5-1)}$$

luego, ahora si existiría evidencia bajo este nivel de significancia para rechazar H_0.

EJERCICIO 52

Un investigador desea saber si es posible concluir que hay relación entre el grado de liberalismo y la posición en la universidad en una población de estudiantes universitarios. Para estos efectos se seleccionó una muestra de 500 estudiantes. La tabla siguiente muestra la clasificación de los datos según sus respuestas:

	Grado de Liberalismo			
Clase	Ligero	Moderado	Alto	Total
1er. año	30	83	37	150
2o. año	19	56	50	125
3er. año	16	46	63	125
4o. año	10	38	52	100
Total	75	223	202	500

¿Qué se puede concluir respecto al problema del investigador?

SOLUCIÓN

Las hipótesis son:

H_0 : Existe independencia entre el grado de liberalismo y el año universitario. ($n_{ij} = \frac{n_{i\cdot}n_{\cdot j}}{n_{ij}}$).

H_1 : No existe independencia entre el grado de liberalismo y el año universitario.($n_{ij} \neq \frac{n_{i\cdot}n_{\cdot j}}{n_{ij}}$).

Para testear tales hipótesis, se ocupa el estadístico

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(o_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}$$

donde $\hat{e}_{ij} = \frac{n_{i\cdot}n_{\cdot j}}{n_{\cdot\cdot}}$, el cual rechaza H_0 cuando $\chi^2 \geq \chi^2_{1-\alpha,(I-1)(J-1)}$.

Luego la tabla de valores esperados es:

	Grado de Liberalismo			
Clase	Ligero	Moderado	Alto	Total
1er. año	22.50	66.90	60.60	150
2o. año	18.75	55.75	50.50	125
3er. año	18.75	55.75	50.50	125
4o. año	15.00	44.60	40.40	100
Total	75	223	202	500

Por lo tanto el estadístico de prueba queda

$$\chi^2 = \frac{(30-22{,}5)^2}{22{,}5} + \frac{(83-66{,}9)^2}{66{,}9} + \frac{(37-60{,}6)^2}{60{,}6} + \frac{(19-18{,}75)^2}{18{,}75} + \frac{(56-55{,}75)^2}{55{,}75} + \frac{(50-50{,}5)^2}{50{,}5}$$

$$+ \frac{(16-18{,}75)^2}{18{,}75} + \frac{(46-55{,}75)^2}{55{,}75} + \frac{(63-50{,}5)^2}{50{,}5} + \frac{(10-15)^2}{15} + \frac{(38-44{,}6)^2}{44{,}6} + \frac{(52-40{,}4)^2}{40{,}4}$$

$$= 26{,}751$$

Como $\chi^2 = 26{,}751 > 12{,}592 = \chi^2_{0{,}95{,}6}$, se rechaza H_0, es decir, con un 95 % de confianza el grado de liberalismo en los estudiantes universitarios no es independiente del año que cursa el alumno.

EJERCICIO 53

Un estudio de la relación entre las condiciones de las instalaciones en gasolineras y la agresividad en el precio de la gasolina reporta los siguientes datos basados en una muestra de $n = 144$ gasolineras.

	Agresividad	Neutral	No agresiva	$n_{i.}$
Anticuada	24	15	17	56
Estándar	52	73	80	205
Moderna	58	86	36	180
$n_{.j}$	134	174	133	441

En el nivel 0.01, ¿la información sugiere que las condiciones de instalaciones y las políticas de precios son independientes entre si?

SOLUCIÓN

La hipótesis a docimar es:

H_0: Las condiciones de las instalaciones con la política de precios son independientes.
vs
H_1: No existe independencia.

La siguiente tabla resumen entrega la información necesaria para calcular el estadístico χ^2.

```
          C1     C2     C3  Total
   1      24     15     17     56
       17,02  22,10  16,89
       2,867  2,278  0,001

   2      52     73     80    205
       62,29  80,88  61,83
       1,700  0,769  5,343

   3      58     86     36    180
       54,69  71,02  54,29
       0,200  3,159  6,159

Total   134    174    133    441
```

luego, bajo un 99 % de confianza

$$\chi^2 = 22{,}476 > 13{,}2767 = \chi^2_{0{,}99;(3-1)\cdot(3-1)}$$

lo que indica que existe suficiente evidencia con este nivel de confianza para rechazar H_0, es decir el conocimiento de la política de precios de una gasolinera proporciona información acerca de la condición de las instalaciones de la gasolinera.

EJERCICIO 54

Se obtuvo una muestra aleatoria de individuos que viajan solos en automóvil al trabajo, en una gran zona metropolitana, y cada individuo fue clasificado de acuerdo con el tamaño de su automóvil y la distancia de recorrido citadino. ¿La siguiente información sugiere que dicha distancia y el tamaño del automóvil están relacionados en la población a la cual se hizo el muestreo? Exprese las hipótesis pertinentes y utilice una prueba Chi-cuadrado con un nivel 0.05.

		Distancia de Recorido		
		$[0, 10)$	$[10, 20)$	$[20, \ldots)$
Tamaño de Automovil	Subcompacto	6	27	19
	Compacto	8	36	17
	Mediano	21	45	33
	Grande	14	18	6

SOLUCIÓN

La hipótesis a docimar es:

H_0: Existe independencia entre la distancia de recorrido y el tamaño del automóvil.

vs

H_1: No existe independencia.

La siguiente tabla resumen entrega la información necesaria para calcular el estadístico χ^2.

	C1	C2	C3	Total
1	6	27	19	52
	10,19	26,21	15,60	
	1,724	0,024	0,741	
2	8	36	17	61
	11,96	30,74	18,30	
	1,309	0,899	0,092	
3	21	45	33	99
	19,40	49,90	29,70	
	0,131	0,480	0,367	
4	14	18	6	38
	7,45	19,15	11,40	
	5,764	0,069	2,558	
Total	49	126	75	250

luego, bajo un 95 % de confianza

$$\chi^2 = 14,158 > 12{,}59159 = \chi^2_{0,95;(4-1)\cdot(3-1)}$$

lo que indica que existe suficiente evidencia con este nivel de confianza para rechazar H_0, es decir, la distancia de recorrido proporciona información acerca el tamaño del automóvil.

EJERCICIO 55

Un ginecólogo analiza la posible relación entre la edad de la monarquía y la aparición de cáncer de mama. Con el fin de estudiarlo clasifica a las mujeres que acuden a su consulta en dos grupos, aquellas que tuvieron la monarquía antes de los 12 años (a las que distingue con el valor cero), y aquellas que la tuvieron después de esta edad (a las que distingue con el valor 1). Se presentan a continuación los resultados obtenidos:

	Cancer de Mama	
Edad de la Menarquia	Sí	No
0	64	53
1	47	139

Determine si existe relación o no entre estas variables.

SOLUCIÓN

Para medir si existe relación entre la edad de la monarquía y el cáncer de mama, realizamos un test de independencia.

$$H_0 : n_i j = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n_{\cdot\cdot}} \quad H_1 : \text{no existe independencia}$$

Para tal hipótesis, ocupamos el estadístico χ^2.

$$\chi^2 = \sum \frac{(obs - esp)^2}{esp}$$

en donde los observados son los valores que aparecen en la tabla y los esperados los calculamos mediante H_0, por ejemplo, el esperado para la casilla

$$n_{11} = \frac{n_{1\cdot} \cdot n_{\cdot 1}}{n_{\cdot\cdot}} = \frac{117 \cdot 111}{303} = 42{,}8613$$

Luego para cada casilla, los esperados serían los que se muestran a continuación:

		Cancer		
		Si	No	Total
Edad	0	64 42,86	53 74,14	117
	1	47 68,14	139 117,86	186
Total		111	192	303

Luego el estadístico nos queda de la siguiente manera:

$$\chi^2 = \frac{(64 - 42{,}86)^2}{42{,}86} + \frac{(53 - 74{,}14)^2}{74{,}14} + \frac{(47 - 68{,}17)^2}{68{,}14} + \frac{(139 - 117{,}86)^2}{117{,}86}$$

$$= 10{,}425 + 6{,}027 + 6{,}558 + 3{,}791$$

$$= 26{,}801$$

Ahora, rechazamos H_0 si $\chi^2 > \chi^2_{(filas-1)(columnas-1),1-\alpha}$ donde filas en este caso tenemos 2 y columnas 2 y el α lo escogemos como 0.05. Por lo tanto tenemos $\chi^2_{1,0,95} = 3{,}84$ buscado en una tabla de la distribución Chi-Cuadrado.

Luego, como $\chi^2 = 26{,}801 > \chi^2_{1,0,95} = 3{,}84$ se rechaza la hipótesis de que ambas variables sean independientes con un 95 % de confianza.

EJERCICIO 56

Un mecánico analiza la posible relación entre la edad de la maquina y la aparición de una falla grave. Con el fin de estudiarlo clasifica a las maquinas en dos grupos, aquellas que tuvieron una falla grave antes de los 12 años (a las que distingue con el valor 0), y aquellas que la tuvieron después de esta edad (a las que distingue con el valor 1). Se presentan a continuación los resultados obtenidos:

		Falla Grave	
		Si	No
Edad Maquina	0	64	53
	1	47	139

(a) Calcule el Test χ^2 de Pearson.

(b) Determine si existe relación o no entre la variables

SOLUCIÓN

Completamos la tabla dada con los valores esperados

		Falla Grave		
		Si	No	Total
Edad Máquina	0	64 42,861	53 74,139	117
	1	47 68,139	139 117,861	186
Total		111	192	303

(a) Dada la tabla completa con los los valores esperados calculamos el estadístico como sigue:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

$$= \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

$$= \frac{(64-42{,}8613)^2}{42{,}8613} + \frac{(53-74{,}138)^2}{74{,}138} + \frac{(47-68{,}138)^2}{68{,}138} + \frac{(139-117{,}861)^2}{117{,}861}$$

$$= 10{,}425 + 6{,}027 + 6{,}558 + 3{,}791$$

$$= 26{,}801$$

(b) Se rechaza H_0 : $\exists$ independencia entre la edad de la máquina y si la falla es grave, si

$$\chi^2 > \chi_{(1-\alpha;(r-1)\cdot(c-1))}$$

Como

$$\chi^2 = 26{,}801 > 3{,}841459 = \chi_{0{,}95;1}$$

Se rechaza la hipótesis de independencia entre las fallas graves y la edad de las máquinas.

EJERCICIO 57

Una muestra de 500 personas responde dos preguntas: filiación política y actitud hacia una reforma de impuestos, los resultados son los siguientes:

	Actitud hacia Reforma			
Filiación	A favor	Indiferente	En contra	Total
Demócrata	138	83	64	285
Republicano	64	67	84	215
Total	202	150	148	500

¿Existe relación entre la tendencia política y la actitud hacia la reforma de impuestos?. Plantee la hipótesis necesaria y concluya.

SOLUCIÓN

Las hipótesis son:

H_0 : Existe independencia entre la tendencia política y la actitud hacia la reforma. ($n_{ij} = \frac{n_{i\cdot}n_{\cdot j}}{n_{ij}}$).

H_1 : Existe asociación entre la tendencia política y la actitud hacia la reforma.($n_{ij} \neq \frac{n_{i\cdot}n_{\cdot j}}{n_{ij}}$).

Para testear tales hipótesis, se ocupa el estadístico

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(o_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}$$

donde $\hat{e}_{ij} = \frac{n_{i\cdot}n_{\cdot j}}{n_{\cdot\cdot}}$, el cual rechaza H_0 cuando $\chi^2 \geq \chi^2_{1-\alpha,(I-1)(J-1)}$.

Luego la tabla de valores esperados es:

	Actitud hacia Reforma			
Filiación	A favor	Indiferente	En contra	Total
Demócrata	115.14	85.5	84.36	285
Republicano	86.86	64.5	63.64	215
Total	202	150	148	500

Por lo tanto el estadístico de prueba queda

$$\chi^2 = \frac{(138-115{,}14)^2}{115{,}14} + \frac{(83-85{,}5)^2}{85{,}5} + \frac{(64-84{,}36)^2}{84{,}36}$$

$$+ \frac{(64-86{,}86)^2}{86{,}86} + \frac{(67-64{,}5)^2}{64{,}5} + \frac{(84-63{,}64)^2}{63{,}64}$$

$$= 22{,}51$$

Como $\chi^2 = 22{,}51 > 5{,}99 = \chi^2_{0{,}95,2}$, se rechaza H_0, es decir, con un 95 % de confianza la tendencia política influye en la actitud hacia la reforma.

EJERCICIO 58

En una muestra aleatoria de 100 universitarios se clasificó cada uno de ellos según si había consumido alguna vez droga o no y el promedio de notas. A partir de los datos tabulados en la tabla ¿Proporcionan estos datos evidencia suficiente como para concluir que hay una relación entre las dos variables? Use $\alpha = 0{,}05$.

	¿Ha consumido Drogas?		
Promedio notas	Si	No	$Total$
$\leq 4{,}0$	10	29	39
$> 4{,}0$	20	41	61
Total	30	70	100

SOLUCIÓN

Las hipótesis son:

H_0 : Existe independencia entre el consumo de drogas y el promedio de notas ($n_{ij} = \frac{n_{i\cdot}.n_{\cdot j}}{n_{ij}}$).

H_1 : Existe asociación entre el consumo de drogas y el promedio de notas.($n_{ij} \neq \frac{n_{i\cdot}.n_{\cdot j}}{n_{ij}}$).

Para testear tales hipótesis, se ocupa el estadístico

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(o_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}$$

donde $\hat{e}_{ij} = \frac{n_{i\cdot}.n\cdot j}{n_{\cdot\cdot}}$, el cual rechaza H_0 cuando $\chi^2 \geq \chi^2_{1-\alpha,(I-1)(J-1)}$.

Luego la tabla de valores esperados es:

Promedio notas	¿Ha consumido Drogas? *Si*	*No*	*Total*
$\leq 4{,}0$	11,7	27,3	39
$> 4{,}0$	18,3	42,7	61
Total	30	70	100

Por lo tanto el estadístico de prueba queda

$$\chi^2 = \frac{(10-11{,}7)^2}{11{,}7} + \frac{(29-27{,}3)^2}{27{,}3} + \frac{(20-18{,}3)^2}{18{,}3} + \frac{(41-42{,}7)^2}{42{,}7}$$

$$= 0{,}578$$

Como $\chi^2 = 0{,}578 < 3{,}841 = \chi^2_{0{,}95,1}$, no se rechaza H_0, es decir, con un 95 % de confianza el consumo de droga no influye en el promedio de notas de los estudiantes.

4.2. Ejercicios Propuestos

1. Suponga que se tienen dos muestras aleatorias independientes $X_1, \ldots, X_n \sim Exp(\theta)$ y $Y_1, \ldots, Y_m \sim Exp(\mu)$.

   a) Derive el TRV de $H_0 : \theta = \mu$ vs $H_1 : \theta \neq \mu$.

   b) Muestre que el test derivado en (a) puede estar basado en el estadístico

   $$T = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} X_i + \sum_{i=1}^{m} Y_i}$$

   c) Encuentre la distribución de T bajo H_0.

2. Se utilizan dos máquinas para llenar botellas de plástico con un volumen neto de 16.0 onzas. Las distribuciones de los volúmenes de llenado pueden suponerse normales, con desviaciones estándar $\sigma_1 = 0{,}020$ y $\sigma_2 = 0{,}025$ onzas. Un miembro del grupo de ingeniería de calidad sospecha que el volumen neto de llenado de ambas máquinas es el mismo, sin importar si éste es o no de 16 onzas. De cada máquina se toma una muestra aleatoria de 10 botellas.

Máquina 1:	16,03	16,04	16,05	16,05	16,02	16,01	15,96	15,98	16,02	15,99
Máquina 2:	16,02	15,97	15,96	16,01	15,99	16,03	16,04	16,02	16,01	16,00

   a) ¿ Se encuentra el ingeniero en lo correcto? Utilice $\alpha = 0{,}05$.

   b) ¿Cuál es el valor P de esta prueba?

   c) Si se supone que el tamaño de las muestras es el mismo, ¿qué tamaño de muestra debe utilizarse para asegurar que $\beta = 0{,}05$ si la diferencia verdadera entre las medias es 0.08? Suponga que $\alpha = 0{,}05$.

   d) ¿Cuál es la potencia de la prueba del inciso a) si la diferencia verdadera entre las medidas es 0.08?

3. Existen dos tipos de plásticos apropiados para su uso por un fabricante de componentes electrónicos. La tensión de ruptura de este plástico es un parámetro importante. Se sabe que $\sigma_1 = \sigma_2 = 1{,}0$ psi. De una muestra aleatoria de tamaño $n_1 = 10$ y $n_2 = 12$, se tiene que $\overline{x}_1 = 162{,}5$ y $\overline{x}_2 = 155{,}0$. La compañía no adoptará el plástico 1 a menos que la tensión de ruptura de éste exceda a la del plástico 2 al menos por 10 psi. Con base a la información contenida en la muestra, ¿La compañía deberá utilizar el plástico 1? Utilice $\alpha = 0{,}05$ para llegar a una decisión.

4. En una industria se desea verificar si la productividad media de los operarios del período diurno es igual a la productividad media de los operarios del período nocturno. Se supone que las productividades de los operarios de los diferentes períodos son independientes y normalmente distribuidas. Se seleccionan muestran de igual tamaño para cada uno de los períodos obteniéndose los siguientes resultados:

Período	N de operarios	Media	Varianza
Diurno	15	12	35.71
Nocturno	15	10	36.43

a) Verifique igualdad de varianzas con nivel $\alpha = 0{,}1$

b) Verifique si las productividades medias son iguales con nivel $\alpha = 0{,}05$.

c) Determine la probabilidad de aceptar la igualdad de medias en (b), si la realidad es que la diferencia entre las productividades medias es de una unidad. Aproxime los valores t_α por z_α.

5. Para alcanzar la máxima eficiencia en una operación de ensamblaje en una fábrica los obreros nuevos requieren alrededor de un mes de capacitación. Se sugiere un nuevo método de capacitación y se realiza una prueba para compararlo con el tradicional. Para este fin se capacitan dos grupos de nueve obreros durante un período de tres semanas; uno de los grupos aplica el nuevo método y el otro el tradicional. Al final de las tres semanas de capacitación se mide el tiempo (en minutos) que le toma a cada obrero ensamblar el dispositivo. Los resultados obtenidos de las muestras son los siguientes (asuma normalidad de los datos)

	Método	
Estadístico	Tradicional	Nuevo
n	9	9
Promedio	35	31
Varianza	25	20

¿Hay suficiente evidencia que indique que las medias de los tiempos reales son diferentes con los dos métodos? Realice una prueba con el nivel $\alpha = 5\,\%$. Determine el valor-p de la prueba. (sea explícito, dé hipótesis, test y conclusión).

6. Se tomaron muestras independientes de alumnos del curso de Estadística EYP2113 sección 1 ($n_1 = 8$) y sección 2 ($n_2 = 13$) los cuales fueron evaluados mediante un test de habilidades. Los resultados fueron los siguientes:

	n	Media	Desv. Estándar
Sección 1	8	13.5	5.0
Sección 2	13	9.5	5.0

a) Lleve a cabo un test bajo el supuesto de normalidad, que permita decidir si los rendimientos medios difieren. Use $\alpha = 5\,\%$.

b) Lleve a cabo un test no paramétrico. Use $\alpha = 5\,\%$.

7. El contenido de nicotina de dos marcas de cigarros, medidas en miligramos, es la siguiente:

A	2.1	4.0	6.3	5.4	4.8	3.7	6.1	3.3		
B	4.1	0.6	3.1	2.5	4.0	6.2	1.6	2.2	1.9	5.4

¿Los contenidos de nicotina de las dos marcas serán diferentes?. Considere $\alpha = 5\,\%$.

a) Desarrolle el test bajo el supuesto de normalidad de las observaciones.

b) Suponga ahora que no hay evidencia que permita asumir normalidad.

8. Un experimento quiere determinar la eficacia de un nuevo elemento versus la dieta actual (control) para reducir la cantidad de grasa en los cerdos. Los datos se encuentran en la siguiente tabla:

Nuevo	676	206	230	256	280	433	337	466	497	512	794	428	452	512
Control	88	570	605	617	653	2913	924	286	1098	982	2346	321	615	519

a) Utilice la teoría normal para testear la hipótesis de que existe diferencia entre los dos tipos de producción.

b) Docime la misma hipótesis usando el método no paramétrico.

9. Para la elaboración de un neumático se utilizan dos métodos. A dichos neumáticos se les mide el desgaste. Se seleccionan 12 neumáticos de cada tipo y siendo sus mediciones de desgaste, las siguientes:

Proceso 1	329	436	457	463	477	479	1297	1319	1340	1385	1398	1440
Proceso 2	313	563	670	940	1002	1261	1305	1531	1614	1694	1701	1708

a) Utilice la teoría de normalidad para testear la hipótesis de que no existe diferencia entre los métodos de elaboración.

b) Docime la misma hipótesis usando el método no paramétrico.

10. El nivel de hidrocarburos (HC) de un automóvil puede ser evaluado por dos métodos FTP (norma USA) y ECE (norma europea). Interesa probar que existe una diferencia significativa entre los métodos utilizados. Para ello se tomaron 10 automóviles, los cuales son evaluados por cada uno de los métodos obteniendo lo siguiente:

Auto	1	2	3	4	5	6	7	8	9	10	Prom.	D.E.
FTP	0.23	0.22	0.13	0.13	0.13	0.39	0.09	0.16	0.29	0.12	0.189	0.0937
ECE	0.03	0.11	0.09	0.08	0.05	0.33	0.08	0.10	0.33	0.04	0.123	0.1095

Lleve a cabo test bajo normalidad y test no paramétrico. ¿Qué concluye?.

11. En la fabricación de semiconductores, a menudo se utiliza una sustancia química para quitar el silicio de la parte trasera de las obleas antes de la metalización. En este proceso es importante la rapidez con la que actúa la sustancia. Se han comparado dos

soluciones químicas, utilizando para ello dos muestras aleatorias de 10 obleas para cada solución. La rapidez de acción observada es la siguiente (en mils/min):

Solución 1:	9,9	9,4	9,3	9,6	10,2	10,6	10,3	10,0	10,3	10,1
Solución 2:	10,2	10,6	10,7	10,4	10,5	10,0	10,2	10,7	10,4	10,3

a) ¿Los datos apoyan la información que la rapidez promedio de acción es la misma para ambas soluciones? Para obtener sus conclusiones, utilice $\alpha = 0{,}05$ y suponga que las varianzas de ambas poblaciones son iguales.

b) Calcule el valor P para la prueba del inciso a).

c) Construya diagramas de caja para las dos muestras. ¿Estas gráficas apoyan la hipótesis de que las varianzas son iguales? Escriba una interpretación práctica de estas gráficas.

12. Una muestra aleatoria de 200 hombres casados, todos retirados, se clasifico de acuerdo a la educación y el número de hijos de cada uno de ellos:

		Cantidad de hijos		
		$0-1$	$2-3$	más de 3
Educación	Primaria	14	37	32
	Secundaria	19	42	17
	Bachillerato	12	17	10

Pruebe la hipótesis, con un nivel de significancia del 5 %, que el tamaño de una familia es independiente del nivel de educación del padre.

13. Una compañía opera cuatro máquinas tres turnos al día. De los registros de producción, obtienen los datos siguientes sobre el número de fallas:

	Máquinas			
Turno	A	B	C	D
1	41	20	12	16
2	31	11	9	14
3	15	17	16	10

Pruebe la hipótesis (con $\alpha = 0{,}05$) de que el número de fallas es independiente del turno. Encuentre el valor P de esta prueba.

14. Un artículo publicado en el *Journal of Marketing Ressearch* (1970, pág. 36-42) contiene un estudio de la relación entre las condiciones de las instalaciones de la gasolinerías y la dinámica de la política de mercadotecnia seguida por ellas. Para ello se investigó una muestra de 441 gasolinerías, y se obtuvieron los resultados siguientes. ¿Existe evidencia de que la política de mercadotecnia y las condiciones de la gasolinería son independientes? Utilice $\alpha = 0{,}05$.

	Condición		
Política	Subestándar	Estándar	Moderna
Dinámica	24	52	58
Neutral	15	73	86
No dinámica	17	80	36

15. Cada uno de 325 individuos que participan en cierto programa de medicamentos, se clasifico con respecto a la presencia o ausencia de hipoglucemia y con respecto a la dosis media diaria de insulina. ¿Apoyan los datos siguientes lo dicho de que la presencia o ausencia de hipoglucemia es independiente de las dosis de insulina? Pruebe usando $\alpha = 0{,}05$.

		Dosis diaria de insulina				
		$< 0{,}25$	$0{,}25 - 0{,}49$	$0{,}5 - 0{,}74$	$0{,}75 - 0{,}99$	$> 1{,}0$
Condición de Hipoglucemia	Presente	4	21	28	15	12
	Ausente	40	74	59	26	46

16. Los siguientes datos corresponden a combinaciones de sexo de los recombinantes que resultan de seis diferentes genotipos masculinos. ¿Soportan los datos la hipótesis de que la distribución de frecuencia entre las tres combinaciones de sexo es homogénea con respecto a los diferentes genotipos? Defina los parámetros de interés, exprese H_0 y H_1 pertinentes, y realice el análisis.

		Combinación de sexo		
		M/M	M/F	F/F
Genotipo Masculino	1	35	80	39
	2	41	84	45
	3	33	87	31
	4	8	26	8
	5	5	11	6
	6	30	65	20

17. La siguiente tabla muestra mediciones independientes del flujo de agua (en mts/seg) de dos canales de regadío:

Canal El Moro	114	109	280	80	180	95	106	
Canal La Diuca	180	303	325	272	97	275	251	190

a) Realice una prueba no paramétrica que permita probar la hipótesis "en el segundo canal de regadío se observa un flujo mayor". Utilizando aproximación normal para el estadígrafo, dé el valor-p de la prueba. Sea explícito (Hipótesis, test, regla, etc.).

b) Debido a la imprecisión de las mediciones, un especialista propone dividir el flujo en tres categorías: Bajo (< 115), Medio (115-250) y Alto (> 250). Construya la tabla adecuada y lleve a cabo el test más apropiado. Sea explícito en cuanto al tipo de prueba, valor del test y entregue el valor-p.

18. Se diseña un generador de números seudoaleatorios de modo que los enteros 0 a 9 tengan la misma probabilidad de ocurrencia. Los primeros 10 mil números son:

0	1	2	3	4	5	6	7	8	9
967	1008	975	1022	1003	989	1001	981	1043	1011

a) ¿El generador trabajó de manera apropiada? Utilice $\alpha = 0{,}01$.

b) Calcule el valor P de esta prueba.

Capítulo 5

Regresión Clásica

5.1. Ejercicios Resueltos

EJERCICIO 59

Los siguientes valores corresponden a las edades de un grupo de personas que realizan un determinado deporte y al peso en kg. de los mismos.

Edades	18	26	28	34	42	48	52	54	60	36
Peso	54	64	54	62	70	76	66	76	74	68

(a) Encuentre la recta de regresión.

(b) Determine un intervalo de 95 % de confianza para β_0 y β_1.

(c) Complete la tabla ANOVA.

(d) Calcule el R^2.

SOLUCIÓN

(a) La recta de regresión tiene la forma $y_i = \beta_0 + \beta_1 x_i + e_i$ y sabemos que

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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Evaluando y reemplazando los valores que se entregaron, se obtiene que

$$\hat{\beta}_1 = 0{,}499 \qquad \hat{\beta}_0 = 46{,}5$$

Luego la recta de regresión estimada es

$$\hat{y}_i = 46{,}5 + 0{,}499 x_i$$

(b) Los intervalos de confianza para los coeficientes de la recta de regresión están dado por

$$IC(\hat{\beta}_i) : \hat{\beta}_i \pm t_{n-2,\frac{\alpha}{2}} S \sqrt{(X'X)^{-1}_{ii}}$$

donde $S^2 = \frac{\sum e_i^2}{n-2} = 21{,}31$ y

$$(X'X)^{-1} = \frac{1}{n\sum x_i^2 - \left(\sum x_i\right)^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix}$$

Luego los intervalos de confianza quedan de la siguiente manera:

$$IC(\hat{\beta}_0) : \hat{\beta}_0 \pm 4{,}61\sqrt{10} t_{8;0{,}025} \rightarrow (12{,}883; 80{,}117)$$
$$IC(\hat{\beta}_1) : \hat{\beta}_1 \pm 4{,}61\sqrt{0{,}000593} t_{8;0{,}025} \rightarrow (0{,}2399; 0{,}7580)$$

(c) La forma que tiene la tabla ANOVA es como sigue:

Tabla ANOVA

Fuente	g.l.	SS	MS	F
Regresión	$k-1$	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2/k-1$	$\frac{\left(\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2/k-1\right)}{\left(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/n-k\right)}$
Error	$n-k$	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/n-k$	
Total	$n-1$	$\sum_{i=1}^{n}(y_i - \bar{y})^2$	$\sum_{i=1}^{n}(y_i - \bar{y})^2/n-1$	

Luego, haciendo los cálculos pertinentes se obtiene para estos datos, la siguiente tabla.

Tabla ANOVA

Fuente	g.l	SS	MS	F
Regresión	1	419,9	419,9	19,70
Error	8	170,5	21,31	
Total	9	590,4	65,6	

(d) Por definición, se tiene que $R^2 = \frac{SSR}{SST} = 0{,}71$, lo que significa que el modelo explica un 71 % de la variabilidad presente en los datos.

EJERCICIO 60

En el cultivo de cultivos in vitro se ha observado que si se colocan dos núcleos de crecimiento en el mismo cultivo, a menudo el crecimiento se organiza formando un patrón alrededor del eje de los núcleos. A esto se le llama un campo de atracción. Se ha observado que los campos de atracción se forman con mayor frecuencia si los núcleos están cercanos. En un experimento de 1951 se colocaron 20 núcleos a distancias diferentes y se midió la incidencia de campos de atracción (Y) para las diferentes distancias (X). Lamentablemente se borró parte del análisis de regresión y se le solicita completarlo.

(a) Complete la tabla ANOVA que se entrega a continuación:

Tabla ANOVA

Fuente	g.l	SS	MS	F
Regresión	1	2,0559		301,08
Error				
Total				

(b) Calcule el R^2 e interprete.

(c) Utilizando la siguiente información realice test de hipótesis para β_1. Concluya.

Ecuación de Regresión:	$Y = \beta_0 + \beta_1 X$	
Predictor	Coef.	Des. Estándar
Constant	1.176232	0.038939
Distancia	-0.27801	0.01602

SOLUCIÓN

Para trabajar con esta regresión, se deben cumplir los siguientes supuestos

1. $E(e_i) = 0$.
2. $Var(e_i) = \sigma^2$.
3. $Cov(e_i, e_j) = 0, \quad \forall i \neq j$.

(a) Recordemos que la tabla ANOVA tiene la siguiente forma

Tabla ANOVA

Fuente	g.l	SS	MS	F
Regresión	$k-1$	SS_R	$SS_R/(k-1)$	MS_R/MS_E
Error	$n-k$	SS_E	$SS_E/(n-k)$	
Total	$n-1$	SS_T		

Del enunciado tenemos que $n = 20$, luego comparando tenemos que $k = 2$ (n^o de parámetros a estimar) y así completamos los g.l.
Con el SS_R y sus g.l. se calcula el $MS_R = 2{,}0559$ lo que nos permite, despejando del F, encontrar el valor de MS_E y así el valor de SS_E.
Finalmente, recordado que $SS_R + SS_E = SS_T$, reemplazamos y obtenemos la siguiente tabla:

Tabla ANOVA

Fuente	g.l	SS	MS	F
Regresión	1	2,0559	2,0559	301,08
Error	18	0,123	0,0068	
Total	19	2,1789		

(b) Rescatando información de la tabla ANOVA, se tiene que

$$R^2 = \frac{SS_R}{SS_T}$$

$$= \frac{2{,}0559}{2{,}1789}$$

$$= 0{,}94$$

El R^2 indica el porcentaje o proporción de variabilidad presente en los datos, explicada por el modelo, el cual en este caso es muy adecuado.

(c) El test para β_1 es:

$$H_0 : \beta_1 = 0 \quad vs \quad H_1 : \beta_1 \neq 0$$

Tal hipótesis se testea con el estadístico

$$t_c = \frac{\hat{\beta}_1 - \beta_{H_0}}{s.e.(\hat{\beta}_1)}$$

el cual rechaza H_0 si $|t_c| > t_{(n-2);\frac{\alpha}{2}}$.

Luego en este caso resulta:

$$|t_c| = \left| \frac{-2{,}27801}{0{,}01602} \right|$$

$$= 17{,}35$$

Como $|t_c| = 17{,}35 > 2{,}101 = t_{18;0,025}$, se rechaza H_0, es decir, existe regresión en el modelo.

EJERCICIO 61

Es sabido que la potencia de un vehículo se relaciona directamente con el número de pistones. Sea Y_i la potencia del vehículo i (miles r.p.m.) y X_i el número de pistones del vehículo.

(a) Postule el modelo y los supuestos.

(b) Derive el EMCO de β y obtenga una expresión para la varianza.

(c) Estime la ecuación de regresión si una muestra de 5 vehículos entrega

X	2	2	3	4	4
Y	5	6	9	11	13

SOLUCIÓN

(a) Como la relación es directa (no hay intercepto) el modelo es de la forma

$$y_i = \beta x_i + e_i \qquad i = 1, ..., n$$

Los supuestos a cumplir son:

1. $E(e_i) = 0$.
2. $Var(e_i) = \sigma^2$.
3. $Cov(e_i, e_j) = 0, \quad \forall i \neq j$.

(b) Estimador de mínimos cuadrados para β

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta x_i)^2$$

luego derivando con respecto al parámetro tenemos

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \beta} = -2 \sum_{i=1}^{n} (y_i - \beta x_i) \cdot x_i = 0$$

$$\Rightarrow \sum_{i=1}^{n} (y_i - \beta x_i) \cdot x_i = 0$$

$$\Rightarrow \sum_{i=1}^{n} (x_i y_i - \beta x_i^2) = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i - \beta \sum_{i=1}^{n} x_i^2 = 0$$

luego despejando nos queda

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

Ahora para la varianza de $\hat{\beta}$ tenemos

$$Var(\hat{\beta}) = Var\left(\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \right)$$

$$= \frac{1}{\left(\sum_{i=1}^{n} x_i^2 \right)^2} Var\left(\sum_{i=1}^{n} x_i y_i \right)$$

$$\overset{ind}{=} \frac{1}{\left(\sum_{i=1}^{n} x_i^2\right)^2} \sum_{i=1}^{n} x_i^2 Var(y_i)$$

pero $Var(y_i) = Var(\beta x_i + e_i) = Var(e_i) = \sigma^2$

Luego, reemplazando obtenemos que

$$Var(\hat{\beta}) = \frac{1}{\left(\sum_{i=1}^{n} x_i^2\right)^2} \sum_{i=1}^{n} x_i^2 \sigma^2 = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$$

(c) Considerando el estimador que encontramos para β, se tiene:

$$\begin{aligned}
\hat{\beta} &= \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \\
&= \frac{10 + 12 + 27 + 44 + 52}{4 + 4 + 9 + 16 + 16} \\
&= \frac{145}{49} \\
&= 2{,}96
\end{aligned}$$

Así la recta de regresión estimada es

$$\hat{y}_i = 2{,}96 x_i$$

.

EJERCICIO 62

Considere $Y_1 = \theta_1 + \theta_2 + e_1, \quad Y_2 = 2\theta_2 + e_2, \quad Y_3 = -\theta_1 + \theta_2 + e_3$.

(a) Encuentre el EMCO para θ_1 y θ_2 en función de los Y.

(b) Derive expresiones para evaluar la significancia de los parámetros.

SOLUCIÓN

(a) Para encontrar tal estimador escribiremos el modelo en forma matricial, como se muestra a continuación:

$$\underbrace{\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix}}_{\mathbf{Y}} = \underbrace{\begin{pmatrix} 1 & 1 \\ 0 & 2 \\ -1 & 1 \end{pmatrix}}_{X} \cdot \underbrace{\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix}}_{\mathbf{e}}$$

y sabemos que para tal modelo, el estimador $\hat{\beta} = (X'X)^{-1}X'Y$. Por lo tanto reemplazando, se obtiene:

$$\hat{\beta} = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{6} \end{pmatrix} \begin{pmatrix} Y_1 - Y_3 \\ Y_1 + 2Y_2 + Y_3 \end{pmatrix}$$

$$= \begin{pmatrix} \frac{Y_1 - Y_3}{2} \\ \\ \frac{Y_1 + 2Y_2 + Y_3}{6} \end{pmatrix}$$

$$= \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix}$$

(b) Los test de significancia individual de parámetros son

$$H_0 : \theta_1 = 0 \quad vs \quad H_1 : \theta_1 \neq 0$$

El estadístico de prueba para este test es

$$t = \frac{\hat{\theta}_i}{\sqrt{Var(\hat{\theta}_i)}}$$

el cual rechaza H_0 si $|t| > t_{n-2,\frac{\alpha}{2}}$.

Sabemos que

$$Var(\hat{\boldsymbol{\beta}}) = \sigma^2(X'X)^{-1} = \sigma^2 \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{6} \end{pmatrix}$$

Luego obtenemos:

- Para θ_1:

$$t = \frac{\frac{Y_1 - Y_2}{2}}{\sigma\sqrt{0{,}5}}$$

- Para θ_2:

$$t = \frac{\frac{Y_1+2Y_2+Y_3}{6}}{\sigma\sqrt{0{,}1666}}$$

En donde la decisión de rechazar o no, dependerá del valor de $\hat{\sigma} = \frac{SSR}{n-2} = \frac{\sum(y_i-\hat{y})^2}{n-2}$.

EJERCICIO 63

Suponga que se tiene interés en ajustar un modelo de regresión lineal simple

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, ..., n$$

donde

$$\varepsilon_i \sim N(0, \sigma^2)$$

y β_0 y σ^2 son conocidos.

(a) Encuentre el estimador de mínimos cuadrados de β_1.

(b) ¿Cuál es la varianza del estimador encontrado en el inciso (a)?

(c) Encuentre una expresión para el intervalo de confianza del $100(1-\alpha)\,\%$ para la pendiente β_1. ¿Este intervalo es mayor que el intervalo correspondiente al caso donde tanto β_0 como β_1 son desconocidos?

SOLUCIÓN

(a) Estimador de mínimos cuadrados para β_1

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_o - \beta_1 x_i)^2$$

luego derivando con respecto al parámetro tenemos

$$\frac{\partial \sum_{i=1}^{n} \varepsilon_i^2}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \cdot x_i = 0$$

$$\Rightarrow \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \cdot x_i = 0$$

$$\Rightarrow \sum_{i=1}^{n} (x_i y_i - \beta_0 x_i - \beta_1 x_i^2) = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0$$

luego despejando y recordando que β_0 es conocido, nos queda

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2}$$

(b) Se pide calcular la $Var(\hat{\beta}_1)$.

$$Var(\hat{\beta}_1) = Var\left(\frac{\sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2} \right)$$

$$= \frac{1}{\left(\sum_{i=1}^{n} x_i^2 \right)^2} Var\left(\sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i \right)$$

$$= \frac{1}{\left(\sum_{i=1}^{n} x_i^2 \right)^2} Var\left(\sum_{i=1}^{n} x_i y_i \right)$$

$$= \frac{1}{\left(\sum_{i=1}^{n} x_i^2 \right)^2} \sum_{i=1}^{n} x_i^2 Var(y_i)$$

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

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

$$= \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$$

(c) Cuando ambos parámetros son desconocidos el intervalo para β_1 es de la siguiente forma:

$$\beta \in \left[\hat{\beta} \mp t_{(n-2);\frac{\alpha}{2}} \cdot s.e(\hat{\beta})\right]$$

donde

$$s.e(\hat{\beta}) = \sqrt{\frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}}$$

considerando σ^2 conocido.

EJERCICIO 64

Suponga que se especifica un modelo lineal simple sin intercepto

$$y_i = \mu x_i + \varepsilon_i, \quad i = 1, \ldots, n$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

(a) Encuentre el estimador de mínimos cuadrados de $\mu, \hat{\mu}$ y de $\sigma^2, \hat{\sigma}^2$.

(b) Calcule $E(\hat{\mu})$ y $Var(\hat{\mu})$.

(c) Estime la ecuación de regresión a partir del siguiente conjunto de datos

x	2	2	3	4	4
y	5	6	9	11	13

SOLUCIÓN

(a) Estimador de mínimos cuadrados para μ

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \mu x_i)^2$$

luego derivando con respecto al parámetro tenemos

$$\frac{\partial \sum_{i=1}^{n} \varepsilon_i^2}{\partial \mu} = -2 \sum_{i=1}^{n} (y_i - \mu x_i) \cdot x_i = 0$$

$$\Rightarrow \sum_{i=1}^{n} (y_i - \mu x_i) \cdot x_i = 0$$

$$\Rightarrow \sum_{i=1}^{n} (x_i y_i - \mu x_i^2) = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i - \mu \sum_{i=1}^{n} x_i^2 = 0$$

luego despejando nos queda

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

Ahora para σ^2 tenemos que el estimador es

$$\hat{\sigma}^2 = \frac{SS_E}{n-2}$$

$$= \frac{\sum_{i=1}^{n} \varepsilon_i^2}{n-2}$$

$$= \frac{\sum_{i=1}^{n}(y_i - \hat{\mu}x_i)^2}{n-2}$$

$$= \frac{\sum_{i=1}^{n} y_i^2 - 2\hat{\mu}\sum_{i=1}^{n} x_i y_i + \hat{\mu}^2 \sum_{i=1}^{n} x_i^2}{n-2}$$

$$= \frac{\sum_{i=1}^{n} y_i^2 - \frac{2\left(\sum_{i=1}^{n} x_i y_i\right)^2}{\sum_{i=1}^{n} x_i^2} + \frac{\left(\sum_{i=1}^{n} x_i y_i\right)^2 \cancel{\sum_{i=1}^{n} x_i^2}}{\left(\sum_{i=1}^{n} x_i^2\right)^{\cancel{2}}}}{n-2}$$

$$= \frac{\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} x_i y_i\right)^2}{\sum_{i=1}^{n} x_i^2}}{n-2}$$

$$= \frac{\sum_{i=1}^{n} y_i^2}{n-2} - \frac{\left(\sum_{i=1}^{n} x_i y_i\right)^2}{(n-2)\left(\sum_{i=1}^{n} x_i^2\right)}$$

(b)

$$E(\hat{\mu}) = \frac{1}{\sum_{i=1}^{n} x_i^2} E\left(\sum_{i=1}^{n} x_i y_i\right)$$

$$= \frac{1}{\sum_{i=1}^{n} x_i^2} \sum_{i=1}^{n} x_i E\left(y_i\right)$$

$$= \frac{1}{\sum_{i=1}^{n} x_i^2} \sum_{i=1}^{n} x_i \cdot \mu x_i$$

$$= \frac{1}{\sum_{i=1}^{n} x_i^2} \cdot \mu \cdot \sum_{i=1}^{n} x_i^2$$

$$= \mu$$

$\therefore$ $\hat{\mu}$ es un estimador insesgado.

(c) Para estimar la recta debemos solo calcular $\hat{\mu}$ en base a los datos

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} = \frac{10 + 12 + 27 + 44 + 52}{4 + 4 + 9 + 16 + 16} = 2{,}959$$

luego la recta pedida es de la forma:

$$\hat{y} = 2{,}959 \cdot x_i$$

EJERCICIO 65

Se presentan los siguientes datos sobre $x =$ % de absorción de luz a $5800A$ e $y =$ pico de fotovoltaje:

x	4.0	8.7	12.7	19.1	21.4	24.6	28.9	29.8	30.5
y	0.12	0.28	0.55	0.68	0.85	1.02	1.15	1.34	1.29

(a) Construya una gráfica de dispersión de estos datos. ¿Qué sugiere?

(b) Obtenga la ecuación de la recta de regresión estimada suponiendo que el modelo de regresión lineal simple es apropiado.

(c) ¿Qué proporción de la variación observada en pico de fotovoltaje se puede explicar por el modelo de regresión?

(d) Pronostique el pico de fotovoltaje cuando el % de absorción sea 19.1 y calcule el valor del residuo correspondiente.

(e) Se piensa que hay una regresión lineal útil entre % de absorción y pico de fotovoltaje. ¿Esta de acuerdo?. Realice una prueba formal.

SOLUCIÓN

(a) Observando el gráfico de dispersión siguiente se sugiere que hay una asociación lineal entre el % de absorción de luz y el pico de fotovoltaje

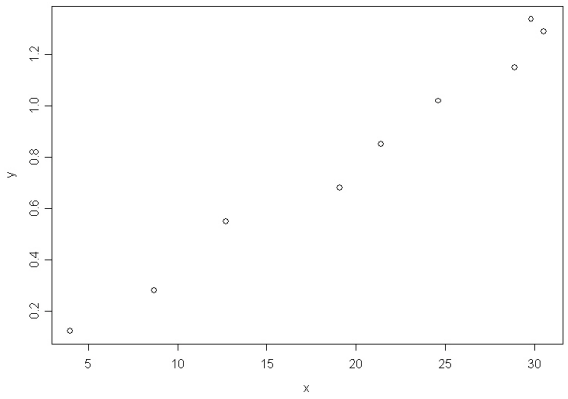

Figura 5.1: Gráfico de dispersión

(b) La ecuación de la recta estimada es la siguiente:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

con

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \qquad \text{y} \qquad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

de los datos obtenemos que

$$\overline{y} = 0{,}8088889 \qquad \text{y} \qquad \overline{x} = 19{,}96667$$

además se obtienen también

$$S_{xx} = \sum_{i=1}^{9}(x_i - \overline{x})^2 = 746{,}4$$

$$S_{yy} = \sum_{i=1}^{9}(y_i - \overline{y})^2 = 1{,}514089$$

$$S_{xy} = \sum_{i=1}^{9}(x_i - \overline{x})(y_i - \overline{y}) = 33{,}32567$$

luego, reemplazando tenemos que

$$\hat{\beta}_0 = -0{,}08259353 \qquad \text{y} \qquad \hat{\beta}_1 = 0{,}04464854$$

quedando la ecuación de la recta estimada como sigue

$$\hat{y}_i = -0{,}08259353 + 0{,}04464854 \cdot x_i$$

(c) Lo que se pide corresponde a la definición de R^2

$$R^2 = \frac{SS_R}{S_{yy}} = 1 - \frac{SS_E}{S_{yy}}$$

donde

$$\begin{aligned} SS_E &= S_{yy} - \hat{\beta}_1 \cdot S_{xy} \\ &= 1{,}514089 - 0{,}04464854 \cdot 33{,}32567 \\ &= 0{,}02614669 \end{aligned}$$

reemplazando tenemos que

$$\begin{aligned} R^2 &= 1 - \frac{0{,}02614669}{1{,}514089} \\ &= 0{,}9827311 \\ &\approx 98{,}27\,\% \end{aligned}$$

(d) El pronostico cuando el % de absorción es de 19.1 es

$$\begin{aligned} \hat{y} &= -0{,}08259353 + 0{,}04464854 \cdot 19{,}1 \\ &= 0{,}7701936 \end{aligned}$$

El residuo será la diferencia entre el verdadero valor observado para $x = 19{,}1$ y el calculado por medio de la recta de estimación

x	y	$\hat{y}$
19,1	0,68	0,7701936

luego el residuo es

$$e = 0{,}68 - 0{,}7701936 = -0{,}0901936$$

(e) Dado que se pide una verificación para la regresión lineal, tenemos que probar si el coeficiente β_1 es significativo, es decir distinto de cero.

La hipótesis a docimar es la siguiente:

$$\mathrm{H}_0 : \beta_1 = 0 \qquad \text{vs} \qquad \mathrm{H}_1 : \beta_1 \neq 0$$

El estadístico de prueba es

$$t_c = \frac{\hat{\beta}_1 - 0}{s.e(\hat{\beta}_1)}$$

Se rechaza H_0 si $|t_c| > t_{(n-2);\frac{\alpha}{2}}$

Tenemos que

$$s.e(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

y

$$\hat{\sigma}^2 = \frac{SS_E}{n-2}$$

$$= \frac{0{,}02614669}{7}$$

$$= 0{,}003735241$$

luego el estadístico T_c queda

$$T_c = \frac{0{,}04464854 - 0}{\sqrt{\frac{0{,}003735241}{746{,}4}}} = 19{,}95877$$

al comparar tenemos que $|T_c| = 19{,}95877 > t_{7;1-\frac{\alpha}{2}} = 2{,}364624$.

Existe suficiente evidencia para rechazar H_0 con un 95 % de confianza, es decir, hay una relación lineal útil entre el % de absorción de luz y el pico de fotovoltaje.

EJERCICIO 66

Se ha observado que los campos de atracción se forman con mayor frecuencia si los núcleos están cercanos. En un experimento se colocaron 20 núcleos a distancias diferentes y se midió la incidencia de campos de atracción (Y) para las diferencias distancias (X). Lamentablemente se borro parte del análisis de regresión y se le solicita completarlo.

(a) Completa la tabla ANOVA que se entrega a continuación:

Tabla ANOVA

Fuente	g.l	SS	MS	F
Regresión	1	2.0559		301.08
Error				
Total				

(b) ¿Qué porcentaje de la variable total esta siendo explicada por el modelo?

(c) Utilizando la siguiente información realice test de hipótesis para los parámetros del modelo. Concluya.

```
Ecuaci{\'o}n de regresi{\'o}n Y = 1.18 - 0.278 X

Predictor  Coef     Stdev    t-ratio Const      1.176232 0.03839
30.64 Distancia -0.278010 0.01602 -17.35
```

SOLUCIÓN

(a) La tabla ANOVA esta compuesta de los siguientes elementos.

Tabla ANOVA

Fuente	g.l	SS	MS	F
Regresión	p	SS_R	SS_R/p	MS_R/MS_E
Error	$n-1-p$	SS_E	$SS_E/n-1-p$	
Total	$n-1$	SS_T		

luego la tabla queda

Tabla ANOVA

Fuente	g.l	SS	MS	F
Regresión	1	2.0559	301.08	0.1229
Error	18	0.1229	0.0068	
Total	19	2.1788		

(b) Lo que se pide es el R^2.

$$R^2 = \frac{SS_R}{S_{yy}} = 1 - \frac{SS_E}{Syy}$$

reemplazando tenemos que

$$R^2 = 1 - \frac{0{,}1229}{2{,}1788}$$

$$= 0{,}9435928$$

$$\approx 94{,}36\,\%$$

(c) Mediante el test T docimaremos las siguientes hipótesis:

$$\mathrm{H}_o : \beta_0 = 0 \qquad \text{vs} \qquad \mathrm{H}_1 : \beta_0 \neq 0$$

y

$$\mathrm{H}_o : \beta_1 = 0 \qquad \text{vs} \qquad \mathrm{H}_1 : \beta_1 \neq 0$$

donde la región de rechazo para este caso esta definida por

$$R : \left\{ |T_c| > t_{n-2;\frac{\alpha}{2}} \right\}$$

Como $t_{7;0,025} = 2{,}365$ tenemos que $|T_0| = 30{,}64 > t_{7;0,025}$ y $|T_1| = 17{,}35 > t_{7;0,025}$, en ambos casos se rechaza H_0 con un 95 %, es decir, los parámetros son significativos.

EJERCICIO 67

Se ha comprobado que las aleaciones amorfas tienen una excelente resistencia a la corrosión. En *Corrosion Science*(septiembre de 1993) se informó de la resistividad de una aleación amorfa de hierro, boro y silicio después de la cristalización. Se reconocieron cinco especímenes de la aleación a 700°C, cada uno durante un intervalo de tiempo distinto. Después se midió el potencial de pasivación -una medida de la resistividad de la aleación cristalizada- para cada especímenes. Los datos experimentales son los siguientes:

Tiempo de Recorrido (minutos) x	Potencial de Pasivación (m V) y
10	-408
20	-400
45	-392
90	-379
120	-385

(a) Construya un diagrama de dispersión para los datos.

(b) Suponiendo que la mejor forma de describir la relación entre las variables es con una línea recta, utilice el método de mínimos cuadrados para estimar la ordenada al origen y la pendiente de la línea recta. Interprete estos valores.

(c) Trace la línea de mínimos cuadrados sobre el diagrama de dispersión.

(d) Según la línea de mínimos cuadrados. ¿Cuál es el potencial de pasivación esperado y, cuando el tiempo de recocido es de $x = 30$ minutos?.

(e) Calcule el R^2 para este modelo. Proporcione una interpretación de esta cantidad.

(f) Realice los test individuales con $\alpha = 0{,}05$, $H_0 : \beta_i = 0$ vs $H_1 : \beta_i \neq 0$, $i = 0, 1$.

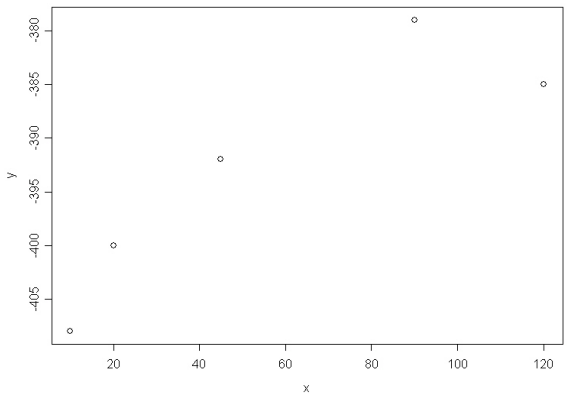

Figura 5.2: Gráfico de Dispersión

SOLUCIÓN

(a) En la figura se muestra el gráfico de dispersión de los datos.

(b) Lo estimadores de mínimos cuadrados para la ordenada de origen ($\hat{\beta}_0$) y la pendiente de la línea recta ($\hat{\beta}_1$) son:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \qquad \text{y} \qquad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

Para poder estimarlos necesitamos

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{5} (x_i - \overline{x})^2 &&= 8780 \\ S_{yy} &= \sum_{i=1}^{5} (y_i - \overline{Y})^2 &&= 534{,}8 \\ S_{xy} &= \sum_{i=1}^{5} (x_i - \overline{x})(y_i - \overline{y}) &&= 1918 \end{aligned}$$

además

$$\overline{y} = -392{,}8 \qquad \text{y} \qquad \overline{x} = 57$$

reemplazando tenemos que

$$\hat{\beta}_1 = 0{,}218451 \qquad \text{y} \qquad \hat{\beta}_0 = -405{,}2517$$

luego la recta es de la forma

$$\hat{y} = -405{,}2517 + 0{,}218451 \cdot x$$

(c) Ahora sobre el gráfico de dispersión se construye la recta de regresión.

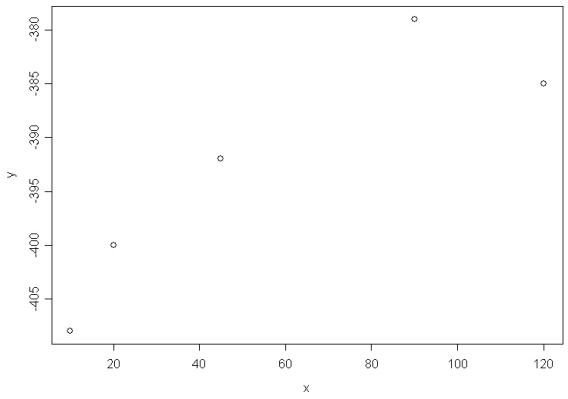

Figura 5.3: Recta de regresión

(d) El potencial de pasivación esperado y, cuando el tiempo de recorrido es de $x = 30$ minutos es

$$\hat{y} = -405{,}2517 + 0{,}218451 \cdot 30 = -398{,}6982$$

(e) Para poder calcular el R^2 necesitamos

$$SS_E = S_{yy} - \hat{\beta}_1 \cdot S_{xx} = 534{,}8 - 0{,}218451 \cdot 1918 = 115{,}811$$

luego reemplazando tenemos

$$\begin{aligned} R^2 &= \frac{SS_R}{S_{yy}} \\ &= 1 - \frac{SS_E}{S_{yy}} \\ &= 1 - \frac{115{,}811}{534{,}8} \\ &= 0{,}78345 \\ &\approx 78{,}35\,\% \end{aligned}$$

(f) Se pide docimar hipótesis para $\beta_0 = 0$ y $\beta_1 = 0$.

Docimemos primero la siguiente:

$$\mathrm{H}_0 : \beta_0 = 0 \qquad \text{vs} \qquad \mathrm{H}_1 : \beta_0 \neq 0$$

el estadístico de prueba es

$$t_c = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{Var(\hat{\beta})}}$$

$$= \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\hat{\sigma}\left\{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right\}}}$$

$$= \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\frac{SS_E}{n-2} \cdot \left\{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right\}}}$$

$$= \frac{-405{,}2517 - 0}{\sqrt{\frac{115{,}811}{5-2} \cdot \left\{\frac{1}{5} + \frac{57^2}{8780}\right\}}}$$

$$= -86{,}38847$$

La región de rechazo esta dada por

$$|t_c| > t_{(n-(k+1));\frac{\alpha}{2}}, \qquad \text{donde } k \text{ es el n}^{\text{o}} \text{ de variables explicativas}$$

considerando un $\alpha = 0{,}05$ (95 % de confianza)

$$t_{(n-(k+1));\frac{\alpha}{2}} = t_{5-2;0{,}025} = 3{,}182446$$

Como $|t_c| = 86{,}38847 > t_{3;0{,}025}(3) = 3{,}182446$, existe evidencia suficiente bajo un 95 % de confianza para rechazar H_0, es decir el parámetro β_0 es significativo, se puede considerar como distinto a cero.

Ahora docimemos la siguiente hipótesis

$$\mathrm{H}_0 : \beta_1 = 0 \qquad \text{vs} \qquad \mathrm{H}_1 : \beta_1 \neq 0$$

el estadístico de prueba es el estadístico de prueba es

$$t_c = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{Var(\hat{\beta}_1)}}$$

$$= \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\frac{\hat{\sigma}}{S_{xx}}}}$$

$$= \frac{0{,}218451 - 0}{\sqrt{\frac{\frac{115{,}811}{5-2}}{8780}}}$$

$$= 3{,}294481$$

La región de rechazo al igual que el caso anterior esta dada por

$$|t_c| > t_{(n-(k+1));\frac{\alpha}{2}}, \quad \text{donde } k \text{ es el n}^{\text{o}} \text{ de variables explicativas}$$

considerando nuevamente $\alpha = 0{,}05$ (95 % de confianza)

$$t_{(n-(k+1));\frac{\alpha}{2}} = t_{3;0{,}025} = 3{,}182446$$

Como $|t_c| = 3{,}294481 > t_{3;0{,}025} = 3{,}182446$, existe evidencia suficiente bajo un 95 % de confianza para rechazar H_0, es decir el parámetro β_1 es significativo y se puede considerar como distinto a cero.

EJERCICIO 68

La presencia de carburos duros en aleaciones de hierro blanco con alto cromo da como resultado una excelente resistencia a la abrasión, por lo mismo son adecuados para el manejo de materiales en la industria minera. Los datos de $y =$ pérdida por desgaste abrasivo (mm^3) y $x =$ contenido de austenita retenida (%), en pruebas de desgaste de pernos con granete como abrasivo, fueron analizados con un modelo de regresión lineal simple. Utilice el resultado que se presenta de MINITAB para contestar las siguientes preguntas:

- ¿Cuál es la ecuación de la recta de regresión estimada?
- Complete la tabla de análisis de varianza (tabla ANOVA).
- ¿Qué proporción de la variación observada de pérdida de desgaste se puede atribuir al modelo de regresión lineal simple para esa relación?
- Pruebe la utilidad del modelo de regresión lineal simple, use $\alpha = 0{,}01$.

- Estime la pérdida real promedio por desgaste cuando el contenido es 50 % ofreciendo información acerca de la confiabilidad y la precisión.

- ¿Qué valor de pérdida por desgaste pronosticaría cuando el contenido es 30 %, y cuál es el valor del residuo correspondiente, sabiendo que el valor observado fue de 0.80?

Otros datos relevantes:

$$\sum_{i=1}^{n} x_i^2 = 41574{,}84 \qquad \text{y} \qquad \overline{x} = 42{,}32941$$

```
Regression Analysis: y versus x

Predictor Coef         SE Coef     T      P
Constant  0.787218  0.09525879  8.264  0.0001
x         0.007570  0.00192626  3.930  0.0013

Analysis of Variance (tabla ANOVA)

Source          DF     SS       MS       F
Regression           0.63690
Residual
Error           15
Total                1.25551
```

SOLUCIÓN

(a) Con los datos entregados por la salida de Minitab la recta de regresión estimada es:

$$\hat{y} = 0{,}787218 + 0{,}007570x$$

(b) La tabla Anova queda como sigue:

Tabla ANOVA

Fuente	g.l	SS	MS	F
Regresión	1	0.6369	0.6369	15.443
Error	15	0.6186	0.04124	
Total	16	1.2555		

(c) Se pide el R^2.

$$R^2 = \frac{SS_R}{S_{yy}}$$

$$= \frac{0{,}6369}{1{,}255}$$

$$= 0{,}5072$$

$$\approx 50{,}72\,\%$$

(d) Observando el valor-p de x, tenemos que

$$\text{valor-p} = 0{,}0013 < 0{,}01 = \alpha$$

Por lo tanto de rechaza la hipótesis $\mathrm{H}_0 : \beta_1 = 0$.

(e) La estimación para la perdida real promedio por desgaste cuando el contenido es 50 % es:

$$\hat{y} = 0{,}787218 + 0{,}00757 \cdot 50 = 1{,}165718$$

(f) El valor de perdida por desgaste que pronosticaría cuando el contenido es 30 % es:

$$\hat{y} = 0{,}787218 + 0{,}00757 \cdot 30 = 1{,}014318$$

Sabiendo que le verdadero valor observado fue 0.8, el residuo es

$$e = y - \hat{y} = 0{,}8 - 1{,}014318 = -0{,}214318$$

EJERCICIO 69

Se ha observado que para predecir la demanda (consumo) de combustible para la calefacción, resulta ser más preciso el pronóstico a largo plazo de las temperaturas y el uso de la relación temperatura-consumo que el tratar de pronosticar directamente analizando las ventas de combustible. Un distribuidor de combustible mantiene un registro de ventas mensuales de combustible y de temperaturas máximas en esos meses. A continuación aparecen los datos de nueve de estos meses seleccionados al azar.

Ventas (y)	26.2	17.4	7.8	12.3	35.9	42.1	26.4	19.0	10.1
Temperaturas (x)	46.5	54.6	65.2	62.3	41.9	38.6	43.7	52.0	59.8

(a) Encuentre la recta de mínimos cuadrados para estos datos.

(b) Grafique los puntos y la recta como una verificación de sus cálculos.

(c) Utilice la ecuación de la recta ajustada para predecir la venta observada cuando la temperatura es de 50º F.

(d) Estime σ^2.

(e) Pruebe la significancia de la regresión con $\alpha = 0{,}5$. ¿A qué conclusiones puede llegarse?

(f) Encuentre un intervalo de confianza del 90 % para las ventas mensuales esperadas (medias) en aquellos meses en que el promedio de la temperatura máxima sea de 45º F.

(g) Calcule e interprete el R^2.

SOLUCIÓN

(a) Para poder calcular los estimadores de la recta de regresión, necesitamos los siguientes resultados:

	x	y	x^2	y^2	$x \cdot y$
	46.5	26.2	2162.25	686.44	1218.30
	54.6	17.4	2981.16	302.76	950.04
	65.2	7.8	4251.04	60.84	508.56
	62.3	12.3	3881.29	151.29	766.29
	41.9	35.9	1755.61	1288.81	1504.21
	38.6	42.1	1489.96	1772.41	1625.06
	43.7	26.4	1909.69	696.96	1153.68
	52.0	19.0	2704.00	361.00	988.00
	59.8	10.1	3576.04	102.01	603.98
Total	464.6	197.2	24711.04	5422.52	9318.12

de la tabla anterior se extraen los siguientes resultados:

$$\sum_{i=1}^{9} x_i = 464{,}6 \Rightarrow \overline{x} = 51{,}62$$

$$\sum_{i=1}^{9} y_i = 197{,}6 \Rightarrow \overline{y} = 21{,}91$$

$$\sum_{i=1}^{9} x_i^2 = 24711{,}04$$

$$\sum_{i=1}^{9} y_i^2 = 5422{,}52$$

$$\sum_{i=1}^{9} x_i y_i = 9318{,}12$$

Con esto podemos calcular

$$S_{xx} = \sum_{i=1}^{9} x_i^2 - 9\overline{x}^2 = 24711{,}04 - 9 \cdot (51{,}62)^2 = 729{,}4204$$

$$S_{yy} = \sum_{i=1}^{9} y_i^2 - 9\overline{y}^2 = 5422{,}52 - 9 \cdot (21{,}91)^2 = 1102{,}0871$$

$$S_{xy} = \sum_{i=1}^{9} x_i y_i - 9\overline{x} \cdot \overline{y} = 9318{,}12 - 9 \cdot (51{,}62)(21{,}91) = -860{,}8278$$

Luego

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-860{,}8278}{729{,}4204} = -1{,}180$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 21{,}91 + 1{,}180 \cdot 51{,}62 = 82{,}822$$

Donde la recta de regresión es:

$$\hat{y} = 82{,}822 - 1{,}180x$$

(b) El gráfico de los puntos y la recta de regresión se presenta a continuación

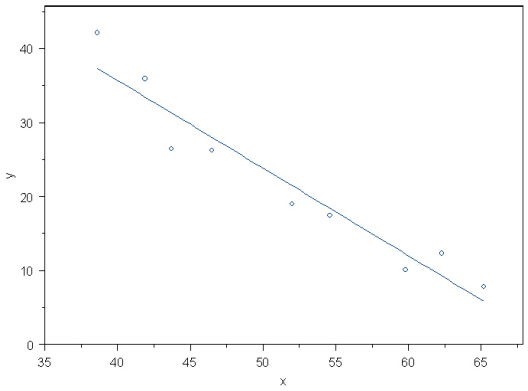

Figura 5.4: Recta de regresión

(c) La predicción de las ventas para una temperatura de 50º F es:

$$\hat{y} = 82{,}822 - 1{,}180 \cdot 50 = 23{,}822$$

(d) Tenemos que

$$\hat{\sigma}^2 = \frac{SS_E}{n-2} \qquad \text{y} \qquad SS_E = S_{yy} - \hat{\beta}_1 S_{xy}$$

Con los resultados obtenidos en (a) se calcula SS_E.

$$\begin{aligned} SS_E &= S_{yy} - \hat{\beta}_1 S_{xy} \\ &= 1102{,}0871 + 1{,}180 \cdot (-860{,}8278) \\ &= 86{,}31 \end{aligned}$$

Luego

$$\hat{\sigma}^2 = \frac{86{,}31}{7} = 12{,}33$$

(e) La hipótesis de significancia para la regresión es:

$$\mathrm{H}_o : \beta_1 = 0 \qquad \text{vs} \qquad \mathrm{H}_1 : \beta_1 \neq 0$$

Se rechaza H_o si $|T_1| > t_{n-2;\frac{\alpha}{2}}$, donde $T_1 = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$.

Para este caso tenemos que:

$$T_1 = \frac{-1{,}180}{se(\hat{\beta}_1)} \qquad \text{con} \qquad se(\hat{\beta}_1) = \sqrt{\frac{12{,}33}{729{,}4204}} = 0{,}13$$

$$T_1 = -\frac{1{,}180}{0{,}13} \approx -9{,}1$$

como $t_{7;0{,}025} = 2{,}365$ tenemos que $|T_1| > t_{7;0{,}025}$.

Luego rechazamos la hipótesis nula, es decir, los datos presentan suficiente evidencia de que las ventas de combustible están relacionadas linealmente con la temperatura.

(f) El intervalo pedido es el siguiente:

$$\mathrm{IC}(\mu_{y/|x_0}) = \hat{\mu}_{y|x_0} \mp t_{n-2;\frac{\alpha}{2}} \sqrt{\hat{\sigma}^2 \left\{ \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right\}}$$

Reemplazando

$$\begin{aligned} \mathrm{IC}(\mu_{y|45}) &= (82{,}822 - 1{,}180 \cdot (45)) \mp 1{,}895 \sqrt{12{,}33 \left\{ \frac{1}{10} + \frac{(45 - 51{,}62)^2}{729{,}4204} \right\}} \\ &= 29{,}72 \mp 2{,}66 \\ &= (27{,}06; 32{,}38) \end{aligned}$$

(g) $R^2 = 1 - \frac{SS_E}{S_{yy}} = 1 - \frac{86{,}31}{1102{,}0871} = 0{,}9268$

Existe un 92.68 % de variación en los datos mensuales que se explica por la temperatura máxima promedio.

EJERCICIO 70

El grabado con plasma para transferir figuras de líneas finas en procesos actuales de semiconductores. Los siguientes datos se refieren al flujo de cloro (X, en cm^3 normales por minuto) por una boquilla, utilizada en el mecanismo de grabado, y la rapidez de grabado (Y, en 100 A/min).

X	1.5	1.5	2.0	2.5	2.5	3.0	3.5	3.5	4.0
Y	23.0	24.5	25.0	30.0	33.5	40.0	40.5	47.0	49.0

Los estadísticos de resumen son: $\sum x_i = 24{,}0$, $\sum x_i^2 = 70{,}50$, $\sum y_i = 312{,}5$, $\sum y_i^2 = 11626{,}75$, $\sum x_i y_i = 902{,}25$, $\hat{\beta}_0 = 6{,}448718$, $\hat{\beta}_1 = 10{,}602564$.

(a) ¿El modelo de regresión lineal simple especifica una relación útil entre el flujo de cloro y la rapidez de grabado?

(b) Estime el cambio real promedio de rapidez de grabado asociado con un aumento de 1 cm^3 normal por minuto en el flujo, con un intervalo de confianza del 95 %, e interprete el intervalo.

(c) Calcule el intervalo de confianza de 95 % de confianza para $\mu_{Y|x=3}$, la rapidez real promedio de grabado cuando el flujo es igual a 3. ¿Se estimó con precisión este promedio?

(d) Calcule el intervalo de predicción de 95 % de confianza para una sola observación futura sobre la rapidez de grabado cuando el flujo es igual a 3. ¿Es probable que la predicción sea exacta?

(e) ¿Recomendaría calcular un intervalo de predicción de 95 % para un flujo de 6? Explique.

SOLUCIÓN

(a) Hay que realizar el test:

$$\mathrm{H}_0 : \beta_1 = 0 \qquad \text{vs} \qquad \mathrm{H}_1 : \beta_1 \neq 0$$

El estadístico de prueba es

$$t_c = \frac{\hat{\beta}_1}{s.e(\hat{\beta}_1)}$$

Se rechaza H_0 si $|t_c| > t_{(n-2);\frac{\alpha}{2}}$

Tenemos que

$$s.e(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

Necesitamos los siguientes resultados

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{9} x_i^2 - 9\overline{x}^2 = 70{,}5 - 9 \cdot (\tfrac{24}{9})^2 = 6{,}5 \\ S_{yy} &= \sum_{i=1}^{9} y_i^2 - 9\overline{y}^2 = 11626{,}75 - 9 \cdot (\tfrac{312{,}5}{9})^2 = 776{,}06 \\ S_{xy} &= \sum_{i=1}^{9} x_i y_i - 9\overline{x} \cdot \overline{y} = 902{,}25 - 9 \cdot (\tfrac{24}{9})(\tfrac{312{,}5}{9}) = 68{,}92 \end{aligned}$$

luego reemplazando obtenemos el valor de $\hat{\sigma}^2$ como sigue:

$$\begin{aligned} \hat{\sigma}^2 &= \frac{SS_E}{n-2} \\ &= \frac{S_{yy} - \hat{\beta}_1 S_{xy}}{n-2} \\ &= \frac{776{,}06 - 10602564 \cdot 68{,}92}{9-2} \\ &= \frac{45{,}33}{7} \\ &= 6{,}48 \end{aligned}$$

Luego

$$s.e(\hat{\beta}_1) = \sqrt{\frac{6{,}48}{6{,}2}} = 0{,}998$$

por lo tanto es estadístico de prueba queda como

$$t_c = \frac{10{,}602564}{0{,}998} \approx 10{,}62$$

y considerando $\alpha = 5\,\%$ tenemos que

$$t_{(n-2);\frac{\alpha}{2}} = t_{7;0,025} = 2{,}365$$

como $|t_c| > 2{,}365$ se rechaza H_0

Por lo tanto el modelo de regresión lineal especifica un regresión útil entre X e Y.

(b) Hay que encontrar un I.C. para β_1:

$$\beta_1 \in \left[\hat{\beta}_1 \mp t_{(n-2);\frac{\alpha}{2}} \cdot s.e(\hat{\beta}_1)\right]$$

$$\beta_1 \in [10{,}602564 \mp 2{,}365 \cdot 0{,}998]$$

$$\beta_1 \in [8{,}2422; 12{,}9628]$$

Con un 95 % de confianza, estimamos el cambio real promedio de rapidez de grabado entre 8.2422 y 12.9628 asociado con un aumento de 1 cm^3 normal por minuto en el flujo.

(c) El intervalo pedido es el siguiente:

$$\text{IC}(\mu_{y|x_0}) = \hat{\mu}_{y|x_0} \mp t_{n-2;\frac{\alpha}{2}} \sqrt{\hat{\sigma}^2 \left\{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right\}}$$

luego necesitamos

$$\overline{x} = \frac{24}{9} = 2{,}67$$

además

$$\hat{\mu}_{y|x_0=3} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

$$= 6{,}448718 + 10{,}602564 \cdot 3$$

$$= 38{,}25641$$

ahora reemplazando tenemos que

$$\mu_{y/x_0=3} \in \left[38{,}25641 \mp 2{,}365\sqrt{6{,}48\left\{\frac{1}{9} + \frac{(3-2{,}67)^2}{6{,}5}\right\}}\right]$$

$$\mu_{y/x_0=3} \in [38{,}25641 \mp 2{,}365 \cdot 0{,}9102]$$

$$\mu_{y/x_0=3} \in [36{,}10; 40{,}41]$$

Se aprecia que si se estimo con precisión este promedio, ya que si observamos la tabla de datos cuando $x = 3$ el valor de $y = 40$ y este valor pertenece al I.C.

(d) El intervalo pedido es el siguiente:

$$\text{IC}(y_0) \in y_0 \mp t_{n-2;\frac{\alpha}{2}} \sqrt{\hat{\sigma}^2 \left\{ 1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right\}}$$

necesitamos

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

$$= 6{,}448718 + 10{,}602564 \cdot 3$$

$$= 38{,}25641$$

Luego reemplazando

$$y_0 \in \left[38{,}25641 \mp 2{,}365 \sqrt{6{,}48 \left\{ 1 + \frac{1}{9} + \frac{(3 - 2{,}67)^2}{6{,}5} \right\}} \right]$$

$$y_0 \in [38{,}25641 \mp 2{,}365 \cdot 2{,}70]$$

$$y_0 \in [31{,}87; 44{,}64]$$

Por lo tanto es probable que la predicción sea exacta.

(e) No cambia ya que el nivel de confianza es el mismo.

(f) Como el valor 6.0 está muy alejado del rango en el cual varía x no sería recomendable calcular un I.C.

EJERCICIO 71

Es difícil determinar la resistencia al corte de puntos de soldadura, mientras que es relativamente sencillo medir el diámetro de soldadura de puntos. Seria ventajoso si se pudiera predecir la resistencia al corte de una medición del diámetro de soldadura. Los datos son:

Y: Resistencia al corte (psi)	X: Diámetro de soldadura (0.0001 pulg.)
370	400
780	800
1210	1210
1560	1600
1980	2000
2450	2500
3070	3100
3550	3600
3940	4000
3950	4000

(a) ¿Existe evidencia para pensar que el ajuste de una regresión lineal es adecuada?

(b) Docime si la correlación entre ambas variables es nula.

(c) Determine la recta por mínimos cuadrados.

(d) Calcule las varianzas de los parámetros encontrados.

(e) Docime la las hipótesis que $\beta_1 = 1$ y $\beta_0 = 0$, usando un nivel de significación igual a 0.01.

(f) Rectifique el punto anterior usando intervalos de confianza adecuados.

SOLUCIÓN

(a) La evidencia se puede obtener al graficar los puntos o calcular el coeficiente de correlación r.

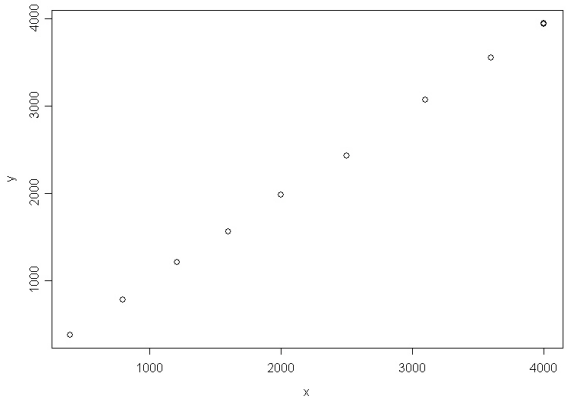

Figura 5.5: Gráfico de puntos

del gráfico se aprecia una fuerte asociación lineal entre la resistencia al corte y el diámetro de la soldadura.

Calculemos ahora el coeficiente de correlación para verificar esta apreciación.

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

$$= \frac{n \cdot \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right] \cdot \left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}$$

$$= \frac{10 \cdot 68674100 - 23210 \cdot 22860}{\sqrt{(10 \cdot 69644100 - 538704100) \cdot (10 \cdot 67719400 - 522579600)}}$$

$$= \frac{156160400}{\sqrt{15773690 \cdot 154614400}}$$

$$= \frac{156160400}{156167846}$$

$$= 0{,}9999$$

Por lo tanto, como $r = 0{,}9999 \approx 1$ hay una fuerte asociación lineal entre el diámetro de soldadura la resistencia al corte, misma conclusión obtenida observando el gráfico. Hay evidencia empírica para pensar que el ajuste de la regresión lineal es adecuado.

(b) La hipótesis que se pide docimar es la siguiente:

Sea ρ: correlación

$$\mathrm{H}_0 : \rho = \rho_0 \qquad vs \qquad \mathrm{H}_1 : \rho > \rho_0$$

esta prueba de hipótesis tiene una región de rechazo dada por:

$$R = \{Z_c > z_{1-\alpha}\}$$

donde

$$Z_c = \frac{\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) - \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right)}{\frac{1}{\sqrt{n-3}}}$$

$$= \frac{\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)}{\frac{1}{\sqrt{n-3}}}$$

$$= \frac{\frac{1}{2}\ln\left(1999\right)}{\frac{1}{\sqrt{7}}}$$

$$= 10{,}05439$$

considerando $\alpha = 0{,}05$ tenemos que $z_{1-0{,}005} = 1{,}64$.

Luego, como $Z_c = 10{,}05439 > z_{1-\alpha} = 1{,}64$, existe evidencia suficiente para rechazar H_0, esto implica que la correlación entre ambas variables no es nula.

(c) La estimación de la recta por mínimos cuadrados está dada por

$$\hat{y} = \hat{\alpha} + \hat{\beta}x$$

donde $\hat{\alpha}$ y $\hat{\beta}$ son los estimadores de mínimos cuadrados.

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$= \frac{n \cdot \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$= \frac{10 \cdot 68674100 - 23210 \cdot 2280}{10 \cdot 69644100 - 538704100}$$

$$= 0{,}99$$

$$\hat{\alpha} = \overline{y} - \hat{\beta} \cdot \overline{x}$$

$$= 22{,}86 - 0{,}99 \cdot 2321$$

$$= 2286 - 2297{,}79$$

$$= -11{,}79$$

Así la recta es

$$\hat{y} = -11{,}79 + 0{,}99 \cdot x$$

(d) $Var(\hat{\beta}) = \left[\dfrac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{\left(\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^2} \right] Var(y_i)$

como $Var(y_i) = \sigma^2$ tenemos que

$$Var(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

ahora para $\hat{\alpha}$

$$Var(\hat{\alpha}) = Var(\overline{y}) + \overline{x}^2 Var(\hat{\beta}) - 2\overline{x} Cov(\overline{y}, \hat{\beta})$$

$$= \frac{\sigma^2}{n} + \overline{x}^2 \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} - 2\overline{x} \cdot 0$$

$$= \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \right]$$

$$= \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n \cdot \sum_{i=1}^{n}(x_i - \overline{x})^2}$$

pero nótese que σ^2 hay que estimarlo.

$$\hat{\sigma}^2 = \frac{SCE}{n-1}, \quad \text{donde} \quad SCE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

luego reemplazando tenemos que

$$\begin{aligned}\sum_{i=1}^{10} e_i &= (-14{,}19942)^2 + (-0{,}2016205)^2 + (23{,}89612)^2 + (-12{,}20603)^2 + (11{,}79177)^2 \\ &\quad + (-13{,}21099)^2 + (12{,}78571)^2 + (-2{,}217046)^2 + (-8{,}21925)^2 + (1{,}78075)^2 \\ &= 1474{,}369\end{aligned}$$

luego la estimación de σ^2 es

$$\hat{\sigma}^2 = \frac{1474{,}369}{10-1} = 163{,}8187$$

por lo tanto, como tenemos que $\sum_{i=1}^{10} (x_i - \overline{x})^2 = 15773690$ y al reemplazar en las igualdades obtenidas para $\hat{\alpha}$ y $\hat{\beta}$ se calculan las varianzas para estos parámetros:

$$\begin{aligned}Var(\hat{\beta}) &= \frac{163{,}8187}{15773690} \\ &= 0{,}00001038557\end{aligned}$$

$$\begin{aligned}Var(\hat{\alpha}) &= \frac{163{,}8187 \cdot 69644100}{10 \cdot 15773690} \\ &= \frac{11409007637}{157736900} \\ &= 72{,}32935\end{aligned}$$

(e) Se pide docimar hipótesis para $\alpha = 0$ y $\beta = 1$.

Docimemos primero la siguiente:

$$\mathrm{H}_0 : \alpha = 0 \qquad \text{vs} \quad \mathrm{H}_1 : \alpha \neq 0$$

el estadístico de prueba es

$$t_c = \frac{\hat{\alpha} - \alpha}{\sqrt{Var(\hat{\alpha})}}$$

$$= \frac{-11{,}79 - 0}{\sqrt{72{,}32935}}$$

$$= -1{,}386298$$

La región de rechazo esta dada por

$$|t_c| > t_{(n-(k+1));\frac{\alpha}{2}}, \quad \text{donde } k \text{ es el n}^{\text{o}} \text{ de variables explicativas}$$

considerando un $\alpha = 0{,}05$ (95 % de confianza)

$$t_{(n-(k+1));\frac{\alpha}{2}} = t_{10-2;0{,}995} = 3{,}355387$$

Como $|t_c| = 1{,}386298 \not> t_{8;0{,}005}(8) = 3{,}355387$, no existe evidencia suficiente bajo un 99 % de confianza para rechazar H_0, es decir el parámetro α no sería significativo se puede considerar como igual a cero.

Ahora docimemos la siguiente hipótesis

$$H_0 : \beta = 1 \qquad \text{vs} \qquad H_1 : \beta \neq 1$$

el estadístico de prueba es

$$t_c = \frac{\hat{\beta} - \beta}{\sqrt{Var(\hat{\beta})}}$$

$$= \frac{0{,}99 - 1}{\sqrt{0{,}00001038557}}$$

$$= -3{,}103022$$

La región de rechazo al igual que el caso anterior esta dada por

$$|t_c| > t_{(n-(k+1));\frac{\alpha}{2}}, \quad \text{donde } k \text{ es el n}^{\text{o}} \text{ de variables explicativas}$$

considerando nuevamente $\alpha = 0{,}05$ (95 % de confianza)

$$t_{(n-(k+1));\frac{\alpha}{2}} = t_{8;0{,}005} = 3{,}355387$$

Como $|t_c| = 3{,}103022 \not> t_{0{,}995}(8) = 3{,}355387$, no existe evidencia suficiente bajo un 99 % de confianza para rechazar H_0, es decir el parámetro β se puede considerar como igual a uno.

(f) Haremos I.C al 99 % para los parámetros α y β.

El I.C(α) esta dado por

$$\alpha \in \left[\hat{\alpha} \mp t_{(n-2);\frac{\alpha}{2}} \cdot s.e(\hat{\alpha})\right]$$

$$\alpha \in \left[-11{,}79 \mp 3{,}355387 \cdot \sqrt{72{,}32935}\right]$$

$$\alpha \in [-40{,}32645; 16{,}74645]$$

como en el I.C se encuentra el cero, se ratifica lo obtenido en (e)

El I.C(β) esta dado por

$$\beta \in \left[\hat{\beta} \mp t_{(n-2);\frac{\alpha}{2}} \cdot s.e(\hat{\beta})\right]$$

$$\beta \in \left[0{,}99 \mp 3{,}355387 \cdot \sqrt{0{,}00001038557}\right]$$

$$\beta \in [0{,}9791867; 1{,}000813]$$

como en el I.C se encuentra el uno, también se ratifica lo obtenido en (e) para β.

EJERCICIO 72

En su tesis para obtener el grado de ph.d. en Cs. de la Ingeniería, un individuo estudió el efecto de la variación de la razón agua/cemento en la resistencia del concreto de 200 lbs/yarda cúbica, obteniendo los siguientes resultados:

Razón agua/cemento	Resistencia
1.21	1.302
1.29	1.231
1.37	1.061
1.46	1.400
1.62	0.803
1.79	0.710

(a) Calcule el valor de r.

(b) Ajuste el modelo $y = \alpha + \beta x + e$.

(c) Pruebe las hipótesis $H_0 : \beta \geq 0$ versus $H_1 : \beta < 0$ con $\alpha = 0{,}05$. Obtenga el nivel de significación alcanzado correspondiente.

(d) Encuentre un IC al 90 % para la resistencia esperada de concreto para una razón agua/cemento de 1.5. ¿Qué pasaría con el IC si se trata de estimar promedios de resistencia para razones de agua/concreto de 0.3 y 2.7 %?.

SOLUCIÓN

(a) Tenemos que por fórmula la correlación es:

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

$$= \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}$$

$$= \frac{6 \cdot 8{,}709 - 8{,}74 \cdot 6{,}148}{\sqrt{(6 \cdot 12{,}965 - 8{,}74^2)(6 \cdot 6{,}569 - 6{,}148^2)}}$$

$$= -\frac{1{,}47952}{\sqrt{1{,}4024 \cdot 1{,}62}}$$

$$= -\frac{1{,}47952}{1{,}51}$$

$$= -0{,}9798$$

El valor del estadístico de correlación es muy cercano a -1, lo que sugiere fuerte asociación negativa entre la razón agua/cemento y la resistencia.

(b) Utilizando las estimaciones por el método de Mínimos Cuadrados, se obtiene que:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$= \frac{6 \cdot 8{,}709 - 8{,}74 \cdot 6{,}148}{6 \cdot 12{,}9625 - 8{,}74^2}$$

$$= -1{,}056$$

$$\alpha = \bar{y} - \hat{\beta}\bar{x}$$

$$= \frac{6{,}148}{6} - (-1{,}056)\frac{8{,}74}{6}$$

$$= 2{,}563$$

Por lo tanto, el modelo de linea recta que mejor ajusta a los datos es

$$\hat{y} = 2{,}563 - 1{,}056x$$

(c) Nótese que si se rechaza H_0, se concluye que $\beta < 0$, lo que significa que la resistencia tiende a disminuir con un incremento en la razón agua/cemento.

El estadístico de prueba es

$$t = \frac{\hat{\beta} - 0}{S_{\hat{\beta}}} = \frac{\hat{\beta} - 0}{S\sqrt{\frac{1}{\sum(x_i - \bar{x})^2}}}$$

donde $S = \sqrt{\frac{SCE}{n-2}}$, con $SCE = \sum(y_i - \bar{y})^2 - \hat{\beta}\sum(x_i - \bar{x})(y_i - \bar{y})$. Por lo tanto si descomponemos por partes, tenemos que:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2$$

$$= 6{,}569 - \frac{1}{6}6{,}148^2$$

$$= 0{,}269$$

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

$$= 8{,}709 - \frac{1}{6}8{,}74 \cdot 6{,}148$$

$$= -0{,}247$$

Luego se obtiene que $SCE = 0{,}008$ y por tanto $S = \sqrt{\frac{0{,}008}{4}} = 0{,}045$.

Además tenemos que

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 = 0{,}234$$

Finalmente, evaluando se obtiene:

$$t = \frac{-1{,}056 - 0}{0{,}045\sqrt{\frac{1}{0{,}234}}} = -11{,}355$$

Este test de hipótesis, rechaza H_0 si $t = -11{,}355 < -2{,}132 = t_{n-2;\alpha}$, lo cual se cumple en este caso, por lo tanto se rechaza H_0 en favor de H_1 con una 5 % de significancia.

Como la prueba es de una cola e inferior, se tiene que

$$\text{valor-p} = P(t < -11{,}355) = 0{,}0001715 < 0{,}005$$

Luego para los valores de α que se usan comúnmente, se concluye que hay evidencia para indicar que la resistencia decrece con un incremento en la razón agua/cemento (en el intervalo de datos contemplado).

(d) El intervalo de confianza está dado por la fórmula

$$\hat{y} \pm t_{(n-2);\frac{\alpha}{2}} S \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Lo que se pide es un intervalo de confianza cuando $x = 1{,}5$, luego evaluando se tiene

$$\hat{y} = 2{,}563 - 1{,}056 \cdot 1{,}5 = 0{,}979$$

Aplicando cálculos se obtiene, al 90 % de confianza

$$0{,}979 \pm 2{,}132 \cdot 0{,}045 \sqrt{\frac{1}{6} + \frac{(1{,}5 - 1{,}457)^2}{0{,}234}}$$

Por lo tanto $\hat{y} \in (0{,}938; 1{,}020)$, lo que significa que si se tiene una razón agua/cemento de 1.5, la resistencia estimada media está en ese intervalo con un 90 % de confianza.

Puede verse de la expresión de la Varianza que el IC se hace más grande cuando x se aleja más de $\bar{x} = 1{,}457$. Los valores de $x = 0{,}3$ y $x = 2{,}7$ se encuentran lejos de los valores que se usaron en el experimento, lo que produciría estimaciones lejanas a la realidad y probablemente un concreto completamente inservible.

EJERCICIO 73

Demuestre que el modelo de regresión lineal simple

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ pueden ser escritos como combinaciones lineales de las respuestas y_i. Encuentre explícitamente las constantes en la combinación lineal.

SOLUCIÓN

Sabemos que al

$$\min_{\beta_0 \beta_1} \sum_{i=1}^{n} \varepsilon_i^2$$

se obtiene:

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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

en el caso de $\hat{\beta}_1$ se tiene que

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) &= \sum_{i=1}^{n}(x_i y_i - x_i\overline{y} - \overline{x}y_i + \overline{xy}) \\
&= \sum_{i=1}^{n}(x_i y_i - \overline{y}(x_i - \overline{x}) - \overline{x}y_i \\
&= \sum_{i=1}^{n} y_i(x_i - \overline{x}) - \overline{y}(x_i - \overline{x}) \\
&= \sum_{i=1}^{n} y_i(x_i - \overline{x}) - \overline{y}\sum(x_i - \overline{x}) \\
&= \sum_{i=1}^{n} y_i(x_i - \overline{x}) - \overline{y}\cdot 0 \\
&= \sum_{i=1}^{n} y_i(x_i - \overline{x})
\end{aligned}$$

luego

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})}{S_{xx}} y_i = \sum_{i=1}^{n} d_i y_i$$

donde $d_i = \frac{(x_i - \overline{x})}{S_{xx}}$ y $S_{xx} = \sum(x_i - \overline{x})^2$.

Para $\hat{\beta}_0$ tenemos que

$$\hat{\beta}_0 = \overline{y} - \frac{\sum_{i=1}^{n}(x_i - \overline{x})}{S_{xx}} y_i \overline{x}$$

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

$$= \sum_{i=1}^{n} c_i y_i$$

con $c_i = \frac{1}{n} - \frac{\sum_{i=1}^{n}(x_i - \overline{x})}{S_{xx}} \overline{x}$

EJERCICIO 74

Demuestre que

$$\hat{\beta}_0 \sim N(\beta_0, \frac{\sigma^2}{nS_{xx}} \sum_{i=1}^{n} x_i^2) \quad y \quad \hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$$

con

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2 \overline{x}}{S_{xx}}$$

SOLUCIÓN

Como $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ y $\hat{\beta}_0$ es combinación lineal de y_i entonces $\hat{\beta}_0 \sim N(\cdot, \cdot)$ donde los parámetros son:

$$E(\hat{\beta}_0) = E(\sum_{i=1}^{n} c_i E(y_i))$$

$$= \sum_{i=1}^{n} c_i(\beta_0 + \beta_1 x_i)$$

$$= \sum_{i=1}^{n} \left\{ \frac{1}{n} - \frac{(x_i - \overline{x})}{S_{xx}\overline{x}} \right\} (\beta_0 + \beta_1 x_i)$$

$$= \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i$$

$$= \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} \left\{ \frac{x_i}{n} - \frac{(x_i - \overline{x})x_i\overline{x}}{S_{xx}} \right\}$$

$$= \beta_0 + \beta_1 \left\{ \overline{x} - \frac{\overline{x}}{S_{xx}} \sum_{i=1}^{n} (x_i - \overline{x})x_i \right\}$$

$$= \beta_0 + \beta_1 \left\{ \overline{x} - \frac{\overline{x}}{S_{xx}} S_{xx} \right\}$$

$$= \beta_0$$

$$Var(\hat{\beta}_0) = Var\left(\sum_{i=1}^{n} c_i y_i\right)$$

$$\overset{ind}{=} \sum_{i=1}^{n} Var(c_i y_i)$$

$$= \sigma^2 \sum_{i=1}^{n} c_i^2$$

$$= \sigma^2 \left[\frac{1}{n\, S_{xx}} \sum_{i=1}^{n} x_i^2\right]$$

para $\hat{\beta}_1$

$$E(\hat{\beta}_1) = E\left(\sum_{i=1}^{n} d_i y_i\right)$$

$$= \sum_{i=1}^{n} d_i(\beta_0 + \beta_1 x_i)$$

$$= \beta_0 \sum_{i=1}^{n} d_i + \beta_1 \sum_{i=1}^{n} d_i x_i$$

$$= 0 + \beta_1 \frac{\sum_{i=1}^{n} (x_i - \overline{x}) x_i}{S_{xx}}$$

$$= \beta_1$$

$$Var(\hat{\beta}_1) = Var\left(\sum_{i=1}^{n} d_i y_i\right)$$

$$\stackrel{ind}{=} \sigma^2 \sum_{i=1}^{n} d_i^2$$

$$= \sigma^2 \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{S_{xx}^2}$$

$$= \frac{\sigma^2}{S_{xx}}$$

Finalmente la covarianza es

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = Cov(\sum_{i=1}^{n} c_i y_i, \sum_{j=1}^{n} d_j y_j)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} c_i d_j Cov(y_i, y_j)$$

$$\stackrel{ind}{=} \sigma^2 \sum_{j=i=1}^{n} c_i d_i$$

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

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

$$= \sigma^2 \left(0 - \frac{\overline{x}}{S_{xx}}\right)$$

$$= -\frac{\sigma^2 \overline{x}}{S_{xx}}$$

EJERCICIO 75

Montgomery y Peck (1992) describen el uso de un modelo de regresión para relacionar la cantidad de tiempo que requiere un vendedor para dar servicio a una maquina expendedora de refrescos, con el número de envases contenidos en la maquina (X_1) y la distancia del vehículo de servicio al sitio donde se encuentra la máquina (X_2). Los datos se presentan a continuación:

Obs.	Y	X_1	X_2
1	9,95	2	50
2	24,45	8	110
3	31,75	11	120
4	35,00	10	550
5	25,02	8	295
6	16,86	4	200
7	14,38	2	375
8	9,60	2	52
9	24,35	9	100
10	27,50	8	300
11	17,08	4	412
12	37,00	11	400
13	41,95	12	500
14	11,66	2	360
15	21,65	4	205
16	17,89	4	400
17	69,00	20	600
18	10,30	1	585
19	34,93	10	540
20	46,59	15	250
21	44,88	15	290
22	54,12	16	510
23	56,23	17	590
24	22,13	6	100
25	21,15	5	400

(a) Construya el modelo.

(b) Determine paso a paso la tabla ANOVA y concluya.

SOLUCIÓN

(a) El modelo es el siguiente:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

para calcular el modelo ajustado

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

necesitamos encontrar los estimadores de mínimos cuadrados a partir de

$$\hat{\beta} = (X'X)^{-1} X'Y$$

luego tenemos que

$$(X'X) = \begin{bmatrix} n & \sum_{i=1}^{25} X_{i1} & \sum_{i=1}^{25} X_{i2} \\ \sum_{i=1}^{25} X_{i1} & \sum_{i=1}^{25} X_{i1}^2 & \sum_{i=1}^{25} X_{i1}X_{i2} \\ \sum_{i=1}^{25} X_{i2} & \sum_{i=1}^{25} X_{i1}X_{i2} & \sum_{i=1}^{25} X_{i2}^2 \end{bmatrix} = \begin{bmatrix} 25 & 206 & 8294 \\ 206 & 2396 & 77177 \\ 8294 & 77177 & 3531848 \end{bmatrix}$$

Invirtiendo (X'X) queda

$$(X'X)^{-1} = \begin{bmatrix} 0{,}2146526166 & -0{,}00749091422 & -3{,}403891e-004 \\ -0{,}0074909142 & 0{,}00167076313 & -1{,}891781e-005 \\ -0{,}0003403891 & -0{,}00001891781 & 1{,}495876e-006 \end{bmatrix}$$

y

$$X'Y = \begin{bmatrix} \sum_{i=1}^{25} Y_i \\ \sum_{i=1}^{25} X_{i1}Y_i \\ \sum_{i=1}^{25} X_{i2}Y_i \end{bmatrix} = \begin{bmatrix} 725{,}42 \\ 8001{,}67 \\ 274580{,}71 \end{bmatrix}$$

por lo tanto los estimadores de mínimos cuadrados son:

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 0{,}2146526166 & -0{,}00749091422 & -3{,}403891e-004 \\ -0{,}0074909142 & 0{,}00167076313 & -1{,}891781e-005 \\ -0{,}0003403891 & -0{,}00001891781 & 1{,}495876e-006 \end{bmatrix} \cdot \begin{bmatrix} 725{,}42 \\ 8001{,}67 \\ 274580{,}71 \end{bmatrix}$$

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 2{,}30920043 \\ 2{,}74036942 \\ 0{,}01243958 \end{bmatrix}$$

luego el modelo ajustado es

$$\hat{Y} = 2{,}30920043 + 2{,}74036942 \cdot X_1 + 0{,}01243958 \cdot X_2$$

(b) La tabla ANOVA tiene la siguiente forma

Tabla ANOVA

Fuente	g.l	SS	MS	F
Regresión	$k-1$	SS_R	$SS_R/(k-1)$	MS_R/MS_E
Error	$n-k$	SS_E	$SS_E/(n-k)$	
Total	$n-1$	SS_T		

para rellenarla necesitamos

$$SS_T = Y'Y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

$$= 27133{,}39 - 21049{,}37$$

$$= 6084{,}021$$

$$SS_R = \hat{\beta}'X'Y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

$$= 27018{,}34 - 21049{,}37$$

$$= 5968{,}974$$

$$SS_E = SS_T - SS_R$$

$$= 115{,}0465$$

nótese que $\hat{\sigma}^2 = S^2 = \frac{SS_E}{n-k}$, donde k es la cantidad de parámetros a estimar. Luego

$$\hat{\sigma}^2 = \frac{115{,}0465}{25-3} = 5{,}229388$$

Ahora la tabla rellenada queda como sigue:

Tabla ANOVA

Fuente	g.l	SS	MS	F
Regresión	2	5968.974	2984.487	570.7144
Error	22	115.0465	5.229388	
Total	24	6084.021		

Para docimar la hipótesis:

$$\mathrm{H}_0 : \beta_1 = \beta_2 = 0 \qquad \text{vs} \qquad \mathrm{H}_1 : \text{ Al menos un } \beta_i \neq 0 \text{ para } i = 1, 2$$

Se compara el F_{ANOVA} con un $F_{k-1;n-k}(1-\alpha)$ de tabla. Si

$$F_{ANOVA} > F_{k-1;n-k}(1-\alpha) \quad \Rightarrow \quad \text{se rechaza } \mathrm{H}_0$$

como

$$F_{ANOVA} = 570{,}7144 > 3{,}443357 = F_{2;22}(0{,}95)$$

Se rechaza H_0, es decir, la regresión es significativa.

Al calcular el R^2 tenemos que

$$R^2 = \frac{SS_R}{SS_T} = 0{,}9810904$$

luego el porcentaje de variabilidad presente en los datos es de 98.11 % aproximadamente.

5.2. Ejercicios Propuestos

1. Un artículo publicado en *Concrete Research* (Ñear Surface Characteristics of Concrete: Intrinsic Permeability', vol. 41, 1989) presenta datos sobre la resistencia a la compresión x y la permeabilidad intrínsica y de varias mezclas y tratamientos de concreto. El resumen de cantidades es el siguiente: $n = 14$, $\sum y_i = 572$, $\sum y_i^2 = 23530$, $\sum x_i = 43$, $\sum x_i^2 = 157{,}42$ y $\sum x_i y_i = 1697{,}80$. Suponga que las dos variables están relacionadas de acuerdo con el modelo de regresión lineal simple.

   a) Calcule las estimaciones de mínimos cuadrados de la pendiente y la ordenada al origen.

   b) Utilice la ecuación de la recta ajustada para predecir la permeabilidad que será observada cuando la resistencia a la compresión sea $x = 4{,}3$.

   c) Proporcione una estimación puntual de la permeabilidad promedio cuando la resistencia a la compresión para $x = 3{,}7$.

   d) Suponga que el valor observado de la permeabilidad para $x = 3{,}7$ es $y = 46{,}1$. Calcule el valor del residuo correspondiente.

2. Un artículo publicado en *Wear* (vol. 152, 1992, págs. 171-181) presenta datos sobre el desgaste del acero dulce y la viscosidad del aceite. A continuación aparecen datos representativos, con x = viscosidad del aceite y y = volumen de desgaste ($10^{-4} mm^3$).

y	240	181	193	155	172	110	113	75	94
x	1.6	9.4	15.5	20.0	22.0	35.5	43.0	40.5	33.0

   a) Construya una gráfica de dispersión de los datos. ¿Parece plausible el uso de un modelo de regresión lineal simple?

   b) Ajuste un modelo de regresión lineal simple utilizando la técnica de mínimos cuadrados.

   c) Haga una predicción sobre el desgaste cuando la viscosidad es $x = 30$.

   d) Obtenga el valor ajustado de y cuando $x = 22{,}0$ y calcule el residuo correspondiente.

3. Un artículo publicado en el *Journal of Environmental Engineering* (vol. 115, núm. 3, 1989, págs. 608-619) informa los resultados de un estudio sobre la aparición de sodio y cloro en los arroyos de la parte central de Rhode Island. los datos siguientes muestran la concentración de cloro y (en mg/l) y el área que rodea a la cuenca x (en porcentaje).

y	4.4	6.6	9.7	10.6	10.8	10.9	11.8	12.1	14.3
x	0.19	0.15	0.57	0.70	0.67	0.63	0.47	0.70	0.60
y	14.7	15.0	17.3	19.2	23.1	27.4	27.7	31.8	39.5
x	0.78	0.81	0.78	0.69	1.30	1.05	1.06	1.74	1.62

a) Dibuje un diagrama de dispersión de los datos. En este caso, ¿parece apropiado el uso de un modelo de regresión lineal simple?

b) Ajuste un modelo de regresión lineal simple utilizando el método de mínimos cuadrados.

c) Estime la concentración promedio de cloro para una cuenca que tiene un área que sea el 1 % de la superficie circunvecina.

d) Encuentre el valor ajustado que corresponde a $x = 0{,}47$ así como el residuo correspondiente.

4. Considere los datos del ejercicio 1. para x = resistencia a la compresión y y = permeabilidad intrínseca del concreto.

a) Pruebe la significancia de la regresión utilizando $\alpha = 0{,}05$. Encuentre el valor P de esta prueba. ¿Puede concluirse que el modelo especifica una relación lineal útil entre las dos variables?

b) Estime σ^2 y la desviación estándar de $\widehat{\beta}_1$.

c) En este modelo, ¿cuál es el error estándar de la ordenada al origen?

5. El ejercicio 3. contiene datos para y = concentración de cloro y x = área que rodea la cuenca.

a) Pruebe la hipótesis $H_0 : \beta_1 = 0$ contra $H_1 : \beta_1 \neq 0$ utilizando el procedimiento del análisis de varianza con $\alpha = 0{,}01$.

b) Encuentre el valor P de la prueba del inciso a).

c) Estime σ^2 y los errores estándar de $\widehat{\beta}_1$ y $\widehat{\beta}_0$.

d) Pruebe que $H_0 : \beta_0 = 0$ contra $H_1 : \beta_0 \neq 0$ con $\alpha = 0{,}01$. ¿Qué conclusiones pueden obtenerse? ¿Parece que el modelo ajustaría mejor los datos si se eliminase la ordenada al origen?

6. Con los datos del ejercicio 1. para x = resistencia a la compresión y y = permeabilidad intrínseca del concreto:

a) Encuentre un intervalo de confianza del 95 % para la pendiente.

b) Encuentre un intervalo de confianza del 95 % para la ordenada al origen.

c) Encuentre un intervalo de confianza del 95 % para la permeabilidad promedio cuando $x = 2{,}5$.

d) Encuentre un intervalo de confianza del 95 % para la permeabilidad cuando $x = 2{,}5$. Explique por qué este intervalo es mayor que el calculado en el inciso c).

7. Con respecto a los datos del ejercicio 2. sobre y = desgaste del acero dulce y x = viscosidad del aceite:

a) Encuentre un intervalo de confianza del 95 % para la ordenada al origen.

b) Encuentre un intervalo de confianza del 95 % para la pendiente.

c) Encuentre un intervalo de confianza del 95 % para el desgaste promedio del acero dulce cuando la viscosidad del aceite es $x = 30$.

8. El ejercicio 3. presenta datos sobre y = concentración de cloro y x = área de la cuenca en la región de la cuenca en la central de Rhode Island.

   a) Encuentre un intervalo de confianza del 99 % para β_1.

   b) Encuentre un intervalo de confianza del 99 % para β_0.

   c) Encuentre un intervalo de confianza del 99 % para la concentración promedio de cloro cuando el área es $x = 1{,}0\,\%$.

   d) Encuentre un intervalo de predicción del 99 % para la concentración de cloro cuando el área es $x = 1{,}0\,\%$.

9. El ejercicio 2. presenta datos sobre el volumen de desgaste y y viscosidad del aceite x.

   a) Calcule R^2 para este modelo. Proporcione una interpretación de esta cantidad.

   b) Haga una gráfica de los residuos de este modelo contra $\widehat{y}$ y contra x. Interprete estas gráficas.

   c) Prepare una gráfica de probabilidad normal de los residuos. ¿Parece ser que se satisface la hipótesis de normalidad?

10. Con respecto al ejercicio 3.:

    a) ¿Qué proporción de la variabilidad total en la concentración de cloro está explicada por el modelo de regresión?

    b) Utilice las observaciones repetidas en $x = 70$ y $x = 78$ para obtener una estimación del error puro con dos grados de libertad.

    c) Utilice el error puro calculado en el inciso b) para probar la falta de ajuste del modelo de regresión. Utilice $\alpha = 0{,}05$. ¿Qué conclusioón puede obtenerse sobre lo adecuado del modelo?

    d) Haga una gráfica de los residuos contra $\widehat{y}$ y contra x. Interprete las gráficas.

    e) Prepare una gráfica de probabilidad normal de los residuos. ¿Parece que se satisface la hipótesis de normalidad?

11. A continuación se proporcionan los resultados obtenidos en la prueba final y los exámenes de 20 estudiantes seleccionados al azar, que tomaron un curso de estadística para ingenieros y otro en investigación de operaciones. Supóngase que los promedios finales tienen una distribución conjunta normal.

Estadística	86	75	69	75	90	94	83	86	71	65
IO	80	81	75	81	92	95	80	81	76	72

Estadística	84	71	62	90	83	75	71	76	84	97
IO	85	72	65	93	81	70	73	72	80	98

a) Encuentre la recta de regresión que relaciona el promedio final en estadística con el promedio final en IO.

b) Pruebe la significancia de la regresión con $\alpha = 0{,}05$.

c) Estime el coeficiente de correlación.

d) Pruebe la hipótesis de que $\rho = 0$, utilizando para ello $\alpha = 0{,}05$.

e) Pruebe la hipótesis de que $\rho = 0{,}5$ utilizando $\alpha = 0{,}05$.

f) Construya un intervalo de confianza del 95 % para el coeficiente de correlación.

12. Se toma una muestra aleatoria de 50 observaciones sobre el diámetro de puntos de soldadura y el valor correspondiente de la resistencia al esfuerzo cortante.

a) Dado que $r = 0{,}62$, pruebe la hipótesis de que $\rho = 0$ utilizando $\alpha = 0{,}01$. ¿Cuál es el valor P de esta prueba?

b) Encuentre un intervalo de confianza del 99 % para ρ.

c) Con base en el intervalo de confianza del inciso b), ¿puede concluirse que $\rho = 0{,}5$ con un nivel de significancia de 0.01?

13. Los ingenieros civiles a menudo utilizan la ecuación de línea recta $E(y) = \widehat{\beta}_0 + \widehat{\beta}_1 x$ para modelar la relación entre la resistencia de corte media $E(y)$ de las juntas de albañilería y el esfuerzo de precompresión x. Con objeto de probar esta teoría, se realizó una serie de pruebas de esfuerzo con tabiques sólidos dispuestos en tripletas y unidos con mortero (*Proceedings of the Institute of Civil Engineers*, marzo de 1990). Se varió el esfuerzo de compresión para cada tripleta y se registró la carga de corte máxima justo antes de la ruptura (llamada resistencia de corte). En la tabla se indican los resultados de esfuerzo para 7 tripletas (medidos en $\mathrm{N/mm^2}$).

Prueba de tripleta	1	2	3	4	5	6	7
Resistencia al corte, y	1.00	2.18	2.24	2.41	2.59	2.82	3.06
Esfuerzo de compresión, x	0	.06	1.20	1.33	1.43	1.75	1.75

a) Grafique los siete puntos de datos en un diagrama de dispersión. ¿Parece ser lineal la relación entre la resistencia de corte y el esfuerzo de precompresión?

b) Utilice el método de mínimos cuadrados para estimar los parámetros del modelo lineal.

c) Interprete los valores de $\widehat{\beta}_0$ y $\widehat{\beta}_1$.

14. El artículo "*Some Field Experience in the Use of an Accelerated Method in Estimating 28–Day Strength of Concrete*" (**J. Amer. Concrete Institute, 1969, p. 895**) consideró la regresión de la resistencia estándar de curado $y = 28$ días (en $\mathrm{lb/pulg^2}$) contra $x =$ resistencia acelerada (en $\mathrm{lb/pulg^2}$). Suponga que la ecuación de la verdadera recta de regresión es $y = 1800 + 1{,}3x$.

(a) ¿Cuál es el valor esperado de la resistencia de 28 días cuando la resistencia acelerada = 2500?.

(b) ¿Cuánto podemos esperar que cambie la resistencia de 28 días cuando la resistencia acelerada aumenta en 1 lb/pulg^2.

(c) Conteste el inciso (b) para un aumento de 100 lb/pulg^2.

(d) Conteste el inciso (b) para una disminución de 100 lb/pulg^2.

15. Refiérase al estudio de Vietnam expuestos al agente Naranja (y la dioxina 2,3,7,8-TCDD). La tabla de datos, que se reproduce a continuación, proporciona las cantidades de 2,3,7,8-TCDD (medidas en partes por millón) tanto en plasma sanguíneo como un tejido graso extraídos de cada uno de los 20 veteranos estudiados. Un objetivo de los investigadores es determinar el grado de asociación lineal entre el nivel de dioxina observado en plasma sanguíneo y en tejido graso. Si se puede establecerse una asociación lineal entre las dos variables, los investigadores querrán construir modelos para: (1) predecir el nivel de 2,3,7,8-TCDD observado en tejido graso y (2) predecir el nivel en tejido graso a partir del nivel en plasma sanguíneo.

Veterano	Niveles de TCDD en plasma	Niveles de TCDD en tejido graso
1	2.5	4.9
2	3.1	5.9
3	2.1	4.4
4	3.5	6.9
5	3.1	7.0
6	1.8	4.2
7	6.0	10.0
8	3.0	5.5
9	36.0	41.0
10	4.7	4.4
11	6.9	7.0
12	3.3	2.9
13	4.6	4.6
14	1.6	1.4
15	7.2	7.7
16	1.8	1.1
17	20.0	11.0
18	2.0	2.5
19	2.5	2.3
20	4.1	2.5

a) Encuentre las ecuaciones de predicción que necesitan los investigadores. Interprete los resultados.

b) Pruebe la hipótesis de que el nivel en tejido graso (x) sirve para predecir linealmente el nivel en plasma sanguíneo (y). Utilice $\alpha = 0{,}05$.

c) Pruebe la hipótesis de que el nivel en plasma sanguíneo (x) sirve para predecir linealmente el nivel en tejido graso (y). Utilice $\alpha = 0{,}05$.

d) Intuitivamente, ¿por qué deben coincidir los resultados de los incisos b) y c)?

16. Se realizó un experimento con objeto de estudiar el agrietamiento por esfuerzos de corrosión de acero inoxidable tipo 304 en un entorno simulado de reactor con agua en ebulición (*Transactions of the ASME*, enero de 1986). Seis especímenes de acero inoxidable se recocieron y se sensibilizaron en agua a 289°C con oxígeno y sulfato disueltos, sometiéndolos a diversos factores de intensidad de esfuerzo (es decir, cargas). La tabla presenta la carga máxima y la rapidez de crecimiento de grietas resultante (en metros por segundo) para los seis especímenes.

Carga máxima x, $MPa \cdot m^{\frac{1}{2}}$	30.0	35.6	41.5	50.2	55.5	61.1
Rapidez de crecimiento de grietas y, $m/s \times 10^{10}$	1.0	2.2	3.9	5.8	5.0	14.0

a) ¿Hay suficientes pruebas que indiquen que la rapidez de crecimiento de grietas aumenta linealmente con la carga máxima? Pruebe con $\alpha = 0{,}10$.

b) Estime el incremento medio en la rapidez de crecimiento de grietas por cada incremento unitario en la carga máxima, empleando un intervalo de confianza de 90 %. Interprete el resultado.

17. Un modelo robusto y muy utilizado para el movimiento humano es la Ley de Fitts. Según esta ley, el tiempo T necesario para moverse y seleccionar un objetivo de anchura W que está a una distancia (o amplitud) A es: $T = a + b\,log_2(2A/W)$ donde a y b son constantes que se estiman mediante regresión lineal simple. La cantidad $log_2(2A/W)$ se denomina índice de dificultad (ID) y representa la variable independiente (medida en *bits*) del modelo. Ciertas investigaciones de las que se informó en el *Special Interest Group on Computer − Human Interaction Bulletin* (julio de 1993) utilizaron la Ley de Fitts para modelar el tiempo (en milisegundos) necesario para realizar cierta tarea en una computadora. Con base en datos obtenidos de $n = 160$ ensayos (empleando diferentes valores de A y W). se obtuvo la siguiente predicción de mínimos cuadrados: $\widehat{T} = 175{,}4 + 133{,}2(ID)$

a) Interprete las estimaciones, 175.4 y 133.2.

b) El coeficiente de correlación para el anális es $r = 0{,}951$. Interprete este valor.

c) Realice una prueba para determinar si el modelo de la Ley de Fitts es estadísticamente adecuado para predecir el tiempo de realización de las tareas. Utilice $\alpha = 0{,}05$.

d) Calcule el coeficiente de determinación, r^2. Interprete el resultado.

18. Refiérase al experimento, informado en *Combustion and Flame*, de difusividad del oxígeno. Los datos para las nueve muestras de mezcla de nitrógeno y óxigeno se reproducen en la siguiente tabla.

Temperatura x	Difusividad de oxígeno y
1,000	1.69
1,100	1.99
1,200	2.31
1,300	2.65
1,400	3.01
1,500	3.39
1,600	3.79
1,700	4.21
1,800	4.64

a) Calcule r y r^2. Interprete sus valores.

b) Realice una prueba para determinar si la temperatura y la difusibidad del oxígeno exhiben una correlación positiva. Utilice $\alpha = 0{,}05$.

19. La exposición pasiva al humo de tabaco en el ambiente se ha asociado a la supresión del crecimiento y a un incremento en la frecuencia de infecciones de las vías respiratorias en niños normales. ¿Esta asociación es más pronunciada en un niños que padecen fibrosis cística? Con el fin de contestar esta pregunta, se estudiaron 43 niños (18 niñas y 25 niños) que asistieron a un campamento de verano de dos semanas para pacientes con fibrosis cística (*New England Journal of Medicine*, 20 de septiembre de 1990). Entre las diversas variables que se midieron estuvieron el percentil de peso del niño (y) y el número de cigarrillos fumados por día en el hogar del niño (x).

a) Para las 18 niñas, el coeficiente de correlación entre y y x se informó como $r = -0{,}50$. Interprete este resultado.

b) Refiérase al inciso a). El valor P para probar $H_0 : \rho = 0$ contra $H_1 : \rho \neq 0$ se informó como $p = 0{,}03$. Interprete este resultado.

c) Para los 25 niños, el coeficiente de correlación entre y y x se informó como $r = -0{,}12$. Interprete este resultado.

d) Refiérase al inciso c). El valor P para probar $H_0 : \rho = 0$ contra $H_1 : \rho \neq 0$ se informó como $p = 0{,}57$. Interprete este resultado.

20. Los siguientes estadísticos de resumen se obtuvieron de un estudio que utilizó el análisis de regresión para investigar la relación entre la flexión de un pavimento y la temperatura superficial del pavimento de varios lugares de una carretera estatal. Aquí $x =$ temperatura (^{o} F) e $y =$ factor de ajuste de flexión ($y \geq 0$):

$$n = 15 \quad \sum x_i = 1425 \quad \sum y_i = 10{,}68$$

$$\sum x_i^2 = 139037{,}25 \quad \sum x_i y_i = 987{,}645 \quad \sum y_i^2 = 7{,}85183$$

(a) Calcule $\hat{\beta}_1$, $\hat{\beta}_0$ y la ecuación de la recta de regresión estimada.

(b) ¿Cuál es la estimación de cambio esperado en el factor de ajuste de flexión cuando la temperatura aumenta 1^o F?.

(c) Suponga que la temperatura se midió en ^{o} C en lugar de ^{o} F. ¿Cuál sería la recta de regresión estimada?.

21. El concreto sin finos, preparado con un agregado grueso clasificado uniformemente y una pasta de cemento y agua, es bueno en zonas de lluvia excesiva por sus excelentes propiedades de drenado. El artículo **"Pavement Thickness Design for No-Fines Concrete Parking Lots".** (*J. of Transporting Engr.*, 1995, pp. 476–484) describe el empleo de un análisis de mínimos cuadrados para estudiar la forma como $y =$ porosidad (%) se relaciona con $x =$ peso unitario (lb/pie^3) en especímenes de concreto. Utilice el resultado que se presenta de MINITAB para contestar las siguientes preguntas:

(a) ¿Cuál es la ecuación de la recta de regresión estimada?.

(b) Interprete el valor estimado de β_1.

(c) Construya un intervalo de confianza de 95 % para β_1. A partir del intervalo de confianza ¿Puede concluir que la variable x es significativa en el modelo de regresión simple?.

(d) ¿Cuál es la estimación de σ?.

(e) ¿Cuál es el valor de la variación total que es explicada por el modelo?.

(f) Encuentre una estimación puntual para la porosidad promedio real de todos los especímenes, cuyo peso unitario sea 110 lb/pie^3.

```
Regression Analysis: y versus x

Predictor  Coef       SE        Coef T    P
Constant    118,910   4,499      26,43    0,000
x          -0,90473   0,04109   -22,02    0,000

Analysis of Variance

Source            DF       SS          MS        F         P
Regression         1       426,62      426,62    484,84    0,000
Residual Error    13        11,44        0,88
Total             14       438,06
```

22. Se sabe que la duración de una linterna tiene relación con la cantidad de veces que se usa. Sea Y_i la duración de la linterna (en dias) y X_i el numero de dias que es usada. Dado el modelo

$$Y_i = \beta * X_i + \epsilon_i$$

a) Derive el EMCO de β y obtenga una expresión para la varianza.

b) Estime la ecuación de regresión para la siguiente muestra

X	3	3	4	4	5
Y	7	10	12	14	13

23. La siguiente tabla contiene la altura y numero de calzado de 10 primos.

altura	183	170	165	180	177	159	172	164	168	181
calzado	44	40	41	43	42	39	41	39	42	43

a) Encuentre la recta de regresion.

b) Complete la tabla ANOVA.

Fuente	gl	SC	MC	F
Regresion				
Error				
Total				

c) Calcule el R^2

24. Complete la tabla ANOVA, sabiendo que $n = 16$ y demuestre que:

$$SCT = SCM + SCR$$

Fuente	gl	SC	MC	F
Regresion			4.18	
Error		1.82		
total		6		

25. Se quiere determinar si existe una relación lineal entre el nivel de ingreso de educación y las predictoras $X_1 =$ años transcurridos desde el egreso y $X_2 =$ nivel de especialización. Una muestra de $n = 43$ egresados de educación son analizados, algunos de cuyos resultados se presentan a continuación:

SCRegresión= 600, SCTotal= 800

$$\hat{\beta} = \begin{pmatrix} 20 \\ -2 \\ 3 \end{pmatrix}, \Sigma_{\hat{\beta}} = \begin{pmatrix} 15 & 2 & 1 \\ 2 & 0{,}6 & 4 \\ 1 & 4 & 0{,}5 \end{pmatrix}$$

a) Plantee el correspondiente modelo, dé los supuestos necesarios. Complete la tabla ANOVA y de R^2.

b) Construya un I. de C. del 95 % para β_1.

c) Es posible simplificar el modelo?.

26. La exposición a la contaminación del aire,? resulta en menor esperanza de vida?. esta pregunta se examinó en el artículo "Does air pollution shorten lives?. statistics and public Policy (1977) Reading Mass. Addiso - Wesley. Se registró información sobre :

$$\begin{aligned} Y &= \text{Tasa total de mortalidad(fallecimientos por 10.000)} \\ X_1 &= \text{Lectura media de partículas suspendidas}(\mu * g/m^3) \\ X_2 &= \text{mínima lectura de sulfatos}((\mu * g/m^3) * 10) \\ X_3 &= \text{Densidad de población}(Hab/Km^2) \\ X_4 &= (\text{Porcentaje de personas de razas no blancas}) * 10 \\ X_5 &= (\text{Porcentaje de personas de más de 65 años de edad}) * 10 \end{aligned}$$

Para 1970 se consideraron $n = 106$ áreas seleccionadas al azar. La ecuación de regresión estimada fue

$$Y = 19{,}7 + 0{,}04X_1 + 0{,}071X_2 + 0{,}001X_3 + 0{,}042X_4 + 0{,}68X_5$$

a) Para este modelo se obtuvo un R^2 =0.825 Existe regresión?. Use $\alpha = 5\,\%$.

b) Si la desviación estándar (o error estándar) de $\hat{\beta}_4$ fue 0.007, determine si el porcentaje de personas que no son de raza blanca es una variable importante en el modelo. Use $\alpha = 1\,\%$.

Capítulo 6

Ejercicios Resueltos de Interrogaciones

6.1. Interrogaciones III

1. El contenido de nicotina de dos marcas de cigarros, medido en miligramos, es el siguiente:

Marcas	Contenido de nicotina									
A	2.1	4.0	6.3	5.4	4.8	3.7	6.1	3.3		
B	4.1	0.6	3.1	2.5	4.0	6.2	1.6	2.2	1.9	5.4

¿Los contenidos de nicotina de las dos marcas serán diferentes?. Considere $\alpha = 5\,\%$, plantee las hipótesis, haga el test correspondiente en los siguientes incisos e interprete el resultado.

a) Desarrolle el test bajo el supuesto de normalidad de las observaciones.

b) Suponga ahora que no hay evidencia que permita asumir normalidad.

2. Con el fin de evaluar la habilidad de los estudiantes para memorizar conceptos en un curso dado, se efectúa el siguiente experimento: En una muestra de $n = 11$ estudiantes, se aplica una prueba de diagnóstico al inicio de éste, y al final del curso se repite la misma prueba. La hipótesis de los investigadores, es que los estudiantes deben incrementar su puntaje en más de dos puntos (para una escala de cero a 10 puntos), lo que permitiría afirmar que los estudiantes poseen la habilidad. Los datos se presentan en la tabla siguiente

	Estudiantes										
Evaluación	1	2	3	4	5	6	7	8	9	0	11
Inicial	2	5	4	3	6	4	2	5	4	3	6
Final	6	7	8	8	4	7	6	9	8	–	8
Diferencia $(F - I)$	4	2	4	5	-2	3	4	4	4	–	2

Lleve a cabo el test correspondiente para probar la hipótesis de los investigadores con un nivel de significación de $\alpha = 0{,}01$.

3. Un inspector que controla la contaminación de las aguas de un río, sospecha que en una ciudad pequeña, por donde pasa el río están botando desperdicios al río, lo que haría disminuir la cantidad media de oxígeno disuelto en el agua. Para confirmar esta afirmación, el inspector obtiene durante 9 días muestras de agua antes de que el río llegue a la ciudad y después de la ciudad. La cantidad de oxígeno en partes por millón se indica en la siguiente tabla:

Día	1	2	3	4	5	6	7	8	9
Antes	5.3	5.2	5.0	4.9	5.5	5.6	4.5	5.7	5.0
Después	5.0	5.1	5.0	4.7	5.2	5.4	4.6	5.3	4.5
Diferencia	0.3	0.1	0.0	0.2	0.3	0.2	-0.1	0.4	0.5

¿Qué conclusión puede obtener el inspector?. Utilice $\alpha = 0{,}01$.

4. Una encuesta realizada en Santiago a 500 mujeres con respecto a sus preferencias entre dos revistas femeninas, A y B, y su sector de residencia, mostró los siguientes resultados: La tabla muestra el número de personas que prefieren las revistas A y B, y su sector de residencia.

	A	B
Centro	34	166
Oriente	78	122
Otro	48	52

   a) ¿De qué tamaño debió haber sido la muestra del sector centro si se quiere estimar la verdadera proporción de preferencias de B con un error máximo (diferencia absoluta entre el estimador y el parámetro) de 0.02 con confianza del 95 %?

   b) ¿Depende la preferencia por la revista A del sector de residencia? Considere $\alpha = 0{,}05$.

5. Se han registrado las duraciones, en horas, de 100 ampolletas, tiempos que han sido tabulados. ¿Puede asumirse que la duración de las ampolletas sigue una distribución exponencial de media 100 horas?

Categoría	Frecuencia
$<$ 50 horas	10
50-100 horas	30
100-200 horas	40
$>$ 200 horas	20

6. Las ampolletas pueden clasificarse según su potencia (Watts) y se piensa que de alguna forma existe una relación entre la duración y la portencia. Para verificar lo anterior se tabulan los datos, obteniendo lo que sigue:

		Duración superior a 200 horas	
		Si	No
Potencia	$<$ 100 w	30	20
	$\geq$ 100 w	20	30

¿Qué diría usted?, justifique.

7. La siguiente tabla presenta la clasificación de dos muestras de alumnos independientes, obtenida de cada una de las secciones del curso EYP2113, según su estatus final.

Etapa	Sección X	Sección Y
Eximido	2	8
Examen	7	5
Examen Recuperativo	11	7

¿Existe evidencia para afirmar que hay un efecto profesor?

8. Un avezado alumno del curso EYP2113 afirma que la tasa de eximición varía según la sección, y para demostrar lo anterior, toma muestras independientes de tamaño 20 en cada una de las secciones (1, 2 y 3) registrando el número de eximidos, que resultaron ser 13, 8 y 12 respectivamente. ¿Hay suficientes evidencias ($\alpha = 5\,\%$) para confirmar la apreciación de este avezado alumno?

9. Un famoso experimento médico fue conducido por Joseph Lister a finales del siglo XIX. La mortalidad asociada con las cirugías era muy alta y Lister conjeturó que el uso de un desinfectante, ácido carbólico, podría ayudar. Durante un período de varios años Lister llevó a cabo 100 amputaciones, donde él decidía si utilizaba o no ácido carbólico. Los datos son:

		¿Usó ácido carbólico?	
		Si	No
¿El paciente vivió?	Si	40	10
	No	20	30

   a) Plantee las hipótesis correspondientes (en palabras y paramétricamente) que le permitan probar la conjetura de Lister.

   b) Lleve a cabo el test. ¿En base a estos datos, Lister estaba en lo cierto?. Utilice $\alpha = 0{,}05$.

10. Se observó la duración en horas de 100 pilas de una determinada marca, obteniendo los siguientes resultados:

Horas	< 20	20-40	40-60	60-80	≥ 80
Frecuencia	5	26	34	22	13

¿Hay evidencia suficiente para rechazar la hipótesis de que los datos siguen una distribución normal de parámetros $\mu = 50$ y $\sigma = 20$ (redondee al entero más cercano los valores esperados)?

11. Sea el modelo $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i = 1, \ldots, k+l+1$ donde $\mathbb{E}(\varepsilon_i) = 0$ para todo i,

$$Cov(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma^2 \text{ si } i = j \\ 0 \text{ si } i \neq j \end{cases} \quad \text{y } x_i = \begin{cases} -1 \text{ si } i = 1, \ldots, k \\ 0 \text{ si } i = k+1, \ldots, k+l \\ k \text{ si } i = k+l+1 \end{cases}.$$

Demuestre que

$$\mathbb{V}\left(\hat{Y}_{k+l+1}\right) = \mathbb{V}\left(\hat{Y}_1\right) + \sigma^2\left(1 - k^{-1}\right) = \mathbb{V}\left(\hat{Y}_{k+1}\right) + \sigma^2\frac{k}{k+1}.$$

12. Se ajustó un modelo de regresión lineal simple, y se tiene un nuevo punto x_0, si denotamos el valor estimado de Y_0 como $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$, determine un I.C. al 95 % para Y_0.

13. Se ajusta el modelo de regresión $y = x\beta + \varepsilon$ para los puntos (x_i, y_i), con $i = 1, \ldots, n$. Si asumimos normalidad de los errores con media cero y varianza σ^2 desconocida, determine un I.C. al nivel $(1-\alpha)\,100\,\%$ para β utilizando el estimador EMCO de β.

14. Se postula que el gasto militar (GM) es una función lineal directa de las exportaciones de los recursos naturales (RN), índice de inestabilidad regional (IR) e inversa a las importaciones de crudo (IC). Para una muestra de $n = 34$ países en vías de desarrollo se considera el GM (en porcentaje del Producto Interno Bruto), $X_1 : RN$ (en millones de US$), $X_2 : IR$ (en porcentaje) y $X_3 : IC$ (en millones de US$). Se ajusta el siguiente modelo $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$, obteniéndose los siguientes resultados

$$\begin{aligned} \boldsymbol{\beta} &= (30, 75, 30, -24)^t,\ R^2 = 80\,\%,\ \hat{\sigma}^2 = 140 \text{ y} \\ \left(\mathbf{X}^t\mathbf{X}\right)^{-1} &= \begin{pmatrix} 12 & 3 & 2 & -1 \\ & 2 & 2 & -3 \\ & & 1 & -1 \\ & & & 1 \end{pmatrix}. \end{aligned}$$

   a) Construya la respectiva tabla ANOVA, ¿existe regresión?. ¿Tienen razón los investigadores cuando plantean que la relación del GM con las IC es inversa, una vez que se controla la RN e IR?

   b) Se sospecha que el modelo $Y_i = \beta_0 + \beta_2 X_{2i} + \varepsilon_i$, es decir, sólo basado en IR es suficiente para explicar el GM. Si para este nuevo modelo $\hat{\sigma}^2 = 200$, ¿qué dice usted?, justifique.

15. Sea Y : gasto en consumo per-cápita (en miles de $), X_1 : Ingreso disponible (en miles de $), X_2 : Tamaño familiar. Para una muestra de $n = 15$ familias se ajusta el siguiente modelo $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$, obteniéndose los siguientes resultados

$$\begin{aligned} \boldsymbol{\beta} &= (300, 70, -85)^t,\ R^2 = 90\,\%,\ \hat{\sigma}^2 = 150 \text{ y} \\ \left(\mathbf{X}^t\mathbf{X}\right)^{-1} &= \begin{pmatrix} 37 & -2 & 1 \\ -2 & 2 & 1 \\ 1 & 1 & 5 \end{pmatrix}. \end{aligned}$$

   a) Construya la respectiva tabla ANOVA, ¿existe regresión?.

   b) Se sospecha que el modelo $Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$, es suficiente para predecir bien a Y. Si para este nuevo modelo $\hat{\sigma}^2 = 250$, ¿qué dice usted?, justifique.

16. En un estudio agrícola se reportan los siguientes datos sobre producción (Y), temperatura media sobre el período entre la fecha de floración y fecha de levantar la cosecha (x_1), y el porcentaje medio de luz solar durante el mismo período (x_2), para un cierto tipo de fruto. Se ajustó un modelo de regresión lineal con $n = 12$ datos y se obtuvieron los siguientes resultados, $\hat{\beta}_0 = 415{,}11$, $S_{\hat{\beta}_0} = 82{,}52$, $\hat{\beta}_1 = -6{,}6$, $S_{\hat{\beta}_1} = 4{,}86$, $\hat{\beta}_2 = -4{,}5$, $S_{\hat{\beta}_2} = 1{,}1$, $SCT = 597{,}8$ y $R^2 = 0{,}768$.

    a) Pruebe $H_0 : \beta_1 = \beta_2 = 0$ contra $H_1 : \{\beta_1 \neq 0$ ó $\beta_2 \neq 0\}$ al nivel 0.05.

    b) Dado que x_2 está en el modelo, ¿se retendría x_1?

    c) Cuando se ajusta el modelo $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \varepsilon$ se tiene $R^2 = 0{,}861$. Pruebe las hipótesis $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$ contra $H_1 : \beta_j \neq 0$ para algún $j \in \{3, 4, 5\}$.

17. Suponga que cada punto fijo x_i, $i = 1, \ldots, n$, de la variable explicativa en un modelo de regresión lineal simple es duplicado (x_{i_1} y x_{i_2}), obteniéndose dos observaciones independientes Y_{i_1} y Y_{i_2}, una para cada valor de x_{i_1} y x_{i_2}, respectivamente. ¿Es cierto que los estimadores EMCO del intercepto y la pendiente del modelo pueden ser encontrados si hiciésemos una regresión con los valores x_i y $\bar{Y}_i = \frac{1}{2}(Y_{i_1} + Y_{i_2})$?

6.2. Soluciones

1. Debemos comparar las hipótesis $H_0 : \mu_A = \mu_B$ contra $H_1 : \mu_A \neq \mu_B$.

   a) Bajo el supuesto de normalidad e independencia de las dos muestras debemos aplicar el test t para dos muestras independientes, en donde

   $$S_p^2 = \frac{(m-1)\,S_A^2 + (n-1)\,S_B^2}{n+m-2} = \frac{7 \times 2{,}25 + 9 \times 3{,}24}{16} = 2{,}81.$$

   Luego,

   $$t = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{m^{-1}+n^{-1}}} = \frac{4{,}5 - 3{,}2}{1{,}68\sqrt{8^{-1}+0{,}1}} = \frac{1{,}3}{1{,}68 \times 0{,}47} = 1{,}65.$$

   Pero como $t_{(m+n-2),1-\alpha/2} = t_{(16),0{,}975} = 2{,}12 > 1{,}65 = t$, entonces no hay evidencias en los datos para rechazar H_0.

   b) Si no suponemos normalidad de las observaciones, debemos aplicar el test de Mann-Whitney, donde los rangos de ambas muestras son,

Observ.	0.6	1.6	1.9	2.1	2.2	2.5	3.1	3.3	3.7	4.0	4.0	4.1	4.8	5.4	5.4	6.1	6.2	6.3
Rango	1	2	3	4	5	6	7	8	9	10.5	10.5	12	13	14.5	14.5	16	17	18

   Tomando la muestra de menor tamaño: $m = 8$, $R = 4 + 8 + 9 + 10{,}5 + 13 + 14{,}5 + 16 + 18 = 93$, $R' = n_1\,(n+m+1) - R = 8\,(8+10+1) - 93 = 59$ y $R^* = \text{mín}\,\{R, R'\} = 59$. De la tabla de valores críticos para el test de Mann-Whitney se tiene que para $\alpha = 0{,}05$ y un test de dos colas el valor crítico es $R_c = 53$. Por tanto, como $R^* = 59 > 53 = R_c$, no se rechaza H_0, y se llega a la misma conclusión que en el inciso (a). Si utilizamos la aproximación a la distribución normal y tomamos la muestra de menor tamaño, entonces $T_Y = 93$, $\mathbb{E}\,(T_Y) = \frac{m(n+m+1)}{2} = 76$ y $\mathbb{V}\,(T_Y) = \frac{nm(n+m+1)}{12} = 126{,}67$.
   Luego,

   $$Z_Y = \frac{T_Y - \mathbb{E}\,(T_Y)}{\sqrt{\mathbb{V}\,(T_Y)}} = 1{,}51 < 1{,}96 = z_{0{,}975} = z_{1-\alpha/2}.$$

   Por tanto, no se rechaza H_0, llegando a la misma conclusión que en el inciso (a).

2. Las hipótesis a probar son $H_0 : \mu_F \leq \mu_I + 2$ contra $H_1 : \mu_F > \mu_I + 2$, las cuales son equivalentes a $H_0 : \mu_F - \mu_I - 2 \leq 0$ contra $H_1 : \mu_F - \mu_I - 2 > 0$, pero como las muestras están apareadas y no debemos asumir normalidad de las observaciones debemos aplicar el test de Wilcoxon asignando rangos a las diferencias restándole 2. Así tenemos

$F - I - 2$	2	0	2	3	-4	1	2	2	2	–	0
Rangos	4	–	4	7	8	1	4	4	4	–	–

   Observemos que los valores cero no tienen asignado rangos ya que éstos valores no influyen en este test, de esta manera se tiene $W_- = 8$ y $W_+ = 28$, lo que implica

$W = \text{mín}\{W_-, W_+\} = 8$. De la tabla de valores críticos para el test de Wilcoxon se tiene que para $n = 8$, $\alpha = 0{,}01$ y un test de una cola el valor crítico es $W_c = 2$. Por tanto, como $W = 8 > 2 = W_c$, no se rechaza H_0, por lo que no se puede concluir que los estudiantes tenderán a aumentar su puntaje en más de dos puntos después del curso.

Si utilizamos la aproximación a la distribución normal, entonces $\mathbb{E}(W_-) = \frac{n(n+1)}{4} = 18$, $\mathbb{V}(W_-) = \frac{n(n+1)(2n+1)}{24} = \frac{8(8+1)(16+1)}{24} = 51$ y

$$Z = \frac{W_- - \mathbb{E}(W_-)}{\sqrt{\mathbb{V}(W_-)}} = \frac{8-18}{\sqrt{51}} = -1{,}4 > -2{,}32 = -z_{0{,}99} = -z_{1-\alpha/2}.$$

Por tanto, no se rechaza H_0 coincidiendo con la conclusión a la que ya habíamos arrivado.

3. Las hipótesis a probar son $H_0 : \mu_A \leq \mu_D$ contra $H_1 : \mu_A > \mu_D$, pero como no debemos asumir normalidad y las muestras están a pareadas, debemos aplicar el test de Wilcoxon. Luego,

Diferencia	0.3	0.1	0.0	0.2	0.3	0.2	-0.1	0.4	0.5
Rangos	5.5	1.5	–	3.5	5.5	3.5	1.5	7	8

De aquí se tiene que $W_- = 1{,}5$ y $W_+ = 34{,}5$, lo que implica $W = \text{mín}\{W_-, W_+\} = 1{,}5$. De la tabla de valores críticos para el test de Wilcoxon se tiene que para $n = 8$, $\alpha = 0{,}01$ y un test de una cola el valor crítico es $W_c = 2$. Por tanto, como $W = 1{,}5 < 2 = W_c$, se rechaza H_0, por lo que se puede concluir que la cantidad media de oxígeno en el agua tiende a disminuir después del paso del río por la ciudad.

Si utilizamos la aproximación a la distribución normal, entonces $\mathbb{E}(W_-) = \frac{n(n+1)}{4} = 18$, $\mathbb{V}(W_-) = \frac{n(n+1)(2n+1)}{24} = \frac{8(8+1)(16+1)}{24} = 51$ y

$$Z = \frac{W_- - \mathbb{E}(W_-)}{\sqrt{\mathbb{V}(W_-)}} = \frac{1{,}5-18}{\sqrt{51}} = -2{,}31 > -2{,}32 = -z_{0{,}99} = -z_{1-\alpha/2}.$$

Por tanto, no se rechaza H_0 contradiciendo la conclusión a la que ya habíamos arrivado. Debemos tener en cuenta que el último procedimiento es aproximado.

4. Observemos que los tamaños de muestras encuentados en cada sector fueron planificados de antemano.

   a) Notemos que la proporción de mujeres que prefirió la revista B en el sector centro fue de $\hat{p} = \frac{166}{200} = .83$. Además si X es la variable aleatoria que toma el valor 1 si la mujer encuestada prefiere la revista B y cero si no, entonces $X \sim Bin(1, p)$, donde la varianza estimada para X sería $\hat{\sigma}^2 = \hat{p}(1-\hat{p}) = \frac{166}{200}\left(1 - \frac{166}{200}\right) = 0.141\,1$. Por tanto, debemos buscar n tal que

$$0{,}95 = \mathbb{P}(|\hat{p} - p| < 0{,}02).$$

Luego, aplicando el teorema de límite central se tiene,

$$0{,}95 = \mathbb{P}\left(|\hat{p}-p| < 0{,}02\right) = \mathbb{P}\left(\left|\frac{(\hat{p}-p)\sqrt{n}}{\sqrt{\hat{p}\,(1-\hat{p})}}\right| < \frac{0{,}02\sqrt{n}}{\sqrt{\hat{p}\,(1-\hat{p})}}\right) \approx \mathbb{P}\left(|Z| < \frac{0{,}02\sqrt{n}}{\sqrt{\hat{p}\,(1-\hat{p})}}\right),$$

donde $Z \sim N(0,1)$. Por tanto

$$\frac{0{,}02\sqrt{n}}{\sqrt{\hat{p}\,(1-\hat{p})}} = z_{0{,}975},$$

luego, $\frac{0{,}02\sqrt{n}}{\sqrt{0.\,141\,1}} = 1{,}96$, de donde $n = 1355.\,1$. Pero como n debe ser un número entero, entonces debemos tomar $n = 1356$.

b) Como de antemano se sabía la cantidad de personas a las que se iba a encuestar en cada sector, entonces debemos aplicar un test de homogeneidad en donde las hipótesis a comparar son $H_0 : p_{ij} = p_{\cdot j}, i = 1,2,3, j = 1,2$ contra $H_1 : p_{ij} \neq p_{\cdot j}$, para algún $i = 1,2,3, j = 1,2$. La tabla de valores esperados en este caso es

	A	B
Centro	64	136
Oriente	64	136
Otro	32	68

Por tanto,

$$\begin{aligned}\chi^2 &= \frac{(34-64)^2}{64} + \frac{(78-64)^2}{64} + \frac{(48-32)^2}{32} + \frac{(166-136)^2}{136} + \frac{(122-136)^2}{136} + \frac{(52-68)^2}{68} \\ &= \frac{5025}{136} = 36.\,949.\end{aligned}$$

Pero como $\chi^2 = 36{,}949 > 5{,}99 = \chi^2_{(2);0{,}95}$, se rechaza H_0 concluyendo que los datos muestran evidencias de que la preferencia por la revista A depende del sector de residencia.

5. Esto es un problema de bondad de ajuste en donde la hipótesis nula es que se cumplan las siguientes igualdades $p_1 = \mathbb{P}(X < 50)$, $p_2 = \mathbb{P}(50 \leq X < 100)$, $p_3 = \mathbb{P}(100 \leq X < 200)$ y $p_4 = \mathbb{P}(X \geq 200)$, donde $X \sim Exp(0{,}01)$; y la alternativa es que alguna de éstas no se cumpla. De aquí tenemos que los valores esperados son

$$\begin{aligned}E_1 &= n\mathbb{P}(X < 50) = 100\int_0^{50} 0{,}01e^{-0{,}01x}dx = 39.\,347 \\ E_2 &= n\mathbb{P}(50 \leq X < 100) = 100\int_{50}^{100} 0{,}01e^{-0{,}01x}dx = 23.\,865 \\ E_3 &= n\mathbb{P}(100 \leq X < 200) = 100\int_{100}^{200} 0{,}01e^{-0{,}01x}dx = 23.\,254 \\ E_4 &= n\mathbb{P}(X \geq 200) = 100\int_{200}^{\infty} 0{,}01e^{-0{,}01x}dx = 13.\,534,\end{aligned}$$

y por tanto

$$\begin{aligned}\chi^2 &= \frac{(10-39.347)^2}{39.347}+\frac{(30-23.865)^2}{23.865}+\frac{(40-23.254)^2}{23.254}+\frac{(20-13.534)^2}{13.534}\\ &= 38.614\end{aligned}$$

Pero como $\chi^2 = 38{,}614 > 7{,}81 = \chi^2_{(3);0,95}$, se rechaza H_0 concluyendo que los datos muestran evidencias de que la duración de las ampolletas no sigue una distribución exponencial de media 100.

6. Evidentemente aquí se tomaron 50 ampolletas con menos de 100w, y 50 con más de 100w. Luego, debemos aplicar un test de homogeneidad. Los valores esperados en este caso son todos igual a 25, luego

$$\chi^2 = \frac{(30-25)^2}{25}+\frac{(30-25)^2}{25}+\frac{(20-25)^2}{25}+\frac{(20-25)^2}{25} = 4,$$

pero como $4 > 3{,}84 = \chi^2_{(1);0,95}$. Entonces se rechaza la hipótesis nula, afirmando que existe relación entre la potencia y la duración de la ampolleta.

7. Al parecer, intencionalmente, se tomaron 20 alumnos en cada sección, por tanto debemos aplicar un test de homogeneidad, en donde los valores esperados para cada celda son

Etapa	Sección X	Sección Y
Eximido	5	5
Examen	6	6
Examen Recuperativo	9	9

Por tanto,

$$\begin{aligned}\chi^2 &= \frac{(2-5)^2}{5}+\frac{(8-5)^2}{5}+\frac{(7-6)^2}{6}+\frac{(5-6)^2}{6}+\frac{(11-9)^2}{9}+\frac{(7-9)^2}{9}\\ &= \frac{217}{45} = 4.822.\end{aligned}$$

Pero como $\chi^2 = 4.822 < 5{,}99 = \chi^2_{(2);0,95}$, no se rechaza H_0 concluyendo que los datos no muestran evidencias para afirmar que existe un efecto profesor.

8. Los datos que el alumno obtuvo se pueden poner de la siguiente forma

Sección	Eximido	No eximido
01	13	7
02	8	12
03	12	8

Nuevamente, como el alumno tomó 20 alumnos de cada sección, debemos aplicar un test de homogeneidad con los siguientes valores esperados

Sección	Eximido	No eximido
01	11	9
02	11	9
03	11	9

Por tanto,

$$\begin{aligned}\chi^2 &= \frac{(13-11)^2}{11}+\frac{(8-11)^2}{11}+\frac{(12-11)^2}{11}+\frac{(12-9)^2}{9}+\frac{(8-9)^2}{9}+\frac{(7-9)^2}{9}\\ &= \frac{280}{99}=2.828.\end{aligned}$$

Pero como $\chi^2 = 2.828 < 5{,}99 = \chi^2_{(2);0,95}$, no se rechaza H_0 concluyendo que los datos no muestran evidencias para afirmar que la tasa de eximisión depende de la sección.

9. Como Lister decidía si se usaba el ácido carbólico entonces el margen del total horizontal de la tabla no es el resultado de un proceso aleatorio por lo que se debe aplicar un test de homogeneidad.

   a) Las hipótesis son $H_0 : p_{ij} = p_{i\cdot}, i = 1,2, j = 1,2$ que significa que la proporción de pacientes que vivieron después de la operación no depende del uso del ácido carbólico. Por otra parte la hipótesis alternativa es $H_1 : p_{ij} \neq p_{i\cdot}$, para algún $i = 1,2, j = 1,2$ que significa que el uso del ácido carbólico sí influye en la proporción de sobrevivientes después de la operación.

   b) En este caso los valores esperados son

		Duración superior a 200 horas	
		Si	No
¿El paciente vivió?	Si	30	20
	No	30	20

   y por tanto,

$$\chi^2 = \frac{(40-30)^2}{30}+\frac{(20-30)^2}{30}+\frac{(10-20)^2}{20}+\frac{(30-20)^2}{20}=\frac{50}{3}=16.667,$$

   pero como 16,667 > 3,84 $= \chi^2_{(1);0,95}$. Entonces se rechaza la hipótesis nula, afirmando que existe relación entre el uso del ácido carbólico y la posibilidad de vivir después de la operación.

10. Esto es un problema de bondad de ajuste en donde la hipótesis nula es que se cumplan las siguientes igualdades $p_1 = \mathbb{P}(X < 20)$, $p_2 = \mathbb{P}(20 \leq X < 40)$, $p_3 = \mathbb{P}(40 \leq X < 60)$, $p_4 = \mathbb{P}(60 \leq X < 80)$ y $p_5 = \mathbb{P}(X \geq 80)$, donde $X \sim N(50, (20)^2)$; y la alternativa es que alguna de éstas no se cumpla. Luego, como $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$, tenemos que los

valores esperados son

$$\begin{aligned}
E_1 &= n\mathbb{P}(X<20) = n\mathbb{P}\left(Z<\frac{20-\mu}{\sigma}\right) = 100\mathbb{P}(Z<-1{,}5) = 6{,}68\\
E_2 &= n\mathbb{P}(20\le X<40) = n\mathbb{P}\left(\frac{20-\mu}{\sigma}\le Z<\frac{40-\mu}{\sigma}\right) = 100\mathbb{P}(-1{,}5\le Z<-0{,}5) = 24{,}17\\
E_3 &= n\mathbb{P}(40\le X<60) = n\mathbb{P}\left(\frac{40-\mu}{\sigma}\le Z<\frac{60-\mu}{\sigma}\right) = 100\mathbb{P}(-0{,}5\le Z<0{,}5) = 38{,}29\\
E_4 &= n\mathbb{P}(60\le X<80) = n\mathbb{P}\left(\frac{60-\mu}{\sigma}\le Z<\frac{80-\mu}{\sigma}\right) = 100\mathbb{P}(0{,}5\le Z<1{,}5) = 24{,}17\\
E_5 &= n\mathbb{P}(X\ge 80) = n\mathbb{P}\left(Z\ge\frac{80-\mu}{\sigma}\right) = 100\mathbb{P}(Z\ge 1{,}5) = 6{,}68,
\end{aligned}$$

y por tanto

$$\begin{aligned}
\chi^2 &= \frac{(5-7)^2}{7}+\frac{(26-24)^2}{24}+\frac{(34-38)^2}{38}+\frac{(22-24)^2}{24}+\frac{(13-7)^2}{7}\\
&= \frac{2581}{399} = 6.4687
\end{aligned}$$

Pero como $\chi^2 = 6.4687 < 9{,}45 = \chi^2_{(4);0,95}$, no se rechaza H_0 concluyendo que los datos no muestran evidencias de que la duración de las pilas no sigue una distribución $N(50,(20)^2)$.

11. Veamos que

$$\mathbb{V}\left(\hat{Y}_j\right) = \mathbb{V}\left(\hat{\beta}_0+\hat{\beta}_1 x_j\right) = \mathbb{V}\left(\hat{\beta}_0\right)+x_j^2\mathbb{V}\left(\hat{\beta}_1\right)+2x_j Cov\left(\hat{\beta}_0,\hat{\beta}_1\right),$$

pero como $Cov\left(\hat{\beta}_0,\hat{\beta}_1\right) = -\sigma^2\frac{\bar{x}}{\sum_{i=1}^n (x_i-\bar{x})^2}$ y

$$\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i = \frac{1}{n}\left[\sum_{i=1}^k(-1)+\sum_{i=k+1}^{k+l}0+k\right] = \frac{-k+k}{n} = 0,$$

entonces,

$$\begin{aligned}
\mathbb{V}\left(\hat{Y}_j\right) &= \mathbb{V}\left(\hat{\beta}_0\right)+x_j^2\mathbb{V}\left(\hat{\beta}_1\right) = \sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^n(x_i-\bar{x})^2}\right)+x_j^2\frac{\sigma^2}{\sum_{i=1}^n(x_i-\bar{x})^2}\\
&= \sigma^2\left(\frac{1}{k+l+1}+\frac{x_j^2}{\sum_{i=1}^{k+l+1}x_i^2}\right).
\end{aligned}$$

Por tanto, como $\sum_{i=1}^{k+l+1}x_i^2 = \sum_{i=1}^k(-1)^2+\sum_{i=k+1}^{k+l}0+k^2 = k(k+1)$, se cumple

$$\mathbb{V}\left(\hat{Y}_1\right) = \sigma^2\left(\frac{1}{k+l+1}+\frac{1}{k(k+1)}\right), \tag{6.1}$$

$$\mathbb{V}\left(\hat{Y}_{k+1}\right) = \frac{\sigma^2}{k+l+1} \tag{6.2}$$

y

$$\mathbb{V}\left(\hat{Y}_{k+l+1}\right) = \sigma^2\left(\frac{1}{k+l+1} + \frac{k^2}{k\,(k+1)}\right). \tag{6.3}$$

Luego, de (6.1) y (6.3), se tiene

$$\mathbb{V}\left(\hat{Y}_{k+l+1}\right) = \mathbb{V}\left(\hat{Y}_1\right) + \sigma^2\left(1-k^{-1}\right).$$

Por otra parte, de (6.1) y (6.2), se tiene

$$\begin{aligned}\mathbb{V}\left(\hat{Y}_1\right) + \sigma^2\left(1-k^{-1}\right) &= \frac{\sigma^2}{k+l+1} + \frac{\sigma^2}{k\,(k+1)} + \sigma^2\left(1-k^{-1}\right)\\ &= \mathbb{V}\left(\hat{Y}_{k+1}\right) + \sigma^2\frac{k}{k+1}.\end{aligned}$$

Concluyendo la demostración.

12. Suponiendo normalidad de las observaciones se tiene que $Y_0 \sim N\left(\beta_0+\beta_1 x_1, \sigma^2\right)$ y $\hat{Y}_0 \sim N\left(\beta_0+\beta_1 x_1, \mathbb{V}\left(\hat{Y}_0\right)\right)$, donde

$$\begin{aligned}\mathbb{V}\left(\hat{Y}_0\right) &= \mathbb{V}\left(\hat{\beta}_0+\hat{\beta}_1 x_0\right) = \mathbb{V}\left(\hat{\beta}_0\right) + x_0^2\mathbb{V}\left(\hat{\beta}_1\right) + 2x_0 Cov\left(\hat{\beta}_0,\hat{\beta}_1\right)\\ &= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n\left(x_i-\bar{x}\right)^2}\right) + \frac{x_0^2\sigma^2}{\sum_{i=1}^n\left(x_i-\bar{x}\right)^2} - 2x_0\frac{\sigma^2\bar{x}}{\sum_{i=1}^n\left(x_i-\bar{x}\right)^2}\\ &= \sigma^2\left(\frac{1}{n} + \frac{\left(x_0-\bar{x}\right)^2}{\sum_{i=1}^n\left(x_i-\bar{x}\right)^2}\right).\end{aligned}$$

Como Y_0 será una nueva observación, entonces ésta será independiente de las observaciones anteriores anteriores $Y_1,\ldots,Y_n$, y por tanto de $\hat{\beta}_0$ y $\hat{\beta}_1$; y por tanto también se tendrá que Y_0 y $\hat{Y}_0$ son independientes. De esta manera tenemos que $Y_0-\hat{Y}_0 \sim N\left(0,\sigma^2+\mathbb{V}\left(\hat{Y}_0\right)\right)$. Luego, de aquí tenemos que un I.C. para Y_0 al nivel $(1-\alpha)\,100\,\%$ es

$$\left[\hat{Y}_0 - z_{1-\alpha/2}\sigma\sqrt{1+\frac{1}{n}+\frac{\left(x_0-\bar{x}\right)^2}{\sum_{i=1}^n\left(x_i-\bar{x}\right)^2}}, \hat{Y}_0 + z_{1-\alpha/2}\sigma\sqrt{1+\frac{1}{n}+\frac{\left(x_0-\bar{x}\right)^2}{\sum_{i=1}^n\left(x_i-\bar{x}\right)^2}}\right].$$

Noten que este intervalo se hace más estrecho mientras más cerca esté x_0 de la media $\bar{x}$. El I.C. al $95\,\%$ se obtiene sustituyendo $z_{1-\alpha/2}$ por 1.96.

13. El estimador EMCO de β es el valor de β que minimiza la función $L(\beta) = \sum_{i=1}^n\left(Y_i-\beta x_i\right)^2$. Luego, de $0 = L'(\beta)$, se tiene $0 = -2\sum_{i=1}^n x_i\left(Y_i-\beta x_i\right)$, y por tanto

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}.$$

Como este estimador es una combinación lineal de variables aleatorias normales e independientes entonces $\hat{\beta} \sim N\left[\mathbb{E}\left(\hat{\beta}\right), \mathbb{V}\left(\hat{\beta}\right)\right]$, donde

$$\mathbb{E}\left(\hat{\beta}\right) = \frac{\sum_{i=1}^{n} x_i \mathbb{E}\left(Y_i\right)}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} \beta x_i^2}{\sum_{i=1}^{n} x_i^2} = \beta$$

y

$$\mathbb{V}\left(\hat{\beta}\right) = \left(\sum_{i=1}^{n} x_i^2\right)^{-2} \sum_{i=1}^{n} x_i^2 \mathbb{V}\left(Y_i\right) = \left(\sum_{i=1}^{n} x_i^2\right)^{-2} \sum_{i=1}^{n} x_i^2 \sigma^2 = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}.$$

Por tanto un I.C. para β al nivel $(1-\alpha)\,100\,\%$ es

$$\left[\hat{\beta} - z_{1-\alpha/2}\sigma\sqrt{\frac{1}{\sum_{i=1}^{n} x_i^2}}, \hat{\beta} + z_{1-\alpha/2}\sigma\sqrt{\frac{1}{\sum_{i=1}^{n} x_i^2}}\right].$$

14. Este es un problema de regresión lineal múltiple.

a) Para completar la tabla ANOVA debemos tener en cuenta que $140 = \hat{\sigma}^2 = \frac{SCE}{n-p} = \frac{SCE}{30}$, por lo que $SCE = 4200$ y además $0{,}8 = R^2 = 1 - \frac{SCE}{SCT} = 1 - \frac{4200}{SCT}$, por lo que $SCT = 21000$.

Fuente de variación	Grados de Libertad	S.C.	S.C.M.	F
Regresión	$p-1=3$	16800	5600	40
Error	$n-p=30$	4200	$\hat{\sigma}^2 = 140$	
Total	$n-1=33$	21000		

Pero como $F = 40 > 2{,}92 = F_{(3,30);0,95}$, podemos decir que sí hay regresión. Por otra parte para probar la relación inversa entre GM y IC debemos comparar las hipótesis $H_0 : \beta_3 \geq 0$ contra $H_1 : \beta_3 < 0$. Luego,

$$t = \frac{\hat{\beta}_3}{\sqrt{S^2_{\hat{\beta}_3}}} = \frac{-24}{\sqrt{140 \times 1}} = -2{,}03.$$

Pero como $|t| = 2{,}03 > 1{,}697 = t_{(30);0,95} = t_{(n-p);1-\alpha}$, se rechaza H_0, y por tanto los datos muestran evidencias para afirmar que la relación entre GM y IC es inversa.

b) En este caso debemos comparar las hipótesis $H_0 : \beta_1 = \beta_3 = 0$ contra $H_1 : \{\beta_1 \neq 0 \text{ ó } \beta_3 \neq 0\}$ dado que β_0 y β_2 ya se encuentran en el modelo. Para esto debemos calcular

$$\begin{aligned} F &= \frac{\left(SCE_0 - SCE_1\right)/\left(p_1 - p_0\right)}{SCE_1/\left(n - p_1\right)} = \frac{\left[\left(n-p_0\right)\hat{\sigma}_0^2 - \left(n-p_1\right)\hat{\sigma}_1^2\right]/\left(p_1-p_0\right)}{\left(n-p_1\right)\hat{\sigma}_1^2/\left(n-p_1\right)} \\ &= \frac{\left[\left(34-2\right)200 - \left(34-4\right)140\right]/\left(4-2\right)}{140} = \frac{\left[\left(32\right)200 - \left(30\right)140\right]/2}{140} = \frac{55}{7} = 7.8571, \end{aligned}$$

pero como $F = 7{,}8571 > 3{,}32 = F_{(2,30);0,95}$, se rechaza H_0 concluyendo que el modelo reducido $Y_i = \beta_0 + \beta_2 X_{2i} + \varepsilon_i$, no es suficiente para explicar la variabilidad de los datos Y_i.

15. Este es un problema de regresión lineal múltiple.

a) Para completar la tabla ANOVA debemos tener en cuenta que $150 = \hat{\sigma}^2 = \frac{SCE}{n-p} = \frac{SCE}{12}$, por lo que $SCE = 1800$ y además $0{,}9 = R^2 = 1 - \frac{SCE}{SCT} = 1 - \frac{1800}{SCT}$, por lo que $SCT = 18000$.

Fuente de variación	Grados de Libertad	S.C.	S.C.M.	F
Regresión	$p-1=2$	16200	8100	54
Error	$n-p=12$	1800	$\hat{\sigma}^2 = 150$	
Total	$n-1=14$	18000		

Pero como $F = 54 > 3{,}89 = F_{(2,12);0,95}$, podemos decir que sí hay regresión, o sea, las variables explicativas X_1 y X_2 son útiles para explicar el comportamiento de los datos Y_i.

b) Una vía de solución es de forma análoga a la hecha en el inciso (b) del ejercicio anterior, pero teniendo en cuenta que las hipótesis que queremos probar son $H_0 : \beta_2 = 0$ contra $H_1 : \beta_2 \neq 0$, otra vía de solución es usando el estadístico t,

$$t = \frac{\hat{\beta}_2}{\sqrt{S^2_{\hat{\beta}_2}}} = \frac{-85}{\sqrt{150 \times 5}} = -3{,}1,$$

pero como $|t| = 3{,}1 > 2{,}179 = t_{(12);0,975} = t_{(n-p);1-\alpha/2}$, se rechaza H_0, y por tanto los datos muestran evidencias de que el modelo reducido $Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$, no es suficiente para explicar la variabilidad de los datos Y_i. Observemos que con esta solución no fue necesario utilizar la información de que $\hat{\sigma}^2 = 250$ para el modelo reducido.

16. Este es un problema de regresión lineal múltiple.

a) En otras palabras debemos responder si existe regresión. Para esto debemos calcular el estadístico F de la tabla ANOVA. Pero como

$$0{,}768 = R^2 = \frac{SCR}{SCT} = \frac{SCR}{597{,}8},$$

se tiene que $SCR = 459{,}11$. Además, $SCE = SCT - SCR = 597{,}8 - 459{,}11 = 138.\,69$. Por tanto,

$$F = \frac{SCR/\,(p-1)}{SCE/\,(n-p)} = \frac{459{,}11/\,(3-1)}{138.\,69/\,(12-3)} = 14{,}9.$$

Por otra parte, de la tabla de distribuciones de Fisher se tiene $F_{(2,9);0,95} = 4{,}26$. Luego, debido a que $F = 14{,}9 > 4{,}26 = F_{(2,9);0,95}$, se rechaza H_0 y podemos concluir que existe regresión.

b) Para esto hay que probar $H_0 : \beta_1 = 0$ contra $H_1 : \beta_1 \neq 0$, pero como

$$t = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}} = \frac{-6{,}6}{4{,}86} = -1{,}36$$

y

$$t_{(n-p);1-\alpha/2} = t_{(9);0{,}975} = 2{,}262,$$

se tiene que $|t| = 1{,}36 < t_{(9);0{,}975}$. Por lo que no se rechaza H_0. Luego, no hay necesidad de retener a la variable x_1 en el modelo cuando la variable x_2 ya está incluida en éste.

c) Nuevamente tenemos,

$$0{,}861 = R^2_* = 1 - \frac{SCE_*}{SCT} = 1 - \frac{SCE_*}{597{,}8}.$$

Lo que implica $SCE_* = 83{,}09$. Por tanto el estadístico F en este caso es

$$F = \frac{(SCE - SCE_*)/3}{SCE_*/(12-6)} = \frac{2\,(138.\,69 - 83{,}09)}{83{,}09} = 1{,}34,$$

pero como $F = 1{,}34 < 4{,}76 = F_{(3,6);0{,}95}$, no se rechaza H_0, concluyendo que cuando las variables x_1 y x_2 están en el modelo, las variables x_1^2, x_2^2 y x_1x_2 no son muy importantes.

17. Sean $\hat{\beta}_1$ y $\hat{\beta}_0$ los EMCO usuales para cuando se usan las observaciones $\left(x_{i_j}, Y_{i_j}\right)$ con $i = 1, \ldots, n$ y $j = 1, 2$, o sea

$$\hat{\beta}_1 = \frac{\sum_{j=1}^{2}\sum_{i=1}^{n}\left(x_{i_j} - \tilde{x}\right)Y_{i_j}}{\sum_{j=1}^{2}\sum_{i=1}^{n}\left(x_{i_j} - \tilde{x}\right)^2} \text{ y } \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\tilde{x},$$

donde $\bar{Y} = \frac{1}{2n}\sum_{j=1}^{2}\sum_{i=1}^{n} Y_{i_j}$ y $\tilde{x} = \frac{1}{2n}\sum_{j=1}^{2}\sum_{i=1}^{n} x_{i_j}$.

Sean $\tilde{\beta}_1$ y $\tilde{\beta}_0$ los EMCO usuales para cuando se usan las observaciones $\left(x_i, \bar{Y}_i\right)$ con $i = 1, \ldots, n$, o sea

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\bar{Y}_i}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \text{ y } \tilde{\beta}_0 = \tilde{Y} - \tilde{\beta}_1\bar{x},$$

donde $\tilde{Y} = \frac{1}{n}\sum_{i=1}^{n}\bar{Y}_i$ y $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Debemos ver si se cumple $\hat{\beta}_1 = \tilde{\beta}_1$ y $\hat{\beta}_0 = \tilde{\beta}_0$.

Teniendo en cuenta $x_{i_1} = x_{i_2} = x_i$ para todo $i = 1, \ldots, n$, se cumple,

$$\tilde{x} = \frac{1}{2n}\sum_{j=1}^{2}\sum_{i=1}^{n} x_{i_j} = \frac{1}{2n}2\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \tag{6.4}$$

y

$$\bar{Y} = \frac{1}{2n}\sum_{j=1}^{2}\sum_{i=1}^{n} Y_{i_j} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\sum_{j=1}^{2} Y_{i_j} = \frac{1}{n}\sum_{i=1}^{n}\bar{Y}_i = \tilde{Y}. \tag{6.5}$$

Además, teniendo en cuenta $x_{i_1} = x_{i_2} = x_i$ para todo $i = 1, \ldots, n$, (6.4) y (6.5), se cumple,

$$\sum_{j=1}^{2}\sum_{i=1}^{n}\left(x_{i_j} - \tilde{x}\right)^2 = \sum_{j=1}^{2}\sum_{i=1}^{n}\left(x_{i_j} - \bar{x}\right)^2 = 2\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \tag{6.6}$$

y

$$\begin{aligned}\sum_{j=1}^{2}\sum_{i=1}^{n}\left(x_{i_j} - \tilde{x}\right)Y_{i_j} &= \sum_{i=1}^{n}\sum_{j=1}^{2}\left(x_{i_j} - \bar{x}\right)Y_{i_j} = \sum_{i=1}^{n}\left[\left(x_{i_1} - \bar{x}\right)Y_{i_1} + \left(x_{i_2} - \bar{x}\right)Y_{i_2}\right] \\ &= \sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left[Y_{i_1} + Y_{i_2}\right] = 2\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\bar{Y}_i. \end{aligned} \tag{6.7}$$

Luego, teniendo en cuenta (6.6) y (6.7), se tiene,

$$\hat{\beta}_1 = \frac{\sum_{j=1}^{2}\sum_{i=1}^{n}\left(x_{i_j} - \tilde{x}\right)Y_{i_j}}{\sum_{j=1}^{2}\sum_{i=1}^{n}\left(x_{i_j} - \tilde{x}\right)^2} = \frac{2\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\bar{Y}_i}{2\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = \tilde{\beta}_1. \tag{6.8}$$

Finalmente, de (6.4), (6.5) y (6.8), se tiene,

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\tilde{x} = \tilde{Y} - \tilde{\beta}_1\bar{x} = \tilde{\beta}_0,$$

por lo que podemos decir que ambos estimadores son iguales.

Apéndice A

Formulario de Distribuciones

	$P(X=x) \quad \vert \quad f_X(x)$	$E(X)$	$V(X)$	$M_X(t)$	$R_X(x)$
$X \sim \text{Bernoulli}(p)$	$p^x(1-p)^{1-x}$	p	$p(1-p)$	$(1-p)+pe^t$	$x=0,1.$
$X \sim \text{Binomial}(n,p)$	$\binom{n}{x}p^x(1-p)^{n-x}$	np	$np(1-p)$	$((1-p)+pe^t)^n$	$x=0,1,\ldots,n$
$X \sim \text{Geom}(p)$	$p(1-p)^{x-1}$	$\frac{1}{p}$	$\frac{(1-p)}{p^2}$	$\frac{pet}{1-(1-p)e^t}$, si $(1-p)e^t<1$	$x=1,2,\ldots$
$X \sim \text{BN}(r,p)$	$\binom{x-1}{r-1}p^r(1-p)^{x-r}$	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$	$\left(\frac{e^tp}{1-(1-p)e^t}\right)^r$	$x=r,r+1,r+2,\ldots$
$X \sim \text{Hiper}(M,N,n)$	$\frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$	np	$n\frac{M}{N}(\frac{N-M}{N})(\frac{N-n}{N-1})$		$x=0,1,\ldots,\text{mín}(M,n)$
$X \sim \text{Poisson}(\mu)$	$\frac{\mu^x e^{-\mu}}{x!}$	μ	μ	$e^{\mu(e^t-1)}$	$x=0,1,\ldots$
$X \sim \text{Unif}(a,b)$	$\frac{1}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{tb}-e^{at}}{t(b-a)}$, $t\neq 0$	$a\leq x\leq b$
$X \sim \text{Exp}(\lambda)$	$\lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda-t}$,$t<\lambda$	$x>0$
$X \sim \text{Normal}(\mu,\sigma^2)$	$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2	$e^{\mu t+\frac{\sigma^2t^2}{2}}$	$-\infty<x<\infty$
$X \sim \text{Gamma}(\alpha,\beta)$	$\frac{1}{\Gamma(\alpha)}\beta^\alpha x^{\alpha-1}e^{-\beta x}$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$	$\left(\frac{\beta}{\beta-t}\right)^\alpha$, $t<\beta$	$x>0$
$X \sim \text{Beta}(\alpha,\beta)$	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$		$0\leq x\leq 1$

Apéndice B

Formulario de Análisis de Regresión Simple

1. **Modelo de Regresión Estimado**

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

2. **Suma de cuadrados**

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

$b)\ S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}.$

$c)\ S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}$

d) $SSE = S_{yy} - \hat{\beta}_1 S_{xy}$.

e) $SSR = \hat{\beta}_1 S_{xy}$.

f) $SST = SSR + SSE = S_{yy}$

3. **Varianzas y Desviaciones Estándar**

a) $\hat{\sigma}^2 = \frac{SSE}{n-2}$

b) $se(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)}$

c) $se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$

4. **Test de Hipótesis para los coeficientes**

a) $H_0 : \beta_0 = 0$ $H_1 : \beta_0 \neq 0$

$$T_0 = \frac{\hat{\beta}_0}{se(\hat{\beta}_0)}$$

b) $H_0 : \beta_1 = 0$ $H_1 : \beta_1 \neq 0$

$$T_1 = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$$

En ambos caso se rechaza la hipótesis nula si $|T_i| > t_{n-2,1-\alpha/2}$

5. **Intervalos de Confianza**

a) Intervalos de Confianza para los coeficientes

$$IC(\beta_0) = \hat{\beta}_0 \mp t_{n-2,1-\alpha/2} \cdot se(\hat{\beta}_0)$$

$$IC(\beta_1) = \hat{\beta}_1 \mp t_{n-2,1-\alpha/2} \cdot se(\hat{\beta}_1)$$

b) Intervalo de Confianza para la Predicción y_0 en el valor x_0, donde $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$

$$IC(y_0) = \hat{y}_0 \mp t_{n-2,1-\alpha/2} \sqrt{\hat{\sigma}^2 \left\{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right\}}$$

c) Intervalo de Confianza para la respuesta media, donde $\hat{\mu}_{y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$

$$IC(\mu_{y|x_0}) = \hat{\mu}_{y|x_0} \mp t_{n-2,1-\alpha/2} \sqrt{\hat{\sigma}^2 \left\{ \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right\}}$$

6. **Coeficiente de Determinación** R^2

$$R^2 = \hat{\beta}_1 \frac{S_{xy}}{S_{yy}} = 1 - \frac{SSE}{S_{yy}}$$

Apéndice C

Tablas de distribución

C.1. Distribución t de Student

	Magnitud de α en una cola							
gl	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.0005**
1	1.38	1.96	3.08	6.31	12.71	31.82	63.66	636.58
2	1.06	1.39	1.89	2.92	4.30	6.96	9.92	31.60
3	0.98	1.25	1.64	2.35	3.18	4.54	5.84	12.92
4	0.94	1.19	1.53	2.13	2.78	3.75	4.60	8.61
5	0.92	1.16	1.48	2.02	2.57	3.36	4.03	6.87
6	0.91	1.13	1.44	1.94	2.45	3.14	3.71	5.96
7	0.90	1.12	1.41	1.89	2.36	3.00	3.50	5.41
8	0.89	1.11	1.40	1.86	2.31	2.90	3.36	5.04
9	0.88	1.10	1.38	1.83	2.26	2.82	3.25	4.78
10	0.88	1.09	1.37	1.81	2.23	2.76	3.17	4.59
11	0.88	1.09	1.36	1.80	2.20	2.72	3.11	4.44
12	0.87	1.08	1.36	1.78	2.18	2.68	3.05	4.32
13	0.87	1.08	1.35	1.77	2.16	2.65	3.01	4.22
14	0.87	1.08	1.35	1.76	2.14	2.62	2.98	4.14
15	0.87	1.07	1.34	1.75	2.13	2.60	2.95	4.07
16	0.86	1.07	1.34	1.75	2.12	2.58	2.92	4.01
17	0.86	1.07	1.33	1.74	2.11	2.57	2.90	3.97
18	0.86	1.07	1.33	1.73	2.10	2.55	2.88	3.92
19	0.86	1.07	1.33	1.73	2.09	2.54	2.86	3.88
20	0.86	1.06	1.33	1.72	2.09	2.53	2.85	3.85
21	0.86	1.06	1.32	1.72	2.08	2.52	2.83	3.82
22	0.86	1.06	1.32	1.72	2.07	2.51	2.82	3.79
23	0.86	1.06	1.32	1.71	2.07	2.50	2.81	3.77
24	0.86	1.06	1.32	1.71	2.06	2.49	2.80	3.75
25	0.86	1.06	1.32	1.71	2.06	2.49	2.79	3.73
26	0.86	1.06	1.31	1.71	2.06	2.48	2.78	3.71
27	0.86	1.06	1.31	1.70	2.05	2.47	2.77	3.69
28	0.85	1.06	1.31	1.70	2.05	2.47	2.76	3.67
29	0.85	1.06	1.31	1.70	2.05	2.46	2.76	3.66
30	0.85	1.05	1.31	1.70	2.04	2.46	2.75	3.65
∞	0.84	1.04	1.28	1.64	1.96	2.33	2.58	3.29

C.2. Distribución χ^2

	Proporción del Area hasta $+\infty$						
gl	**0.995**	**0.99**	**0.975**	**0.95**	**0.90**	**0.75**	**0.50**
1	0.00	0.00	0.00	0.00	0.02	0.10	0.45
2	0.01	0.02	0.05	0.10	0.21	0.58	1.39
3	0.07	0.11	0.22	0.35	0.58	1.21	2.37
4	0.21	0.30	0.48	0.71	1.06	1.92	3.36
5	0.41	0.55	0.83	1.15	1.61	2.67	4.35
6	0.68	0.87	1.24	1.64	2.20	3.45	5.35
7	0.99	1.24	1.69	2.17	2.83	4.25	6.35
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34
11	2.60	3.05	3.82	4.57	5.58	7.58	10.34
12	3.07	3.57	4.40	5.23	6.30	8.44	11.34
13	3.57	4.11	5.01	5.89	7.04	9.30	12.34
14	4.07	4.66	5.63	6.57	7.79	10.17	13.34
15	4.60	5.23	6.26	7.26	8.55	11.04	14.34
16	5.14	5.81	6.91	7.96	9.31	11.91	15.34
17	5.70	6.41	7.56	8.67	10.09	12.79	16.34
18	6.26	7.01	8.23	9.39	10.86	13.68	17.34
19	6.84	7.63	8.91	10.12	11.65	14.56	18.34

	Proporción del Area hasta $+\infty$						
gl	**0.25**	**0.10**	**0.05**	**0.03**	**0.01**	**0.005**	**0.001**
1	1.32	2.71	3.84	5.02	6.63	7.88	10.83
2	2.77	4.61	5.99	7.38	9.21	10.60	13.82
3	4.11	6.25	7.81	9.35	11.34	12.84	16.27
4	5.39	7.78	9.49	11.14	13.28	14.86	18.47
5	6.63	9.24	11.07	12.83	15.09	16.75	20.51
6	7.84	10.64	12.59	14.45	16.81	18.55	22.46
7	9.04	12.02	14.07	16.01	18.48	20.28	24.32
8	10.22	13.36	15.51	17.53	20.09	21.95	26.12
9	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	12.55	15.99	18.31	20.48	23.21	25.19	29.59
11	13.70	17.28	19.68	21.92	24.73	26.76	31.26
12	14.85	18.55	21.03	23.34	26.22	28.30	32.91
13	15.98	19.81	22.36	24.74	27.69	29.82	34.53
14	17.12	21.06	23.68	26.12	29.14	31.32	36.12
15	18.25	22.31	25.00	27.49	30.58	32.80	37.70
16	19.37	23.54	26.30	28.85	32.00	34.27	39.25
17	20.49	24.77	27.59	30.19	33.41	35.72	40.79
18	21.60	25.99	28.87	31.53	34.81	37.16	42.31
19	22.72	27.20	30.14	32.85	36.19	38.58	43.82

C.3. Distribución F ($\alpha = 0{,}05$)

Grados de libertad	**Grados de libertad para el numerador**									
denominador	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	161	199	216	225	230	234	237	239	241	242
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83

(Continuación)

Grados de libertad denominador	Grados de libertad para el numerador								
	12	**15**	**20**	**24**	**30**	**40**	**60**	**120**	∞
1	244	246	248	249	250	251	252	253	254
2	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5
3	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
30	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

C.4. Distribución Normal

	Segunda cifra decimal en z									
z	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998