SOLUCIÓN PROBLEMA 1

Se desea estudiar el tiempo que funciona cierta componente de un sistema hasta que comienza a presentar fallos sistemáticos. A fin de estimar el tiempo medio de funcionamiento a pleno rendimiento, se observaron 10 de estas componentes obteniéndose los siguientes tiempos –en miles de horas– de funcionamiento sin fallos:
1 ,1.5 ,0.8 ,1.2 ,0.9 ,1.1 ,1.4 ,1.3 ,0.7 ,0.1

a) (2 puntos) Dar una estimación por intervalo para el tiempo medio de funcionamiento con una confianza del 90%.

- Estadístico, premisas y distribución

X normal, m.a.s con varianza desconocida, $\hat{t} = \frac{\overline{x} - \mu}{s/\sqrt{n}} \sim t_{n-1}$

- Cálculos, resultado e interpretación.

$\overline{x} = 1$

$s^2 = 0{,}1667$

$$IC(\mu, 90\%) = \overline{x} \pm t_{9,0.95}\sqrt{\frac{s^2}{n}} = 1 \pm 1.8331 * \sqrt{\frac{0{,}1667}{10}} = (0.7633,\ 1.2367)$$

Con una confianza del 90%, el tiempo medio de funcionamiento de la componente está entre 0.763 y 1.237 miles de horas.

b) (2 puntos) Dar una estimación por intervalo para la desviación estándar poblacional con una confianza del 90%.

- Estadístico, premisas y distribución

X normal, m.a.s, $\hat{X} = \frac{s^2(n-1)}{\sigma^2} \sim \chi^2_{n-1}$.

- Cálculos, resultado e interpretación.

$s^2 = 0.1667$

$$IC(\sigma^2, 90\%) = \left(\frac{(n-1)s^2}{\chi^2_{9,1-\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{9,\alpha/2}}\right) = \left(\frac{9 \cdot 0.1667}{16.91898}, \frac{9 \cdot 0.1667}{3.325113}\right) = (0.0887,\ 0.4511) \Rightarrow$$

$$IC(\sigma, 90\%) = \left(\sqrt{\frac{(n-1)s^2}{\chi^2_{9,1-\alpha/2}}}, \sqrt{\frac{(n-1)s^2}{\chi^2_{9,\alpha/2}}}\right) = (0.2978,\ 0.6716)$$

Con una confianza del 90%, la desviación tipo poblacional del tiempo de funcionamiento está 0.298 y 0.672 miles de horas.

c) (2 puntos) ¿Podemos decir que el tiempo de funcionamiento medio está por debajo de 1.200 horas? Asumid varianza conocida igual a 0,12 y α=0,05.

- Hipótesis, estadístico, premisas y distribución

Sea X la v.a. tiempo de funcionamiento de una componente, en miles de horas.

H_0: μ =1,2 H_1: μ <1,2 Unilateral	$\hat{z} = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$ bajo H_0.	X normal, m.a.s, varianza conocida.

- Cálculos, resultado e interpretación.

$$\hat{z} = \frac{1 - 1.2}{\sqrt{0.12/10}} = -1.825742 \Rightarrow \quad \text{Unilateral: rechazamos si } \hat{z} < z_{0.05} = -1{,}644854 \text{ (sin valor absoluto)}$$

p-valor=$P(\hat{z} < Z) = P(-1.83 < Z) = P(Z > 1.83) = 1 - P(Z \le 1.83) = 1 - 0.9664 = 0.0336$

Con un error del 5%, hay evidencia para rechazar la hipótesis nula y demostrar que el tiempo medio de funcionamiento está por debajo de 1200 horas: el valor observado del estadístico está en la región de rechazo (-1.8226< -1,6449), y el p-valor<0.05 (basta uno de los dos).

Una vez las componentes observadas fallaron, se realizó una intervención para arreglarlas y volver a hacerlas funcionar. El mecánico nos ha garantizado que obtendremos un mayor rendimiento del sistema. Así, se volvieron a poner en funcionamiento, obteniéndose los siguientes nuevos tiempos –en miles de horas– de funcionamiento sin fallos:
0.1, 0.3, 0.4, 1.3, 1.4, 1.2, 0.4, 1.4, 0.7, 1.3

d) (2 puntos) ¿Se ha alargado efectivamente el tiempo medio de funcionamiento al arreglar la componente?

- Hipótesis, estadístico, premisas y distribución (tomar α=0.05)

SOLUCION ÓPTIMA

Definimos la nueva variable diferencia del tiempo de funcionamiento entre antes y después de arreglar la componente D=después-antes:

Se desea contrastar

$H_0 : \mu_D = E(D) = 0$

$H_1 : \mu_D = E(D) > 0$ (el tiempo medio al arreglar la componente es mayor al anterior)

Estadístico

$\hat{t} = \dfrac{\bar{D} - \mu_0}{S_D / \sqrt{n}}$ que bajo H0 sigue una distribución t_{n-1}.

Premisas: muestras apareadas, normalidad de la variable diferencia D, varianza desconocida.

SOLUCION ACEPTADA:

Sea Y el tiempo de funcionamiento de una componente después de ser arreglada.

Hipótesis

$H_0 : \mu_Y = 1$

$H_1 : \mu_Y > 1$ (al arreglar la componente, el tiempo medio es superior a antes de arreglar $\bar{x} = 1$)

Estadístico:

$\hat{t} = \dfrac{(x - \mu)}{\sqrt{s^2 / n}}$ que bajo H0 sigue una distribución t_{n-1}

Premisas: m.a.s de Y, Y sigue una distribución normal, varianza desconocida.

- Cálculos, resultado e interpretación.

SOLUCION ÓPTIMA

D=Y-X:
-0.9 -1.2 -0.4 0.1 0.5 0.1 -1.0 0.1 0.0 1.2

$\bar{D} = -0.15$

$s_D^2 = 0.545$

$\hat{t} = \dfrac{-0.15 - 0}{\sqrt{0.545/10}} = \text{-0.6425294} \Rightarrow$ Unilateral: rechazamos si $\hat{t} > t_{9,1-\alpha} = t_{9,0.95} = 1.833113$ (sin valor absoluto)

Con un error del 5%, no hay evidencia para rechazar la hipótesis nula y concluir que el arreglo alarga efectivamente el tiempo medio de funcionamiento: el valor observado NO está en la región de rechazo (-0.64<1.833113). De hecho, con la muestra observada se obtiene un funcionamiento medio inferior (0.85) después de arreglar las componentes.

SOLUCION ACEPTADA

$\bar{y} = 0.85$

$s_Y^2 = 0.26944$

$\hat{t} = \dfrac{0.85 - 1}{\sqrt{0.26944/10}} = \text{-0.91382} \Rightarrow$ Unilateral: rechazamos si $\hat{t} > t_{9,1-\alpha} = t_{9,0.95} = 1.833113$ (sin valor absoluto)

Con un error del 5%, no hay evidencia para rechazar la hipótesis nula y concluir que el arreglo alarga efectivamente el tiempo medio de funcionamiento: el valor observado NO está en la región de rechazo (-0.91382<1.833113). De hecho, con la muestra observada se obtiene un funcionamiento medio inferior (0.85) después de arreglar las componentes.

Datos posteriores recogidos por la empresa, muestran que de las 1000 componentes fabricadas en el primer mes del año, 216 duraron menos de 900 horas.

- (2 puntos) Estimar por intervalo la probabilidad de que una componente dure menos de 900 horas, con una confianza del 95%.
- Estadístico, premisas y distribución.

$\hat{z} = \dfrac{p - \pi}{\sqrt{\pi(1-\pi)/n}} \sim N(0,1)$

Premisas: n grande y probabilidad no extrema: $(1-\pi)n \geq 5, \pi n \geq 5$

- Cálculos, resultado e interpretación.

$$\hat{P} = \frac{216}{1000} = 0.216$$

$$IC(\pi, 95\%) = \hat{P} \pm z_{0.975}\sqrt{\frac{\pi(1-\pi)}{n}} = 0.216 \pm 1.96\sqrt{\frac{\pi(1-\pi)}{n}}$$

Podemos aproximar la probabilidad dentro de la raíz con

$\pi = \hat{P}$ o π=0.5 (màxim valor possible):

$$IC(\pi, 95\%) = 0.216 \pm 1.96\sqrt{\frac{0.216(1-0.216)}{1000}} = (0.1905, 0.2415)$$

$$IC(\pi, 95\%) = 0.216 \pm 1.96\sqrt{\frac{0.5(1-0.5)}{1000}} = (0.1850, 0.2470)$$

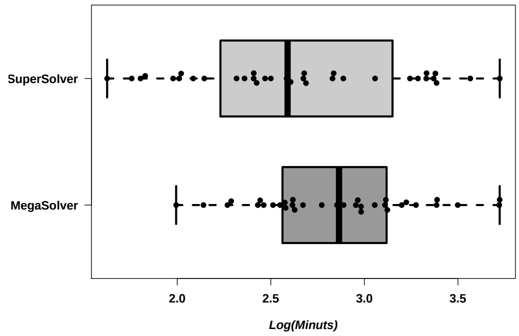
Con una confianza del 95%, la probabilidad de que una componente dure menos de 900 horas está entre 19.0% y 24.2%.

Problema 2 (B5)

Per comparar la velocitat amb la qual resolen dos servidors diferents, *SuperSolver* i *MegaSolver*, problemes d'optimització s'envia un total de 70 problemes de maximització diferents als dos servidors, 35 a cadascun. Pel fet que el temps que triguen els servidors per resoldre els problemes, és asimètrica cap a la dreta, treballem a continuació amb els logaritmes dels temps. Siguin X el logaritme del temps que triga el *SuperSolver* i Y el del *MegaSolver* .

Els valors descriptius a cada mostra són els següents i a més a més es mostra una representació gràfica:

	Mitjana	**Mediana**	**Desv. est.**	**Mínim**	**Màxim**
SuperSolver	2,63	2,59	0,57	1,63	3,72
MegaSolver	2,85	2,86	0,44	1,99	3,72



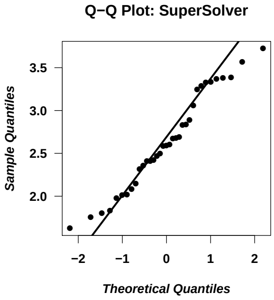
a) (0.5 punts)
Es tracta de dues mostres independents o aparellades? Raoneu la resposta.

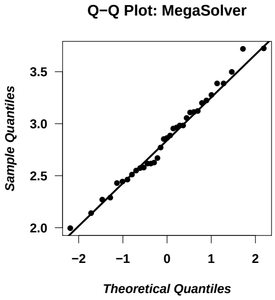
Solució:
Es tracta de mostres independents, ja que els problemes d'optimització enviats a un servidor són independents dels enviats a l'altre.

b) (0.75 punts)
Només basant-vos en els valors descriptius i en el gràfic, creieu que es pot parlar de superioritat d'un dels dos servidors?

Solució:
Tant la mitjana com la mediana dels temps son inferiors en cas del *SuperSolver* el que podria indicar un millor rendiment d'aquest. Per altra banda, observem una major variabilitat dels temps del *SuperSolver* i per tant no es pot fer cap afirmació contundent.

c) (0.75 punts)
Veient els següents *Q-Q plots*, sembla raonable suposar que les variables X i Y segueixen una distribució normal? Raoneu la resposta.





Solució:
En cas del *Megasolver* els punts formen gairebé una recta i per tant podem suposar que provenen d'una distribució normal. En cas del *SuperSolver*, la cosa no està tan clar. No obstant, la suposició que les dades provenen d'una distribució normal sembla raonable, ja que els punts tampoc no s'allunyen massa de la recta.

d) (2 punts)
Estudiem primer la variabilitat dels (logaritmes de) temps d'ambdós servidors. Es pot suposar que són iguals? Responeu aquesta pregunta plantejant i resolent la hipòtesi adient (amb $\alpha = 0.1$) i explicant quines són les premisses que cal fer per realitzar aquesta prova.

Solució:

Hipòtesi: $\mathrm{H}_0\colon \sigma_X^2 = \sigma_Y^2$ vs. $\mathrm{H}_1\colon \sigma_X^2 \neq \sigma_Y^2$

Estadístic de contrast: $F = s_X^2/s_Y^2 \overset{H_0}{\sim} F_{34,\,34}$

Premises: Distribució normal d'X i d'Y i mostres independents.

Càlcul: $F = 0.57^2\big/0.44^2 = 1.68$

Regió crítica: Es rebutja H_0 si $F > F_{34,\,34;0.95} = 1.77$.

Decisió: Sent el valor d'F inferior al valor crític, no rebutgem H_0 i podem suposar igualtat de variàncies.

e) (0.3+0.6+0.5+1+0.6=3 punts)
Per saber si es pot suposar que hi ha diferències entre ambdós servidors, es vol calcular l'interval de confiança (al 95%) per a la diferència de les mitjanes dels logaritmes del temps $\big(\mathrm{IC}(\mu_X - \mu_Y;\, 0.95)\big)$.

- Quines són les premisses?

  Solució:

  $X \sim \mathcal{N}(\mu_X, \sigma)$, $Y \sim \mathcal{N}(\mu_Y, \sigma)$, és a dir distribucions normals i homocedasticitat, i dades independents.

- Quina és la distribució de $\bar{X} - \bar{Y}$ (suposant igualtat de variances)?

  Solució:

  $$\bar{X} \sim \mathcal{N}(\mu_X, \sigma/\sqrt{35}),\ Y \sim \mathcal{N}(\mu_Y, \sigma/\sqrt{35}) \Longrightarrow \bar{X} - \bar{Y} \sim \mathcal{N}(\mu_X - \mu_Y, \sigma\sqrt{1/35 + 1/35})$$

- Calculeu la variància conjunta (*pooled variance*).

  Solució:

  $$s^2 = \frac{1}{35 + 35 - 2}(34 \cdot 0.57^2 + 34 \cdot 0.44^2) = 0.26.$$

- Calculeu $\mathrm{IC}(\mu_X - \mu_Y;\, 0.95)$.

  Solució:

  $$\mathrm{IC}(\mu_X - \mu_Y;\, 0.95) = \bar{x} - \bar{y} \pm \underbrace{t_{35+35-2;\,0.975}}_{=2} \cdot s \cdot \sqrt{1/35 + 1/35} = [-0.46,\, 0.02].$$

- Quina és la interpretació d'aquest interval?

  Solució:
  Aquest interval conté la diferència dels valors esperats d'X i d'Y amb una probabilitat 0.95. Com que l'interval inclou el valor 0, amb un nivell de confiança del 95% no podem afirmar que cap dels servidors es superior a l'altre pel que fa el rendiment mitjà.

f) (2 punts)
En una altre experiment s'han enviat els mateixos 31 problemes d'optimització a ambdós servidors i s'han recollit les diferencies dels logaritmes dels temps. La mitjana d'aquestes diferències ha estat igual a 0.22 (a favor del *SuperSolver*), la desviació estàndard igual a 0.37.
Calculeu l'interval de confiança (al 99%) per a la mitjana de les diferències. Es pot suposar la superioritat d'un dels dos servidors? Per què?

Solució:
L'interval de confiança (al 99%) de la mitjana de les diferències (μ_D) es:

$$\mathrm{IC}(\mu_D;\, 0.99) = 0.22 \pm \underbrace{t_{31-1;\,0.995}}_{=2.75} \cdot 0.37/\sqrt{31} = [0.04,\, 0.4].$$

Pel fet que l'IC no conté el valor 0, podem afirmar, amb un nivell de confiança del 99% que el rendiment mitjà del *SuperSolver* és superior que el del *MegaSolver*.

g) (1 punt)
Mantenint la mitjana igual, a partir de quina desviació estàndard ja no es podria parlar de superioritat (amb un nivell de confiança igual a 99%)?

Solució:
Seria el cas si el limit inferior fos inferior a 0, és a dir si

$$0.22 - 2.75 \cdot s/\sqrt{31} < 0 \iff s > \frac{0.22}{2.75} \cdot \sqrt{31} = 0.45.$$

NOM:____

(Poseu el nom i contesteu cada pregunta en el seu lloc reservat. Expliciteu i justifiqueu els càlculs en les respostes)

Problema 3 (B6)

Les variables Y i X mesuren una puntuació de 0 a 10 en un esforç físic i hores connectat a Facebook, respectivament. Hem observat els següents 9 valors: Y= (8,4,5,5,2,3,3,2,0) i X= (8,9,11,14,15,16,19,20,21) I alguns resultats intermedis són :

$\sum_{i=1}^{9} X_i = 133$ $\sum_{i=1}^{9} X_i^2 = 2145$ $\sum_{i=1}^{9} Y_i = 32$ $\sum_{i=1}^{9} Y_i^2 = 156$ $\sum_{i=1}^{9} X_i Y_i = 400$

$\overline{X} = 14.778$ $\overline{Y} = 3.556$ $s_X = 4.738$ $s_Y = 2.297$

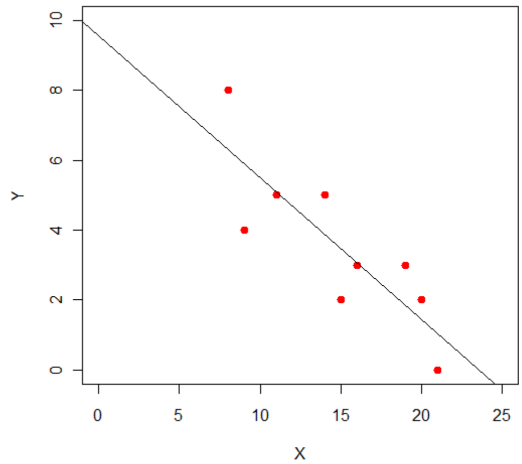
Volem explicar la variabilitat de Y en funció dels valors de X Aplicant regressió lineal, en R alguns dels resultats que obtenim són:

```
Coefficients:                 Estimate  Std.Error   t value    Pr(>|t|)
              (Intercept)     .........   1.5478     6.173     0.000457
              X               .........   0.1003    -4.049     0.004877
```

(1 punt) Quina és la recta de regressió? Y = 9.61 – 0.41 X

Indiqueu el càlcul dels coeficients i dibuixeu la recta

S_{XY}=(400-((133*32)/9))/8 = -9.11 b_1 = -9.11 / $(4.738)^2$ = -0.41 b_0 =3.556 – (-0.41)*14.778 = 9.61



(1 punt) Comenteu el significat dels coeficients de la recta:

Ordenanda a l'orígen: a 0 hores de Facebook correspon una puntuació de 9.61

Pendent (negativa): per cada hora més de connexió a Facebook la puntuació baixa en 0.41 punts

(1 punt) Completeu la taula de descomposició de la variabilitat. Comenteu aquesta descomposició i què indica?

Font Variabilitat	SQ	GdL DF	QM	Rati
Explicada	30.19	1	30.19	17.55
Residual	12.02	7	1.72	
Total	42.21	8		

SQ Explicada = $(-0.41)^2$ 8 $(4.738)^2$ = 30.19

SQ Total = 8 $(2.297)^2$ = 42.21

La variabilitat total (42.21) queda explicada en bona part pel model (30.19) i queda una part residual (12.02)

(1 punt) Resoleu la prova d'hipòtesi per acceptar o no una recta com a model:

O bé prova de pendent zero (H_0:β_1=0, bilateral): t-statistic = -4.049 p-value = 0.004877

O bé prova global de F-statistic: 17.55 on 1 and 7 DF (p-value aprox 0.004877)

Amb confiança del 95% no podem acceptar que no hi hagi relació lineal o que la recta sigui plana o de pendent 0.

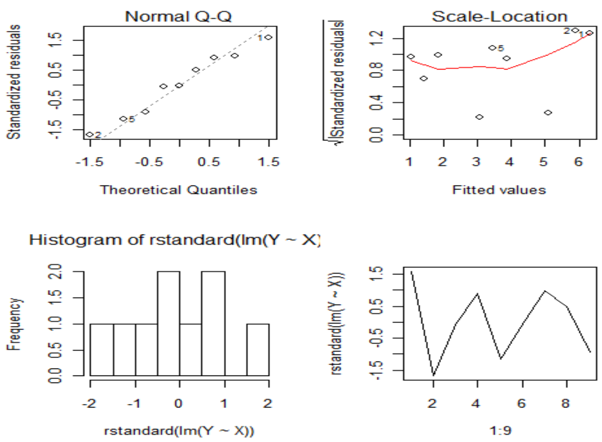
Per tant, una recta de pendent -0.41 ajusta a les dades.

(1 punt) Calculeu el coeficient de determinació i feu una interpretació global del model

R-squared: 30.19 / 42.21 = 0.72

La variabilitat explicada pel model és del 72%. Les hores connectades a Facebook expliquen un 72% de la variabilitat de la puntuació Y segons una recta de pendent -0.41

(1 punt) Feu l'anàlisi de les premisses.



No hi ha greus problemes de falta de normalitat (gràfic NormalQ-Q alineat i Histograma no lluny de campana), ni de falta d'independència (cap patró en els gràfics de la dreta), ni de heterocedasticitat (no zones amb diferents variabilitats en els gràfics de la dreta). Per tant es pot assumir normalitat, independencia i homocedasticitat.

(1 punt) Feu la prova d'hipòtesi adient per si es pot acceptar o no que el pendent de la recta és -0.5

H_0: β_1=-0.5 (bilateral)

$(b_1-(-0.5))/S_{b1}$ = ((-0.41) – (-0.5)) / 0.1003 = 0.90 en t_7

Podem acceptar que el pendent és -0.5, és a dir que per cada hora més de connexió la puntuació baixa mig punt perquè -2.365 < 0.90 <2.365 ($t_{0.975,7}$=2.365)

(sense taules cal saber que en una t el punt crític per no acceptar H_0 està al voltant de 2 i 0.9 és prou inferior)

Ara considerem la variable hores de connexió només observada en 3 valors o categories: 10, 15 o 20 hores. Volem relacionar la mateixa variable Y anterior amb la nova variable X_categories=(10,10,10,15,15,15,20,20,20).

En R alguns dels resultats que obtenim són:

```
Response: Y
              Df    Sum_Sq    Mean_Sq   F_value    Pr(>F)
X_categories  ......  24.222    ..........   ..........   0.07748
Residuals     ......  ..........  ..........
```

(1 punt) Completeu la taula de descomposició de la variabilitat. Comenteu aquesta descomposició i què indica?

Font Variabilitat	SQ	GdL DF	QM	Rati
Explicada	24.222	2	12.111	4.04
Residual	17.99	6	3.000	
Total	42.21			

SQ Total = 42.21 calculat anteriorment. SQResidual = 42.21 – 24.222 = 17.99

La variabilitat total (42.21) queda explicada una part pel model (24.222) i queda una part residual (17.99).

Respecte el model anterior de regressió amb X, la part explicada és inferior i la residual superior.

(1 punt) Resoleu la prova d'hipòtesi de si hi ha relació o no entre la puntuació Y i X_categories

Prova global (H_0: $\mu_1=\mu_2=\mu_3$) de F-statistic: 4.04 on 2 and 6 DF, p-value: 0.07748

Amb risc del 5% podem acceptar que no hi ha relació entre Y i X_categories. Per tant, no podem demostrar/afirmar que hi ha relació entre Y i X_categories.

Per tant, les mitjanes de puntuació no les considerem diferents entre les categories

(1 punt) Calculeu el coeficient de determinació i feu una interpretació global del model

R-squared: 24.222 / 42.21 = 0.57

La variabilitat explicada pel model és del 57%, i el model de diferenciar les mitjanes de Y segons les 3 categories no s'ha pogut acceptar

(el model d'una recta de pendent -0.41 per explicar Y en funció de X, no de X_categories, explicava fins un72%)