NOM:____

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

Problema B4.

1) (2,5 punts) La implementació d'e-status en l'assignatura PE ha millorat el rendiment acadèmic dels alumnes que l'utilitzen. Se sap que la proporció d'execucions de problemes no aprovades és del 10%, tot i que s'espera disminuir aquest percentatge. Durant el Q1 del curs 2010-2011 es va prendre una mostra de 115 execucions i es va observar que en 10 execucions no es va aconseguir l'aprovat. Hi ha evidència de canvi en la proporció de problemes aprovats? Per respondre plantegeu una PS.

- (0,5 punts) Hipòtesi (i indicar si la prova és bilateral o unilateral):

H_0: $\pi = 0{,}10$ (unilateral, a l'esquerra)

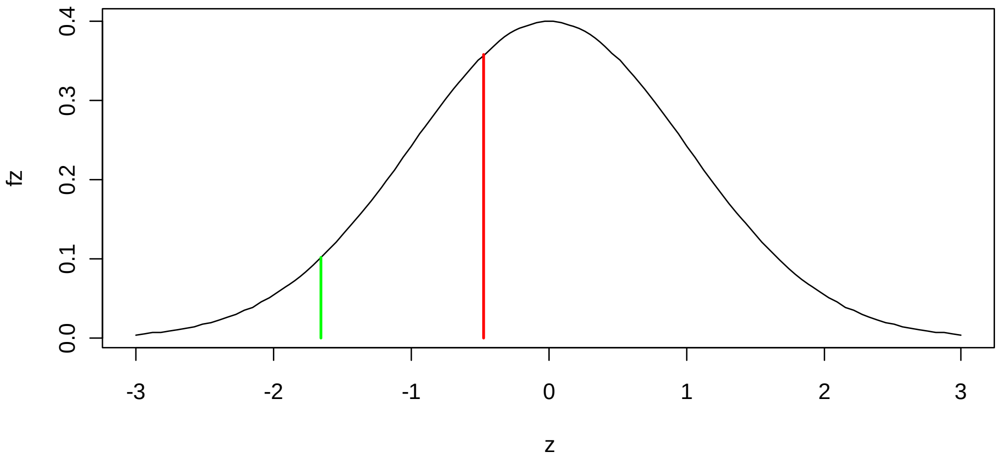
- (0,5 punts) Càlcul del valor de l'estadístic:

z = (0,087-0,10)/sqrt(0,10(1-0,10)/115) = -0,01304348/ 5,521441 = -0.47

- (0,5 punts) Càlcul del P-valor:

pvalor = 0,3205174

- (0,5 punts) Representeu gràficament el valor obtingut i el punt crític sobre la distribució de l'estadístic:



- (0,5 punts) Decisió de la PS i interpretació:

No hi ha evidencia per rebutjar la hipòtesis de que la proporció d'execucions en les que no s'aconsegueix l'aprovat és 10%.

Per conèixer l'ús que fan els estudiants d'e-status i així avaluar la seva dedicació a l'assignatura es va decidir estudiar el nombre de vegades que els alumnes resolen els problemes del programa durant el quadrimestre. Es va prendre una mostra aleatòria de 20 estudiants i es va registrar el nombre d'execucions de problemes que van realitzar al llarg del curs, obtenint els següents resultats.

$$\sum x = 1228 \qquad \sum x^2 = 86984$$

Si es considera que el nombre d'execucions és una variable amb distribució normal:

2) (1 punt) Amb aquesta mostra calculeu una estimació puntual del nombre mitjà d'execucions realitzades durant el quadrimestre.

$\hat{\mu} = \bar{x} = 61{,}4$

3) (1 punt) Doneu una estimació de l'error de la mitjana o error típic.

error tipus = 24,69264/sqrt(20) = 5,521441

4) (1,5 punts) Estimeu el nombre mitjà d'execucions amb un IC al 90%.

$61,4 \pm t_{19,0.975}$ * s/sqrt(20) = 61,4 ± 1,729 * 24,69264/sqrt(20) = [51.85269 , 70.94731]

5) (1,5 punts) Un altre aspecte a considerar per a una millor avaluació de l'ús d'e-status és estudiar les variacions en el nombre d'execucions. Obtingui un IC al 95% per a la variància del nombre d'execucions per a la mostra disponible.

$IC(\sigma^2, 1\text{-}\alpha) = (24,69264^2\ (20\text{-}1)/32,852\ ,\ 24,69264^2\ (20\text{-}1)/8,907)$
$IC(\sigma^2, 1\text{-}\alpha) = (18.77851^2\ ,\ 36.06537^2)$

6) (2,5 punts) Estudis realitzats en assignatures similars a PE d'altres estudis indiquen que el nombre mitjà d'execucions realitzades en un quadrimestre és de 80. Utilitzeu la mostra disponible per posar a prova amb un CH si els alumnes de PE formen part de la mateixa població que els d'altres estudis (en quant a la mitjana). Preneu $\alpha = 5\%$.
- (0,5 punts) Hipòtesis:

H_0: $\mu = 80$
H_1: $\mu \neq 80$

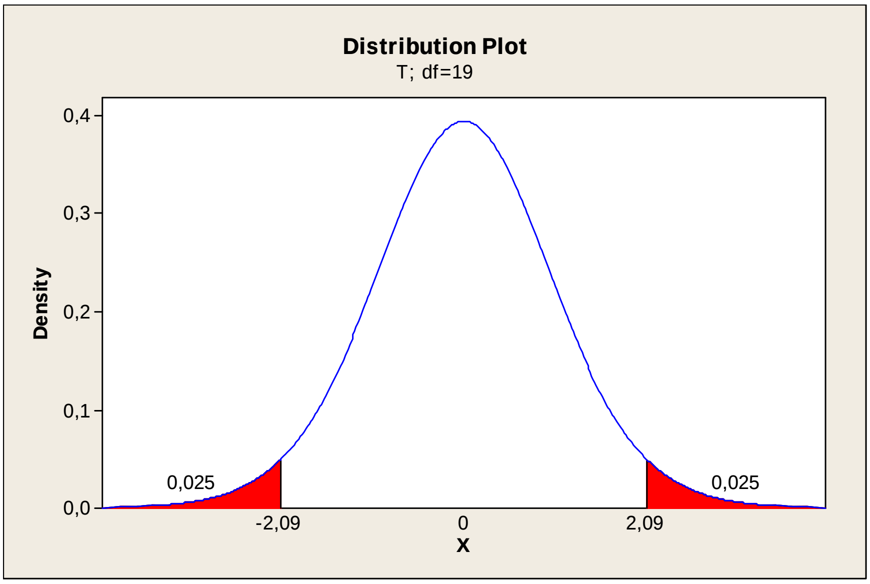
- (0,5 punts) Càlcul del valor de l'estadístic:

t = (61,4 – 80) / (24,69264/sqrt(20)) = -3,3687

- (0,5 punts) Quin és el valor que utilitzarà per decidir si l'estadístic anterior permet rebutjar la hipòtesi nul·la?

tcrític = 2,093

- (0,5 punts) Representeu gràficament el valor obtingut i el punt crític sobre la distribució de l'estadístic.



- (0,5 punts) Decisió del CH i interpretació:

Hi ha suficient evidència per rebutjar la hipòtesis de que el número mig d'execucions en PE és 80.

NOM:____

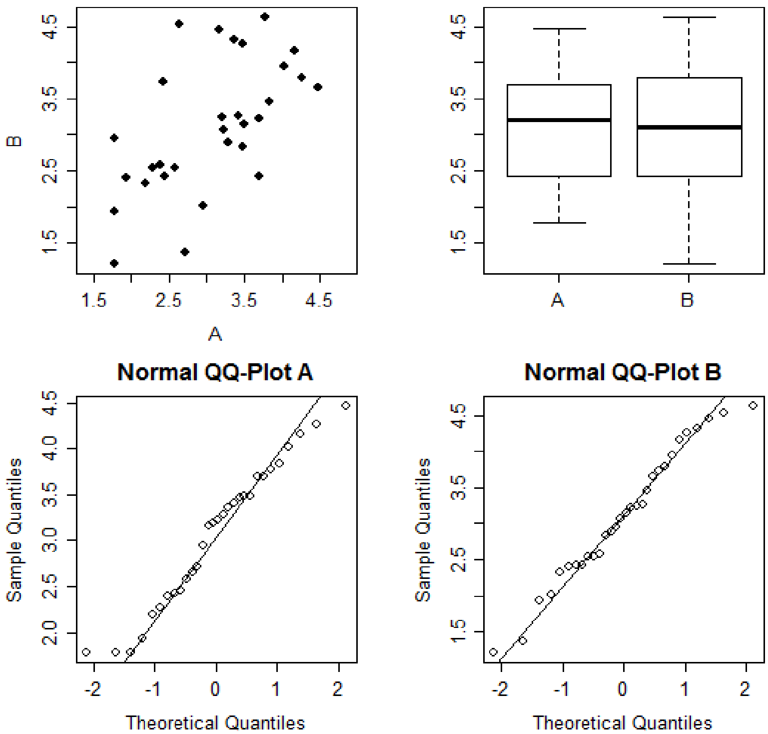
(Poseu el nom i contesteu cada pregunta en el seu lloc reservat. Expliciteu i justifiqueu els passos en les respostes).

Problema B5.

Un enginyer informàtic esta interessat en la comparació de la eficiència de dos algorismes SortA i SortB per a ordenar vectors amb 10 milions de números reals. Per a tal propòsit, l'enginyer genera 10 milions de números aleatoris, i els ordena amb SortA i SortB dins l'entorn estadístic R, fent server la funció `system.time` per mesurar el temps d'execució. Les mesures del temps d'execució es repeteixen 30 vegades pels dos algorismes, amb el codi:

```
n <- 30
m <- 10000000
A <- NULL
B <- NULL
for (i in 1:n) {
  z <- rnorm(m)
  a <- system.time(SortA(z))
  A <- c(A,a[3])
  b <- system.time(SortB(z))
  B <- c(B,b[3])
}
D <- B-A
```

La funció `system.time` retorna una estructura amb 3 elements, essent el tercer element el temps (en segons) gastat per la execució corresponent. Amb les dades obtingudes, l'enginyer fa estadística descriptiva gràfica i numèrica, obtenint els resultats a continuació.



Descriptiva gràfica

Descriptiva numèrica

	N	Mean	Variance
A	30	3.063	0.608
B	30	3.108	0.843
D	30	0.045	0.604

1. (1p) El disseny emprat en aquest estudi és de dues mostres independents o de dades aparellades? Argumenteu la resposta.

   És un disseny de dades aparellades. Els dos algorismes s'apliquen sempre a *un mateix joc de dades*. En un disseny amb dades aparellades s'espera una correlació entre el temps emprat pels dos algorismes, i aquesta correlació efectivament s'observa en el diagrama bivariant B versus A.

2. (3p) Feu el contrast de hipòtesi H0: $\mu_A = \mu_B$ versus H1 $\mu_A \neq \mu_B$, detallant el càlcul de l'estadístic utilitzat, i indicant la seva distribució de referència, premisses del contrast, la regla de decisió, el P-valor (encara que sigui aproximadament) i la decisió.

$H_0: \delta = \mu_B - \mu_A = 0$

$H_1: \delta \neq 0$

$d = \overline{x_B} - \overline{x_A} = 3.108 - 3.063 = 0.045$

$T = \frac{d}{\frac{s_d}{\sqrt{n}}} = \frac{\sqrt{30} \cdot 0.045}{\sqrt{0.604}} =$ **0.317**

Distribució de referència: t_{29}

Premisses: observacions (aquí diferències) independents, normalitat de les diferències (per a n petit).

$T < T_{29;0.975} = 2.045$ **no refusem H0.**

$\text{valor p} = 2 \cdot P(T_{29} > 0.317) = 2 \cdot 0.377 = 0.754$

(es pot aproximar per la Normal: valor p ≈ 2·P(Z>0.32) = 0.749)

Valor p >> 0.05, no refusem H0.

3. (2p) Feu un interval de confiança de 95% per a la diferència en eficiència entre els dos algorismes. Doneu una interpretació del interval obtingut.

$$IC(\delta) = \bar{x}_B - \bar{x}_A \pm t_{n-1}\frac{s_d}{\sqrt{n}} = 0.045 \pm 2.045\frac{\sqrt{0.604}}{\sqrt{30}} = (-0.245;\ 0.335)$$

El valor 0 es troba dins el interval, i per tant no es refusa la H0. Amb un confiança del 95%, el valor de δ es troba entre -0.245 i 0.335. Si l'experiment es repetís moltes vegades, en 95% dels casos el interval contindria el valor de δ.

4. (2p) Podem considerar que l'ordre de les execucions és aleatori? Si haguéssiu de fer l'experiment vosaltres mateixos, ho faríeu tal com s'ha fet o diferent? Argumenteu la resposta.

Sistemàticament s'executa algorisme A abans que B en cada iteració, produint una cadena d'execucions A-B-A-B-A, etc. Aquest ordre no és aleatori. Es podria considerar aleatoritzar l'ordre del les execucions.

5. (2p) Hi ha correlació entre el temps de les execucions fetes amb A i B? Raoneu la resposta, i calculeu el coeficient de correlació mostral entre les variables A i B.

Sí, es veu al gràfic de les observacions A vs B, i es lògic, pel Disney utilitzat.

$V(D) = V(A) + V(B) - 2Cov(A, B)$

$$Cov(A, B) = 0.5 * (V(D) - (V(A) + V(B))) = -0.5 * (0.604 - (0.608 + 0.843)) = 0.424$$

$r(A, B) = \frac{Cov(A, B)}{\sqrt{V(A) * V(B)}} = \frac{0.424}{\sqrt{0.608 * 0.843}} = 0.592$ **(aquí, V() i Cov() es refereixen als indicadors mostrals)**

NOM:____

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

Problema B6

Es vol estudiar si el temps de CPU (en milisegons) d'un programa ve afectat pel nombre d'operacions d'entrada i sortida (E/S) amb el disc. Es pren una mostra de 7 observacions i s'obtenen les següents dades:

nombre E/S (X)	CPU (Y)
14	2
16	5
27	7
42	9
39	10
50	13
83	20

Amb aquestes dades tenim que:

$\sum X = 271$,

$\sum X^2 = 13855$,

$\sum Y = 66$,

$\sum Y^2 = 828$,

$\sum XY = 3375$

a) Calculeu els estimadors de la constant i del pendent de la recta de regressió i representeu gràficament la recta estimada (*2 punts*)

```
x=c(14,16,27,42,39,50,83)
y=c(2,5,7,9,10,13,20)
n=7
reg=lm(y~x)
summary(reg)


Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.008282   0.831105   -0.01    0.992
x            0.243756   0.018681   13.05 4.72e-05 ***

Residual standard error: 1.083 on 5 degrees of freedom
Multiple R-squared: 0.9715,     Adjusted R-squared: 0.9658
F-statistic: 170.3 on 1 and 5 DF,  p-value: 4.716e-05
```

En l'estimació dels paràmetres:

```
b1=(n*sum(x*y)-sum(x)*sum(y))/(n*sum(x*x)-(sum(x))^2)
b1= 0.2437564
b0=mean(y)-b1*mean(x)
b0= -0.008282365
```

b) Calculeu la taula de descomposició de la variància (ajut: la variància residual val 1.173) (*2 punts*)

```
var.resi=1.173
sq.resi=(n-2)*var.resi
sq.resi= 5.865
sq.tot=(n-1)*var(y)
sq.tot= 205.7143
sq.reg=sq.tot-sq.resi
sq.reg= 199.8493
```

Font de variació	SQ	g. de llibertat	SQM
Regressió	199.849	1	199.849
Residual	5.865	5	1.173
Total	205.714	6	34.286

Punts a tenir en compte en aquesta solució:

- **Una SUMA DE QUADRATS MAI pot ser negativa, per definició, ja que estem sumant valors als quadrats.**
- **Fixeu-vos que la variància residual és calcula com $s^2=\sum e^2/(n-2)$i en aquest cas val 1.173 (dada proporcionada a l'enunciat del problema); per tant la SQResidual=$\sum(y-\hat{y})^2= \sum e^2= s^2*(n-2)$=1.173*5=5.865**

c) Calculeu i interpreteu el coeficient de determinació R^2 (*1 punt*)

```
R2=sq.reg/sq.tot
R2= 199.849/205.714 = 0.9715
```

El valor del coeficient de determinació és molt elevat, això ens indica que el 97% de la variabilitat del Temps de CPU està explicada per una relació lineal amb el nombre d'E/S. Les operacions de disc incideixen molt en el temps de CPU

d) Calculeu el interval de confiança al 95% pel pendent de la recta i poseu a prova si aquest és significatiu. Quines premisses s'han de complir per a que tingui validesa aquesta estimació? (*2 punts*)

```
s2b1=var.resi/((n-1)*var(x))
s2b1= 0.0003487513
sb1=sqrt(s2b1)= 0.0187
t=b1/sb1
t= 13.05263
```

En aquest cas estem posant a prova la

$\mathbf{H_0 : \beta_1 = 0}$

$\mathbf{H_1 : \beta_1 \neq 0}$

Treballant amb una confiança del 95%, el punt crític que permet rebutjar o no la H_0 és,

```
p=0.975
t=qt(p,df)
t= 2.571
```

Com que 13.053 > 2.571, rebutgem H_0, per tant el pendent β_1 és significativament diferent de zero, és a dir, que la X està relacionada linealment amb la Y.

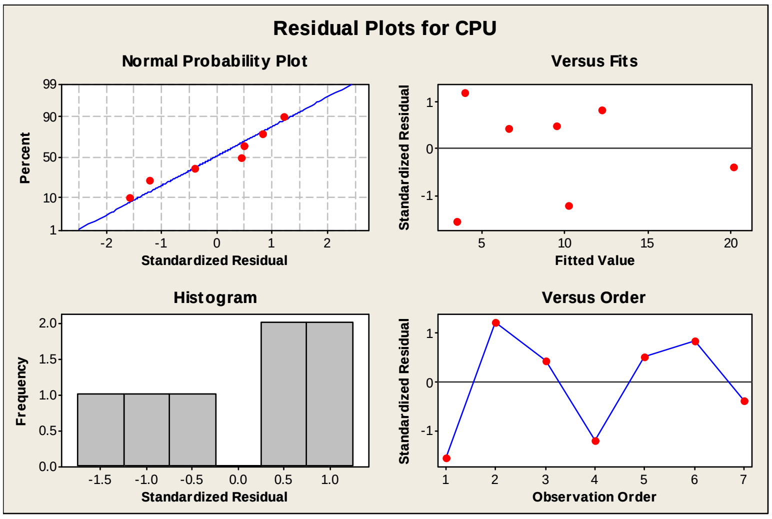
e) Doneu una previsió, així com el seu interval de confiança al 95%, del valor promig del temps de CPU quan el nombre d'E/S és 30 (*2 punts*)

```
xhat=30
yhat=b0+b1*xhat
yhat =  7.304
s2.yhat=var.resi*(1/n+(xhat-mean(x))^2/((n-1)*var(x))
s2.yhat = 0.026
s.yhat=sqrt(s2.yhat)
s.yhat = 0.163
df=n-2 = 5
p=0.975
t=qt(p,df)= 2.571
ic.sup=yhat+t*s.yhat
ic.inf=yhat-t*s.yhat
```

```
ic=c(ic.inf,ic.sup)
ic = (6.886, 7.723)
```

En aquest punt és molt IMPORTANT entendre que el que estem calculant és LA VARIANCIA DE LA PREVISIÓ i no la variància mostral com alguns estudiants han fet. Sempre que calculem una previsió, aquesta te associada una variabilitat, és a dir la variància d'aquesta previsió.

f) La següent figura ens mostra les gràfiques dels residus necessàries per poder validar aquest model de regressió. A partir de la inspecció gràfica d'aquestes, dieu quines premisses es verifiquen i si n'hi ha alguna que no. Justifiqueu la vostre resposta. (*1 punt*)



Finalment, en aquest gràfic, el Normal Probability Plot permet verificar si els residus segueixen una distribució normal i MAI S'HA D'INTERPRETAR COM QUE EXISTEIX RELACIÓ LINEAL ENTRE LES Y I LES X.

```
###### Comprovació i Representació gràfica
reg=lm(y~x)
plot(x,y)
abline(reg,col=2)
```