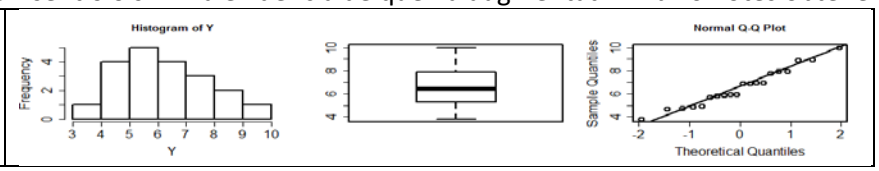
NOM:_____COGNOM:_____

(***Contesteu cada pregunta en el seu lloc. Expliciteu i justifiqueu els càlculs***)

Problema 1 (B4)

En l'estudi del seguiment d'una assignatura al llarg dels quadrimestres i després de canvis aplicats a l'assignatura, els professors volen comprovar si la nota mitjana esperada segueix sent 6 o si hi ha evidència de que ha augmentat. Amb 20 notes obtenen:

$$\sum_{i=1}^{20} Y_i = 131.41 \qquad \sum_{i=1}^{20} Y_i^2 = 915.30$$



(1 punt) **1.-** Primer calculen les estimacions puntuals de la nota mitjana i de la desviació tipus:

$\bar{y}$ = mean(Y) = 131.41/20 = **6.57**

s_Y = sd(Y) = sqrt(((915.3)-(131.41^2 /20))/19) = **1.65** (s_Y^2=2.73)

(3 punts) **2.-** Llavors posen a prova (amb un risc del 5%) si l'esperança poblacional de la nota és 6 o superior, suposant que la variabilitat poblacional s'ha mantingut i és coneguda (assumeixen desviació poblacional igual a 1.5). Indiqueu:

- (0.5) les hipòtesis, premisses, la fórmula de l'estadístic i dir quina distribució segueix sota la hipòtesis nul·la

H_0: $\mu_Y = 6$ Y~N(0,1) $(\bar{y} - 6) / (\sigma /\text{sqrt}(n)) \sim N(0,1)$

H_1: $\mu_Y > 6$ σ coneguda (σ=1.5)

- (0.5) càlcul de l'error tipus (o estàndard error) i del valor de l'estadístic

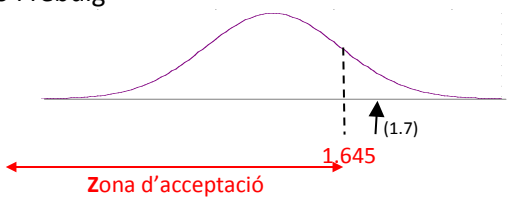
σ /sqrt(n) = **0.335**

$(\bar{y} - 6)/(\sigma /\text{sqrt}(n))$ = (6.57-6.0) / 0.335 = 0.57 / 0.335 = **1.7**

- (0.5) representació gràfica de l'estadístic, el/s punt/s crític/s i les zones d'acceptació i rebuig

Punt crític 1.645

(taules: P(Z<=1.64)=0.9495 i P(Z<=1.65)=0.9505)



- (0.5) càlcul del p-valor

1 – 0.9554 = 0.0446

(taules P(Z<=1.70)=0.9554)

- (1) segons els dos apartats anteriors, a quina conclusió arribeu sobre la prova d'hipòtesi. Interpreteu els resultats

- estadístic (1.7) està més enllà del punt crític (1.645). Està a la zona de rebuig

- p-valor (0.0446) és inferior al risc (0.05)

Per tant, hi ha evidència per no acceptar la hipòtesis de mitjana esperada 6, sinó que és superior

(2 punts) **3.-** Des de la Facultat també es fa la mateixa prova (si l'esperança poblacional de la nota és 6 o superior, amb un risc del 5%) usant les mateixes dades però sense suposar la variabilitat coneguda. Indiqueu:

- (0.5) les hipòtesis, premisses, la fórmula de l'estadístic i dir quina distribució segueix sota la hipòtesis nul·la

H_0: $\mu_Y = 6$ Y~N(0,1) $(\bar{y} - 6) / (s_Y /\text{sqrt}(n)) \sim t_{19}$

H_1: $\mu_Y > 6$

- (0.5) càlcul de l'error tipus (o estàndard error) i del valor de l'estadístic

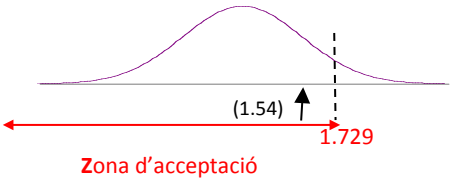
s_Y /sqrt(n) = **0.369**

$(\bar{y}$ - 6)/(s_Y /sqrt(n)) = (6.57-6.0) / 0.369 = 0.57 / 0.369 = **1.54**

- (0.5) representació gràfica de l'estadístic, el/s punt/s crític/s i les zones d'acceptació i rebuig

Punt crític 1.729

(taules: t_{19} acumula 0.95 en el punt 1.729)



- (0.5) en funció de l'apartat anterior, a quina conclusió arribeu sobre la prova d'hipòtesi. Interpreteu els resultats

Estadístic (1.54) és inferior al punt crític (1.729). Està a la zona d'acceptació.

Per tant, NO hi ha evidència per no acceptar la hipòtesis de mitjana esperada 6, res s'oposa a acceptar que l'esperança és 6 i no superior.

(1 punt) **4.-** Compareu els apartats **2** i **3**

A apartat 2 les dades mostren evidència per a no acceptar que l'esperança és 6, sinó que és superior

A apartat 3 les dades no mostren evidència per a rebutjar que l'esperança és 6, res s'oposa a acceptar que l'esperança és 6 i no superior.

És més realista i raonable no assumir valor de sigma; si s'estudia un possible efecte en la nota mitjana també el pot haver tingut en la desviació de les notes com per no assumir el valor "històric" (la desviació mostral és major a la assumida i per tant l'error tipus augmenta fent que res s'oposi a acceptar esperança 6)

(3 punts) **5.-** Finalment el que tenim és una variable dicotòmica (A aprovat, o S suspès) amb 6 suspesos i 14 aprovats. Poseu a prova si el valor esperat de la proporció d'aprovats és del 75% o no. Amb un risc del 5% indiqueu:

- (0.5) les hipòtesis i l'estimació puntual de la proporció d'aprovats

H_0: $\pi = 0.75$

H_1: $\pi \neq 0.75$ P = #A / 20 = 14/20 = 0.70

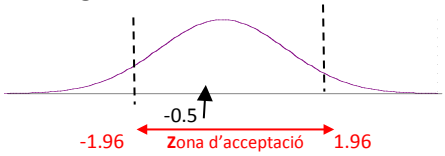
- (0.5) la fórmula i el càlcul de l'estadístic

(0.70 – 0.75) / sqrt(0.75*0.25/20) = -0.05 / 0.10 = - 0.5

- (0.5) representació gràfica de l'estadístic amb el/s punt/s crític/s i les zones d'acceptació i rebuig

Punts crítics -1.96 i 1.96

(taules: P(Z<=1.96)=0.9750 1.96= $z_{0.975}$)



- (0.5) càlcul d'un interval de confiança pel valor esperat de la proporció d'aprovats, i interpretació

$0.70 \pm z_{0.975}$ sqrt(0.7*0.3/20) = 0.70 ± 1.96 0.10 = 0.70 ± 0.196 = [0.504 , 0.896]

Amb un 95% de confiança la proporció esperada d'aprovats està entre 50.4% i 89.6%

- (1) segons els dos apartats anteriors, a quina conclusió arribeu sobre la prova d'hipòtesi. Interpreteu els resultats

- estadístic (-0.5) està entre els punts crítics (-1.96 i 1.96). Està a la zona d'acceptació

- el valor posat a prova (0.75) cau dins el IC [0.504,0.896]

Per tant, NO hi ha evidència per no acceptar la hipòtesis de proporció esperada del 75% d'aprovats, res s'oposa a acceptar que la proporció és del 75%.

NOM:_____COGNOM:_____

Problema 2 (B5)

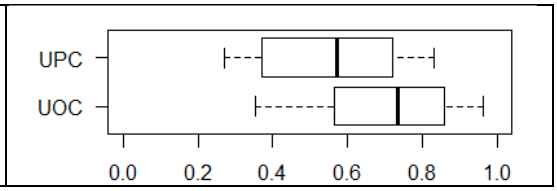
(Contesteu cada pregunta en el seu lloc. Expliciteu i justifiqueu els càlculs)

1. Les universitats UPC i UOC estan preocupades pel plagi dels treballs acadèmics dels alumnes. Ambdues universitats tenen algoritmes per determinar la probabilitat que un treball hagi estat plagiat en base a fer cerques de fragments dels documents a internet a través de diferents cercadors. La UPC sospita que l'algoritme de la UOC dóna millor resultats que el seu propi i decideix comparar les probabilitats que retornen ambdós algoritmes en treballs que es sap amb tota certesa que han estat plagiats. Per fer l'estudi, s'han recollit 10 treballs d'alumnes de la UPC i 21 treballs d'alumnes de la UOC, obtenint la següent descriptiva de resultats respecte a les probabilitats.

```
> summary(UPC)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2696  0.4015  0.5709  0.5548  0.6867  0.8311
> summary(UOC)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.3542  0.5654  0.7350  0.7110  0.8583  0.9614
```

```
> sd(UPC)
[1] 0.1888975

> sd(UOC)
[1] 0.1796934
```



a) Justifiqueu si es tracta de dades aparellades o independents (1 punt)

Independents. Provenen d'unitats mostrals diferents

b) Plantegeu una prova d'hipòtesis per decidir si la mitjana de les probabilitats de l'algoritme de la UOC és superior o no a la de l'algoritme de la UPC en els treballs plagiats. Digueu explícitament si és un contrast bilateral o unilateral. (1 punt)

$$\begin{cases} H_0: \mu_{UOC} = \mu_{UPC} \\ H_1: \mu_{UOC} > \mu_{UPC} \end{cases} \quad \text{(unilateral)}$$

c) Indiqueu l'expressió de l'estadístic, la seva distribució sota H_0 i les premisses necessàries (1 punt)

$t = \frac{(\bar{y}_{UOC} - \bar{y}_{UPC})}{s \cdot \sqrt{\frac{1}{n_{UOC}} + \frac{1}{n_{UPC}}}} \sim t_{29}$ sent $s^2 = \frac{(n_{UOC}-1)\cdot S^2_{UOC} + (n_{UPC}-1)\cdot S^2_{UPC}}{n_{UOC}+n_{UPC}-2}$ Premisses: MAS, $Y_{UOC}, Y_{UPC} \sim N$ i $\sigma_{UOC} = \sigma_{UPC}$

d) Només mirant la descriptiva de l'enunciat, comenteu si us semblen raonables les premisses (1 punt)

MAS. No es pot avaluar ja que depèn de la recollida de les dades. Tots els treballs i les combinacions de n treballs haurien de tenir la mateixa probabilitat de ser escollits de forma aleatòria entre la població de tots els treballs plagiats.
NORMALITAT. Els dos boxplots semblen força simètrics i la mitjana i la mediana en ambdues poblacions s'assemblen bastant. Per tant, aquesta premissa és raonable.
VARIÀNCIES IGUALS. Les desviacions tipus són força semblants. És raonable pensar que són iguals.

e) Calculeu el valor de l'estadístic i el/s punt/s crític/s suposant un α = 0.05 (1 punt)

$$s^2 = \frac{(21-1)\cdot 0.18^2 + (10-1)\cdot 0.19^2}{21+10-2} = 0.0333$$

$$t = \frac{0.71 - 0.55}{0.183 \cdot \sqrt{\frac{1}{21} + \frac{1}{10}}} = 2.23$$

$$Punt\ crític \rightarrow t_{0.95,29} = 1.699$$

f) Indiqueu, justifiqueu i interpreteu quina és la conclusió de la prova d'hipòtesis (1 punt)

Hi ha evidència per rebutjar H_0, ja que el valor de l'estadístic està més enllà del punt crític. Per tant, hi ha prou evidència per dir que la mitjana de les probabilitats de l'algoritme de la UOC és superior al de l'algoritme de la UPC en els treballs plagiats.

2. Evidentment, a les universitats els interessa un algoritme que classifiqui els treballs plagiats com a tal, però també que classifiqui els treballs no plagiats com a originals. Ambdues universitats determinen que hi ha plagi en un treball quan la probabilitat d'algun dels seus algoritmes respectius és superior a 0.5. A més dels 31 treballs previs, s'escullen dues mostres aleatòries de treballs addicionals per cada universitat dels quals es té la certesa que no han estat plagiats. Es recullen un total de 100 treballs per a cada universitat. Es vol saber si la proporció d'encert global de l'algoritme (en tots els treballs, tant en els plagiats com en els no plagiats) és la mateixa en les dues universitats o no. Els resultats estan a les taules següents:

UPC	Prob≤0.5	Prob>0.5	Total
NO Plagiats	76	14	90
Plagiats	6	4	10
Total	82	18	100

UOC	Prob≤0.5	Prob>0.5	Total
NO Plagiats	53	26	79
Plagiats	4	17	21
Total	57	43	100

a) Indiqueu les hipòtesis i si el contrast ha de ser unilateral o bilateral (1 punt)

$$\begin{cases} H_0: \pi_{UOC} = \pi_{UPC} \\ H_1: \pi_{UOC} \neq \pi_{UPC} \end{cases}$$ (bilateral)

b) Indiqueu l'expressió de l'estadístic, la seva distribució sota H_0 i les premisses assumides. (1 punt)

Opció 1

$\chi^2 = \sum \left(\frac{(f_{ij} - e_{ij})^2}{e_{ij}} \right) \sim \chi^2_1$ →Premisses: MAS, $e_{ij}>5$

Opció 2

$Z = \frac{(p_{UPC} - p_{UOC}) - (\pi_{UPC} - \pi_{UOC})}{\sqrt{\frac{p \cdot (1-p)}{n_{UPC}} + \frac{p \cdot (1-p)}{n_{UOC}}}} \sim N(0,1)$ →Premisses: MAS, n_{UPC}, n_{UOC} grans

c) Calculeu el valor de l'estadístic i el punt/s crític/s (1 punt)

Opció 1

Observats	UPC	UOC	Total
Encerts	80	70	150
Errades	20	30	50
Total	100	100	200

Esperats	UPC	UOC	Total
Encerts	75	75	150
Errades	25	25	50
Total	100	100	200

$\chi^2 = \frac{5^2}{75} + \frac{5^2}{75} + \frac{5^2}{25} + \frac{5^2}{25} = 2.67$ → Punt crític = $\chi^2_{1,0.95} = 3.84$

Opció 2

$p = \frac{100 \cdot 80/100 + 100 \cdot 70/100}{200} = 0.75 \rightarrow Z = \frac{0.8 - 0.7}{\sqrt{2 \cdot \frac{0.75 \cdot 0.25}{100}}} = 1.63$ →Punts crítics = +/- 1.96

d) Indiqueu, justifiqueu i interpreteu quina és la conclusió de la prova d'hipòtesis (1 punt)

No hi ha evidència per rebutjar H_0, ja que el valor de l'estadístic en valor absolut queda entre els punts crítics. Per tant, NO hi ha prou evidència per dir que la proporció d'encerts poblacional de l'algoritme de la UPC i de la UOC siguin diferents.

NOM:_____COGNOM:_____

(*Contesteu cada pregunta en el seu lloc. Expliciteu i justifiqueu els càlculs.*)

Problema 3 (B6)

Es vol estudiar la relació lineal entre els resultats de dos exàmens per aprovar unes oposicions. S'han recollit les notes del primer examen (X) i del segon examen (Y) de 50 aspirants a la plaça, i obtenim els següents resultats:

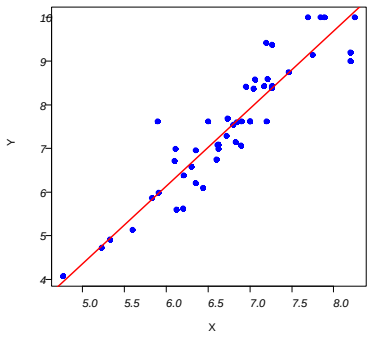
$\sum x_i = 337.49$

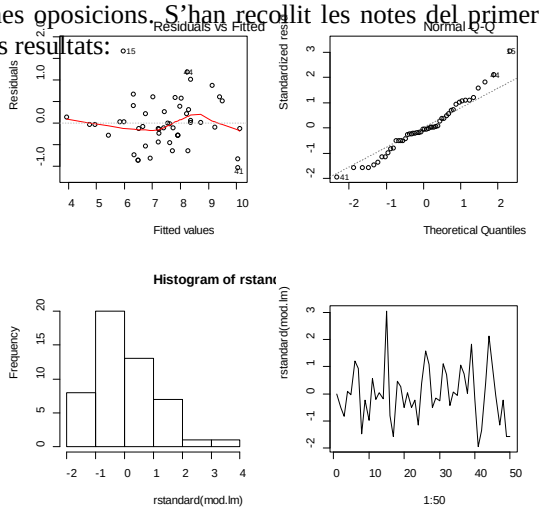
$\sum y_i = 373.12$

$\sum x_i^2 = 2306.08$

$\sum y_i^2 = 2888.158$

$\sum x_i\, y_i = 2568.386$






```
Coefficients:
                  Estimate     Std. Error   t value      Pr(>|t|)
(Intercept)       ---          0.7197       ---          ---
X                 ---          0.1060       16.764       < 2e-16
Residual standard error: 0.5616 on 48 degrees of freedom
Multiple R-squared:  ---,    Adjusted R-squared:  ---
F-statistic:   281 on 1 and 48 DF,  p-value: < 2.2e-16
```

1.- Calculeu la covariància i la correlació entre X i Y (2 punts).

Mean(x)=337.49/50= 6.7498 s_x^2 = (2306.08 - 50*(6.7498)^2) / 49 = 0.5732591 s_x =0.7571388

Mean(y)= 373.12/50=7.4624 s_y^2 = (2888.158 - 50*(7.4624)^2) / 49 =2.118108 s_y=1.455372

S_{XY} = (2568.386-(337.49*373.12)/50)/49= 1.01838

r = 1.01838/(0.7571388*1.455372)= 0.924188

2.- Calculeu la recta de regressió tot estimant els valors de β_0 i β_1. Com interpreteu el valor de b_1? (2 punts).

b_1= 1.01838 / 0.5732591 = 1.776474

(ó 16.764*0.1060 = 1.776984)

b_0= 7.4624 - (1.776474)*6.7498 = -4.528444

Y = -4.5284 + 1.7765 X

La pendent b_1 = 1.7765 és positiva, indicant una relació positiva entre les dues notes. Per cada punt de més en la prova 1, esperem 1.78 punts més en la segona prova.

3.- Poseu a prova si la recta de regressió passa per l'origen de coordenades amb un risc $\alpha = 0.05$ (1 punt).

H_0: β_0=0 (H_1: β_0<>0)

t = (b_0-0) / Sb_0 = (-4.5284 -0)/ 0.7197 = -6.292

amb risc del 5% l'estadístic (-6.292) està més enllà dels punts crítics (-2.021 i 2.021 ja que $t_{40,0.975}$=2.021). Per tant, hi ha evidència per rebutjar que β_0 sigui 0. És a dir que la recta no passi per l'origen de coordenades (0,0).

4.- Calculeu un interval del 95% de confiança pel paràmetre β_1 i interpreteu els resultats (1 punt).

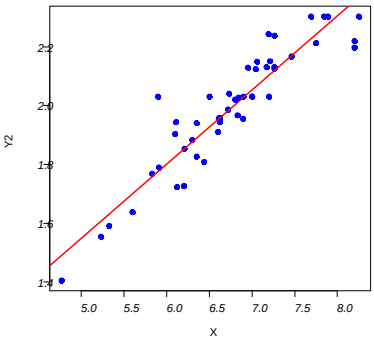
$b_1 \pm t_{40,0.975}\, Sb_1$ = 1.7765 +- 2.021 * 0.1060 = [1.562274, 1.990726]

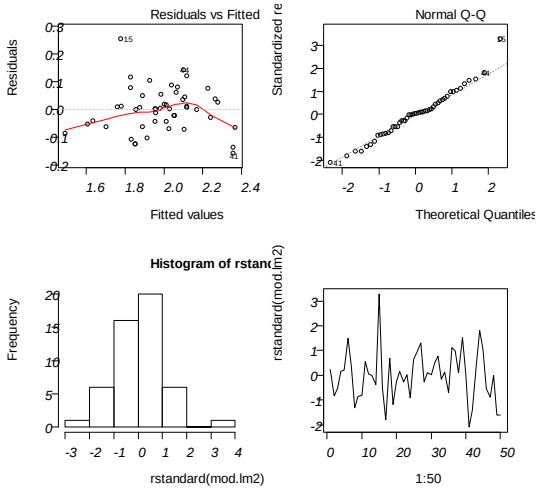
IC de β_1 : amb una confiança del 95% es pot afirmar que β_1 està entre 1.562274 i 1.990726. Cada unitat que augmenta X implica que Y augmenta entre 1.562274 i 1.990726.

5.- Interpreteu els valors del coeficient de determinació i de correlació (1 punt).

r = 0.9241753 relació lineal positiva i força intensa (proper a 1).
R^2 = 0.8541 bona part (> 85%) de la variabilitat de Y queda explicada per X.

Hem aplicat una transformació logarítmica als resultats del segon examen (Y) obtenint unes noves dades (Y2) pels 50 aspirants i n'estudiem la relació lineal amb X:






```
Call:
lm(formula = Y2 ~ X, data = datos)

Residuals:
      Min        1Q    Median        3Q       Max
-0.157903 -0.050662  0.002815  0.043129  0.254894

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2900     0.1019   2.847  0.00648 **
X             0.2519     0.0150  16.790  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0795 on 48 degrees of freedom
Multiple R-squared:  0.8545,    Adjusted R-squared:  0.8515
F-statistic: 281.9 on 1 and 48 DF,  p-value: < 2.2e-16
```

6.- Indiqueu el nou model, calculeu una predicció puntual per la nota (Y) pel cas de X=7 i calculeu un interval de confiança per valors individuals per la predicció (2 punts).

Y2 = 0.2519 * X + 0.29
Y2 = 0.2519 * 7 + 0.29 = 2.053
Per tant: Y = exp(2.053) = 7.79124

S= 0.0795

2.053 $\pm$ $t_{40,0.975}$ * 0.0795 * sqrt (1 + 1/50 + ((7 - 6.7498)^2) / (0.5732591*49)) = 2.053 +- 2.021 * 1.011053 = [1.890482, 2.215356].
Per tant, IC per la nota: exp(2.053) $\pm$ [exp(1.890482), exp(2.215356)] => 7.79124 $\pm$ [6.62256, 9.164671].

$$\hat{y}_h \pm t_{n-2,0.975} S \sqrt{1 + 1/n + \frac{(X_h - \bar{X})^2}{\sum (X_i - \bar{X})^2}}$$

7.- Compareu els dos models en quant al coeficient de determinació i les premisses. Creieu que ha valgut la pena aplicar el logaritme a la Y en aquest sentit? (1 punt).

Amb ambdós models, observant els gràfics no apreciem diferències en quant a les premisses (linealitat entre variables, homocedasticitat, independència i normalitat dels residus). A més a més, donat que el coeficient de determinació no millora, no hi ha motius per utilitzar el model aplicant el logaritme a la Y.