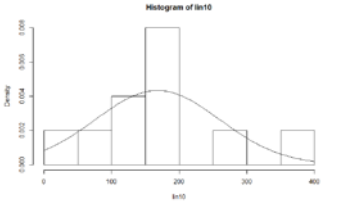
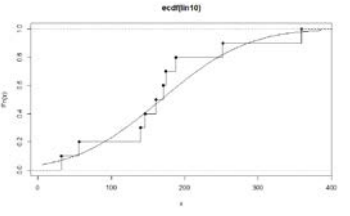
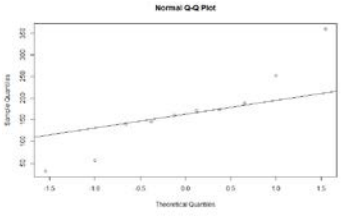


## Problema 1 (B4)

Un grup d'estudiants de la FIB està estudiant els temps de cerca d'un nombre en un vector de  $10^5$  elements de l'algorisme de cerca lineal. Ha generat un únic vector amb  $10^5$  naturals de l'1 al 100000. Escullen deu vegades a l'atzar l'element a buscar entre 1 i  $10^5$ ; i obtenen els resultats següents expressats en ns.

$\sum_{i=1}^{10} x_i = 1679$ $\sum_{i=1}^{10} x_i^2 = 358243$			
Histograma (funció densitat empírica) vs funció densitat teòrica segons model Normal		Funció distribució empírica vs F. distr. teòrica segons model Normal	QQPlot (quantils empírics vs Qs. teòrics segons model Normal)

1. Calculeu les estimacions puntuals del temps mitjà i de la desviació tipus (1 punt)

$$\bar{x} = 167.9 \text{ ns}$$

$$s_x = 92.09832 \text{ ns}$$

2. Calculeu l'error tipus de la mitjana. Interpreteu la desviació i l'error tipus i digueu quina utilitat poden tenir en aquestes dades (p.e., per a què utilitzaríeu cadascú?) (1 punt)

$$s_{\bar{x}} = \frac{s_x}{\sqrt{n}} = \frac{92.09832}{\sqrt{10}} = 29.12405 \text{ ns}$$

$s_x$  mesura la dispersió de les dades (mesura descriptiva). I  $s_{\bar{x}}$  estima l'error esperat d'estimació de  $\mu$ . Utilitzo  $s_x$  per descriure com són els casos, però  $s_{\bar{x}}$  dona el error esperat de  $\bar{x}$  com a estimació de  $\mu$  (error de l'estimació): La  $\bar{x} = 167.9$  està dotada d'un error aleatori de 29.1.

3. A partir dels gràfics argumenteu si podem suposar que el temps segueix una distribució normal. (1 punt)

Segons l'histograma, podria ser, però és un gràfic molt pobre, ja que depèn molt de l'amplada dels intervals. Els 2 gràfics basats en la distribució acumulada són més fins i apunten a una distribució amb cues més pesades, més compatible amb una uniforme que amb una Normal.

Hi ha poques dades i és difícil treure conclusions clares. Convindria estudiar-ho més, amb més dades: Potser estudiar l'ajustament a una uniforme o altres distribucions...

4. Assumint que les dades segueixen la distribució normal (independentment de la resposta a les preguntes anteriors), calculeu un interval bilateral amb 90% de confiança per la mitjana de temps. Interpreteu. (2 punts)

Tenim que  $X \sim \text{Normal}$  i que  $n=10$  i  $\alpha = 0.05$ , per tant,  $t_{n-1, 1-\frac{\alpha}{2}} = t_{9, 0.95} = 1.833$

$$(\bar{x} - t_{n-1, 1-\frac{\alpha}{2}} \cdot \sqrt{\frac{s^2}{n}}, \bar{x} + t_{n-1, 1-\frac{\alpha}{2}} \cdot \sqrt{\frac{s^2}{n}}) = (167.9 \mp 1.833 \cdot 29.1) = (114.52, 221.28)$$

Amb confiança 90%,  $\mu$  pertany a aquest interval

5. Se suposa que la mitjana poblacional del temps de cerca és 200 ns o superior, i els estudiants volen estudiar amb una prova d'hipòtesi al 5% si el temps de cerca ha millorat. Indiqueu:

a) Les hipòtesis, les premisses, la fórmula de l'estadístic i quina és la distribució d'aquest sota la hipòtesi nul·la (1 punt)

$$\begin{cases} H_0: \mu_X \geq 200 \\ H_1: \mu_X < 200 \end{cases}$$

$$\text{O bé: } \begin{cases} H_0: \mu_X = 200 \\ H_1: \mu_X < 200 \end{cases}$$

La premissa és  $X \sim \text{Normal}$

$$\text{L'estadístic és } \hat{t} = \frac{\bar{x} - \mu_0}{\sqrt{\frac{s^2}{n}}} \text{ i } \hat{t} \sim t_9$$

- b) Calculeu el valor de l'estadístic (1 punt)

$$\hat{t} = \frac{\bar{x} - \mu_0}{\sqrt{\frac{s^2}{n}}} = \frac{167.9 - 200}{\sqrt{\frac{29,124}{10}}} = \frac{-32.1}{29,124} = -1.1$$

- c) Representeu gràficament el(s) punt(s) crític(s), les zones d'acceptació i de rebuig i el valor de l'estadístic (1 punt)

El punt crític [-- taules fila 9, columna 0.95] és -1.833 [qt(0.05,9)=-1.833113]

La zona d'acceptació és  $[-1.833, \infty]$

La zona de rebuig és  $(-\infty, -1.833)$

Gràficament.

- d) A partir de l'estudi i dels càlculs realitzats, interpreteu els resultats de la prova d'hipòtesi (1 punt)

El valor de l'estadístic -1.1 pertany a la zona d'acceptació, per tant podem concloure que no tenim evidències significatives (al 5%) per poder garantir que el sistema no es col·lapsarà. Com manca d'evidència no és prova de res, és raonable preguntar-se si això podria ser degut a la reduïda mostra.

6. A l'apartat 4 hem fet un IC bilateral al 10%; i al 5, una PH unilateral al 5%. Té sentit? Compareu els resultats (1 punt)

Deixar un 10% en tots 2 costats equival (assumint costats iguals) a deixar un 5% amb un costat. Té sentit.

El valor 200 ns pertany a l'interval amb 90% de confiança de la mitjana poblacional (apartat 5) de 114.52 a 221.28.

Això és coherent amb el fet de no haver pogut rebutjar en l'apartat 5 que la mitjana poblacional fos 200 (o superior).

**Problema 2 (B5).** Es vol comparar els diners que es gasta el pare Noel i els reis mags d'orient per comprar regals. En una mostra aleatòria de 30 famílies, per a cadascuna d'elles, es recull el cost econòmic dels regals portats pel pare Noel i els regals portats pels reis. Siguin C1 i C2 els diners gastats pel primer i pels segons i D la diferència entre ells ( $D=C2-C1$ ). Aquesta és la descriptiva d'aquestes variables:

*Nota: les unitats són euros de despesa per persona*

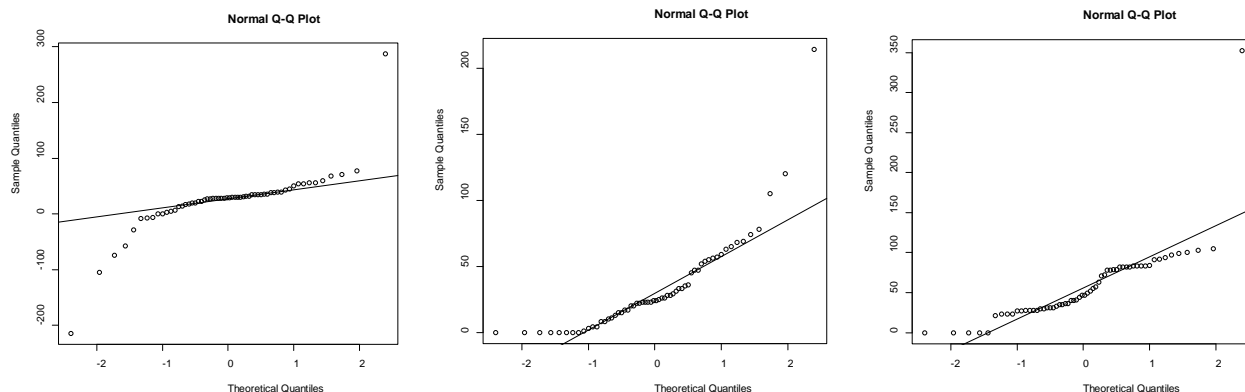
```
> summary(x)
  C1PareNoel      C2ReisMags      Diferencia
Min.   : 0.00   Min.   : 0.00   Min.   : -214.00
1st Qu.: 10.75  1st Qu.: 29.50  1st Qu.: 16.25
Median : 24.00  Median : 47.00  Median : 29.00
Mean   : 33.47  Mean   : 57.55  Mean   : 24.03
3rd Qu.: 48.25  3rd Qu.: 82.00  3rd Qu.: 38.00
Max.   :214.00  Max.   :352.00  Max.   : 287.00

> sd(x$C1PareNoel)
[1] 35.60207
> sd(x$C2ReisMags)
[1] 49.06081
> sd(x$Diferencia)
[1] 55.62525
```

**1) (1 punt).** Es tracta de dues mostres independents o aparellades? Raoneu la resposta.  
**Aparellades.** Cada família es susceptible de rebre presents portats pel pare Noel i pels reis mags. Podem pensar que hi ha una correlació entre les observacions de C1 i C2 que permeti fer una comparació "dins" de cada unitat.

**2) (2 punts).** Fixeu-vos en els següents *qqnorms*.

QQ norms per `x$Diferencia`, `x$C1PareNoel`, `x$C2ReisMags`



- a) La funció `qqnorm` o `qqplot` grafica las funcions quantil d'una mostra versus una distribució normal teòrica. Expliqueu el gràfic i què permet observar. (0.67 punts)
- b) Comenteu si es compleix la premissa de Normalitat en tots ells. N'hi ha algun d'ells en el que la normalitat sembli més creïble? (0.67 punts)
- c) En funció de la resposta de la pregunta 1), Quins són el gràfics rellevants per l'anàlisi que s'ha de realitzar?(0.67 punts) Raoneu les respostes.

a) La idea d'aquest tipus de gràfics és que, si les dos distribucions coincideixen, veurem un gràfic molt semblant a una recta. Per tant, permet observar com de prop està la distribució d'unes dades d'una distribució normal teòrica.

b) Els dos últims no compleixen la premissa de normalitat. Podríem acceptar-la amb una mica de reticència a la variable diferència (D), els punts sobre la recta indiquen similitud entre els quantils observats i els teòrics en cas d'una Normal. Hi ha alguns valors extrems a les cues que no segueixen del tot la línia.

c) Al ser mostres aparellades, ens hem de fixar en el *qqnorm* de la diferència (D)

**3) (4 punts).** Per saber si es pot suposar que hi ha diferències entre els diners que gasten el pare Noel i els reis, es vol plantejar una prova d'hipòtesi d'igualtat de mitjanes:  $H_0: \mu_{C1} = \mu_{C2}$  vs.  $H_1: \mu_{C1} \neq \mu_{C2}$

- a) Calcula l'estadístic per fer la comparació, digues quina distribució segueix sota la hipòtesi nul·la i amb quines premisses.

$$t = \frac{\bar{D}}{S \sqrt{1/n_D}} = \frac{24.03}{55.62525 \sqrt{1/30}} = 2.366$$

Segueix una t-Student amb 29 graus de llibertat sota les premisses de que  $D \sim N$  i m.a. aparellada

b) Digues quin és(són) el(s) punt(s) crític(s) amb un 5% de significació i treu conclusions sobre la prova d'hipòtesi.

$t = 2.366 > t_{29, 0.975} = 2.0452$  (punt crític) → Es rebutja la hipòtesi nul·la de que les mitjanes siguin iguals. Creiem que la mitjana de despesa feta amb els reis mags és major (o diferent) que la mitjana feta amb el pare Noel.

c) Calcula l'interval de confiança del 99% per a la diferència de mitjanes i interpreta'l.

$$IC(\mu_D, 99\%) = \bar{D} \pm t_{29, 0.995} S \sqrt{\frac{1}{n_D}} = 24.03 \pm 2.7564 * 55.62525 \sqrt{\frac{1}{30}} \\ = [-3.96; 52.02]$$

Nota: A les taules podeu trobar el valor per a  $t_{29, 0.995} = 2.7564$

d) Interpreta que vol dir l'interval de confiança del 99% (si no has trobat els valors numèrics, fes servir x i y per a resoldre aquest apartat). Si l'interval de confiança es calcula amb un nivell del 90%. Com creieu que serà aquest: més ample o més estret que el calculat amb el 99% de confiança? Raoneu la resposta.

Amb un 99% de confiança, la diferència de mitjanes es troba entre -3.96 i 52.02 (que donat que es despesa en euros podem prendre l'interval de -3.96 a 52.02 €), sent –segurament– el cost dels regals dels reis mags més gran que els del pare Noel. Però aquest grau de confiança tan ampli no exclou del tot la possibilitat de que sigui a la inversa. L'interval del 90% serà més estret que el del 99% de confiança.

**4) (3 punts).** Ara interessa conèixer si la proporció de famílies que rep regals amb un cost superior a 20 euros per persona és el mateix quan els regals són del pare Noel o són dels reis mags.

a) Es decideix emprar dues mostres independents de grandària 100 (una en cada moment de temps). Les proporcions  $p_1$  i  $p_2$  de despesa superior als 20 euros per persona són 0.65 i 0.90 del pare Noel i dels reis respectivament. Calcula l'estadístic per dur a terme la comparació i digues quina distribució segueix sota la hipòtesi nul·la i sota quines premisses.

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{p_1 + p_2}{2} = 0.775 \\ z = \frac{p_2 - p_1}{\sqrt{p \cdot (1-p)/n_1 + p \cdot (1-p)/n_2}} = \frac{0.25}{\sqrt{2 \cdot \frac{0.775 \cdot 0.225}{100}}} = \frac{0.25}{0.05905} = 4.233$$

Segueix una Normal estàndard sota les premisses de  $n_1$  i  $n_2$  grans i m.a.s. independents.

b) Digues quin és el punt crític amb un 5% de significació i treu conclusions sobre la prova d'hipòtesi.

El punt crític és  $z_{0.975} = 1.96$ . Com que  $z > z_{0.975} = 1.96$ , hi ha prou evidència per dir que les probabilitats de rebre un regal amb un cost superior als 20 euros són diferents si el regal ve del pare Noel o dels reis mags.

c) Calcula l'interval de confiança del 95% per la diferència de proporcions  $\pi_1$  i  $\pi_2$

$$IC(\pi_2 - \pi_1, 95\%) = (p_2 - p_1) \pm z_{0.975} \sqrt{p \cdot (1-p)/n_1 + p \cdot (1-p)/n_2} = 0.25 \pm 1.96 \cdot \sqrt{2 \cdot \frac{0.775 \cdot 0.225}{100}} \\ = [0.134; 0.366]$$

Amb un 95% de confiança, la diferència de proporcions es troba entre 0.134 i 0.366

### Problema 3 (B6)

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

El temps que navega un client dins d'una plana web de venda online de productes d'informàtica es creu que pot estar relacionat amb la seva despesa a l'hora d'adquirir productes. S'han recollit dades de clients que han realitzat alguna compra. A continuació, es pot trobar la descriptiva de les variables Temps (minuts), Despesa (euros) i LogDespesa (logaritme natural de la despesa). A sota, hi ha part de la sortida de R dels 2 models lineals que ajusten la despesa (transformada o no) segons el temps de navegació.

Descriptiva																															
<pre>&gt; summary(dades)</pre> <table border="1"> <thead> <tr> <th>Temps</th> <th>Despesa</th> <th>LogDespesa</th> </tr> </thead> <tbody> <tr> <td>Min. : 0.010</td> <td>Min. : 7.89</td> <td>Min. : 2.066</td> </tr> <tr> <td>1st Qu.: 2.535</td> <td>1st Qu.: 38.48</td> <td>1st Qu.: 3.650</td> </tr> <tr> <td>Median : 5.075</td> <td>Median : 67.39</td> <td>Median : 4.209</td> </tr> <tr> <td>Mean : 5.094</td> <td>Mean : 95.12</td> <td>Mean : 4.205</td> </tr> <tr> <td>3rd Qu.: 7.580</td> <td>3rd Qu.: 118.06</td> <td>3rd Qu.: 4.771</td> </tr> <tr> <td>Max. : 9.900</td> <td>Max. : 397.89</td> <td>Max. : 5.986</td> </tr> </tbody> </table>	Temps	Despesa	LogDespesa	Min. : 0.010	Min. : 7.89	Min. : 2.066	1st Qu.: 2.535	1st Qu.: 38.48	1st Qu.: 3.650	Median : 5.075	Median : 67.39	Median : 4.209	Mean : 5.094	Mean : 95.12	Mean : 4.205	3rd Qu.: 7.580	3rd Qu.: 118.06	3rd Qu.: 4.771	Max. : 9.900	Max. : 397.89	Max. : 5.986	<pre>&gt; sd(dades\$Temps) ; sd(dades\$Despesa) ; sd(dades\$LogDespesa)</pre> <pre>[1] 3.007195</pre> <pre>[1] 86.4399</pre> <pre>[1] 0.8550065</pre> <pre>&gt; cov(dades\$Temps, dades\$Despesa) ; cov(dades\$Temps, dades\$LogDespesa)</pre> <pre>[1] 136.9266</pre> <pre>[1] 1.574313</pre>									
Temps	Despesa	LogDespesa																													
Min. : 0.010	Min. : 7.89	Min. : 2.066																													
1st Qu.: 2.535	1st Qu.: 38.48	1st Qu.: 3.650																													
Median : 5.075	Median : 67.39	Median : 4.209																													
Mean : 5.094	Mean : 95.12	Mean : 4.205																													
3rd Qu.: 7.580	3rd Qu.: 118.06	3rd Qu.: 4.771																													
Max. : 9.900	Max. : 397.89	Max. : 5.986																													
Model 1	Model 2																														
<pre>&gt; mod1 &lt;- lm(Despesa~Temps, dades) ; summary(mod1)</pre> <p>Coefficients:</p> <table border="1"> <thead> <tr> <th></th> <th>Estimate</th> <th>Std. Error</th> <th>t value</th> <th>Pr(&gt; t )</th> </tr> </thead> <tbody> <tr> <td>(Intercept)</td> <td>(1)</td> <td>(2)</td> <td>1.234</td> <td>0.22</td> </tr> <tr> <td>Temps</td> <td>(3)</td> <td>(4)</td> <td>(5)</td> <td>1.81e-08 ***</td> </tr> </tbody> </table> <p>Residual standard error: (6) on 98 degrees of freedom            Multiple R-squared: 0.2775, Adjusted R-squared: 0.2701            F-statistic: 37.64 on 1 and 98 DF, p-value: 1.806e-08</p> <pre>&gt; par(mfrow=c(2, 1)) ; plot(Despesa~Temps, dades) ; &gt; plot(mod1, 1)</pre>		Estimate	Std. Error	t value	Pr(> t )	(Intercept)	(1)	(2)	1.234	0.22	Temps	(3)	(4)	(5)	1.81e-08 ***	<pre>&gt; mod2 &lt;- lm(LogDespesa~Temps, dades) ; summary(mod2)</pre> <p>Coefficients:</p> <table border="1"> <thead> <tr> <th></th> <th>Estimate</th> <th>Std. Error</th> <th>t value</th> <th>Pr(&gt; t )</th> </tr> </thead> <tbody> <tr> <td>(Intercept)</td> <td>(1)</td> <td>(2)</td> <td>24.737</td> <td>&lt; 2e-16 ***</td> </tr> <tr> <td>Temps</td> <td>(3)</td> <td>(4)</td> <td>(5)</td> <td>1.3e-11 ***</td> </tr> </tbody> </table> <p>Residual standard error: (6) on 98 degrees of freedom            Multiple R-squared: 0.3749, Adjusted R-squared: 0.3685            F-statistic: 58.78 on 1 and 98 DF, p-value: 1.296e-11</p> <pre>&gt; par(mfrow=c(2, 1)) ; plot(LogDespesa~Temps, dades) ; &gt; plot(mod2, 1)</pre>		Estimate	Std. Error	t value	Pr(> t )	(Intercept)	(1)	(2)	24.737	< 2e-16 ***	Temps	(3)	(4)	(5)	1.3e-11 ***
	Estimate	Std. Error	t value	Pr(> t )																											
(Intercept)	(1)	(2)	1.234	0.22																											
Temps	(3)	(4)	(5)	1.81e-08 ***																											
	Estimate	Std. Error	t value	Pr(> t )																											
(Intercept)	(1)	(2)	24.737	< 2e-16 ***																											
Temps	(3)	(4)	(5)	1.3e-11 ***																											

a) Amb els darrers gràfics que apareixen al final de cada model, quines premisses es poden avaluar? Quin dels dos models s'ajusta millor a les premisses del model lineal? (1 punt)

Les premisses que es poden avaluar són LINEALITAT i HOMOSCEDASTICITAT. El segon model compleix ambdues premisses, mentre que el primer NO compleix cap de les 2.

b) En el model escollit en l'apartat anterior i amb les dades proporcionades a l'enunciat, calcula els valors que falten a la sortida dins dels quadres 1 a 6 [Pista: el nombre d'observacions es dedueixen a partir dels graus de llibertat del model] (3 punts)

Com que els graus de llibertat són 98, sabem que la "n" és 100.

<pre>&gt; b1 &lt;- cov(dades\$Temps, dades\$LogDespesa) / var(dades\$Temps)</pre>	# (3)
<pre>&gt; b0 &lt;- mean(dades\$LogDespesa) - b1*mean(dades\$Temps)</pre>	# (1)
<pre>&gt; sb0 &lt;- b0/24.737</pre>	# (2)
<pre>&gt; s2 &lt;- (100 - 1) * (sd(dades\$LogDespesa) ^2 - b1*cov(dades\$Temps, dades\$LogDespesa)) / (100 - 2)</pre>	# (6)
<pre>&gt; s &lt;- sqrt(s2)</pre>	# (4)
<pre>&gt; sb1 &lt;- sqrt(s2/((100 - 1)*var(dades\$Temps)))</pre>	# (5)
<pre>&gt; t1 &lt;- b1/sb1</pre>	# (3)

```
> cat(' (1): ', b0, ' (2): ', sb0, ' (3): ', b1, ' (4): ', sb1, ' (5): ', t1, ' (6): ', s, '\n')
(1): 3.318486
(2): 0.1341507
(3): 0.1740877
(4): 0.02270744
(5): 7.66655
(6): 0.6794339
```

En els següents apartats, si no has pogut resoldre l'apartat anterior emprant els valors de la següent taula

(1)	(2)	(3)	(4)	(5)	(6)
3	0.2	0.2	0.02	8	0.7

c) Interpreta que suposa un increment d'un minut més navegant respecte a la despesa en compres (2 punts)

$x_0 = \text{temps en minuts qualsevol dins del rang d'estudi}$

$x_1 = x_0 + 1$

$$\left. \begin{aligned} \hat{y}_0 &= e^{b_0 + b_1 \cdot x_0} \\ \hat{y}_1 &= e^{b_0 + b_1 \cdot x_1} = e^{b_0 + b_1 \cdot (x_0 + 1)} = e^{b_0 + b_1 \cdot x_0} \cdot e^{b_1} \end{aligned} \right\} \rightarrow \frac{\hat{y}_1}{\hat{y}_0} = e^{b_1}$$

Amb dades de b)

$$e^{b_1} = e^{0.1740} = 1.19$$

Amb dades de taula

$$e^{b_1} = e^{0.2} = 1.22$$

Un minut més navegant suposa un increment estimat en la despesa del **19%** (o del **22%** si s'empran dades de la taula).

d) Calcula un interval de confiança del **90%** per la pendent del model escollit (2 punts)

Amb dades de b)

```
> (IC <- b1 + c(-1, 1) * 1.66 * sb1)
```

```
[1] 0.1363934 0.2117820
```

Amb dades de taula

```
> (IC <- 0.2 + c(-1, 1) * 1.66 * 0.02)
```

```
[1] 0.1668 0.2332
```

El quantil 0.95 de la t de Student amb 98 graus de llibertat (1.66) s'obté interpolant els valors de les t de Student amb 60 i 120 graus de llibertat, respectivament. El IC90% pel pendent de la recta va de **0.14 a 0.21** (o de 0.16 a 0.23 amb dades de taula).

e) Fes la predicció puntual i per interval de confiança del 95% per un usuari que navega 1 minut. (2 punts)

Amb dades de b)

```
> mi nuts <- 1
```

```
> (pr <- b0 + b1*mi nuts)
```

```
[1] 3.492574
```

```
> (IC <- pr + c(-1, 1) * 1.99 * s * sqrt(1 + 1/100 + (mi nuts- 5.094)^2/(99*var(dades$Temps))))
```

```
[1] 2.121221 4.863926
```

```
> exp(c(pr, IC))
```

```
[1] 32.870432 8.341314 129.531791
```

Amb dades de taula

```
> mi nuts <- 1
```

```
> (pr <- 3 + 0.2*mi nuts)
```

```
[1] 3.2
```

```
> (IC <- pr + c(-1, 1) * 1.99 * 0.7 * sqrt(1 + 1/100 + (mi nuts- 5.094)^2/(99*var(dades$Temps))))
```

```
[1] 1.787137 4.612863
```

```
> exp(c(pr, IC))
```

```
[1] 24.53253 5.97233 100.77223
```

Es preveu que es gastin **32.8** euros (24.53 amb dades de taula) amb un interval de confiança de **8.3 a 129.5** euros (de 6.0 a 100.8 amb dades de taula)