GRAUS DE GEI A CATALUNYA

Angella Clérigo, Albert Comas, Noelia López, Álvaro Terrón

Probabilitat i Estadística

Grup 12.6

DESEMBRE 2022

ÍNDEX

RESUM

Objectiu: La comparació de les notes de tall de diferents universitats en el Grau d'Enginyeria Informàtica a Catalunya segons diferents paràmetres com ara la comunitat autònoma, els diferents anys o bé les assignacions.

Mètodes: Per tal de poder realitzar aquest treball hem fet una àmplia búsqueda de les diferents notes de tall en el Grau d'Enginyeria Informàtica. Tenim per comparar 8 universitats públiques diferents de Catalunya, i de cada universitat tenim les notes de tall des del curs 2017-2018 fins al 2021-2022. Per cada any vam posar la nota de la primera assignació i de la segona. En total hem aconseguit reunir 40 notes de tall.

Originalment, voliem treballar tant amb universitats públiques com privades, però degut a que hi ha poques privades i molt poca informació que es pugui extreure d'elles vam decidir que ens perjudicaria en el treball. També voliem afegir la variable "Taxa d'abandonament", però les dades que vam trobar estaven separades per gènere. Com no sabíem ni el total d'alumnes inscrits, ni quants dels alumnes eren dones/homes, vam decidir treure aquesta variable.

Resultats i conclusions:

- Pregunta 1: L'interval de confiança de la diferència en mitjana entre la 1ª i 2ª assignació és de **[0.1834 , 0.3825]**, per tant rebutgem l'hipòtesis de que les notes són iguals.
- Pregunta 2: L'interval de confiança de la diferència en mitjana entre la nota de la 1ª assignació de fa 1 any i la nota de fa 5 anys és de **[1.998,2.847]**, per tant afirmem l'hipòtesis de que les notes són diferents.
- Pregunta 3: Fent la comprovació de linealitat entre les notes de la 1ª assignació i las de la 2ª, podem concloure en que aquestes dues variables segueixen un model lineal amb equació **N2a = -0.20486 + 0.99022*N1a.**
- Pregunta 4: L'interval de confiança de la diferència en mitjana entre les notes de la 1ª assignació de les universitats de Barcelona i les notes de las de fora és de **[*1.654977, 3.139743].*** Per tant, sí que són majors las de Barcelona.

Amb les dades que hem obtingut hem pogut fer un bon estudi dels diferents graus de Catalunya i respondre a cada pregunta que vam formular. Cal aclarir que aquest és un **estudi observacional**, és a dir, estar més orientat a fer prediccions. Hem relacionat una resposta amb altres valors observats (per exemple, hem comparat notes en funció de provincia).

INTRODUCCIÓ

Les notes de tall varien segons l'any en que fas la Selectivitat i la universitat on vulguis estudiar. En l'àmbit de les enginyeries aquestes variacions s'han fet més clares degut a l'esclat que hi ha últimament en la tecnologia de la informació. Aquest esdeveniment fa que sigui susceptible fer un anàlisi estadístic d'aquestes notes, en concret, de les notes en el Grau en Enginyeria Informàtica de les diferents universitats catalanes.

OBJECTIU

Com hem viscut aquesta variació de la nota de tall en aquest grau, ens ha impulsat a realitzar un estudi sobre les possibles variacions i diferències de les notes de les diferents universitats de Catalunya que cursen el Grau d'Enginyeria Informàtica, al llarg del temps o bé segons la localització de la universitat.

MATERIAL I MÈTODES

En primer lloc vam haver d'escollir la mostra a tractar, havíem de decidir quantes notes de tall faríem servir i el nombre d'anys a observar a l'hora de realitzar el nostre treball. I a més a més revisar els diferents centres a Catalunya on es podia realitzar aquesta carrera.
Mitjançant la web de Catalunya "unportal.net" vam trobar les notes de tall dels darrer cinc anys. Aquesta web ens proporciona molta informació sobre diferents universitats i les seves notes de la selectivitat tan a primera assignació com a segona.
I per tots els altres càlculs necessaris hem emprat el mateix programa R.

1. **Prova de significat (bilateral)**

   És bilateral perquè en cada pregunta volem comprobar si és igual o diferent a un número en específic.

   H0: μa = μe
   H1: μa != μe

   pregunta 1:

   μa = mitjana poblacional de N1a
   μe = mitjana poblacional de N2a

   pregunta 2:

   μa = mitjana poblacional de N1a al 5è any
   μe = mitjana poblacional de N1a al 1r any

   N1a = Nota 1ª assignació
   N2a = Nota 2ª assignació

2. **Premises**

   2.1 Normalitat

   Veure normalitat en les variables de resposta donades; notes de tall de la primera assignació i de la segona.

   2.2 Desviacions desconegudes i iguals en mostres independents

2.3 La distribució és una t-student amb diferents graus de llibertat depenent de què ens qüestionem.

2.4 La mostra es aleatoria

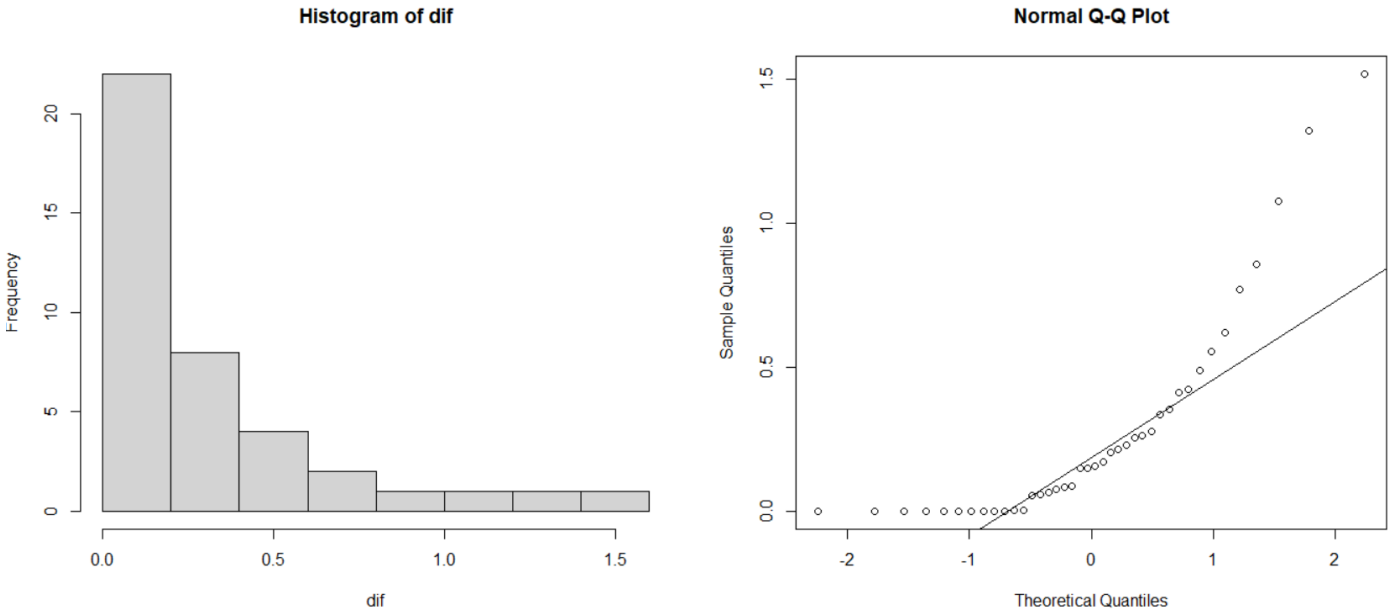
RESULTATS

1- Quant val l'interval de confiança de la diferència en mitjana de les notes de la primera assignació(N1a) i les notes de la segona(N2a)?

Hem de veure si la **diferència entre les notes de tall de la primera assignació i les de la segona** compleixen la **premisa de normalitat.** Per tant hem de veure l'histograma i el qqnorm de la diferència de notes i veure si podem asumir que segueix un model normal. Per això treballarem amb aquestes dues variables com a mostra **aparellada**.

dif = N1a - N2a
dif → c(0.148 0.170 0.256 0.276 0.424 0.230 0.000 0.151 0.064 0.215 0.205 0.004 1.079 0.002 0.078 0.336 0.000 0.000 0.084 0.060 0.000 0.000 0.056 0.488 0.262 0.000 0.414 0.353 1.322 0.088 0.000 0.000 0.158 0.860 0.770 0.000 0.000 0.620 0.556 1.520)

```
par(mfrow=c(1,2))
hist(dif)
qqnorm(dif)
qqline(dif)
```

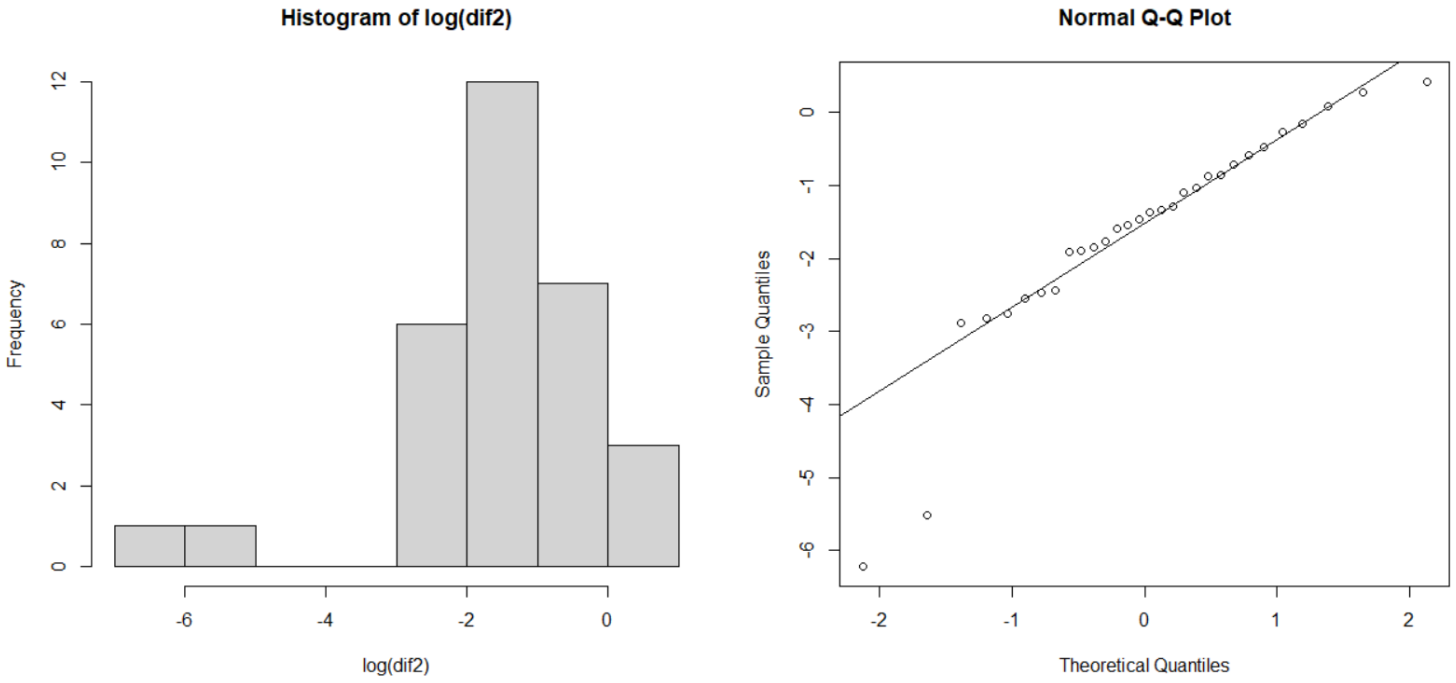


Com surt un model exponencial negatiu, ja que al principi s'acumulen moltes diferències que donen 0, hem de fer el logaritme de les diferències. Com no es pot fer el logaritme de valors nuls, hem hagut de treure els valors on la diferència és 0:

- UAB any 2 (18-19)
- UPC (EPSE) any 2 (18-19)
- UPC (EPSE) any 3 (19-20)
- UPF any 1 (17-18)
- UPF any 2 (18-19)
- UL any 1 (17-18)
- UDG any 1 (17-18)
- UDG any 2 (18-19)
- URV any 1 (17-18)
- URV any 2 (18-19)

dif2 → c(0.148 0.170 0.256 0.276 0.424 0.230 0.151 0.064 0.215 0.205 0.004 1.079 0.002 0.078 0.336 0.084 0.060 0.056 0.488 0.262 0.414 0.353 1.322 0.088 0.158 0.860 0.770 0.620 0.556 1.520)

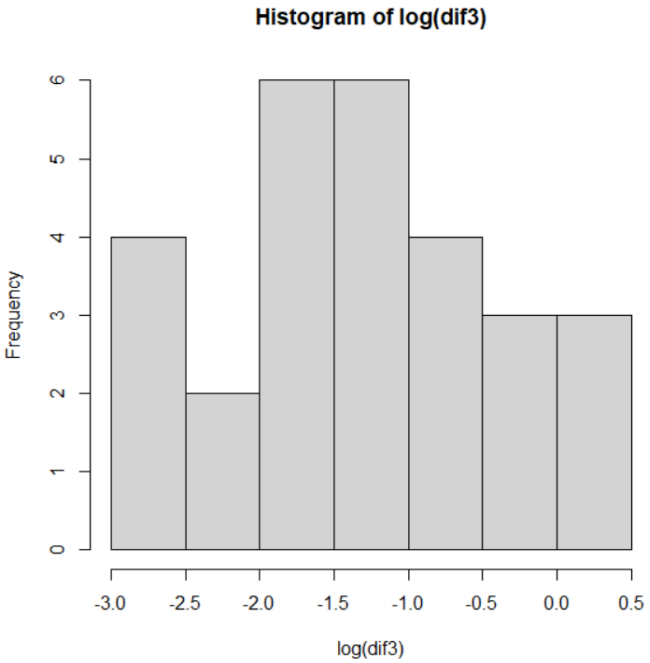
```
par(mfrow=c(1,2))
hist(log(dif2))
qqnorm(log(dif2))
qqline(log(dif2))
```

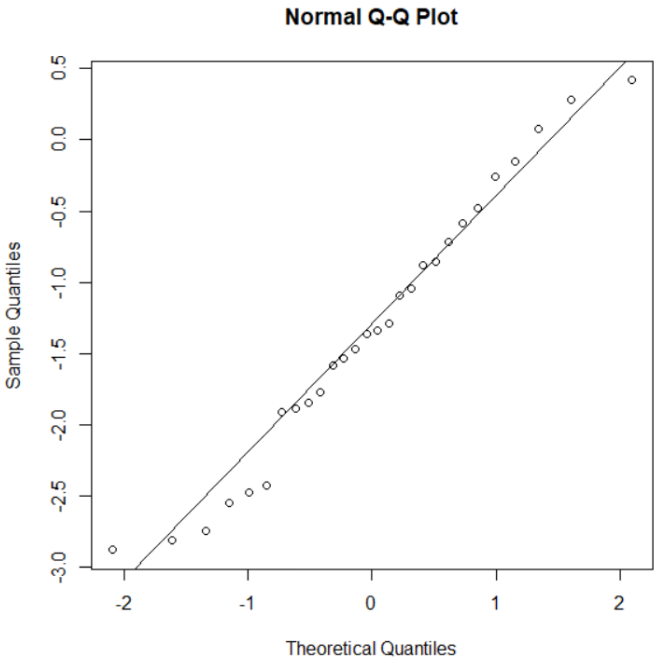


Ara les gràfiques mostren un model normal, a excepció de dos casos aïllats que es poden veure en l'histograma i en el qqnorm. Aquests dos valors tenen una diferència molt propera a 0 comparat amb les altres, per tant a l'hora de fer el logaritme quedaven massa apartats de la resta de diferències. Hem decidit treure aquests dos valors (subratllats del vector de diferències): UPC(FIB) any 2 i 4

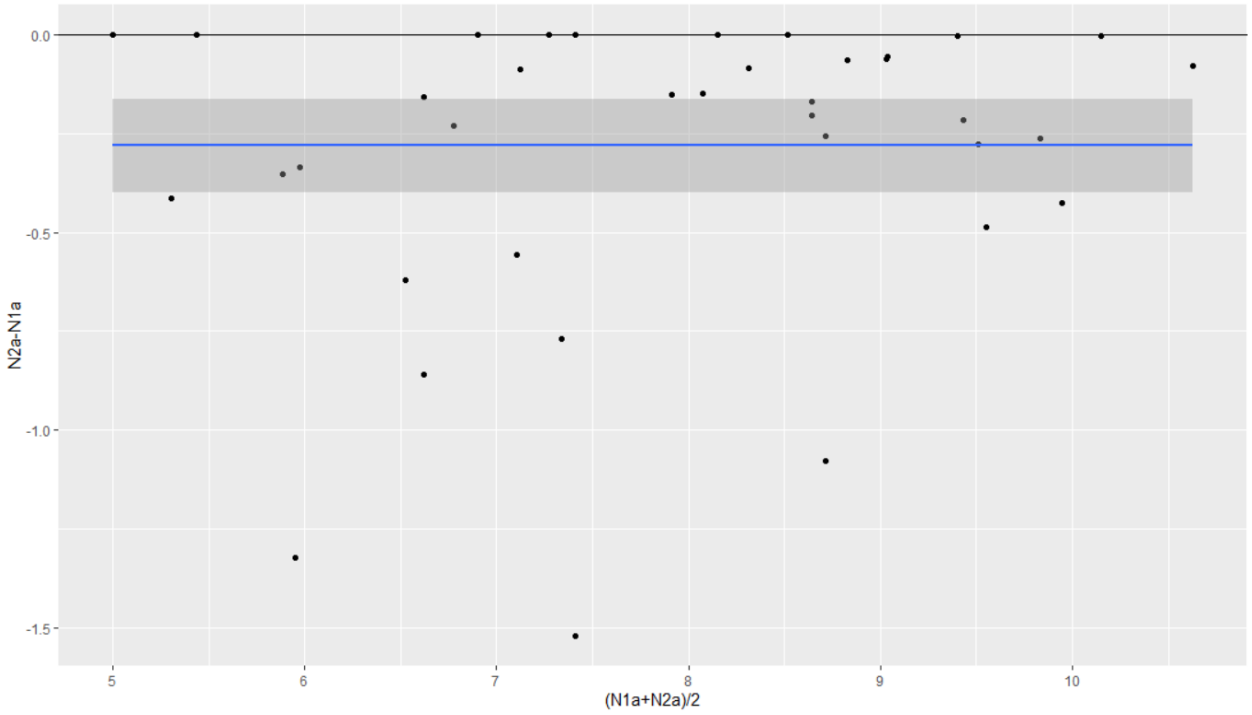
dif3 → c(0.148 0.170 0.256 0.276 0.424 0.230 0.151 0.064 0.215 0.205 1.079 0.078 0.336 0.084 0.060 0.056 0.488 0.262 0.414 0.353 1.322 0.088 0.158 0.860 0.770 0.620 0.556 1.520)

```
par(mfrow=c(1,2))
hist(log(dif3))
qqnorm(log(dif3))
qqline(log(dif3))
```






```
N1a = N1a
N2a = N2a
p = paired(N1a,N2a)
plot(p,type='BA') #Gràfic Bland Altman
```



Ara veient les gràfiques i amb l'ajut del Bland Altman, veiem com millora l'anterior normalitat. Tant com en el qqnorm com en el Bland Altman es veu com els punts es reparteixen per sobre i per sota de la línea, així com uns quants també tocant-la. Per tant, **decidim treure els valors anteriors mencionats.**

D → N1-N2
lm(log(D) ~ 1)
summary(lm(log(D) ~ 1))

n1 = dades$N1a
n2 = dades$N2a
dif3 = n1-n2

dif3 → c(0.148 0.170 0.256 0.276 0.424 0.230 0.151 0.064 0.215 0.205 1.079 0.078 0.336 0.084 0.060 0.056 0.488 0.262 0.414 0.353 1.322 0.088 0.158 0.860 0.770 0.620 0.556 1.520)

summary(lm((log(dif3))~1))

```
Call:
lm(formula = (log(dif3)) ~ 1)

Residuals:
    Min      1Q  Median      3Q     Max
-1.55402 -0.56711 -0.02261  0.64356  1.74709

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.3284     0.1791  -7.416 5.61e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9478 on 27 degrees of freedom
```

La mitjana de la diferència entre la primera assignació i la segona assignació és $e^{-1.3284}$ = **0.2649**.

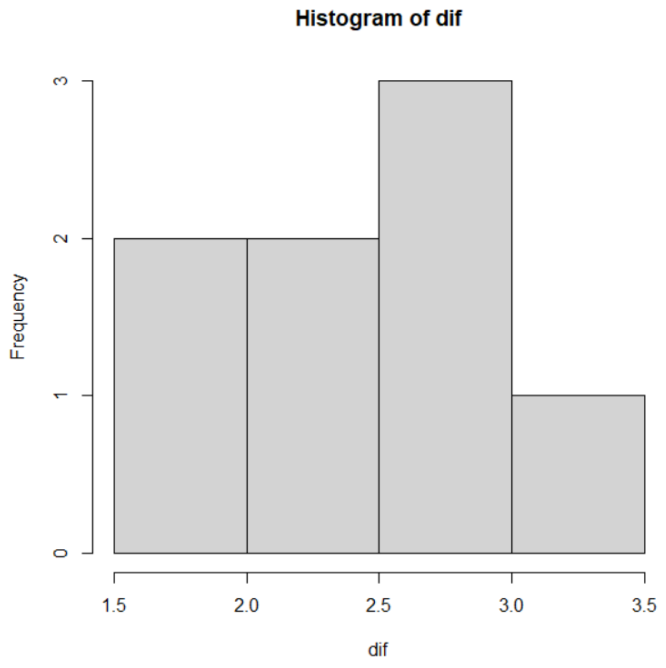
confiança = 0.95
alfa = 1-confiança
t = qt(1-alfa/2,27)
IC(diferència,95%) = [b0 - t*SE , b0 + t*SE] → [-1.3284-t*0.1791,-1.3284+t*0.1791] → [-1.6959 , -0.9609] → com esta en logaritme, hem de fer exponencial → [0.1834 , 0.3825]

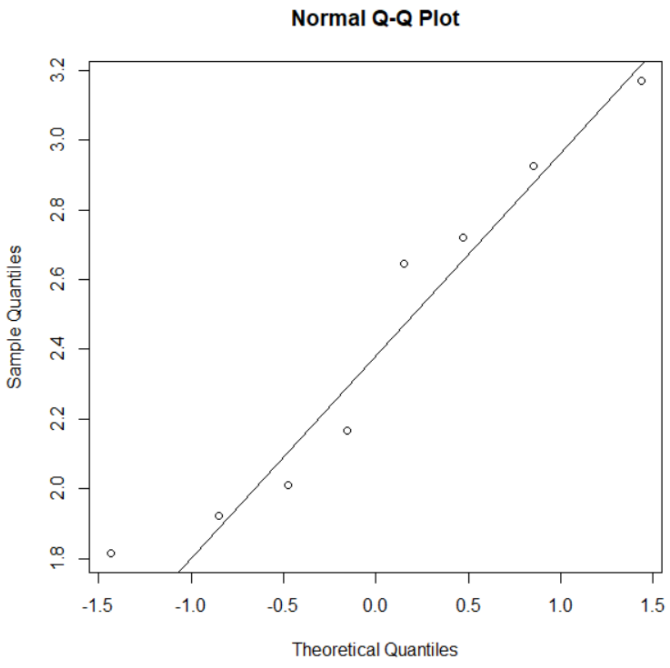
2- Quant val l'interval de confiança de la diferència en mitjana de la nota de la primera assignació de fa 1 any i la nota de fa 5 anys?

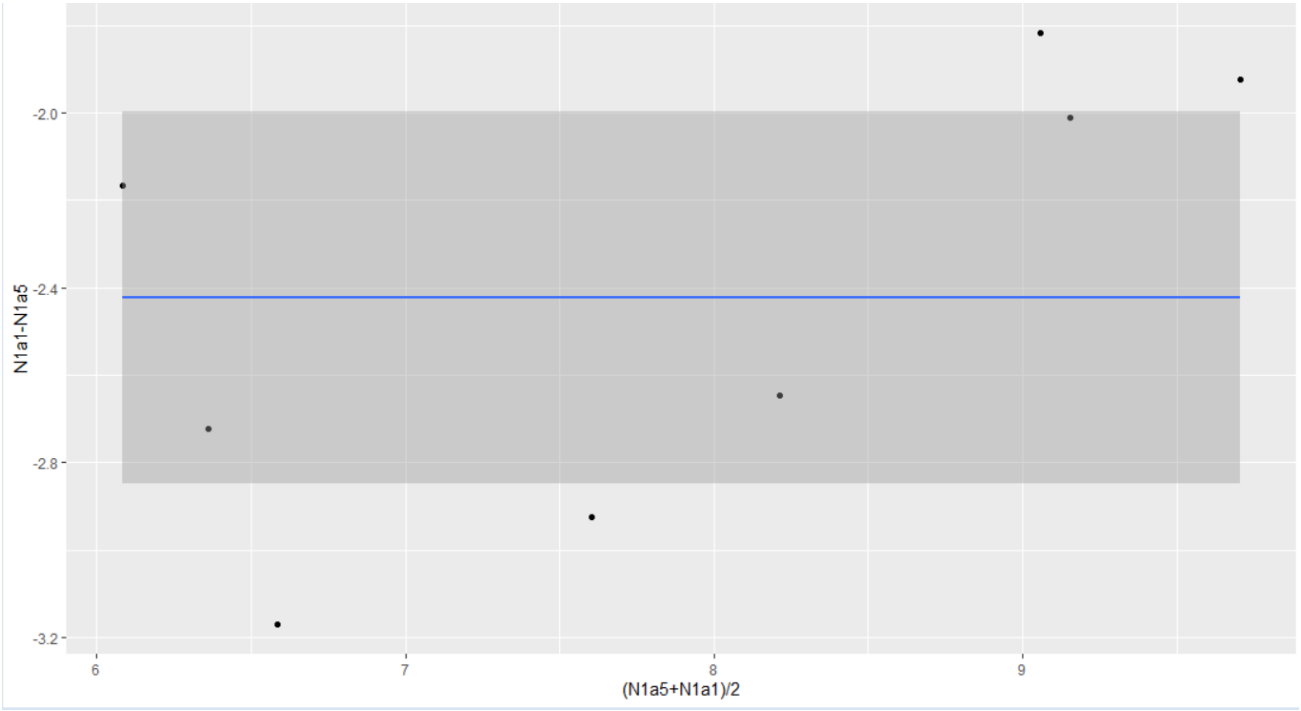
Hem de veure que la **diferència en mitjana de les notes de fa 1 any(N1a5) i de fa 5 anys(N1a1)** de les notes de primera assignació compleixen la **premisa de normalitat**. Per tant hem de veure l'histograma i el qqnorm de la diferència de notes i veure si podem asumir que segueix un model normal. Per això treballarem amb aquestes dues variables com a mostra **aparellada**.

dif = N1a5 - N1a1
dif → c(2.010 2.647 1.923 2.924 1.816 2.166 2.722 3.170)

```
par(mfrow=c(1,2))
hist(dif)
qqnorm(dif)
qqline(dif)
N1a5 = any5$N1a
N1a1 = any1$N1a
p = paired(N1a5,N1a1)
plot(p,type='BA') #Gràfic Bland Altman
```







El problema en aquest cas és que la n és petita, són les n=8 universitats que estudiem. Per tant és difícil treure conclusions de les gràfiques, però no es veu evidència clara en cap d'elles per contradir que segueix un model normal.

```
splita = split(dades,dades$Any)
any5 = splita$'5'
any1 = splita$'1'
dif = any5$N1a - any1$N1a

dif → c(2.010 2.647 1.923 2.924 1.816 2.166 2.722 3.170)

# summary(lm(dif ~ 1))

        Call:
        lm(formula = dif ~ 1)

        Residuals:
            Min       1Q   Median       3Q      Max
        -0.60625 -0.43400 -0.01575  0.35025  0.74775

        Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
        (Intercept)   2.4222     0.1795    13.5 2.88e-06 ***
        ---
        Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

        Residual standard error: 0.5076 on 7 degrees of freedom
```

La mitjana de la diferencia de la nota de la 1ª assignació del 5è any al 1r any és de **2.422**, amb un error estàndard de **0.1795** i una desviació estàndard dels residus de **0.5076**.

conf = 0.95
alfa = 1-conf
t = qt(1-alfa/2,7)
IC → [2.4222-t*0.1795,2.4222+t*0.1795] → [1.998,2.847]

La diferència de *Any5 - Any1* de cada universitat és:

UB → 2.010
UAB → 2.647
FIB → 1.923
EPSEVG → 2.924
UPF → 1.816
UL → 2.166
UDG → 2.722
URV → 3.170

Notes de la primera assignació de l'Any 1 (2017-2018):

Universitat	***N1a***	*N2a*	*Any*	*Prov*
UB	*8.148*	*8.000*	*1*	*B*
UAB	*6.890*	*6.660*	*1*	*B*
UPC(FIB)	***8.745***	*8.540*	*1*	*B*
UPC(EPSEVG)	*6.140*	*5.804*	*1*	*B*
UPF	***8.150***	*8.150*	*1*	*B*
UL	*5.000*	*5.000*	*1*	*F*
UDG	*5.000*	*5.000*	*1*	*F*
URV	*5.000*	*5.000*	*1*	*F*

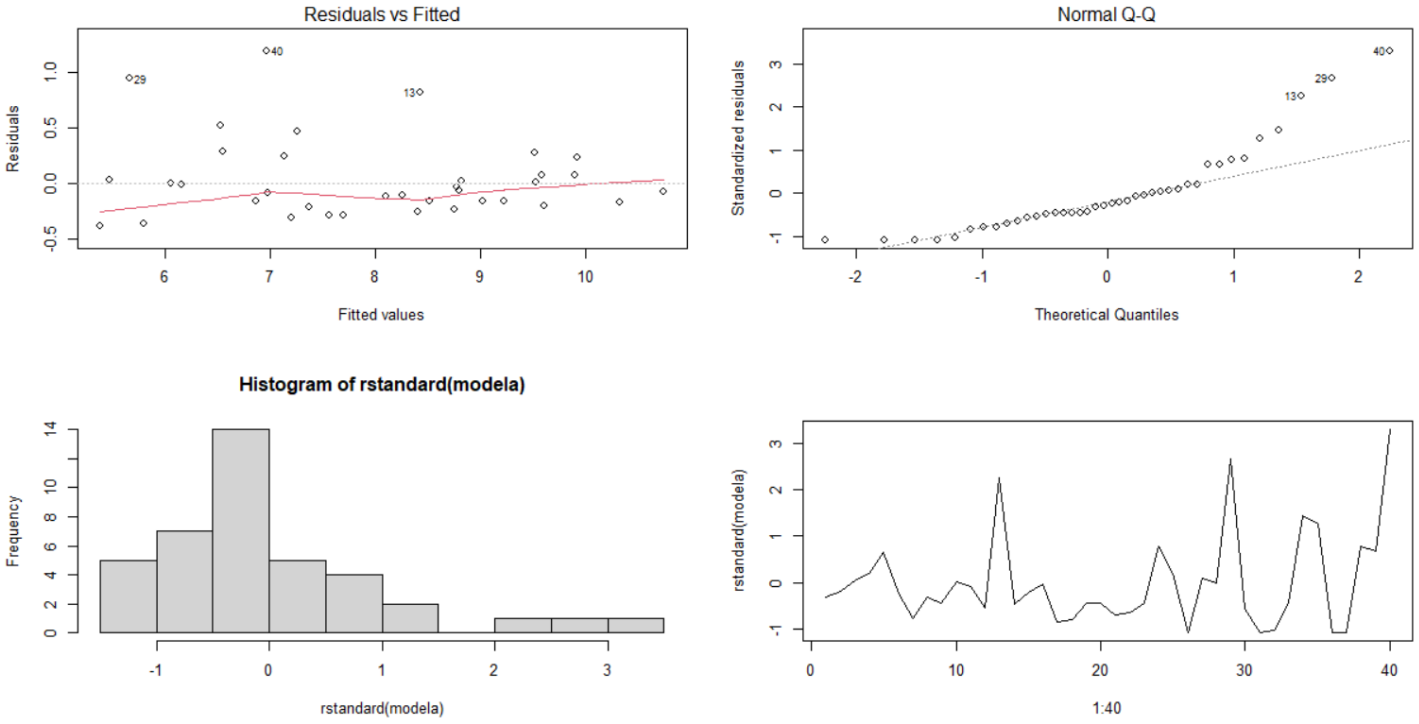
Notes de la primera assignació de l'Any 5 (2021-2022):

Universitat	***N1a***	*N2a*	*Any*	*Prov*
UB	*10.158*	*9.734*	*5*	*B*
UAB	*9.537*	*9.322*	*5*	*B*
UPC(FIB)	*10.668*	*10.590*	*5*	*B*
UPC(EPSEVG)	*9.064*	*9.004*	*5*	*B*
UPF	*9.966*	*9.704*	*5*	*B*
UL	*7.166*	*7.078*	*5*	*F*
UDG	*7.722*	*6.952*	*5*	*F*
URV	*8.170*	*6.650*	*5*	*F*

3- Hi ha relació lineal entre les notes de la primera assignació(N1a) i las de la segona(N2a)?

Hem de fer una **validació de la linealitat** entre les notes de la **primera assignació (N1a) i les de la segona (N2a)**. Per veure-ho, farem un diagnosi del model N1a~N2a amb 4 gràfiques de residus.

modela = lm(dades$N1a~dades$N2a)
par(mfrow=c(2,2))
plot(modela,c(2,1)) #QQ-Norm i Standard Residuals vs. Fitted
hist(rstandard(modela)) # Histograma dels residus estandaritzats
plot (1:40,rstandard(modela),type="l") # Ordre dels residus estandaritzats



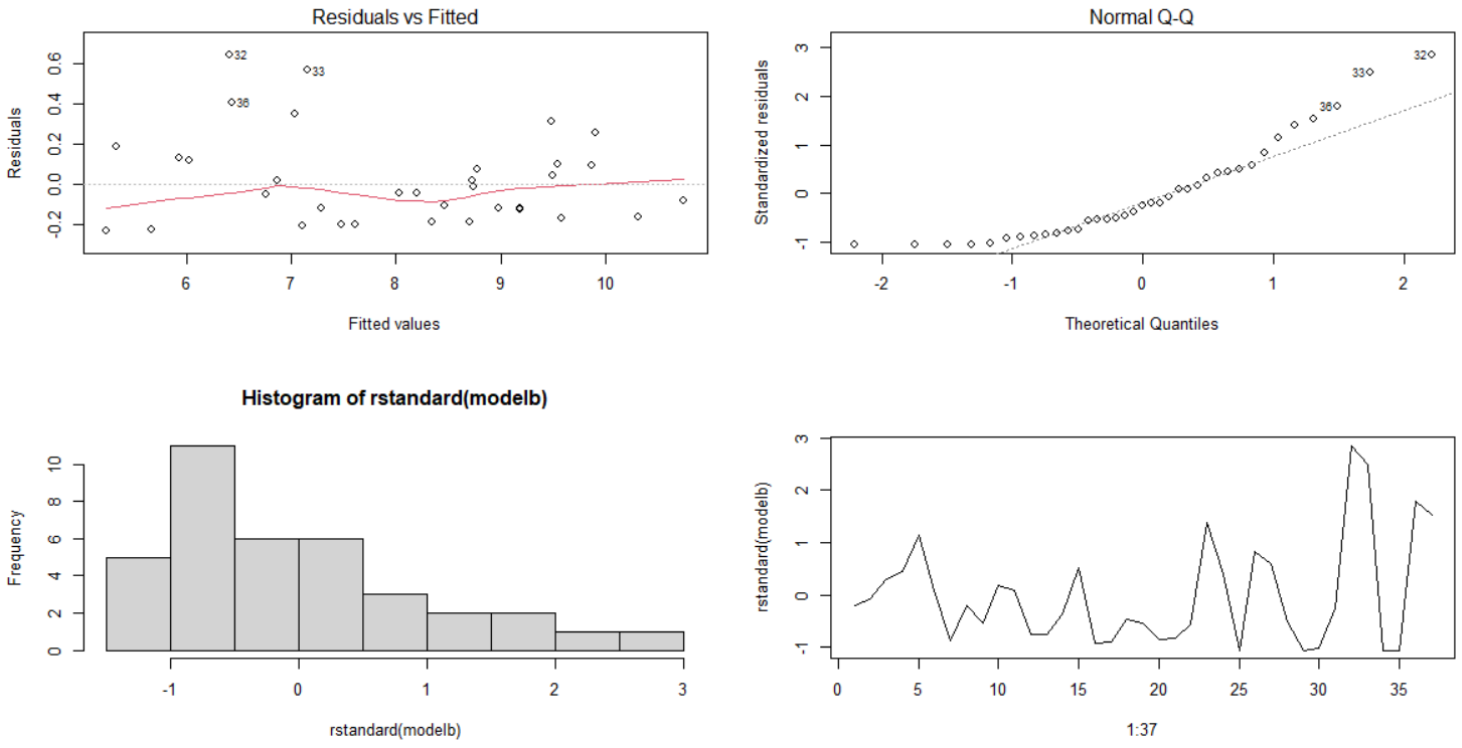
Per la primera gràfica de residus enfront les prediccions veiem que es compleix la **linealitat** ja que els punts estan per sota i per sobre la línea vermella uniformement ***** .
També podem veure en la mateixa gràfica que es distancian de la línea de la mateixa forma, sense zones amb més i menys dispersió ***** , per tant compleix la **homoscedasticitat**.
Per la segona gràfica "Normal Q-Q" es veu com els punts s'ajusten a la recta força bé a excepció per la part més a la dreta ***** . Podem dir que compleix la **normalitat**.
Per l'última gràfica observem els residus enfront l'ordre de recollida, i podem veure que no hi ha dependència ni cap patró en específic ***** . Es compleix la **independència**.

***** Com es pot veure en cada gràfica, i més clarament en l'histograma, hi ha 3 valors apartats de la resta. Per tant vam proposar treure algunes dades, específicament les files 13, 29 i 40 de la taula, per controlar millor la variabilitat per fer les prediccions.

Notes traient aquells 3 valors:
N1a_2 =
c(8.148,8.728,8.844,9.650,10.158,6.890,7.412,7.990,8.860,9.537,8.745,9.406,10.148,10.668,6.140,6.902,7.272,8.356,9.064,8.150,8.518,9.064,9.798,9.966,5.000,5.510,6.059,7.166,5.000,5.436,6.700,7.050,7.722,5.000,5.000,6.834,7.382)

N2a_2 =
c(8.000,8.558,8.588,9.374,9.734,6.660,7.412,7.839,8.796,9.322,8.540,9.402,10.146,10.590,5.804,6.902,7.272,8.272,9.004,8.150,8.518,9.008,9.310,9.704,5.000,5.096,5.706,7.078,5.000,5.436,6.542,6.190,6.952,5.000,5.000,6.214,6.826)

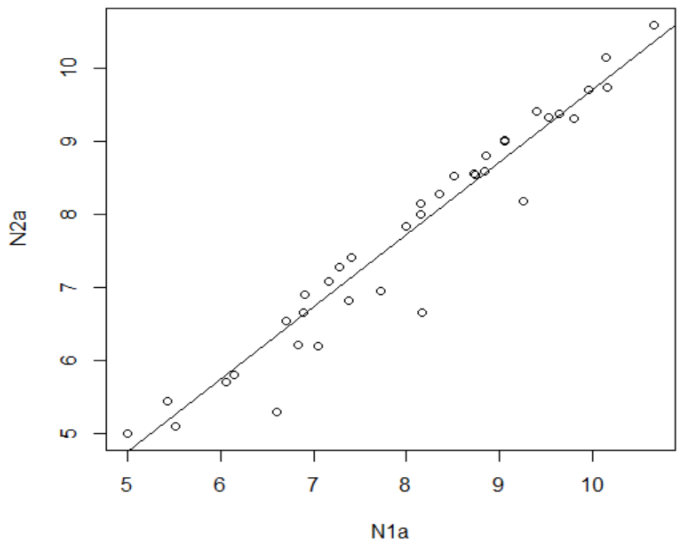
```
modelb = lm(N1a_2~N2a_2)
par(mfrow=c(2,2))
plot(modelb,c(2,1)) #QQ-Norm i Standard Residuals vs. Fitted
hist(rstandard(modelb))  # Histograma dels residus estandaritzats
plot (1:37,rstandard(modelb),type="l") # Ordre dels residus estandaritzats
```



Al comprovar la linealitat amb aquests canvis, obtenim unes gràfiques molt similars als anteriors. Si ens fixem, podem veure una petita millora amb la homoscedasticitat i es pot observar més clarament la campana de Gauss al histograma de rstandard. Malgrat això, no trobem que hi hagin canvis suficientment notables per treure aquestes dades.

En conseqüència, hem decidit **no treure els valors** perquè les gràfiques originals són suficients per dur a terme la hipòtesi.

```
# cor(N1a, N2a)
        0.9741894
# lm(N2a ~ N1a)
        Call:
        lm(formula = N2a ~ N1a)

        Coefficients:
        (Intercept)         N1a
         -0.2049       0.9902

#summary(lm(N2a ~ N1a))

        Call:
        lm(formula = N2a ~ N1a)

        Residuals:
            Min       1Q   Median       3Q      Max
        -1.23523 -0.09661  0.12612  0.25376  0.30212

        Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
        (Intercept) -0.20486    0.29664  -0.691    0.494
        N1a          0.99022    0.03722  26.604   <2e-16 ***
        ---
        Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

        Residual standard error: 0.3764 on 38 degrees of freedom
        Multiple R-squared:  0.949,     Adjusted R-squared:  0.9477
        F-statistic: 707.8 on 1 and 38 DF,  p-value: < 2.2e-16


# par(mfrow=c(1,1))
# plot(N1a,N2a)
# abline(lm(N2a~N1a))
```



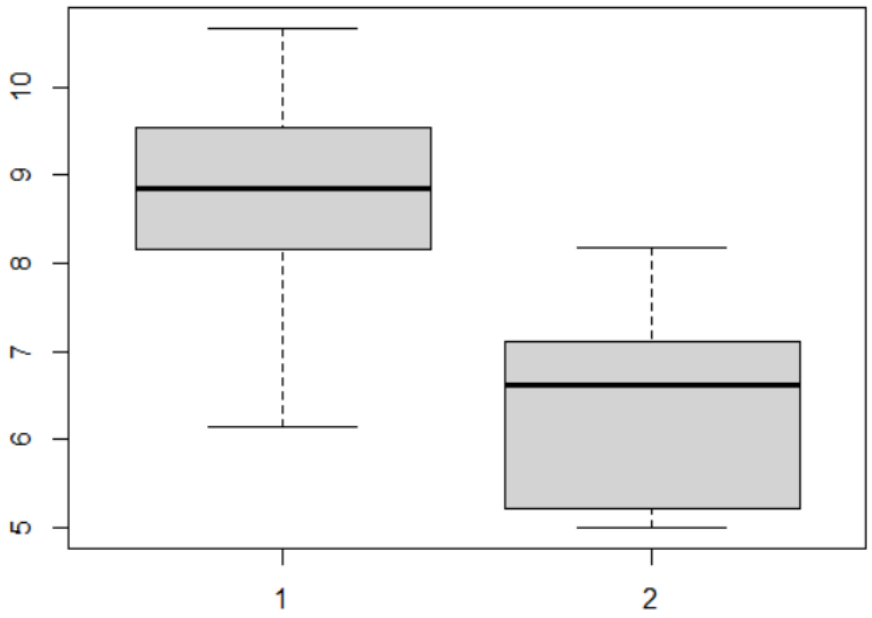
El model recull un **94.9% de la variabilitat** total de la variable resposta. La resta residual amb desviació de **0.3764**
L'ordenada a l'origen (terme independent) és **-0.20486** amb IC del 95%: -0.20486±qt(0.975,38)*0.29664 → [-0.8054, 0.3957]
El pendent (terme lineal) és **0.99022** amb IC del 95%: 0.99022±qt(0.975,38)*0.03722 → [0.9149, 1.0656]

L'equació de la recta estimada és:

N2a = -0.20486 + 0.99022*N1a

4- La nota de tall de la primera assignació de les universitats de Barcelona (Prov = B) son més grans en mitjana que les de fora de Barcelona (Prov = F)?

Primerament, comparem les desviacions amb un boxplot, per saber si són iguals o diferents.
splitp = split(dades,Prov)
nb = splitp$'B'
nf = splitp$'F'
boxplot(nb$N1a,nf$N1a)



Com que les desviacions donen força diferents, enlloc d'aplicar *lm()* per fer els càlculs farem un *t.test(nb$N1a,nf$N1a,var.equal=F)* ja que comparem dues mitjanes.

t.test(nb$N1a,nf$N1a,var.equal=F)

Welch Two Sample t-test

data: nb$N1a and nf$N1a
t = 6.5879, df = 30.799, p-value = 2.405e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
1.654977 3.139743
sample estimates:
mean of x mean of y
8.70676 6.30940

La mitjana de la nota de la 1ª assignació de les universitats de Barcelona és de **8.70676**, i la de fora de Barcelona és de **6.3094**. Hi ha un error estàndard de **0.3685221** en aquest model (càlcul fet en l'annex).

L'interval de confiança IC(diferència,95%) = [*1.654977, 3.139743]* *

* IC → [estimate-qt(0.975,38)*SE, estimate+qt(0.975,38)*SE]

DISCUSSIÓ

Pregunta 1:

Hem vist que la mitjana de la diferència entre les notes de la primera assignació i de la segona és de 0.2649 amb un interval de confiança **[0.1834 , 0.3825]**. Amb un 95% de confiança el 0 no pertany a l'interval, i ho contrastem mirant que el p-value que mostra el model és menor que el risc ($5.61x10^{-8}$ < 0.05). Per tant, hem trobat evidència per contrastar que les mitjanes poblacionals d'aquestes dues notes no son iguals.

Ara ens podríem preguntar la nota de la primera assignació és 0.2 punts major en mitjana que la de la segona assignació. Com que 0.2 punts és dins l'interval de confiança [0.1834 , 0.3825], amb un 95% de confiança és raonable pensar que la mitjana de les notes de la primera assignació son 0.2 punts més altes que la mitjana de les notes de la segona assignació.

Aquestes conclusions estan fetes a partir de treure les dades que feien que no es complís la premisa inicial, per tant ha pogut afectar a la representativitat de la mostra. En aquest cas s'havien de treure les diferències que donaven 0 i dos valors que eren molt propers a 0, per així fer el logaritme i que seguís un model normal.

Pregunta 2:

El resultat dóna una mitjana de la diferència de les notes de primera assignació de fa 1 any i les de fa 5 de 2.422 amb un interval de confiança de **[1.998,2.847]**.
Amb un 5% de risc podem pensar que el 0 no pertany a l'interval. Si mirem el p-value que ens dona la mostra, veiem que és més petit que el risc ($2.88x10^{-6}$ < 0.05) i corrobora la hipòtesi de que les mitjanes poblacionals són diferents.

Ens podem preguntar si la nota de la primera assignació de fa 1 any és major en mitjana en 2 punts que la nota de la primera assignació de fa 5 anys. Els 2 punts és dins de l'interval de confiança [1.998,2.847], aleshores amb un 95% de confiança no hem trobat evidència per contradir la hipòtesi de que la nota de fa 1 any és major en 2 punts de la nota de fa 5 anys. Es pot contrastar mirant que el p-value sigui major que el risc, recalculant el p-value amb valor esperat = 2:
estadístic = (2.422-2)/0.1795 = 2.350975
$p - value(2.350975) = 2 * (pt(-2.350975, 7))$ = 0.05101424
Com que 0.05101424 > 0.05, confirma que és raonable pensar que és major en punts.

Al ser 2 un valor proper al límit inferior de l'interval també podem concloure que la mitjana és lleugerament superior al valor de la hipòtesi.
Si comparem les universitats per separat, veiem que les úniques universitats que no compleixen la hipòtesi son la FIB i la UPF. Si entrem més en detall, aquestes dues universitats son les que tenen la nota més alta fa 5 anys, per tant és raonable pensar que son las que han incrementat menys la nota perquè parteixen d'una nota alta.

Pregunta 3:

El model recull un 94.9% de la variabilitat total queda explicada de la nota de la primera assignació(N1a). És a dir, podem predir amb un 94.9% la nota de la segona assignació(N2a) en base de la nota N1a. És pràcticament determinista.

La resta és residual amb desviació de 0.3764. Podem esperar fluctuacions de 0.3764 punts respecte les previsions de durada en funció de la nota de la primera assignació.

Segons el coeficient b_0, podem saber que si una nota de la primera assignació té valor 0, la nota de la segona assignació serà -0.20486.
I per la b_1, com és positiva, sabem que és ascendent; per cada unitat més a la nota de la primer assignació, esperem 0.99022 unitats més per a la de la segona.

Pregunta 4:

Hem vist que les mitjanes de les notes de tall de Barcelona es 8.70676 i per les de fora la mitjana és 6.3094. Ara hem de veure si aquestes dues notes les podem considerar iguals.

Hem calculat amb un 95% de confiança l'interval de la diferència ***[1.654977, 3.139743]*.** Com que el 0 no pertany a aquest interval, és raonable pensar que les notes de la primera assignació de Barcelona són diferentes en mitjana que les notes de la primera assignació de fora de Barcelona. En concret, les de Barcelona són entre 1.65 i 3.14 més grans. El *p value* dona $2.4x10^{-7}$ que és molt més petit que el risc (0.05), per tant podem afirmar que aquest valor és extrem i per tant que el 0 no pertany al IC i que les dues mitjanes no son iguals.

A més, hem calculat el SE (on s és la variància del "pooled") a mà:

se = s*sqrt(1/nB + 1/nF) = 1.128364*sqrt(1/25 + 1/nF) = 0.3685221

Ara ens podem preguntar si les notes de Barcelona són 3 punts més grans en mitjana que las de fora de Barcelona. Ens trobem que el 3 si pertany a aquest interval i per tant podem concloure amb un 95% de confiança que les notes de dins de Barcelona son 3 punts superior que la de la resta de Catalunya.

<u>LIMITACIONS</u>

- Manca de recursos i dades de les universitats privades de GEI perquè les mantenen confidencials.
  En principi, com diem al protocol, també havíem proposat fer comparacions de notes segons si la universitat era pública o privada. Però a l'hora de la recollida de dades vam veure que no teniem la disponibilitat d'aquestes.

- Hem recollit les notes dels 5 darrer anys, ja que era força complicat trobar aquesta informació a partir de 5 any enrere, i que sigues veredigna.

- És difícil preveure si en els següents 5 anys les notes canviaren molt perquè depenen en major part de l'examen de selectivitat. Un canvi en el model de fer aquest examen afecta directament a aquests tipus d'estudi.

- Hem hagut de treure en alguna pregunta algunes dades per tal de seguir les premises, cosa que afecta als resultats i conclusions que es treuen.

TREBALL FUTUR

- Fer un estudi de la taxa d'abandonament dels graus d'informàtica a Catalunya, fent-ne comparacions segons l'any, si es pública o privada i el gènere.

- Fer prediccions de les notes que esperem en 10 anys, ampliant així en nombre de dades. A una n major, millor previsió es podrà fer.

ANNEX

DADES:

#dades = read.table("clipboard",header=TRUE)

#attach(dades)

Universitat	N1a	N2a	Any	Prov
UB	8.148	8	1	B
UB	8.728	8.558	2	B
UB	8.844	8.588	3	B
UB	9.65	9.374	4	B
UB	10.158	9.734	5	B
UAB	6.89	6.66	1	B
UAB	7.412	7.412	2	B
UAB	7.99	7.839	3	B
UAB	8.86	8.796	4	B
UAB	9.537	9.322	5	B
UPC(FIB)	8.745	8.54	1	B
UPC(FIB)	9.406	9.402	2	B
UPC(FIB)	9.255	8.176	3	B
UPC(FIB)	10.148	10.146	4	B
UPC(FIB)	10.668	10.59	5	B
UPC(EPSEVG)	6.14	5.804	1	B
UPC(EPSEVG)	6.902	6.902	2	B
UPC(EPSEVG)	7.272	7.272	3	B
UPC(EPSEVG)	8.356	8.272	4	B
UPC(EPSEVG)	9.064	9.004	5	B
UPF	8.15	8.15	1	B
UPF	8.518	8.518	2	B
UPF	9.064	9.008	3	B
UPF	9.798	9.31	4	B
UPF	9.966	9.704	5	B
UL	5	5	1	F
UL	5.51	5.096	2	F
UL	6.059	5.706	3	F
UL	6.612	5.29	4	F
UL	7.166	7.078	5	F
UDG	5	5	1	F
UDG	5.436	5.436	2	F
UDG	6.70	6.542	3	F
UDG	7.050	6.19	4	F
UDG	7.722	6.952	5	F
URV	5	5	1	F
URV	5	5	2	F
URV	6.834	6.214	3	F
URV	7.382	6.826	4	F
URV	8.17	6.65	5	F

Any:	Prov:
1 - 2017/18	B - Barcelona
2 - 2018/19	F - Fora
3 - 2019/20	
4 - 2020/21	
5 - 2021/22	

Glossari

Scripts amb R:

lm() de la pregunta 4 que al final es va fer amb t.test() perquè les desviacions eren diferents.

```
# lm(N1a ~ Prov)
        Call:
        lm(formula = N1a ~ Prov)

        Coefficients:
        (Intercept)       ProvF
             8.707      -2.397

# summary(lm(N1a ~ Prov))
        Call:
        lm(formula = N1a ~ Prov)

        Residuals:
           Min      1Q  Median      3Q     Max
        -2.5668 -0.8179  0.1452  0.8368  1.9612

        Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
        (Intercept)   8.7068     0.2257  38.581  < 2e-16 ***
        ProvF        -2.3974     0.3685  -6.505 1.16e-07 ***
        ---
        Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

        Residual standard error: 1.128 on 38 degrees of freedom
        Multiple R-squared:  0.5269,    Adjusted R-squared:  0.5144
        F-statistic: 42.32 on 1 and 38 DF,  p-value: 1.157e-07

# emmeans(lm(N1a~Prov),~Prov)
         Prov emmean    SE df lower.CL upper.CL
         B      8.71     0.226 38     8.25     9.16 → IC[8.25,9.16]
         F      6.31     0.291 38     5.72     6.90 → IC[5.72,6.90]

        Confidence level used: 0.95

# pairs(emmeans(lm(N1a~Prov),~Prov))
        contrast estimate    SE df t.ratio p.value
        B - F         2.4 0.369 38   6.505  <.0001 → IC [1.652999, 3.147001] *
```

I pel càlcul a mà del SE (amb variància “pooled”) de la pregunta 4:

```
# dadessplit <- split(dades, dades$Prov)
# dadesB <- dadessplit$'B'
# dadesF <- dadessplit$'F'
# dadesB <- dadesB$N1a
# dadesF <- dadesF$N1a
# nB <- length(dadesB)
# nF <- length(dadesF)
# sB2 <- var(dadesB);  sB2
        [1] 1.318741
# sF2 <- var(dadesF);  sF2
        [1] 1.195144
# s2 <- ((nB-1)*sB2 + (nF-1)*sF2)/((nB-1)+(nF-1));   s2
        [1] 1.273205
# s <- sqrt(s2); s
        [1] 1.128364
# se <- s*sqrt(1/nB + 1/nF);   se
        [1] 0.3685221
```