NOM:_____

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

Problema 1 (B4). El Servei Català de Trànsit (SCT) ens controla. Des de fa un temps, el SCT es capaç de captar les senyals dels conductors que envien "whatsapps" mentre estan conduint a les autopistes que gestiona. L'aplicació que han construït recull la informació corresponent a 4 variables: identificador de l'autopista, data, hora i punt quilomètric de l'autopista des d'on es rep el senyal. Un hacker amic nostre ha aconseguit accedir a una mostra de 100 deteccions d'una mateixa autopista. Les dades dels punts quilomètrics i la seva descriptiva es mostren a les següents taules:

0.3	0.3	1.4	1.8	2.3	2.7	2.8	2.9	3.3	3.3	3.9	4.9	5.2	6.8	6.9	7.1	7.6	8.0	8.8	9.3
9.4	10.2	10.2	10.6	10.8	11.2	11.5	12.5	13.4	13.5	13.8	15.1	15.2	16.0	16.2	16.4	17.2	17.3	17.7	17.9
18.8	18.8	19.1	19.7	20.0	20.1	20.7	21.0	21.9	22.0	22.1	22.1	22.3	22.8	22.8	24.3	24.5	24.6	25.3	25.5
26.9	27.4	27.7	27.9	28.2	28.3	29.0	29.1	29.4	29.7	29.9	31.4	31.6	31.7	31.9	32.2	32.4	32.8	33.3	34.2
34.8	34.8	34.9	35.1	35.2	35.3	35.3	35.5	35.6	35.8	36.2	36.6	37.0	37.2	37.5	37.9	37.9	37.9	38.6	38.9

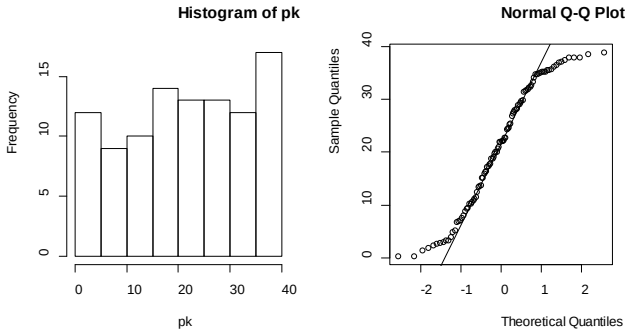
```
> length(pk)
[1] 100
```

```
> summary(pk)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.30   11.10   22.05   21.37   31.98   38.90
```

```
> sd(pk)
[1] 11.66476
```

```
> sum(pk<10)
[1] 21
```

1) (1 punt). Proposa una distribució que et sembli raonable per aquestes dades i justifica-ho.



Uniforme continua. L'histograma mostra una distribució pràcticament plana (segurament el volum de trànsit ha de ser similar al llarg de l'autopista i la gent ha d'enviar *whatsapps* independentment d'on estigui i del trànsit)

2) (1 punt). Ens interessa saber de quina autopista provenen les dades però l'identificador de l'autopista és un codi que no permet esbrinar de quina via es tracta. Si sabéssim la longitud total de l'autopista de la qual hem recollit les dades, la podríem comparar amb les longituds reals de les autopistes de Catalunya. Independentment de la resposta del primer apartat, suposa que les dades es distribueixen com una uniforme continua. Dóna almenys 2 estimacions puntuals raonables pel punt quilomètric màxim d'aquesta autopista (inclou la forma analítica, és a dir la fórmula, i el càlcul dels teus estimadors).

Possibilitats:

$\widehat{X_{(n)}} = 2 \cdot \bar{x} = 42.7$ $\widehat{X_{(n)}} = 2 \cdot med = 44.1$ $\widehat{X_{(n)}} = \sqrt{12} \cdot S = 40.4$ $\widehat{X_{(n)}} = (n+1) \cdot \frac{x_{(n)}}{n} = 39.3$ $\widehat{X_{(n)}} = x_{(n)} = 38.9$

3) (4 punts. 0.8 per apartat) Coneixem que l'autopista C-58 té exactament 40.0 km. Posem a prova la hipòtesi de que l'autopista de la qual hem recollit les dades tingui aquesta longitud suposant que segueix una distribució uniforme continua. Tingués present que una forma equivalent és contrastar si el punt quilomètric mitjà és 20 amb un α=0.05.

a) Hipòtesis

$\begin{cases} H_0: X_{(n)} = 40 \\ H_1: X_{(n)} \neq 40 \end{cases}$ → Suposant que segueix una uniforme, és equivalent a $\begin{cases} H_0: \mu = 20 \\ H_1: \mu \neq 20 \end{cases}$ prova bilateral

b) Estadístic, distribució, premisses

$Z = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} \sim N(0,1)$ suposant (1) m.a.s; i (2) n≥100

c) Valor de l'estadístic i punt crític

Estadístic → $Z = \frac{21.37-20}{11.67/\cdot\sqrt{100}} = 1.175$

Punt crític → $Z_{0.975} = 1.96$

d) Conclusió. Veient els resultats, ¿creus que hem identificat l'autopista com la C-58?

No podem rebutjar la hipòtesi que l'autopista tingui 40 km (equivalent a dir que la μ=20). És, a dir, no tenim prou evidència per descartar que l'autopista sigui la C-58, però tampoc podem estar-ne segurs.

[Deixem de banda el tema que pugui haver-hi més autopistes amb 40 km de longitud]

e) Calcula el valor mínim de la mitjana de la mostra amb el qual acceptaríem la hipòtesi alternativa de que l'autopista té més de 40 km amb un α=0.05 unilateral.

$\begin{cases} H_0: X_{(n)} = 40 \\ H_1: X_{(n)} > 40 \end{cases}$ → Suposant que segueix una uniforme, és equivalent a $\begin{cases} H_0: \mu = 20 \\ H_1: \mu > 20 \end{cases}$ prova unilateral

Rebutjarem la hipòtesi nul·la si es compleix la desigualtat $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} > Z_{0.95} \rightarrow \bar{x} > \mu_0 + Z_{0.95} \cdot \sigma/\sqrt{n}$

on $\sigma^2 = \frac{(b-a)^2}{12} = \frac{b^2}{12} = \frac{(2\mu_0)^2}{12} = \frac{40^2}{12} = 133.33$ [també es pot emprar la S^2]

Per tant, rebutjarem si $\bar{x} > 20 + 1.645 \cdot \frac{11.55}{\sqrt{100}} = 21.90$

Amb una mitjana mostral superior a 21.90 rebutjaríem la hipòtesi nul·la.

4) (4 punts. 1 per apartat). L'exercici anterior l'hem basat en la premissa de que era una distribució uniforme, però no estem del tot segurs. Per intentar discernir si es tracta d'una uniforme [0,40], fes un contrast d'hipòtesi amb un α=0.05 sobre si la probabilitat de tenir *whatsapps* en els 10 primers quilòmetres és igual a la probabilitat teòrica de tenir-los abans de 10 en una uniforme [0,40]

a) Hipòtesi

$\begin{cases} H_0: \pi = q(10) = 0.25 \\ H_1: \pi \neq q(10) = 0.25 \end{cases}$

b) Estadístic, distribució, premisses

$\hat{Z} = \frac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \sim N(0,1)$ suposant (1) m.a.s; (2) $\pi_0 \cdot n \geq 5$; i (3) $(1-\pi_0)\cdot n \geq 5$

c) Valor de l'estadístic i punt crític

Estadístic → $\hat{Z} = \frac{0.21-0.25}{\sqrt{\frac{0.25(1-0.25)}{100}}} = -0.924$

Punt crític → $Z_{0.975} = 1.96$

d) Conclusió. ¿Podem afirmar que la distribució de referència és una uniforme?

No podem rebutjar la hipòtesi de que la funció de referència sigui una uniforme continua, però tampoc podem estar-ne segurs; de fet, qualsevol distribució que acumuli una probabilitat de 0.25 abans del 10 podria ser candidata.

Bloc 5. El passat 21 d'abril de 2014 *La Vanguardia* va publicar un article titulat "Conducció política temerària" que, basant-se en un sondeig del CIS (Baròmetre de setembre de 2013), donava a entendre que hi ha una relació entre conduir temeràriament i votar a certs partits. Concretament s'assenyalava als votants d'UPyD com els més disposats a infringir les regles de circulació mentre que els de CiU estarien entre els més respectuosos amb elles. A l'hora de buscar possibles causes d'aquesta sorprenent relació entre vot i conducció, l'últim paràgraf de l'article apuntava la possibilitat de que l'edat dels votants hi pogués tenir a veure (l'article deia que els votants d'UPyD són en mitjana més joves que els de PP, PSOE, CiU o CC, per exemple).

Farem servir les dades del Baròmetre de setembre de 2013 del CIS per comprovar si les afirmacions de l'article estaven ben fonamentades. Ens centrarem únicament en les dades corresponents als enquestats que van manifestar que a les eleccions de 2011 (`Recuerdo.de.voto.2011`) van votar o bé a UPyD o bé a CiU. Estudiarem la seva edat (variable `Edad`). Pel que fa a la temeritat en la conducció, ens limitarem a estudiar les respostes a la preguntes de si s'excedeix el límit de velocitat en ciutat, codificada així:

`1:Siempre, 2:Muchas veces, 3:Algunas veces, 4:Pocas veces, 5:Nunca.`

- En primer lloc volem comparar les edats esperades de votants d'UPyD i de CiU, a les que anomenarem μ_1 i μ_2, respectivament. En R definim els vectors `Edad.UPyD` i `Edad.CiU` que contenen les edats dels enquestats que han declarat haver votat el 2011 a UPyD o CiU, (respectivament). Les mides mostrals són, respectivament, $n_1 = 73$ i $n_2 = 65$, respectivament. A continuació oferim unes descriptives numèriques i gràfiques d'aquestes dues variables.

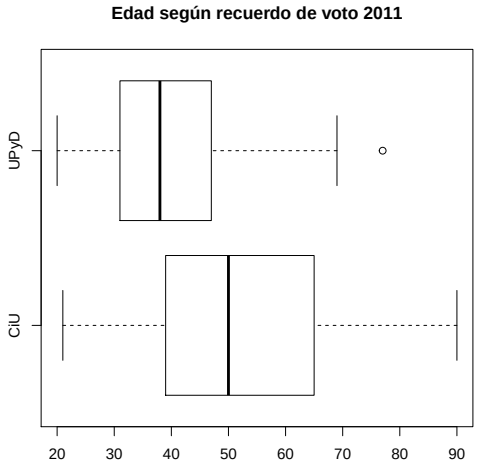
```
> summary(Edad.UPyD)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  20.00   31.00   38.00   39.66   47.00   77.00

> sd(Edad.UPyD)

[1] 13.43836

> summary(Edad.CiU)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  21.00   39.00   50.00   52.31   65.00   90.00

> sd(Edad.CiU)

[1] 17.61917
```

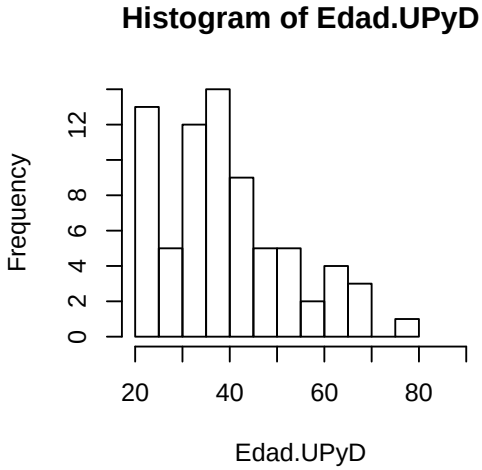
1. **(1 punt)** Es tracta de mostres aparellades o de mostres independents? Per què?
   SOLUCIÓ:
   Es tracta de mostres independents perquè els enquestats que es declaren votants d'UPyD són individus distints dels que es declaren votants de CiU. No hi ha cap relació ú a ú entre els uns i els altres. A més a més, les mides mostrals de les dues mostres són diferents.
2. **(2 punts)** Constrasteu la igualtat de les variàncies, σ_1^2 i σ_2^2, de les edats de votants d'UPyD i de CiU, respectivament. Indiqueu les hipòtesis nul·la i alternativa. Quin estadístic fareu servir per fer la prova? Quina distribució té sota la hipòtesi nul·la? Sota quina condició? Creus que aquesta condició es verifica en aquest cas? Doneu el valor de l'estadístic de la prova, digueu quin és el p-valor de la prova i si rebutjaríeu H_0 amb un risc $\alpha = 0.01$.
   Nota: Feu servir la següent informació (recordeu que `pf(x,g1,g2)` *és la funció de distribució d'una F amb graus de llibertat* `g1` *i* `g2` *avaluada al punt* `x`*):*

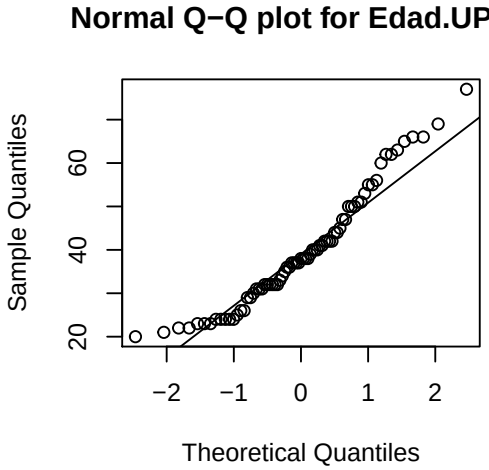
```
> pf(1.71901,72,64)

[1] 0.9858147

> 2*(1-pf(1.71901,64,72))

[1] 0.02610106

> pf(1.874246,64,72)

[1] 0.995

> pf(1.893433,72,64)

[1] 0.995
```

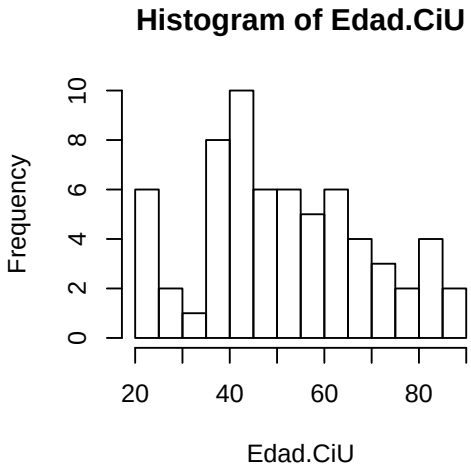
   SOLUCIÓ:
   La prova d'hipòtesi per contrastar igualtat de variàncies és

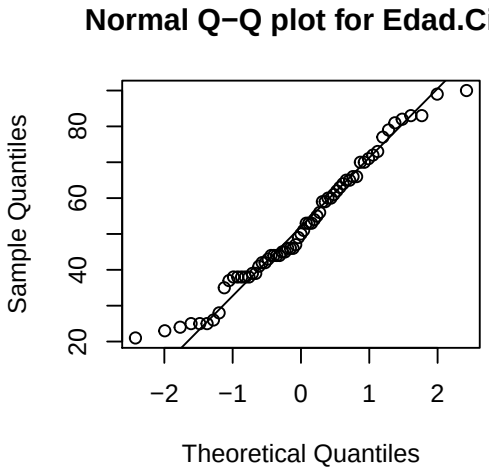
$$\begin{cases} H_0 & : & \sigma_1^2 = \sigma_2^2, \\ H_1 & : & \sigma_1^2 \neq \sigma_2^2. \end{cases}$$


Edad según recuerdo de voto 2011
UPyD
CiU
20
30
40
50
60
70
80
90


Histogram of Edad.UPyD
Frequency
0
2
4
6
8
12
20
40
60
80
Edad.UPyD


Normal Q−Q plot for Edad.UPyD
Sample Quantiles
20
40
60
−2
−1
0
1
2
Theoretical Quantiles


Histogram of Edad.CiU
Frequency
0
2
4
6
8
10
20
40
60
80
Edad.CiU


Normal Q−Q plot for Edad.CiU
Sample Quantiles
20
40
60
80
−2
−1
0
1
2
Theoretical Quantiles

Farem servir l'estadístic del test F: $\hat{F} = S_1^2/S_2^2$, que sota H_0 té distribució F_{n_1-1,n_2-1}, sempre que es donin aquestes condicions: les dues mostres són independents i provenen de v.a. normals. Al nostre cas les mostres són independents i els normal q-q plots indiquen que no hi ha una marcada falta de normalitat.
També podem fer servir $\hat{F} = S_2^2/S_1^2$, que sota H_0 té distribució F_{n_2-1,n_1-1}.
Si fem servir com a $\hat{F}$ el quocient de la variància mostral més gran entre la més petita (al nostre cas $\hat{F} = S_2^2/S_1^2$), la regió crítica (o regió de rebutg d'H_0) per a un risc $\alpha = 0.01$, és el conjunt de mostres per a les quals

$$\hat{F} > F_{n_2-1,n_1-1,1-(0.01/2)} = F_{64,72,0.995} = 1.874246,$$

donat que `pf(1.874246,64,72) = 0.995` o, de forma equivalent, que `2*(1-pf(1.874246,64,72)) = 0.01`.
Al nostre cas, l'estadístic de la prova val

$$\hat{F} = \frac{\texttt{sd(Edad.CiU)}^2}{\texttt{sd(Edad.UPyD)}^2} = \frac{310.4351}{180.5894} = 1.71901,$$

que és més petit que $F_{64,72,0.995} = 1.874246$ i, per tant, no rebutgem H_0 a nivel $\alpha = 0.01$.
Calculem el p-valor: $2\Pr(F_{64,72} > 1.71901) =$ `2*(1-pf(1.71901,64,72))` $= 0.02610106$, que és més gran que $\alpha = 0.01$ i, per tant, tornem a concloure per un altre camí que no rebutgem H_0 a nivel $\alpha = 0.01$.

3. **(3 punts)** Contrasteu ara la igualtat de les edats esperades dels votants d'UPyD i de CiU, μ_1 i μ_2. Indiqueu les hipòtesis nul·la i alternativa, tenint en compte el que diu la notícia del diari per decidir si fareu una prova bilateral o unilateral. Quin estadístic fareu servir per fer la prova? Quina distribució té sota la hipòtesi nul·la? Sota quines condicions? Es verifiquen aquestes condicions? Quina és la regió crítica de la prova amb un risc $\alpha = 0.01$? Doneu el valor de l'estadístic de la prova, digueu quin és el p-valor de la prova i si rebutjaríeu H_0 amb un risc $\alpha = 0.01$.
Nota: Aproximeu la distribució t_n per una $N(0,1)$ si els graus de llibertat n són més grans que 100.
SOLUCIÓ:
La prova d'hipòtesi per contrastar igualtat d'esperances és

$$\begin{cases} H_0 & : \quad \mu_1 = \mu_2, \\ H_1 & : \quad \mu_1 < \mu_2. \end{cases}$$

Com que, per l'apartat anterior, no hem rebutjat que les variàncies són iguals, farem servir l'estadístic del test t per a mostres independents i variàncies iguals:

$$\hat{t} = \frac{\bar{y}_1 - \bar{y}_2}{S\sqrt{(1/n_1) + (1/n_2)}}, \text{ on } S^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2},$$

que sota H_0 té distribució $t_{n_1+n_2-2}$, sempre que es donin aquestes condicions: les dues mostres són independents i provenen de v.a. normals amb la mateixa variància. Al nostre cas les mostres són independents, els normal q-q plots indiquen que no hi ha una marcada falta de normalitat i a l'apartat anterior no hem rebutjat que les variàncies són iguals.
La regió crítica (o regió de rebutg d'H_0) per a un risc $\alpha = 0.01$, és el conjunt de mostres per a les quals

$$\hat{t} < t_{n_1+n_2-2,\alpha} = t_{136,0.01} \approx z_{0.01} = -2.3266, \text{ (fent servir la taula de la } N(0,1)).$$

Al nostre cas,

$$S^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2} = 241.6933,\ S = 15.54649,\ \hat{t} = \frac{\bar{y}_1 - \bar{y}_2}{S\sqrt{(1/n_1) + (1/n_2)}} = -4.771361,$$

que és més petit que -2.3266 i, per tant, rebutgem H_0.
Calculem el p-valor de la prova, $\Pr(t_{136} < -4.771361) \approx \Pr(N(0,1) < -4.771361) \approx 0$ (fent servir la taula de la $N(0,1)$). Per tant, tornem a concloure per un altre camí que no rebutgem H_0 a nivel $\alpha = 0.01$.

4. **(2 punts)** Doneu un interval de confiança 99% per a la diferència de les edats esperades dels votants d'UPyD i de CiU.
   SOLUCIÓ:
   Fem servir la fòrmula

$$\mathrm{IC}_{1-\alpha}(\mu_1-\mu_2) \equiv \left((\bar{y}_1-\bar{y}_2) \pm t_{n_1+n_2-2,\alpha/2} S\sqrt{(1/n_1)+(1/n_2)}\right) \approx \left((\bar{y}_1-\bar{y}_2) \pm z_{\alpha/2} S\sqrt{(1/n_1)+(1/n_2)}\right),$$

   que amb les nostre dades val

$$\left((39.66-52.31) \pm 2.575 \times 15.54649\sqrt{(1/73)+(1/65)}\right) \equiv (-19.478, -5.823).$$

- Ara compararem els estils de conducció dels votants d'UPyD i de CiU. Les preguntes del baròmetre del CIS relatives a aquest tema només les havien de respondre aquells enquestats que habitualment condueixen un vehicle de motor. Hi va haver 63 votants d'UPyD que condueixen, i 54 de CiU.

  Definim la variable `Exc.lv.ciud` que val `TRUE` si un conductor enquestat declara que excedeix el límit de velocitat a la ciutat *sempre*, *moltes vegades* o *algunes vegades*, i val `FALSE` en altres casos. El creuament de les variables `Exc.lv.ciud` i `voto.2011.cond` (que indica quin partit van votar, UPyD o CiU, els votants d'aquests dos partits que condueixen habitualment) ve donat per la següent taula:

	Exc.lv.ciud	
voto.2011.cond	TRUE	FALSE
UPyD	25	38
CiU	12	42

5. **(2 punts)** Sigui p_1 (respectivament, p_2) la probabilitat que un conductor votant d'UPyD (respectivament, de CiU) triat a l'atzar excedeixi el límit de velocitat a la ciutat sempre, moltes vegades o algunes vegades. Contrasteu la igualtat de p_1 i p_2. Indiqueu les hipòtesis nul·la i alternativa, tenint en compte el que diu la notícia del diari per decidir si fareu una prova bilateral o unilateral. Quin estadístic fareu servir per fer la prova? Quina distribució aproximada té sota la hipòtesi nul·la? Quina és la regió crítica de la prova amb un risc $\alpha = 0.05$? Doneu el valor de l'estadístic de la prova, digueu quin és el p-valor de la prova i si rebutjaríeu H_0 amb un risc $\alpha = 0.05$.
   SOLUCIÓ:
   Les hipòtesis nul·la i alternativa són

$$\begin{cases} H_0 & : \quad p_1 = p_2, \\ H_1 & : \quad p_1 > p_2. \end{cases}$$

   Farem servir l'estadístic del test z per a comparació de proporcions en mostres independents:

$$\hat{z} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{(\hat{p}(1-\hat{p})/n_1) + (\hat{p}(1-\hat{p})/n_2)}}, \text{ on } \hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_1}{n_1+n_2},$$

   que sota H_0 té distribució aproximada $N(0,1)$.
   La regió crítica (o regió de rebutg d'H_0) per a risc $\alpha = 0.05$, és el conjunt de mostres per a les quals

$$\hat{z} > z_{1-\alpha} = 1.645, \text{ (fent servir la taula de la } N(0,1)).$$

   Al nostre cas,

$$\hat{p}_1 = \frac{25}{63} = 0.3968, \ \hat{p}_2 = \frac{12}{54} = 0.2222, \ \hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_1}{n_1+n_2} = \frac{37}{117} = 0.3162,$$

$$\hat{z} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{(\hat{p}(1-\hat{p})/n_1) + (\hat{p}(1-\hat{p})/n_2)}} = 2.024726,$$

   que és més gran que 1.645 i, per tant, rebutgem H_0 a nivel $\alpha = 0.05$.
   Calculem el p-valor de la prova, $\Pr(N(0,1) > 2.024726) = 1 - 0.9783 = 0.0217$ (fent servir la taula de la $N(0,1)$). Per tant, tornem a concloure per un altre camí que no rebutgem H_0 a nivel $\alpha = 0.05$.

- *L'últim punt de l'article que hauríem de comprovar és si en efecte és l'edat del votant (i no el sentit del seu vot) la variable que influeix en el seu estil de conduir, però això ja no ens hi cap a aquest examen!. Només us direm que aquesta és la conclusió a la que arribem si fem servir un model de regressió lineal amb resposta codificada d'1 a 5, i dues variables explicatives: edat i sentit del vot (UPyD o CiU).*

NOM:_____COGNOM:_____

(***Contesteu cada pregunta en el seu lloc. Expliciteu i justifiqueu els càlculs. Totes les preguntes valen igual***)

Problema 3 (B6).

El tema de B7 d'un grup d'estudiants consisteix en determinar si la distància al mòdem wifi és determinant en la velocitat de connexió. Per això, dissenyen un estudi en el que prenen la velocitat de descàrrega amb una pàgina estàndard, ubicant l'ordinador a diferents llocs de casa i mesurant la distància al mòdem en línia recta. Aquestes són les dades recollides:

```
D [m.]  0.80 1.100 2.000 2.100 3.300 3.400 4.600 4.800 5.000 5.000 5.000 5.600 5.700 5.800 7.400
V [MBs] 2.98 2.102 2.485 3.016 3.037 1.965 2.704 2.942 3.123 2.683 2.474 2.544 2.412 1.357 2.443
```

1. Amb R han estimat els paràmetres d'un model que relaciona "distància" amb "velocitat" però, ¿quina ha estat la variable resposta i quina la variable explicativa, entenent que han fet un model *sensat*? Justifiqueu si la situació oposada tindria més o menys sentit. El fragment de la sortida que hem inclòs a continuació us pot ajudar a contestar.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.75517    _______   9.048 5.67e-07 ***
x           -0.04968    0.06762  ______    0.476
```

El model *sensat* és la velocitat V com a variable resposta (Y) en funció de la distància D (X). La D és explicativa, perquè podem decidir la distància però no la velocitat, que és observada. La situació oposada és factible com a model per a predir la distància a la que es troba l'ordinador del mòdem segons la velocitat de descàrrega, però la relació causa-efecte que és molt clara a la primera situació aquí no té sentit. A la sortida podem fixar-nos que el terme independent (2.755) és proper a la mitjana dels valors de V, una indicació de que V seria a l'eix d'ordenades.

2. Trobeu els valors mancants de la sortida anterior.

El Std. Error de la primera fila és 2.75517 / 9.048 = 0.3045, perquè el "t value" és el quocient de l'estimador i el seu error tipus. Per la mateixa raó, el valor mancant de la segona fila és -0.73469.

3. Escriviu a continuació quines són les dues proves d'hipòtesis formals que s'estan contrastant a la sortida anterior. Expliqueu també què volen dir cadascuna d'elles:

- H_0: $\beta_0 = 0$ *vs* $\beta_0 \neq 0$. És a dir: posem a prova si la recta no té terme independent; equivalent a dir que la velocitat és proporcional a la distància; equivalent a dir que la recta passa per l'origen de coordenades.
- H_0: $\beta_1 = 0$ *vs* $\beta_1 \neq 0$. És a dir: posem a prova si la recta és horitzontal, o si el seu pendent és zero; equivalent a dir que la velocitat no depèn de la distància.

4. Corregiu/milloreu les afirmacions següents, o confirmeu si és que no contenen cap error:

- Cada metre que ens apropem al mòdem significa que la velocitat baixa 0.05MBs (aprox.)

Error: si ens apropem, **no**. Si *ens allunyem* 1m (la distància augmenta en 1), la velocitat baixa aprox. 0.05MBs.

- Si ens allunyem un metre del mòdem la velocitat augmenta en 2.76MBs (aprox.)

Error: o la frase anterior, o millor dir que 2.76 indicaria la velocitat esperada a 0m del mòdem

- Quan la distància al mòdem disminueix 5cm aconseguim una millora en la velocitat de 1MBs

Error: les unitats del número -0.04968 no són metres, sinó MBs per metre; la millora de velocitat per 5cm seria tan sols de 2.5 KBs.

- No es veu que la velocitat es modifiqui si la distància al mòdem canvia

Correcte: el p-valor 0.476 ens indica que l'alteració de la velocitat per causa de la distància no és significativa (l'error tipus és major que la magnitud de l'efecte).

5. Sabent que la mitjana de D és 4.106667, la de V és 2.551133, les respectives variàncies són 3.656381 i 0.2263794, i la correlació val -0.1996768, reproduïu el càlcul del valor `0.06762` de la taula anterior. Calculeu també l'estimació de la desviació residual i el coeficient de determinació corresponent a aquestes dades. Justifiqueu totes les passes.

El coeficient R^2 és el quadrat de la correlació, per tant: 0.03987

La variància residual val 14/13 (var(V)–b_1 cov(D,V)) . La covariància val –0.1996768 $\sqrt{3.656381 \;\; 0.2263794}$ = –0.181665

S^2 = 14/13 (0.2263794–(–0.04968)(–0.181665)) = 0.2340738

Per tant, la desviació residual S = 0.4838

Finalment, l'error tipus del pendent es calcula com $\sqrt{S^2/(n-1)\text{var}(D)} = \sqrt{\frac{0.2340738}{14\ 3.656381}}$ = 0.06762.

6. Construïu un interval de confiança al 99% per al pendent de la recta, i expliqueu què significa el que heu calculat.

IC(β_1, 99%) = –0.04968 ± $t_{13,\,0.995}$ 0.06762. De les taules, $t_{13,\,0.995}$ = 3.012

IC = (–0.2534, 0.1540)

Amb una confiança molt gran, que només deixaria fora un cas de cada 100, creiem que l'autèntic pendent és un valor entre –0.25 i 0.15 MBs/m. A la vista de les dades disponibles, incrementar la distància al mòdem pot representar un decrement de la velocitat però també un increment.

7. Un amic té molta por de les radiacions electromagnètiques que desprèn el mòdem, però pensa que si s'allunya massa està perdent velocitat. Creieu que aquests resultats (concretament, l'apartat 6) el poden ajudar per decidir on ha de col·locar l'ordinador, i perquè?

Les dades no confirmen que si s'allunya del mòdem estigui perdent velocitat. Però no poder rebutjar una hipòtesi no significa que sigui certa. Potser la prova no és prou potent per detectar un efecte real associat a la relació entre D i V. Actualment no podríem dir què seria millor, si estar lluny o seure aprop.

8. Com tothom diu la seva, un altre amic diu que aquest estudi no conclou res perquè la grandària de la mostra no és l'adequada. Que el que cal és que la precisió de l'estimació sigui alta i, per tant, que l'error estàndard de l'estimador del pendent no sigui superior a 0.025 (*ell sabrà perquè*). Assumint (a) que la desviació residual poblacional és 0.5 i (b) que la variabilitat de les distàncies preses al nou estudi serà més o menys igual: quantes observacions s'haurien d'obtenir?

Amb les suposicions (a) i (b) anteriors es pot afirmar que l'expressió:

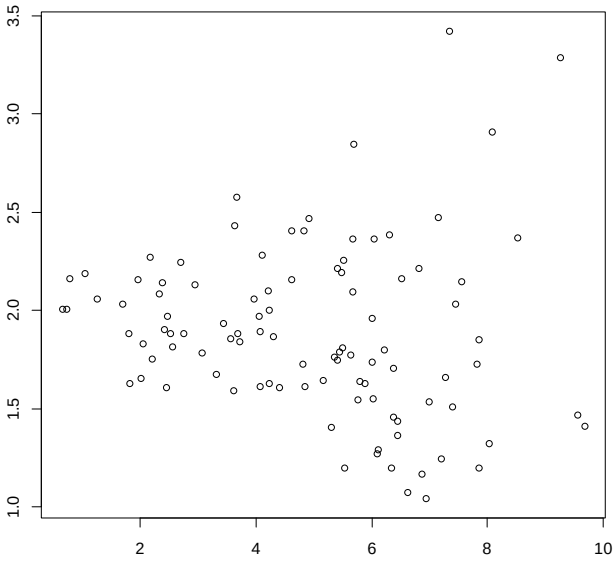
$\sqrt{S^2/(n-1)\mathrm{var}(D)}$ ha de valdre 0.025, tenint en compte que prenem S^2=0.25 i var(D) igual al valor anterior (3.656381).

Llavors, podem treure la n:

$\sqrt{\frac{0.25}{(n-1)\ 3.656381}} = 0.025.\ n = 1 + \frac{0.25}{0.025^2 3.656381} = 110.3978$

Necessitaríem 111 observacions.

9. Imagineu que l'estudi s'ha repetit, i hem obtingut el resultat adjunt. Expliqueu si seria correcte utilitzar el model lineal simple que coneixeu per relacionar les variables. Digueu quines assumpcions ("premisses") del model es podrien donar per vàlides i quines no.



Encara que aquest és el plot normal Y vs X, i no el plot dels residus contra els valors ajustats (o la X), es pot apreciar bé que alguna premissa està fallant.

La més clara és la homoscedasticitat: els valors de velocitat es dispersen molt més a distàncies grans (dreta) que a distàncies petites.

Per la premissa de linealitat, no es pot dir res en contra: aparentment la velocitat baixa quan la distància creix —o varia molt poc—, però no sembla que ho faci (si ho fa) amb pendents variables.

Si ens fixem bé, també es pot criticar la premissa de normalitat: les desviacions de la velocitat respecte el valor esperat poden ser majors en sentit positiu (cap a dalt) que en sentit negatiu (cap a baix). Llavors, aquesta absència de simetria és una prova en contra de la normalitat dels residus.

Respecte a la premissa d'independència no es pot dir res, perquè no coneixem els detalls de l'estudi i com s'ha planificat, i perquè aquest gràfic no ens proporciona cap pista, tal com la seqüència de les observacions en ordre de recollida.