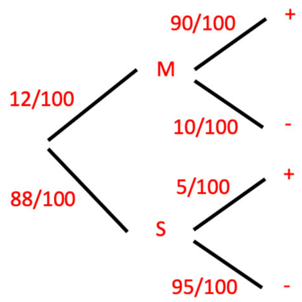
NOM:_________________________COGNOM:____

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

Problema 1 (A)

En una població on la malaltia X és endèmica, se sap que la prevalença de la malaltia és d'un 12% (és a dir, el 12% de la població té la malaltia). Es disposa d'una prova per a detectar la malaltia, però no és totalment fiable, ja que, la sensibilitat del test és del 90% (és a dir, dóna positiu en un 90% dels casos de persones realment malaltes); però dóna un 5% de falsos positius (és a dir, dóna erròniament positiu en un 5% de persones sanes).

1. Dibuixa l'arbre associat a aquest esdeveniment (0.5 punts)



Nota: M: Malalt; S: Sa; +: Positiu; -: Negatiu

2. Si una persona està malalta, quina és la probabilitat que el test li hagi sortit negatiu (fals negatiu)? (0.5 punts)

P(-|M) = 10/100 = 0.1

3. Es demana la taula de probabilitats conjuntes i marginals. A més, també es demana desenvolupar formalment el càlcul de la probabilitat que el test sigui negatiu? (2 punt)

	+	-	Total
Malalt	12/100·90/100=0.108	12/100·10/100=0.012	0.12
Sa	88/100·5/100=0.044	88/100·95/100=0.836	0.88
Total	0.152	**0.848**	1

$$P(-) = P(- \cap M) + P(- \cap S) = P(-|M)P(M) + P(-|S)P(S) = 0.1 \cdot 0.12 + 0.95 \cdot 0.88 = 0.848$$

4. Es demana la taula amb les probabilitats condicionades al resultat del test: P(M|+), P(M|-), P(S|+) i P(S|-). Expliqueu breument el resultat que heu trobat. (2 punt)

	+	-
Malalt	P(M|+) = 0.108/0.152=0.711	P(M|-) = 0.012/0.848=0.014
Sa	**P(S|+) = 0.044/0.152=0.289 (pregunta 6)**	P(S|-) = 0.836/0.848=0.986
Total	1	1

5. L'especificitat del test és la capacitat del test de donar negatiu en persones sanes. Quina és l'especificitat? (0.5 punts)

Especificitat = P(-|S) = 1 - P(+|S) = 1 - falsos positius = 1 - 0.05 = 0.95

6. Quina és la probabilitat que una persona estigui sana si li ha sortit el test positiu? (1 punt)

$$P(S|+) = \frac{P(+|S)P(S)}{P(+)} = \frac{P(+|S)P(S)}{P(+\cap S) + P(+\cap M)} = \frac{P(+|S)P(S)}{P(+|S)P(S) + P(+|M)P(M)}$$
$$= \frac{{}^{5}/_{100} \cdot {}^{88}/_{100}}{{}^{5}/_{100} \cdot {}^{88}/_{100} + {}^{90}/_{100} \cdot {}^{12}/_{100}} = \frac{0.044}{0.152} = 0.2895$$

O anàlogament, veure taula de la pregunta 4.

7. Hi ha dependència o independència entre estar malalt i el resultat del test (justifica la resposta)? (0.5 punts)

Per exemple,

$P(Malalt \cap +) \neq P(Malalt) \cdot P(+)$

$0.108 \neq 0.12 \cdot 0.152 => Dependents$

En aquesta població hi ha 2 hospitals, X i Y. Aquesta és la funció de probabilitat conjunta en relació al nombre de pacients que ingressen setmanalment:

	Y = 0	Y = 1	Y = 2
X = 0	0.02	0.04	0.35
X = 1	0.03	0.00	0.10
X = 2	0.40	0.05	0.01

8. Quant valen les seves esperances i variances? (1.5 punt)

E(XY) = 1·2·0.10 + 2·1·0.05 + 2·2·0.01 = 0.34

E(X) = 1·(0.03+0.00+0.10) + 2·(0.40+0.05+0.01) = 1.05

E(X^2) = 1·(0.03+0.00+0.10) + 4·(0.40+0.05+0.01) = 1.97

E(Y) = 1·(0.04+0.00+0.05) + 2·(0.35+0.10+0.01) = 1.01

E(Y^2) = 1·(0.04+0.00+0.05) + 4·(0.35+0.10+0.01) = 1.93

$$\sigma_X = \sqrt{E(X^2) - E(X)^2} = \sqrt{1.97 - 1.05^2} = 0.9314$$

$$\sigma_Y = \sqrt{E(Y^2) - E(Y)^2} = \sqrt{1.93 - 1.05^2} = 0.9539$$

9. Es demana la correlació entre X i Y i interpretar el resultat? (1 punt)

Cov(X,Y) = E(XY)-E(X)E(Y) = 0.34 – 1.05·1.01 = -0.7205

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{-0.7205}{0.9314 \cdot 0.9539} = -0.8110$$

Hi ha una forta correlació lineal entre el nombre d'ingressos als dos hospitals. La correlació és negativa indicant que quan incrementen els ingressos hospitalaris a l'hospital X, decreixen a l'hospital Y, i viceversa.

10. Quina és la distribució de la probabilitat del nombre d'ingressos a l'hospital Y en una setmana que ha ingressat a X només un pacient? (0.5 punt)

$$P(Y|X=1) = \frac{P(Y \cap X=1)}{P(X=1)} = \frac{P(Y \cap X=1)}{0.13}$$

k	**P(Y|X=1)**
0	0.03/0.13=0.231
1	0
2	0.10/0.13=0.769

NOM:____

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

Problema bloc B

Mentre que mesures sofisticades de control i seguretat en aeroports estan pendents de temes legals, de moment es segueix funcionant, entre altres, amb arcs de seguretat i control del pes dels equipatges. Contesteu les següents preguntes indicant, per a totes les variables necessàries, la definició i model probabilístic (també en les que no es posa explícitament)

Per una part, se sap que al passar per l'arc de seguretat sona una alarma a una mitjana de 5 passatgers per hora. Es demana:

- (1 punt) Definiu unes variables pel nombre de passatgers als que sona l'alarma: una variable de nombre per hora i una altra de nombre per minut. I definiu unes variables pel temps entre passatgers als que sona l'alarma: una pel temps en hores i una altra pel temps en minuts Indiqueu els models probabilístics corresponents i els seus paràmetres

Nh és "nombre de passatgers als que sona l'alarma en una hora" Nh és Poisson amb λ=5

Nm "és nombre de passatgers als que sona l'alarma en un minut" Nm és Poisson amb λ =5/60=1/12=0.0833

Th és "nombre d'hores entre passatgers als que sona l'alarma" Th és Exponencial amb λ =5

Tm és "nombre de minuts entre passatgers als que sona l'alarma" Tm és Exponencial amb λ =5/60=1/12=0.0833

- (1 punt) Calculeu la probabilitat que, en una hora, a cap passatger li soni l'alarma i la probabilitat que li soni com a màxim a 1 passatger

$P(Nh=0) = 5^0 \exp(-5) / 0! = \exp(-5) = 0.0067$

$P(Nh\leq 1) = P(N=0)+P(N=1) = \exp(-5) + (5^1 \exp(-5) / 1!) = \exp(-5)+5\exp(-5) = 0.04$

- (1 punt) Calculeu la probabilitat d'estar més de 10 minuts sense que soni l'alarma, i la probabilitat d'estar-ne més de 10 sense que soni l'alarma quan porta almenys 5 minuts sense sonar

$P(Tm>10) = 1-P(Tm<=10) = 1 – (1-\exp(-10/12)) = \exp(-10/12) = 0.4346$

$P(Tm>10 \mid Tm>5) = P(Tm>10,Tm>5) / P(Tm>5) = P(Tm>10) / P(Tm>5) = \exp(-10/12) / \exp(-5/12) =0.4346/0.66 = 0.66$
$= P(Tm>5) = \exp(-5/12) = 0.66$

- (1 punt) Quin és el temps en minuts que es pot esperar sense que soni l'alarma de l'arc de seguretat? (Justifiqueu el càlcul)

$E(Tm) = 1/(5/60) = 60/5 = 12$ minuts

Per una altra part, se sap que el pes de l'equipatge dels passatgers segueix una distribució Normal de mitjana 24 kg i variància 4 kg^2 Es demana:
- (1 punt) Calculeu la probabilitat que l'equipatge d'un passatger passi de 25 kg

Pes és Normal amb mu=24kg i sigma=sqrt(4)= 2 kg

P(Pes>25) = P(Z>(25-24)/2) = P(Z>0.5) = 1-P(Z<=0.5) = 1-pnorm(0.5) = 1-0.6914625 = 0.3085375 = 0.3

- (1 punt) Calculeu la probabilitat que en un grup de 10 passatgers n'hi hagi 2 amb equipatge amb pes superior a 25 kg

Np és "nombre de passatgers amb equipatge de pes superior a 25 kg" Np és Binomial amb n=10 i p=0.3

P(Np=2) = $\binom{10}{2}$ 0.3^2 0.7^8 = (10! / 2! 8!) 0.3^2 0.7^8 = (10*9/2) 0.3^2 0.7^8 = 45*0,09*0.05765 = 0.233

- (1 punt) Calculeu el nombre esperat de viatgers que passaran fins que un tingui equipatge superior a 25 kg

Pfins és "nombre de passatgers fins un amb equipatge de pes superior a 25 kg" Pfins és Geometrica amb p=0.3

E(Pfins) = 1/p = 1/0.3 = 3.33 (entre 3 i 4 passatgers)

- (1.5 punts) Calculeu la probabilitat que en el grup de 10 passatgers el pes mitjà de l'equipatge superi els 25 kg

PesMig pel TCL és Normal amb mu=24 kg i sigma=2/sqrt(10) = 0.63 kg

P(PesMig>25) = P(Z>(25-24)/0.63) = P(Z>1.6) = 1-P(Z<=1.6) = 1-pnorm(1.6) = 1-0.9452007 = 0.0547993 = 0.055

- (1.5 punts) Calculeu el pes màxim que podem observar amb un error del 10%

PesTotal pel TCL és Normal amb mu=24*10=240 kg i sigma=2*sqrt(10) = 6.3 kg

P(PesTotal > max) = 0.10
P(PesTotal <= max) = 0.9 P(Z<=(max-240)/6.3) = 0.9 (max-240)/6.3 = qnorm(0.9) = 1.281552 max=(1.281552*6.3)+240
max=248.074

pnorm(0.1) = 0.5398278	pnorm(0.5) = 0.6914625	pnorm(0.6) = 0.7257469	pnorm(0.9) = 0.8159399	pnorm(1.0) = 0.8413447	pnorm(1.6) = 0.9452007
qnorm(0.1) = -1.281552	qnorm(0.2) = -0.8416212	qnorm(0.5) = 0	qnorm(0.6) = 0.2533471	qnorm(0.8) = 0.8416212	qnorm(0.9) = 1.281552

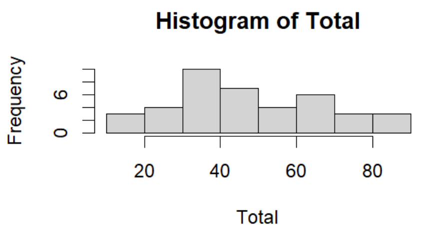
NOM:____________________________COGNOM:____

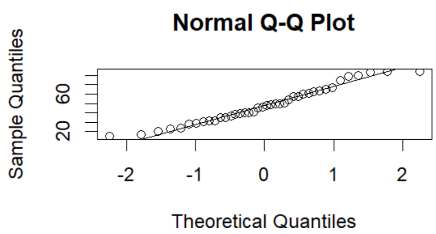
Contesteu cada pregunta en el seu lloc. Expliciteu i justifiqueu els càlculs.

Problema 3 (Bloc C)

A Catalunya cada mes de març més de cent mil alumnes participen a les proves Cangur de resolució de problemes. Hi participa alumnat des de 5è de primària fins a 2n de batxillerat i cicles formatius. Per l'edició de 2023, s'ha escollit una mostra a l'atzar de 40 participants de 5è de primària i s'han obtingut les següents dades sobre la puntuació Total (T):

$$\sum_{i=1}^{40} t_i = 1920{,}5 \;\text{ i }\; \sum_{i=1}^{40} t_i^2 = 107.111{,}5$$





1. A partir de les dades anteriors, doneu una estimació puntual de la mitjana de puntuació Total (T) de la prova de 5è de primària. Quin és l'error estàndard d'aquesta estimació? (1,5 punts)

$\bar{T} = \frac{1920{,}5}{40} = 48{,}0125$ i $s_T^2 = \frac{107111{,}5 - \frac{1920{,}5^2}{40}}{39} = 382{,}1409$, $s_T = 19{,}5484$

I per tant, se = 19,55/ $\sqrt{40}$ = 3,0909

2. A partir dels gràfics, argumenteu si podem suposar que la puntuació Total (T) de la prova de cinquè de primària segueix una distribució normal. (1 punt)

A partir de l'histograma i del qqnorm podem suposar que la puntuació Total segueix una distribució normal. L'histograma s'ajusta prou bé a la campana de Gauss de la distribució normal i en el gràfic qqnorm veiem també un bon ajustament de les dades respecte la qqline que marca els valors teòrics del model normal.

3. A continuació podeu trobar dos outputs d'R. Un correspon a l'IC de la mitjana de la puntuació total amb un 90% de confiança i l'altre amb un 99%. Indiqueu de manera raonada quin és cadascun (1 punt)

OUTPUT A - One Sample t-test	OUTPUT B -One Sample t-test
data: Total t = 15.534, df = 39, p-value < 2.2e-16 alternative hypothesis: true mean is not equal to 0 ?? percent confidence interval: 42.80476 53.22024 sample estimates: mean of x ??	data: Total t = 15.534, df = 39, p-value < 2.2e-16 alternative hypothesis: true mean is not equal to 0 ?? percent confidence interval: 39.64267 56.38233 sample estimates: mean of x ??

IC(μ_T , 90%) = (42.80476, 53.22024)

IC(μ_T , 99%) = (39.64267, 56.38233)

En augmentar la confiança, el valor del quantil augmenta, fent que l'interval tingui més amplitud. Per això l'IC amb 99% correspon a l'output B. Més conceptualment, es pot argumentar que en augmentar la confiança, reduïm el risc i per això augmenta la incertesa i el rang de l'interval.

4. Calculeu un interval de confiança (IC) del 95% per a la mitjana de la puntuació Total de la prova de cinquè de primària. Recordeu justificar les premisses. (1 punt)

A partir de l'apartat anterior, tenim que T~ Normal. Com que la variància poblacional és desconeguda, es té que

IC(μ$_T$, 95%) = ($\bar{T} - t_{n-1,1-\frac{\alpha}{2}} \cdot \sqrt{\frac{s^2}{n}}$, $\bar{T} + t_{n-1,1-\frac{\alpha}{2}} \cdot \sqrt{\frac{s^2}{n}}$) =

En el nostre cas es té que $t_{n-1,1-\frac{\alpha}{2}}$ = $t_{39,0'975}$ = 2,0227

= $\left(48{,}0125 - 2{,}0227 \cdot \sqrt{\frac{19{,}5484^2}{40}}\right.$, $\left.48{,}0125 + 2{,}0227 \cdot \sqrt{\frac{19{,}5484^2}{40}}\right)$ = (41,7606, 54,2644)

Ara s'afegeixen també les dades del nivell de 2n de batxillerat perquè es vol estudiar si hi ha diferències entre els dos nivells respecte la variabilitat de les dades. En concret, s'escull també una mostra de 40 alumnes de 2n de batxillerat en què la seva puntuació total (TB) segueix una distribució normal i que té 172,1458 per variància mostral.

5.- Calculeu l'interval de confiança del 95% per al quocient de variàncies entre el total de puntuació de la prova de cinquè de primària i 2n de batxillerat. Argumenteu si podem concloure que tenen o no la mateix variància. Interpreteu en termes de desviacions tipus. (2 punts)

$$IC\left(\frac{\partial_T^2}{\partial_{TB}^2}, 95\%\right) = \left(\frac{\frac{s_T^2}{s_{TB}^2}}{F_{(nT-1,nTb-1),1-\frac{\alpha}{2}}}, \frac{\frac{s_T^2}{s_{TB}^2}}{F_{(nT-1,nTb-1),\frac{\alpha}{2}}}\right) = \left(\frac{\frac{382{,}1409}{172{,}1458}}{F_{(39,39),0{,}975}}, \frac{\frac{382{,}1409}{172{,}1458}}{F_{(39,39),0{,}025}}\right)$$

$$= \left(\frac{2{,}2199}{1{,}8907}, \frac{2{,}2199}{0{,}5289}\right) = (1{,}1740,\ 4{,}1972)$$

A partir de les mostres recollides, no és versemblant que les puntuacions totals de les proves de cinquè de primària i de 2n de batxillerat tinguin la mateixa variància. En concret, amb confiança del 95%, la desviació tipus a primària seria entre 1.08 i 2.04 vegades més gran que la de batxillerat.

En el nivell de 2n de batxillerat la prova consta de tres parts amb deu problemes a cada part. A més de la puntuació total, per cada alumne es recull la puntuació de cadascuna de les parts sobre un puntuació de 10 punts. Es vol estudiar la diferència de mitjanes entre la puntuació obtinguda a la primera (P1) i a la tercera part (P3).

6.- Justifiqueu que s'ha fet un disseny amb dades aparellades. (1 punt)

Es tracta d'un disseny aparellat perquè per cada alumne s'han recollit les dues puntuacions, la puntuació de la primera part i la de la tercera.

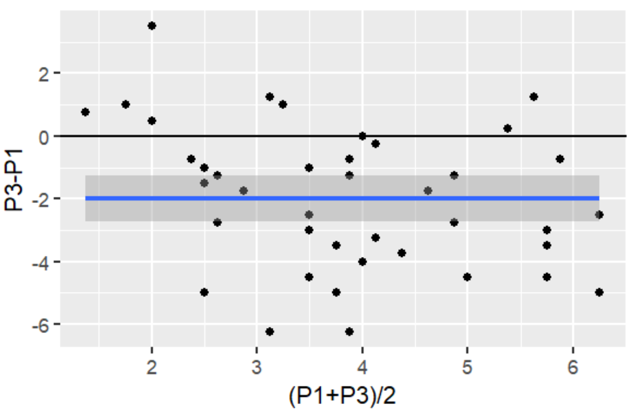
7.- Calculeu l'IC de la diferència de mitjanes (D) entre la puntuació obtinguda a la primera (P1) i a la tercera part (P3) amb D=P3-P1 i una confiança del 95% (1,5 punts)

```
> sum(D)      [1] -79.25              > sum(D^2)     [1] 359.6875
```

$\bar{D} = \frac{-79.25}{40} = -1.9813$ i $s_T^2 = \frac{359.6875 - \frac{79.25^2}{40}}{39} = 5{,}1968$, s_T = 2,2796. I per tant, se = 2,2796/ $\sqrt{40}$ = 0,3604

IC(μ$_D$, 95%)= ($\bar{D} - t_{n-1,1-\frac{\alpha}{2}} \cdot \sqrt{\frac{s^2}{n}}$, $\bar{D} + t_{n-1,1-\frac{\alpha}{2}} \cdot \sqrt{\frac{s^2}{n}}$) = (-1.9813 - 2,0227·0,3604, -1.9813 + 2,0227·0,3604) = (-2.7103, -1.2523)

8. Indica quin gràfic hi ha a continuació i interpreteu-ho en el context de l'estudi amb disseny amb dades aparellades realitzat (1 punt)



És el gràfic de Bland-Altman, o de diferències versus mitjana amb dades aparellades.
La qualificació de la tercera pregunta sembla inferior a la de la primera (-2.7, -1.3) per a qualsevol condició de l'alumne (nota mitjana alta o baixa), encara que a la banda esquerra els resultats podrien estar més igualats. Tampoc es pot descartar que sigui una casualitat de la mostra.

Problema D (Solució)

1) Primer, escriviu una proposta explícita de l'objectiu formal de l'estudi, assenyalant quin és el paper de cadascuna de les variables. Quines són quantitatives, i quines són categòriques?

   L'objectiu és determinar com el rendiment acadèmic, mesurat amb la variable Q, està relacionat amb el temps de trajecte i/o amb el mode de transport (públic o privat). La variable resposta serà Q; M i T seran variables explicatives.

   Q i M són variables quantitatives i T és variable qualitativa o categòrica.

2) És un estudi experimental o observacional? Si és experimental, digueu com s'han assignat les unitats a la intervenció. Si no ho és, justifiqueu la vostra opció.

   És observacional, els autors demanen als companys si volen participar, i els que accepten proporcionen les seves dades, que no poden ser assignades pels investigadors (no decideixen com viatgen ni quant triguen els seus trajectes).

3) Mireu el Quadre 1: expliqueu l'estudi estadístic realitzat i interpreteu els resultats obtinguts.

   És un test estadístic per comparar les mitjanes en el temps de trajecte de dos grups independents, 'pub' i 'priv', amb la mateixa variància. L'interval de confiança al 95% és [20.3, 29.5], és a dir: el temps mitjà en transport públic és de 20 a 30 minuts superior al temps mitjà en privat.

   Observació: En aquesta pregunta no es demanava pel compliment de les premisses. En aquest cas, és una enquesta i no sabem com s'han seleccionat els participants (veure darrera pregunta extra per més detalls sobre les limitacions de l'estudi), per tant no hi ha garanties de que les mostres siguin independents realment, o cada mostra sigui una m.a.s.

4) Quins són els dos models estadístics darrera els resultats presents als quadres 2 i 3?

   El primer model seria $Q = \beta 0 + \beta 1\ M + \varepsilon$, (Model lineal simple)
   i el segon $Q = \beta 0 + \beta 1\ M + \beta 2\ T + \varepsilon$ (Model lineal múltiple)

5) Quin és el significat dels estimadors que surten a la línia "Intercept" ([a] i [b] / [h] i [i])?

   [a] = 4.094 és l'estimació del terme independent del primer model, és a dir, quina és la mitjana de la qualificació quan el temps és nul. [b] = 0.498 és l'estimació de l'error tipus, per tant la seva incertesa és de gairebé mig punt.

   [h] = 5.85 és l'estimació del terme independent del segon model, mitjana de Q amb temps nul i la categoria de referència de T que és 'pub': És a dir, si viatges amb transport públic i el temps és 0 la mitjana de la qualificació és 5.85, amb error tipus 0.87.

   Els IC aproximats respectius serien (3.12, 5.09) i (4.14, 7.56)

6) Comenteu els indicadors de la línea "M" ([c], [d] i [e] / [j], [k] i [l]). Expliqueu les implicacions que se'n deriven, amb referència al grau d'evidència estadística que poden aportar.

   Aquí tenim el pendent del model perquè M és una variable quantitativa, per tant ens diu l'increment esperat de Q amb un minut més de temps de trajecte. Pel primer cas, l'increment és 0.029 amb error tipus 0.01; pel segon, 0.0017 amb ET 0.015.

   Els corresponents valors de "t value" ens diuen que al primer la ràtio senyal/soroll (2.8) és *rellevant*, però no al segon (0.1). Sembla que el primer model ens explica que el temps està relacionat amb la qualificació, i amb el segon no ho podem afirmar.

7) Quina informació porta la línia del Quadre 3 "Tpriv". Segons aquest resultat, quina interpretació fem en el àmbit de l'estudi?

   Aquest punt ens diu que si la categoria fos 'priv' en lloc de 'pub' la qualificació Q disminuiria 1.21 punts (amb ET=0.5), per tant és un canvi important i no atribuible a l'atzar simplement. Amb el segon model, que incorpora les dues variables M i T, veiem que el temps de trajecte no és un factor rellevant però el mode de transport sí que ho pot ser.

8) Expliqueu breument la variació que s'ha produit entre els indicadors [f] i [p], i entre [g] i [q]. Per què l'un ha disminuit i l'altre ha augmentat?

Perquè el segon model afegeix una variable explicativa més, llavors la fracció de la variabilitat de Q que es pot explicar augmenta i per tant disminueix la magnitud de la desviació residual (que és la que no s'explica pels factors del model)

9) Escriviu una breu conclusió global: què creiem que està relacionat amb el rendiment escolar i quins arguments (estadístics) tenim?

Segons hem vist al primer test, els dos modes de transport són diferents en la mitjana del temps (es triga més en públic). Per tant, quan considerem només M al model, els que tenen temps més baixos són probablement els que van en transport privat, que també són els que tenen, per la raó que sigui, pitjors qualificacions. De manera que es podria observar una recta amb un cert pendent. En canvi, quan T s'incorpora al model, aquest factor pot explicar per sí mateix les variacions de Q, mentre que dins d'un grup de T el temps M ja no és un factor explicatiu, ja que el pendent és irrellevant.

EXTRA: Crítica sobre la metodologia científica d'aquest estudi.

La crítica principal és com s'han seleccionat els participants (no es sap quin ha estat el mètode, però es pot suposar que no ha estat a l'atzar). Possiblement molts han rebutjat participar, i potser els elegits no són independents entre sí: imagineu que els autors de l'estudi han demanat a les seves amistats del seu cercle, el qual segurament no és representatiu de l'entorn.

També es pot dir que una anàlisi que inclou dos factors que estan fortament relacionats entre ells (com el tipus de transport i el temps de trajecte) no és una bona opció. Podríem considerar un model de Q en funció només de T, que potser seria el model més simple i complet.