NOM **PE**, 09/01/2012

(Poseu el nom i contesteu cada pregunta en el seu lloc reservat. Expliciteu i justifiqueu els passos en les respostes)

Problema B4

Abans de llençar al mercat la nova polsera *HyperPowerBalance*, un equip d'assessors comercials tracta de veure les oportunitats d'èxit del producte. La campanya publicitària ha divulgat les seves suposades virtuts, i es pretén calcular quina proporció de la població analitzada creu que pot ser eficaç.

Preg. 1. S'admet que aquesta proporció podria ser de fins a un 30%, en el cas més optimista. ¿Quina grandària hauria de tenir la mostra destinada a estimar la proporció real, si volem que el interval al 95% de confiança tingui una precisió del 3% (o una amplitud total de 6%)? **[1pt.]**

p=0.30; p*(1–p)*(1.96/0.03)^2

= 896.3733, solució: **897**

Preg. 2. Després, només es poden recollir 300 enquestes, de les quals 72 persones manifesten que sí creuen en l'eficàcia de la polsera. Amb aquestes dades, feu formalment la prova d'hipòtesis per comprovar si hi ha a la població una proporció del 30% com a possible *target* del producte. Considereu una prova unilateral, amb risc α=5%, i les fases habituals: **[3pt.]**

- Hipòtesis
- Estadístic de la prova
- Distribució de l'estadística anterior sota la hipòtesi nul·la
- Premisses
- Càlculs
- P-valor
- Conclusió i valoració del grau d'evidència present a la mostra
- Interval de confiança (no cal que sigui unilateral) i interpretació global.

- Hipòtesis, $\pi = 0.30$ vs $\pi < 0.30$
- Estadístic de la prova, $z = (p - \pi) / \text{sqrt}(0.30*0.70/300)$
- Distribució de l'estadística anterior sota la hipòtesi nul·la, $z \sim N(0, 1)$
- Premisses, m.a.s., n gran, p no excessivament petita, la aproximació a la normal és raonable
- Càlculs,
  p = 72/300; (p – 0.30) / sqrt(0.30*0.70/300)
  = –2.267787
- P-valor, prenent z = –2.27, P(Z < z) = 0.0116
- Conclusió i valoració del grau d'evidència present a la mostra,
  Podem rebutjar la hipòtesi nul·la, hi ha evidència per creure que a la població la proporció de creients de les virtuts de la polsera és menor del 30%
- Interval de confiança i interpretació global,
  p + c(–1, 1)*1.96*sqrt(0.30*0.70/300)
  = (0.188, 0.292)
  Potser aquesta proporció es mou entre el 18.8% i el 29.2%. Com la mostra és més petita que el càlcul inicial, tenim un rang massa ampli, i no sabem si en tot cas suposa un *target* suficientment atractiu.

Preg. 3. Posem que a la prova anterior s'ha obtingut un P-valor igual a 1/100. Valoreu aquestes afirmacions, i reescriviu -les, si és el cas. **[1pt.]**

- La probabilitat que la hipòtesi alternativa sigui certa és el P-valor = 1/100
- La probabilitat que una altra mostra de 300 persones obtingui 72 respostes o més favorables a la polsera és el P-valor = 1/100.

La primera no és certa en absolut. De fet, no té sentit parlar d'aquesta probabilitat.
La segona no està tan malament, però hauria de dir: "...una altra mostra de 300 persones obtingui 72 respostes o menys favorables a la polsera és 1/100, tenint en compte que la proporció a la població és del 30%" (el P-valor es calcula assumint que la hipòtesi nul·la és certa).

Preg. 4. Una altra pregunta a la enquesta demanava l'edat de la persona, i sabem que el interval per a la mitjana poblacional al 95% de confiança per al grup de dones que treballen fora de casa (n=27) és (33.00, 38.54) anys. Es demana que trobeu el interval al 99% de confiança. **[1pt.]**

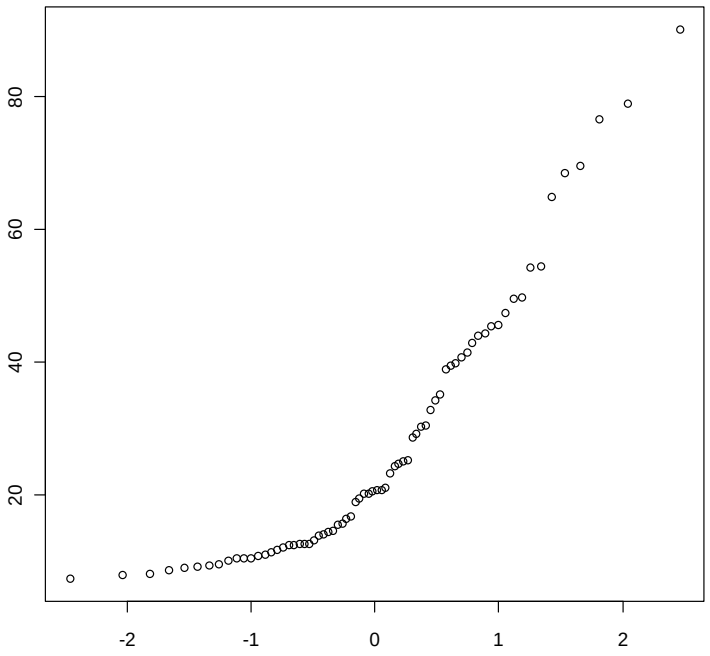
m = 35.77
s = (35.77 – 33) / 2.056 * sqrt(27) = 7.00
IC(μ, 99%) = m ± 2.779 * s/sqrt(27) = (32.03, 39.51)

Preg. 5.. Valoreu la següent afirmació, i corregiu-la si és el cas. **[1pt.]**

- Creiem que 99 de cada 100 dones que treballen fora de casa té una edat entre 32 i 39.5 anys, per a la població d'aquest estudi.

El interval no és para la variable 'edat', és un interval on creiem amb molta confiança que es pot trobar la mitjana real de l'edat d'aquest col·lectiu. Com no sabem com es distribueix l'edat, no és fàcil dir on es troba el 99% de casos individuals (si fos normal, i si equiparem els valors mostrals a paràmetres, llavors el rang aniria de 18 a 54 anys, aproximadament).

Preg. 6. A l'enquesta, una de les preguntes era "quin creus que serà el preu que costarà la nova *HyperPowerBalance*?". L'empresa tenia la idea de vendre-les a 40 euros, i aquesta qüestió tenia el propòsit d'esbrinar si les expectatives de la població estaven molt allunyades de la proposta de preu.
a) A la dreta teniu el `qqnorm()` del preu (considerant únicament les 72 persones que van respondre que podia ser eficaç): què opineu sobre aquesta premissa? b) Ara imaginem que "es sap" que la distribució és normal. El preu mitjà a la mostra de 72 respostes és de 27.70€, i la desviació tipus mostral de 19.70€. Diríeu que en mitjana la gent creu que la polsera hauria de costar menys de 40€? Feu una prova d'hipòtesis formal (resumida), i exposeu la conclusió. **[2pt.]**



a) Doncs que no sembla que la mostra provingui d'una variable normal. Els punts haurien de formar una línia recta, i estan desplaçats a l'esquerra, cap a preus baixos.

b)
μ = 40 vs μ <40
$t = (m - \mu) / (s/\sqrt{n}) \sim t_{71}$, es pot veure com N(0,1).
m.a.s., la variable no és normal, però la mostra sembla prou gran (72 observacions permet assumir que la mitjana mostral serà gairebé normal)
t = –5.30
Es pot rebutjar que μ = 40 de forma contundent (P-valor << 0.0001)
En forma de IC(95%), el preu esperat es troba entre 23.15€ i 32.25€.

Preg 7. Hauríeu aconsellat preguntar sobre el cost de la polsera a les 300 persones enquestades, en lloc de només a aquest grup de 72? Doneu arguments a favor i en contra (penseu sobre tot en arguments estadístics). **[1pt.]**

A favor, podríem pensar que sempre és millor disposar d'una mostra més gran, i que seleccionar només una fracció del total pot esbiaixar el resultat. En contra, si una persona no té interès en adquirir la polsera li donarà un preu arbitrari, potser nul, potser molt gran, de manera que la variable recollida a la totalitat de la mostra tindrà molta més variabilitat, sense cap benefici real per l'anàlisi posterior. Si més del 75% dels valors de la mostra són arbitraris, hi haurà massa soroll pertorbant la informació interessant.

NOM:_____

(Poseu el nom i contesteu cada pregunta en el seu lloc reservat. Expliciteu i justifiqueu els passos en les respostes)

Problema B5

a) Uns estudiants de la FIB volen verificar si dos programes d'ordenació (X, Y) triguen el mateix. Mitjançant base de dades diferents, però de dimensions similars, executen 21 vegades cada programa, cadascun amb la base de dades corresponent.

Els resultats que s'han obtingut en programa són:

Variable	Mitjana	Variància	N
X	10.38	2.52	21
Y	10.95	1.90	21

a.1 (0.5 punts) D'acord amb l'enunciat, es tracta d'un disseny "aparellat" o "independent"? Raoneu la resposta.

Independent, ja que son dues bases de dades independents, sense cap tipus de connexió entre elles.

a.2 (3 punts) Es pot admetre que la variàncies en ambdues casos és la mateixa? Per respondre aquesta pregunta, plantegeu una prova d'hipòtesi, indicant:

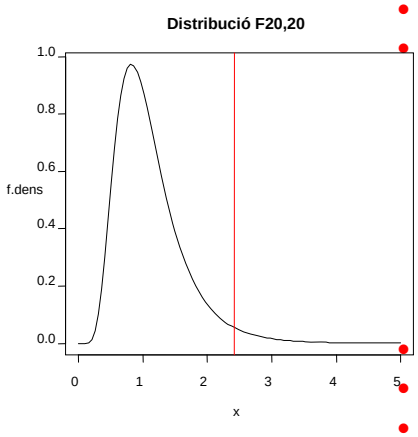
- Hipòtesis (i indiqueu si la prova és bilateral o unilateral)
- Calculeu el valor de l'estadístic
- Digueu quina distribució segueix l'estadístic i quants graus de llibertat té.
- Representeu gràficament el valor obtingut i el punt crític sobre la distribució de l'estadístic
- Doneu la conclusió

- Prova d'hipòtesi bilateral

Ho: $\sigma_X^2 = \sigma_Y^2$

H_1: $\sigma_X^2 \neq \sigma_Y^2$

- Estadístic: $\hat{F} = \frac{S_X^2}{S_Y^2} \sim F_{n1-1,n2-1}$ que en aquest cas serà $F_{20,20}$. Els graus de llibertat son 20 al numerador i 20 al denominador.
- El valor de l'estadístic $\hat{F}$ =2.52/1.90 =1.33 (Recordeu, és preferible posar el valor més gran de la variància al numerador i el més petit, al denominador).
- Treballant amb una probabilitat del 95%, $_{F20,20}(0.975)=2.46$. Com que el valor de l'estadístic és 1.33 < 2.46, no podem rebutjar la hipòtesi nul·la



La recta vermella vertical marca el punt crític amb p=0.975 . Valors de l'estadístic a la dreta d'aquest punt crític, ens fan rebutja la H_o d'igualtat de variàncies.

- Conclusió: Es pot considerar que les dues variàncies poblacionals son iguals, amb un marge d'error del 5%.

Atenció: l'estadístic F mai pot ser negatiu, és quocient de dues variàncies!

a.3 (3 punts) Es pot admetre que els dos programes triguen el mateix? Per respondre aquesta pregunta, plantegeu una prova d'hipòtesis, dient:

- Hipòtesis (i indiqueu si la prova és bilateral o unilateral)
- Especifiqueu les premisses necessàries
- Independentment dels resultats de l'apartat a.2, trobeu l'estimació de la variància comú (*pooled*) dels programes X i Y
- Determineu el valor de l'estadístic, la seva distribució i quants graus de llibertat té.
- Tenen els programes rendiments diferents? Raoneu la vostra resposta.

- Prova d'hipòtesi bilateral

Ho: $\mu_X = \mu_Y$

H_1: $\mu_X \neq \mu_Y$

- Premisses: Les dades provenen d'una distribució normal i son independents entre si.
- $s^2_{pooled} = \frac{(n1-1)s^2_X + (n2-1)s^2_Y}{n1+n2-2} = \frac{20*2.52+20*1.90}{40} = 2.21$
- Estadístic: $\hat{t} = \frac{(\bar{x}-\bar{y})-(\mu_X-\mu_Y)_0}{s_{pooled}\sqrt{1/n1+1/n2}} \sim t_{n1+n2-2}$ que en aquest cas serà t_{40}. Els graus de llibertat son 40
- El valor de l'estadístic és $\hat{t} = \frac{(10.38-10.95)}{\sqrt{2.21}/\sqrt{1/21+1/21}} = \frac{-0.57}{0.46} = -1.23$
- Treballant amb una confiança del 95%, $t_{40}(0.025) = -2.021$. Com que -1.23 > -20.21, per tant, no podem rebutgem la H0 d'igualtat de mitjanes, és a dir es pot considerar que els dos programes tenen rendiments similars.

b) El professor d'aquests estudiants els ha aconsellat que potser és millor executar els programes en una sola base de dades i simultàniament.

Els resultats que s'han obtingut són

Variable	Mitjana	Variància	N
X-Y	-0.57	1.25	20

b.1 (2 punts) Es pot admetre que els dos programes triguen el mateix? Per respondre aquesta pregunta, plantegeu una prova d'hipòtesi que resoldreu mitjançant el càlcul del interval de confiança per a la diferència de rendiments.
Caldrà que:

- Indiqueu la hipòtesi i si la prova és bilateral o unilateral
- Calculeu el interval de confiança al 95% per a la diferència de mitjanes poblacionals
- A quin resultat arribeu?

- Plantegem una prova d'hipòtesi bilateral

Ho: $\mu_{X-Y} = 0$

H_1: $\mu_{X-Y} \neq 0$

- IC(μ_{X-Y}, 95%) = $\bar{m} \pm t_{N-1,1-\alpha/2}\frac{S}{\sqrt{N}} = -0.57 \pm 2.093\sqrt{\frac{1.25}{20}} = (-1.09,-0.05)$
- Com que el 0 no pertany a aquest interval de confiança, podem considerar que els dos programes no tenen rendiments similars, és a dir, el programa X triga més que el programa Y

b.2 (1.5 punt) Compareu els resultats obtinguts a l'apartat a.3 amb els de l'apartat b.1, és a dir, si es pot admetre en ambdós casos que les mitjanes poblacionals són les mateixes o no. Tant si es pot admetre com si no, indiqueu possibles explicacions.

- En el cas a.3) no podem rebutjar la H_0, és a dir, el rendiment del dos programes és similar, mentre que en el cas b.1) , el rendiment dels dos programes és diferent. Si ens fixem, la diferencia de mitjanes és igual en ambdós casos, -0.57. L'estadístic que estem utilitzant en ambdós casos representa el "senyal" proporcionat per la distància entre les mitjanes relativa al "soroll" aleatori que porti aquest senyal. El que ha variat és que al treballar amb dades independents el soroll estimat p $s_{pooled}\sqrt{1/n1+1/n2}$ val 0.46 i l'estimació del soroll pel cas de les dades aparellades és $\frac{S}{\sqrt{N}}$ =0.27, és a dir, el "soroll" en el primer cas és més del doble que en el segon cas. Això fa que la relació senyal/soroll en el primer cas sigui molt més feble que en el segon cas, la qual cosa s'interpreta que en el primer cas no podem distingir el senyal, és a dir que la diferencia de mitjanes no és significativa, mentre en el segon cas sí.

NOM:_____

(Poseu el nom i contesteu cada pregunta en el seu lloc reservat. Expliciteu i justifiqueu els passos en les respostes).

Problema B6

S'ha mesurat la el temps de descarrega de 8 fitxers amb les següents dades:

Grandària arxiu (MB) : **X**	Temps de descarrega (s) : **Y**
32	10.4
64	19.3
96	33.2
128	41.8
160	50.3
192	58.7
224	74.1
256	81.7

$\sum x_i =$ 1152

$\sum x_i^2 =$ 208896

$\sum y_i =$ 369.5

$\sum y_i^2 =$ 21471.61

$\sum x_i y_i =$ 66937.6

Amb aquestes dades tenim que:

n=8

a) Calculeu els estimadors de la constant i del pendent de la recta de regressió i representeu gràficament la recta estimada **(1 punt)**

$$\bar{x} = \frac{\sum_{i=1}^{8} x_i}{n} = \frac{1152}{8} = 144 \qquad S_x^2 = \frac{\sum_{i=1}^{8} x_i^2 - n \cdot \bar{x}^2}{n-1} = \frac{208896 - 8 \cdot (144)^2}{7} = 78.38^2$$

$$\bar{y} = \frac{\sum_{i=1}^{8} y_i}{n} = \frac{369.5}{8} = 46.19 \qquad S_y^2 = \frac{\sum_{i=1}^{8} y_i^2 - n \cdot \bar{y}^2}{n-1} = \frac{21471.6 - 8 \cdot (46.2)^2}{7} = 25.09^2$$

$$S_{xy} = \frac{\sum_{i=1}^{8} x_i \cdot y_i - \frac{\sum_{i=1}^{8} x_i \cdot \sum_{i=1}^{8} y_i}{n}}{n-1} = \frac{66937.6 - \frac{1152 \cdot 369.5}{8}}{7} = 1961.37$$

$$b_1 = \frac{S_{xy}}{S_x^2} = \frac{1961.37}{78.38^2} = 0.3192$$

$$b_0 = \bar{y} - b_1 \cdot \bar{x} = 46.19 - 0.3192 \cdot 144 = 0.2179$$

b) Calculeu la taula de descomposició de la variància **(2 punts)**

	SQ	**Graus de llibertat (Gdl)**	**QM = SQ/Gdl**	**Rati**
Explicada pel model	$7 \cdot 0.3192^2 \cdot 78.38^2 = 4383$	1	4383/1 = 4383	F = 4383/3.73 = 1175.1
Residual	$4405.3 - 4383 = 22.3$	6	22.3/6 = 3.73	
Total	$7 \cdot 25.09^2 = 4405.3$	7		

c) Calculeu i interpreteu el coeficient de determinació R^2 **(1 punt)**

$R^2 = \frac{SQE}{SQT} = \frac{4383}{4405.3} = 0.994$ El percentatge de variabilitat explicada pel model és del 99.4%. És a dir, la grandària del fitxer captura el 99.4% de la variabilitat del temps de descarrega.

d) Calculeu l'interval de confiança al 95% pel pendent de la recta **(1 punt)**

$$S^2_{b_1} = \frac{QM_R}{(n-1)\cdot S^2_x} = \frac{3.73}{7 \cdot 78.38^2} = 0.0093$$

$$IC(\beta_1, 95\%) = b_1 \mp t_{0.975,6} \cdot S_{b_1} = 0.3192 \mp 2.45 \cdot 0.0093 = [0.2964 \, a \, 0.3420]$$

e) Tenim contractat 3Mbps **(**22.5 MB/min**)**. Poseu a prova si aquesta és la velocitat realment disponible (heu d'adaptar les dades per a treballar amb MB/min). Quina diferència té aquesta prova amb contrastar directament si el quocient X/Y té mitjana igual a 22.5? **(2 pts)**

***Nota**: tingueu en compte que la transformació de les unitats afecta tant a l'estimador com a la seva desviació estàndard.*

3Mbps → 22.5MB/min → 22.5/60 BM/seg → 60/22.5 Seg/MB → 2.67Seg/MB
Donat que 2.67Seg/MB està fora de l'interval [0.2964, 0.3420], aquest valor es pot refusar.
Més formalment,

$$t = \frac{b_1' - 2.6667}{S_{b_1}} = \frac{0.3192 - 2.6667}{0.0093} = -252.46$$

Com que $|t|=252.46 > t_{0.975,6} = 2.45$, rebutgem que la velocitat real sigui igual a la contractada.

Si contrastéssim directament si el quocient d'ambdues variables (x/y) és igual a 22.5 Mb/min, estaríem assumint que tot el temps de descarrega si inverteix en baixar el fitxer i que no hi ha un temps fix addicional (= **Intercept**)

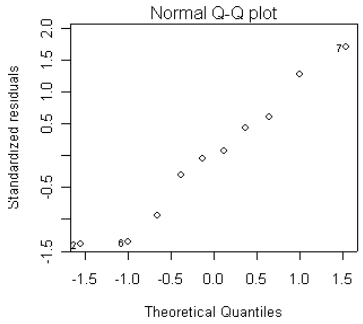
f) Doneu una previsió, així com el seu interval de confiança al 95%, del valor promig del temps de descarrega quan la grandària del fitxer és de 30 MB **(2 punts)**

$$y_{30} = b_0 + b_1 \cdot 30 = 0.2179 + 0.3192 \cdot 30 = 9.79$$

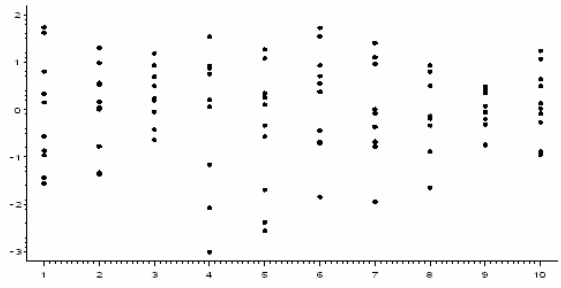
$$IC(Y|X = 30, 95\%) = y_{30} \mp t_{0.975,6} \cdot \sqrt{QM_R} \cdot \sqrt{\frac{1}{n} + \frac{(30-\bar{x})^2}{(n-1)\cdot S_x^2}} = 9.79 \mp 2.45 \cdot \sqrt{3.23} \cdot \sqrt{\frac{1}{8} + \frac{(30-144)^2}{7 \cdot 78.38^2}} = [6.71 \, a \, 12.88]$$

g) Dibuixeu (inventant les dades i les figures) els gràfics NPP (Normal probability plot) i 'residuals versus fitted' (residus en front de les prediccions) de tal manera que mostrin que les premisses **sí** que es compleixen i comenteu quines premisses permeten valorar cada un. **(1 punt)**

NPP: Normalitat si bon ajustament a una recta.



Homoscedasticitat: si les amplituds semblen similars



'Residuals versus fitted':
Linealitat si no mostren tendències d'ordre superior