FIB 2013-14 Q2. FINAL PE, Juny de 2014

login:

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

Problema 1. (Blocs 1 i 2)

Un sistema d'emmagatzematge de fitxers de dades consta de dos servidors anomenats SMALL i LARGE. La mida (en Gb) dels fitxers que arriben al sistema és una variable aleatòria contínua X amb funció de distribució

$$F(x) = \begin{cases} 0 & \text{si} \quad x < 0, \\ 2x - x^2 & \text{si} \quad 0 \geq x < 1, \\ 1 & \text{si} \quad x \geq 1. \end{cases}$$

1. **(1 punt)** Calculeu la funció de densitat d'X.

   Solució:
   Si $x \notin [0, 1]$, $f(x) = F'(x) = 0$. Si $x \in [0, 1]$, $f(x) = F'(x) = 2 - 2x$.

2. **(1 punt)** Calculeu el valor esperat d'X.

   Solució:

$$E(X) = \int_0^1 xf(x)dx = \int_0^1 x(2 - 2x)dx = (x^2 - 2x^3/3)\Big|_0^1 = 1 - 2/3 = 1/3 \text{ Gb}.$$

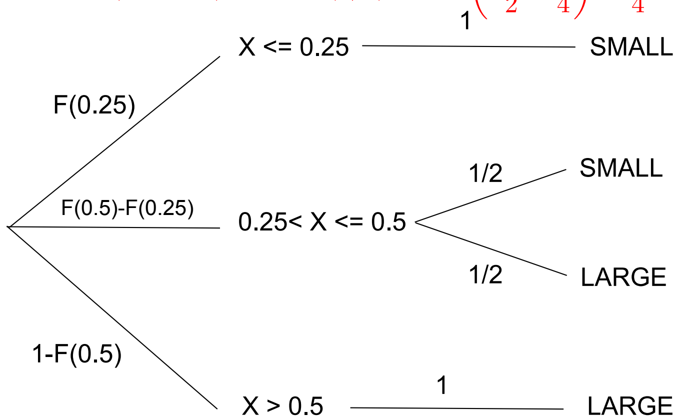
El mòdul de control del sistema fa que tots els fitxers de mida menor o igual que 0.25 Gb es desin a SMALL i tots els que siguin més grans de 0.5 Gb es desin a LARGE. Quan un fitxer té mida entre 0.25 i 0.5 Gb es sorteja, amb probabilitats 1/2 i 1/2, si es desa a un servidor o a l'altre.

3. **(2 punts)** Calculeu la probabilitat que un fitxer que arriba al sistema acabi desat a LARGE.
   Indicació: Expliqueu primer perquè $\Pr(X \leq 0.25) = 7/16$ *i* $\Pr(X > 0.5) = 1/4$.

   Solució:
   Observeu que

$$\Pr(X \leq 0.25) = F(1/4) = 2\frac{1}{4} - \frac{1}{16} = \frac{7}{16},$$

$$\Pr(X > 0.5) = 1 - F(1/2) = 1 - \left(2\frac{1}{2} - \frac{1}{4}\right) = \frac{1}{4}.$$



$$\Pr(\mathsf{LARGE}) = (1 - F(0.5)) + \frac{1}{2}(F(0.5) - F(0.25)) =$$

$$1 - \frac{1}{2}F(1/4) - \frac{1}{2}F(1/2) = 1 - \frac{1}{2}\left(2\frac{1}{4} - \frac{1}{16}\right) - \frac{1}{2}\left(2\frac{1}{2} - \frac{1}{4}\right) = 1 - \frac{7}{32} - \frac{3}{8} = \frac{13}{32} = 0.4063.$$

4. **(2 punts)** Si sabem que un fitxer està a SMALL, quina probabilitat hi ha que la seva mida sigui més petita o igual que 0.25 Gb?

Solució:
Per l'apartat anterior sabem que $\Pr(\mathsf{SMALL}) = 1 - \Pr(\mathsf{LARGE}) = 19/32$. Pel Teorema de Bayes,

$$\Pr(X \leq 0.25|\mathsf{SMALL}) = \frac{\Pr(\mathsf{SMALL}|X \leq 0.25)\Pr(X \leq 0.25)}{\Pr(\mathsf{SMALL})} =$$

$$\frac{1 \cdot F(1/4)}{\Pr(\mathsf{SMALL})} = \frac{7/16}{19/32} = \frac{14}{19} = 0.7368.$$

Ens diuen que el temps d'accés a un fitxer és 1 milisegon si la mida del fitxer és més petita o igual que 0.25 Gb, és 2 milisegons si la seva mida está entre 0.25 Gb i 0.5 Gb i està desat a SMALL, és 3 milisegons si la seva mida está entre 0.25 Gb i 0.5 Gb i està desat a LARGE, i és 4 milisegons si la mida és més gran que 0.5 Gb.

5. **(1 punts)** Doneu la funció de probabilitat de la variable aleatòria T, que mesura el temps d'accés (en milisegons) a un fitxer del sistema.

Solució:

t	1	2	3	4
$\Pr(T = t)$	$14/32$	$5/32$	$5/32$	$8/32$

6. **(2 punts)** Calculeu l'esperança i la desviació estàndard de T.

Solució:

$$E(T) = 1 \cdot \frac{14}{32} + 2 \cdot \frac{5}{32} + 3 \cdot \frac{5}{32} + 4 \cdot \frac{8}{32} = \frac{71}{32} = 2.2188 \text{ milisegons.}$$

$$E(T^2) = 1^2 \cdot \frac{14}{32} + 2^2 \cdot \frac{5}{32} + 3^2 \cdot \frac{5}{32} + 4^2 \cdot \frac{8}{32} = \frac{207}{32}.$$

$$V(T) = E(T^2) - (E(T))^2 = \frac{207}{32} - \frac{71^2}{32^2} = \frac{1583}{1024}.$$

La desviació estàndard de T és

$$\sqrt{V(T)} = \sqrt{\frac{1583}{1024}} = 1.2433 \text{ milisegons.}$$

Definim la variable aleatòria B que val 0 si $X \leq 0.5$ Gb i val 1 si $X > 0.5$ Gb.

7. **(1 punts)** Doneu la funció de probabilitat conjunta de B i T, digueu si són independents i expliqueu per què.

Solució:
Els valors que conjuntament poden prendre (B, T) són

$$(b, t),\ b = 0, 1,\ t = 1, 2, 3, 4.$$

A més a més, quan T val 1, 2 o 3, no pot ser que B valgui 1, i quan T val 4 no pot ser que B valgui 0. Per tant, la distribució conjunta és ve donada per aquestes probabilitats conjuntes:

$\Pr(B = b, T = t)$		t = 1	2	3	4
b	0	$14/32$	$5/32$	$5/32$	0
	1	0	0	0	$8/32$

B i T no són independents perquè no es verifica que $\Pr(B = b, T = t) = \Pr(B = b)\Pr(T = t)$ per a qualsevols b i t. Per exemple,

$$\Pr(B = 4, T = 1) = \frac{8}{32} = \frac{1}{4} \neq \Pr(B = 4)\Pr(T = 1) = \frac{1}{4}\frac{1}{4} = \frac{1}{16}.$$

NOM:____

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

Problema 2 (B3-B4).

El temps en minuts que els estudiants empren per solucionar un problema d'e-status a PE en aquest quadrimestre ha estat:

```
> summary(minu)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.     Sd.
 0.2833  5.6830 10.5200 13.7600 19.1500 85.6800 10.9429
```

Admetem que la variable Y="temps per resoldre un problema" es distribueix com una exponencial.

1. Trobeu les probabilitats que Y sigui inferior a 5 minuts,
   a. suposant que el paràmetre λ el preneu de la mitjana mostral. (1pt)
   b. suposant que el paràmetre λ el preneu de la desviació tipus. (1pt)
   c. suposant que el paràmetre λ el preneu de la mediana. (1pt)

   Recordeu quines són les propietats del model exponencial per derivar el valor de λ a cada cas.

a) $E(Y) = \frac{1}{\lambda} \to \hat{\lambda}_1 = \frac{1}{\bar{x}} = \frac{1}{13.76} = 0.073$

$P(Y < 5) = F_Y(5) = 1 - e^{-0.073 \cdot 5} = \mathbf{0.305}$

b) $V(Y) = \frac{1}{\lambda^2} \to \hat{\lambda}_2 = \frac{1}{s} = \frac{1}{10.94} = 0.091$

$P(Y < 5) = F_Y(5) = 1 - e^{-0.091 \cdot 5} = \mathbf{0.367}$

c) $Med = F_Y^{-1}(0.5) = e^{-0.5\lambda} \to \hat{\lambda}_3 = \frac{\ln(2)}{Med} = \frac{\ln(2)}{10.52} = 0.066$

$P(Y < 5) = F_Y(5) = 1 - e^{-0.066 \cdot 5} = \mathbf{0.281}$

2. La Neus, en canvi, presenta una distribució molt diferent, ja que ella es pren amb calma cada exercici i això suposa que els seus temps segueixin un model Normal. El 35% dels casos els resol en menys de 22 minuts, i per 1 de cada 4 triga més de 33 minuts.
   a. Quin és el temps esperat per la Neus per resoldre un exercici? Quina és la variància del seu temps? (2.5pt)

TN = "Temps que triga la Neus a resoldre un problema"

$$P(TN < 22) = 0.35 \to P\left(Z < \frac{22-\mu}{\sigma}\right) = 0.35 \to (Taules) \to \frac{22-\mu}{\sigma} = -0.385 \to 22 - \mu = -0.385\sigma$$

$$P(TN > 33) = 0.25 \to P\left(Z > \frac{33-\mu}{\sigma}\right) = 0.25 \to (Taules) \to \frac{33-\mu}{\sigma} = 0.675 \to 33 - \mu = 0.675\sigma$$

Restant la 2a equació de la primera:

$11 = 1.06\sigma \to \ \sigma = 10.37 \to \ \sigma^2 = \mathbf{107.73}$

Substituint a la primera equació:

$22 - \mu = -0.385 \cdot 10.37 \to \mu = \mathbf{26}$

   b. Si una tarda de diumenge es posa a fer e-status, resol 5 exercicis, amb una pausa intermèdia després de cada exercici, que es distribueix com N(10min, 2.5min): calcula la probabilitat que el temps total hagi estat de més de tres hores (suposeu que els temps són independents uns d'altres). (2.5pt)

P="Temps de la pausa" ~ N(10 , 2.5)

T="Temps total en resoldre els 5 problemes" ~ N(μ_T , σ_T) = N(170 , 52.81)

$\mu_T = E(T) = E(5X + 4P) = 5E(X) + 4E(P) = 5 \cdot 26 + 4 \cdot 10 = 170$

$\sigma_T^2 = V(T) = V(5X + 4P) = 25V(X) + 16V(P) = 25 \cdot 10.37^2 + 16 \cdot 2.5^2 = 2788.4 \to \sigma_T = 52.81$

$$P(T > 180) = P\left(Z > \frac{180-170}{52.81}\right) = P(Z > 0.19) = (Taules) = 1 - 0.5753 \approx \mathbf{0.425}$$

3. 8 de cada 10 exercicis s'han fet entre les 20h i les 00:00. Si seleccionés de la base de dades (que és molt gran) una mostra a l'atzar de N exercicis, trobeu la probabilitat que almenys un 70% s'hagin resolt després de les 20h:
   a. si N és 20. (1pt)

ND = "Número d'exercicis resolts Després de les 20:00 en una mostra de 20" ~ Bin (n=20,p=0.8)

NA = "Número d'exercicis resolts Abans de les 20:00 en una mostra de 20" ~ Bin (n=20,p=0.2)

$P(ND \geq 0.7 \cdot 20) = P(ND \geq 14) = P(NA \leq 6) = (Taules) = \mathbf{0.913}$

   b. si N és 200 (trobeu una aproximació al resultat exacte). (1pt)

ND_2 = "Número d'exercicis resolts Després de les 20:00 en una mostra de 200"~ Bin (n=200,p=0.8)

Es pot emprar l'aproximació de la Normal:

$ND_2 \approx N(\mu=200\cdot0.8,\ \sigma=\sqrt{200\cdot0.8\cdot0.2}) = N(\mu=160,\ \sigma=5.66)$

$$P(ND_2 \geq 0.7 \cdot 200) = P(ND_2 \geq 140) \approx P\left(Z \geq \frac{140-160}{5.66}\right) = P(Z \geq -3.54) \approx \mathbf{1}$$

```
> round(qchisq(c(0.005,0.01,0.025,0.05,0.10,0.25,0.75,0.9,0.95,0.975,0.99,0.995), 109), 3)
 [1]  74.724  77.616  82.000  85.903  90.558  98.712 118.563 128.298 134.369 139.784 146.257 150.774
```

Ara ja no suposarem que coneixem un model o un paràmetre concret per a cap variable: només disposem de dades.

4. Volem recollir una mostra per estimar la proporció dels exercicis que es fan després de les 20h, i es desitja obtenir un interval de confiança al 95% amb amplada inferior a 6% (és a dir, la precisió serà de ±3%). Quin és el nombre de dades que necessitem recollir? Assumeix màxima incertesa sobre la proporció poblacional (2 pt)

Suposant màxima indeterminació:

$$z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\pi \cdot (1-\pi)}{n}} = 0.03 \rightarrow 1.96 \cdot \sqrt{\frac{0.5 \cdot (1-0.5)}{n}} = 0.03 \rightarrow n = 0.25 \cdot \left(\frac{1.96}{0.03}\right)^2 = 1067.1 \rightarrow n = \mathbf{1068}$$

5. Com podríem aconseguir una mostra més petita (sense canviar l'objectiu, ni modificar la confiança o l'amplada)? (1pt)

Si es sap (amb convicció) que la proporció d'exercicis resolts després de les 20h és substancialment més gran que 0.5 (per exemple, 0.8), **podem utilitzar aquest valor pel càlcul de la grandària**, i la n sortiria més petita:

$$z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\pi \cdot (1-\pi)}{n}} = 0.03 \rightarrow 1.96 \cdot \sqrt{\frac{0.8 \cdot (1-0.8)}{n}} = 0.03 \rightarrow n = 0.16 \cdot \left(\frac{1.96}{0.03}\right)^2 = 682.95 \rightarrow n = 683$$

6. Com és molt raonable que els temps de resolució d'exercicis no siguin una distribució Normal, es vol analitzar la variable X = log(Y), que seria molt menys problemàtica. Una mostra de 110 observacions independents de X es descriu a continuació:

```
> summary(x)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     Sd.
-0.08701 1.70600  2.32100  2.19500  2.75700  4.09200  0.88212
```

a. Creieu que hem fet bé escollint X en lloc de Y? Justifiqueu la resposta. (1.5pt)

Sí, per dos motius:

1) X té una distribució molt més simètrica que Y

	Y	X
Mean/Median	0.95 (1/0.95 = 1.05)	1.31
(Q3-Q2)/(Q2-Q1)	0.71 (1/0.71 = 1.41)	1.78
(Max-Q2)/(Q2-Min)	0.74 (1/0.74 = 1.36)	7.34

2) El màxim i el mínim es troben més o menys a 2 desviacions tipus de la mitjana

$\bar{x} - 2S = 0.43 \rightarrow Pròxim\ a\ -0.09$

$\bar{x} + 2S = 3.96 \rightarrow Pròxim\ a\ 4.09$

b. En qualsevol cas, estimeu E(X) per interval de confiança 90%. Feu també l'estimació de σ^2 per interval de confiança 90%. (1.5pt)

$$IC(\mu, 90\%) = \left[\bar{x} \mp t_{0.95,n-1} \cdot \frac{S}{\sqrt{n}}\right] = \left[2.20 \mp 1.66 \cdot \frac{0.88}{\sqrt{110}}\right] = [2.06, 2.33]$$

$$IC(\sigma^2, 90\%) = \left[\frac{(n-1)\cdot S^2}{\chi^2_{0.95,n-1}}, \frac{(n-1)\cdot S^2}{\chi^2_{0.05,n-1}}\right] = \left[\frac{(110-1)\cdot S^2}{134.369}, \frac{(100-1)\cdot S^2}{85.903}\right] = [0.63, 0.99]$$

c. Volem posar a prova si la desviació de X és 1: desenvolupeu la prova formal per contestar la qüestió, enfront de l'opció que diu que la desviació seria menor que 1. Amb la taula de valors del peu de pàgina procureu afitar el p-valor resultant. (2.5pt)

Hipòtesis:

$\begin{cases} H_0: \sigma = 1 \\ H_1: \sigma < 1 \end{cases}$ és equivalent a $\begin{cases} H_0: \sigma^2 = 1 \\ H_1: \sigma^2 < 1 \end{cases}$

Estadístic, distribució i premisses:

$X = \frac{(n-1)\cdot S^2}{\sigma^2} \sim \chi^2_{n-1}$ si (1) m.a.s i (2) dades provinents de distribució Normal

Càlcul de l'estadístic i punt crític:

Estadístic → $X = \frac{(n-1)\cdot S^2}{\sigma^2} = \frac{(110-1)\cdot 0.88}{1} = 84.82$

Punt crític → $\chi^2_{0.05,109} = 85.903$

P-valor:

Com que $\chi^2_{0.025,109}$ =82.000 < X < 85.903= $\chi^2_{0.05,109}$, el p-valor estarà entre 0.025 i 0.05

Conclusió:

$\left.\begin{array}{r} X < \chi^2_{0.05,109} \\ p-valor < 0.05 \end{array}\right\} \rightarrow$ Rebutgem la hipòtesi de que la desviació poblacional sigui i acceptem que és menor

d. Hauria estat correcte realitzar aquesta prova amb les dades en forma de temps (és a dir, amb una mostra de Y)? Justifiqueu la resposta. (1.5pt)

No, perquè una premissa per a que l'estadístic segueixi una χ^2 sota H_0 és que la distribució de partida sigui Normal.

```
> round(qchisq(c(0.005,0.01,0.025,0.05,0.10,0.25,0.75,0.9,0.95,0.975,0.99,0.995), 109), 3)
 [1]  74.724  77.616  82.000  85.903  90.558  98.712 118.563 128.298 134.369 139.784 146.257 150.774
```

NOM:____

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

Problema 3 (B5-B6).

Un FIBER molt avantatjat ha dissenyat un nou tipus de CPU i vol verificar si el seu prototipus superarà o no al model de CPU actual amb el que està treballant. Per verificar-ho pren una mostra en 32 CPUs, 16 de les antigues que anomenarem A, i 16 de les noves anomenades N, amb els següents resultats de mitjanes i desviacions d'una variable de temps:

$\overline{t_A} = 3.10 \qquad s_{T_A} = 1.66 \qquad \overline{t_N} = 3.07 \qquad s_{T_N} = 1.19$

1.- (3 punts) Poseu a prova si es pot considerar les variances poblacionals iguals o una significativament superior a l'altra:

- hipòtesis (indiqueu si és prova unilateral o bilateral), estadístic i premisses adients

Prova unilateral:

$H_0 : \sigma^2_{T_A} = \sigma^2_{T_N}$

$H_1 : \sigma^2_{T_A} \rangle \sigma^2_{T_N}$

Estadístic:

$\hat{F} = \frac{s^2_{T_A}}{s^2_{T_N}}$

$F_{16-1,16-1} = F_{15,15}$

Premisses:

T_A i T_N normals i mostres independents

- càlcul de l'estadístic, punt crític amb confiança 95%, i representació gràfica (indicant el p_valor i com es calcula per exemple en R)

$$\hat{F} = \frac{1.66^2}{1.19^2} = \frac{2.76}{1.42} = 1.94$$

Punt crític: $F_{15,15,0.95} = 2.4$

P_valor : 1-pf(15,15,1.94)

- es pot rebutjar la hipòtesis nul·la? Justifiqueu-ho i interpreteu-ne les conclusions

No es pot rebutjar la hipòtesis nul·la ja que l'estadístic és menor que el punt crític (o el p_valor seria superior al risc del 5%)

Per tant, no tenim prou evidència per rebutjar la igualtat de variances poblacionals dels temps en les CPU antigues i les noves, tot i que no demostra que siguin iguals.

2.- (3 punts) Poseu a prova si es pot considerar que les dues mitjanes poblacionals són iguals o no amb confiança del 95%:

- hipòtesis (indiqueu si és prova unilateral o bilateral), estadístic i premisses adients

Prova bilateral:

$H_0 : \mu_{T_A} = \mu_{T_N}$

$H_1 : \mu_{T_A} \neq \mu_{T_N}$

Estadístic:

$\hat{t} = \frac{\overline{t_A} - \overline{t_N}}{s\sqrt{1/16 + 1/16}} \qquad \left(s = \sqrt{\frac{15 s^2_{T_A} + 15 s^2_{T_N}}{16+16-2}} \right)$

$t_{16+16-2} = t_{30}$

Premisses:

T_A i T_N normals, mostres independents i

$\sigma_{T_A} = \sigma_{T_N}$

- càlcul de l'estadístic, punts crítics, i representació gràfica (indicant el p_valor i el seu càlcul per exemple en R)

$$\hat{t} = \frac{3.10 - 3.07}{1.45\sqrt{1/16 + 1/16}} = \frac{0.03}{0.51} = 0.06 \qquad s = \sqrt{\frac{15*2.76 + 15*1.42}{16+16-2}} = 1.45$$

Punts crítics: $t_{30,0.975} = 2.042 \quad t_{30,0.025} = -2.042$

P_valor : 2(1-pt(30,0.06))*

- es pot rebutjar la hipòtesis nul·la? Justifiqueu-ho i interpreteu-ne les conclusions

No es pot rebutjar la hipòtesis nul·la ja que l'estadístic (0.06) està entre els punts crítics (-2.4 i 2.4); o el p_valor seria superior al risc del 5%

Per tant, no tenim prou evidència per rebutjar la igualtat de les mitjanes poblacionals dels temps en les CPU antigues i les noves, és a dir no podem rebutjar que tiguin igual funcionament mitjà.

3.- (4 punts) Un professor li comenta a aquest estudiant que pot plantejar el problema com si fossin dades aparellades fent exactament les mateixes proves en els dos casos. Suposem que es fa així i que s'obtenen els resultats indicats a l'inici i, a més a més, es calcula la correlació entre les dades antigues i les noves obtenint $r_{T_A T_N} = 0.77$ amb una variància residual igual a 1.15 en el model de regressió lineal del temps del model nou en funció del temps del model antic.

- calculeu la recta de regressió

$$b_1 = \frac{s_{T_A T_N}}{s^2_{T_A}} = \frac{r_{T_A T_N} s_{T_A} s_{T_N}}{s^2_{T_A}} = \frac{0.77 * 1.66 * 1.19}{1.66^2} = 0.55 \qquad T_N = 1.36 + 0.55 * T_A$$

$$b_0 = \overline{t_N} - 0.55 * \overline{t_A} = 3.07 - 0.55 * 3.1 = 1.36$$

- poseu a prova si la recta es pot considerar que no és plana

$$H_0 : \beta_1 = 0 \qquad H_1 : \beta_1 \neq 0$$

$$\hat{t} = \frac{b_1 - 0}{S_{b_1}} = \frac{0.55}{0.17} = 3.3 \quad \textit{és} \quad t_{16-2} = t_{14} \quad \textit{amb} \quad S_{b_1} = \sqrt{\frac{S^2}{(16-1)s^2_{T_A}}} = \sqrt{\frac{1.15}{15 * 2.76}} = 0.17$$

Punts crítics: $t_{14,0.975} = 2.145 \quad t_{14,0.025} = -2.145$

Podem rebutjar la hipòtesis nul·la ja que l'estadístic (3.3) està fora els punts crítics (-2.145 i 2.145); o el p_valor seria inferior al risc del 5%

Per tant podem considerar que la pendent no és igual a zero i, per tant, la recta no és plana.

- doneu un interval de confiança al 95% de la pendent de la recta

$$b_1 \pm t_{14,0.975} * S_{b_1} = 0.55 \pm 2.145 * 0.17 = 0.55 \pm 0.36 = [0.19, 0.91]$$

- doneu una estimació puntual de T_N quan T_A val 3 amb un interval de confiança al 95% d'aquesta predicció

$$1.36 + 0.55 * 3 = 3.01$$

$$3.01 \pm t_{14,0.975} * \sqrt{1.15}\sqrt{1 + \frac{1}{16} + \frac{(3-3.10)^2}{s^2_{T_A}(16-1)}} = 3.01 \pm 2.145 * \sqrt{1.15}\sqrt{1 + \frac{1}{16} + \frac{(3-3.10)^2}{1.66^2(16-1)}}$$

$$= 3.01 \pm 2.145 * 1.1 = 3.01 \pm 2.37 = [0.64, 5.38]$$