Per calcular aquesta probabilitat utilitzarem que, segons el teorema del límit central, la variable aleatòria

$$\frac{\sum_{i=1}^{100} X_i - 18.000}{100\sqrt{100}} = \frac{\sum_{i=1}^{100} X_i - 18.000}{1.000}$$

es comporta com una normal estàndard. Aleshores:

$$\begin{aligned} P(\sum_{i=1}^{100} X_i \leq 16.800) &= P\left(\frac{\sum_{i=1}^{100} X_i - 18.000}{1.000} \leq \frac{16.800 - 18.000}{1.000}\right) \\ &= P(N(0,1) \leq -1.2) = \texttt{pnorm(-1.2)} = 0.1150. \end{aligned}$$

És a dir, la probabilitat que es puguin arreglar tots els ordinadors és força petita.

Capítol 4

INTERVALS DE CONFIANÇA

En aquest i els següents capítols la situació amb què treballarem serà similar tot i que els mètodes i les proves que utilitzarem en cada cas seran diferents.

La idea és que tindrem unes dades $x_1, \ldots, x_n$ (una mostra) i intentarem a partir d'elles obtenir informació sobre el procés que les ha generat. Quan parlem d'obtenir informació no volem dir descriure les dades, com feiem a estadística descriptiva, sinó que entendrem que les nostres dades són una realització d'una mostra amb una distribució desconeguda. Aquesta distribució no serà desconeguda del tot, sinó que, de fet, dependrà d'un paràmetre desconegut que serà el nostre principal objectiu. Aquest paràmetre podrà ser: la mitjana, la variància, una proporció...

El primer mètode que introduïrem serà el càlcul d'intervals de confiança. L'objectiu és trobar un interval amb un percentatge bastant alt probabilitats, que el valor real del paràmetre que estem estimant estigui dins el nostre interval.

Anem a veure aquestes idees amb exemples i a donar alguns mètodes per construir aquests intervals.

4.1 Exemple

Exemple 44. *Hem observat les jornades laborals de 10 treballadors d'una empresa:*

$$180 \quad 165 \quad 168 \quad 192 \quad 195 \quad 187 \quad 181 \quad 177 \quad 175 \quad 186.$$

Suposem que sabem que la desviació estàndard val 4. Suposem també, que les dades són normals.

Tenim, per tant, una mostra aleatòria simple $X_1, ..., X_n$ *d'una distribució* $N(\mu, 16)$, *és a dir que* $\sigma = 4$. *Volem trobar un interval de confiança al* 95% *per a* μ.

1. Sabem que:
$$Z = \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\overline{X}_n - \mu}{\frac{4}{\sqrt{10}}} \sim N(0,1)$$
(recordem per l'enunciat que: $\sigma = 4$ i $n = 10$).

2. Com que volem un nivell de confiança del 95% hem de determinar u_1 i u_2 tals que
$$P(u_1 \leq Z \leq u_2) = 0.95.$$

Podem agafar per exemple $u_1 = -1.96$ i $u_2 = 1.96$. (Normalment utilitzarem intervals simètrics). Tenim així que

$$-1.96 \leq \frac{\overline{X}_n - \mu}{\frac{4}{\sqrt{10}}} \leq 1.96,$$

que podem reescriure com

$$-1.96\frac{4}{\sqrt{10}} \leq \overline{X}_n - \mu \leq 1.96\frac{4}{\sqrt{10}}$$

o

$$\overline{X}_n - 1.96\frac{4}{\sqrt{10}} \leq \mu \leq \overline{X}_n + 1.96\frac{4}{\sqrt{10}}.$$

És a dir, que obtenim l'interval per a μ

$$\left[\overline{X}_n - 1.96\frac{4}{\sqrt{10}}, \overline{X}_n + 1.96\frac{4}{\sqrt{10}}\right].$$

3. Substituint segons les nostres observacions obtenim l'interval

$$\left[\overline{x} - 1.96\frac{4}{\sqrt{10}}, \overline{x} + 1.96\frac{4}{\sqrt{10}}\right].$$

Com que podem calcular la mitjana $\overline{x} = 180.6$ obtenim l'interval

$$\left[180.6 - 1.96\left(\frac{4}{\sqrt{10}}\right),\ 180.6 + 1.96\left(\frac{4}{\sqrt{10}}\right)\right] = [178.121,\ 183.079].$$

Observació 17. Alguns comentaris sobre la notació i les idees generals dels intervals de confiança

1. Quan parlem d'interval de confiança al 90% serà el mateix que si diem que estem fent un interval de confiança $\gamma = 0.9$

2. L'interval que s'obté depèn de les observacions inicials. És a dir, per a cada conjunt d'observacions –en el nostre exemple, per a cada grup de 10 alumnes– obtindrem un interval diferent! Aleshores, com hem d'interpretar el nivell de confiança? Un interval de confiança per a μ amb un nivell de confiança del 95% ens indica que, per cada 100 mostres que agafem, dels 100 intervals diferents que obtindrem, 95 d'aquests intervals contindran el valor real de μ, desconegut per nosaltres.

3. La tria de u_1 i de u_2 hauria pogut ser diferents, podrien ser qualssevol valors que complissin

$$P(u_1 \leq Z \leq u_2) = 0.95$$

haurien servit, però per comoditat acostumem a agafar valors de manera que l'interval sigui simètric.

4. Si l'interval és simètric, com passa en aquest cas, observem que la longitud de l'interval és:

$$L = 2 \cdot 1.96\frac{4}{\sqrt{10}}$$

5. Si ens interessa estar molt segurs d'on és el paràmetre i augmentem el nivell de confiança, fixeu-vos que a mesura que augmentem la confiança obtindrem intervals més grans. En el nostre exemple el cas extrem seria una confiança del 100%, on hauríem d'agafar $u_1 = -\infty$ i $u_2 = \infty$ i obtindríem un interval amb tots els valors reals.

6. La manera d'obtenir intervals petits –que és el que ens interessa– és augmentar la mida de la mostra. Això és degut al factor $\frac{1}{\sqrt{n}}$ que apareix. Però, el fet d'augmentar la mida de la mostra és car, per tant hem de trobar l'equilibri entre el cost que representa augmentar la mida de la mostra i el nivell de confiança que necessitem.

7. La funció Z que hem utilitzat

$$Z = \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

és una **funció pivotant**. Una funció és una funció pivotant si la seva distribució és coneguda i no depèn del paràmetre μ.

4.2 Intervals de confiança per una població normal

En aquest apartat suposarem que tenim una mostra aleatòria $X_1, \ldots, X_n$ d'una $N(\mu, \sigma^2)$.

4.2.1 Intervals de confiança per a μ, on σ és coneguda

Volem construir un interval de confiança per a μ suposant que el valor de σ és conegut amb un nivell de confiança γ. Estem doncs, en el mateix cas que l'exemple anterior.

Els passos que seguim per calcular-lo són:

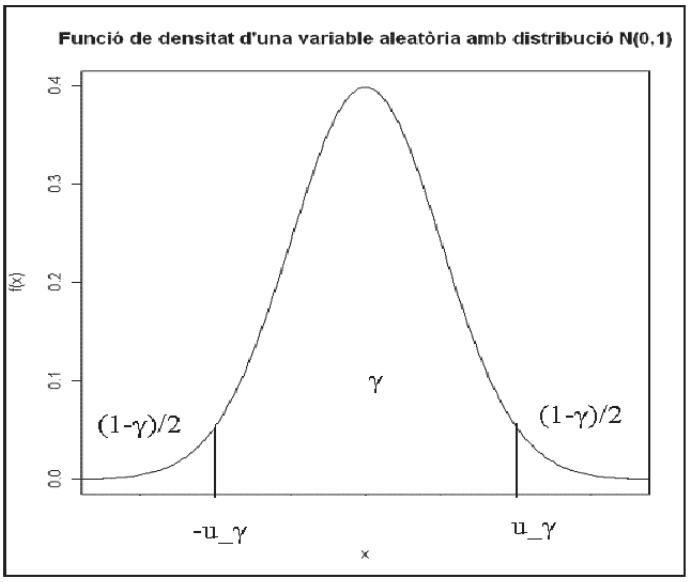
1. Utilizarem que

$$Z = \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1).$$

2. Calculem el valor crític u_γ tal que

$$P(-u_\gamma \leq Z \leq u_\gamma) = \gamma \Rightarrow P(Z < u_\gamma) = 1 - \frac{1-\gamma}{2}$$

en aquest cas $u_\gamma = \texttt{qnorm}(1 - \frac{1-\gamma}{2})$.



3. L'interval serà

$$\overline{X}_n \pm u_\gamma \frac{\sigma}{\sqrt{n}}$$

o el que és el mateix

$$\left[\overline{X}_n - u_\gamma \frac{\sigma}{\sqrt{n}}, \overline{X}_n + u_\gamma \frac{\sigma}{\sqrt{n}}\right].$$

4.2.2 Intervals de confiança per a μ, on σ és desconeguda

Volem construir un interval de confiança per a μ suposant que la σ és desconeguda amb un nivell de confiança γ.

Els passos que seguim per calcular-lo són:

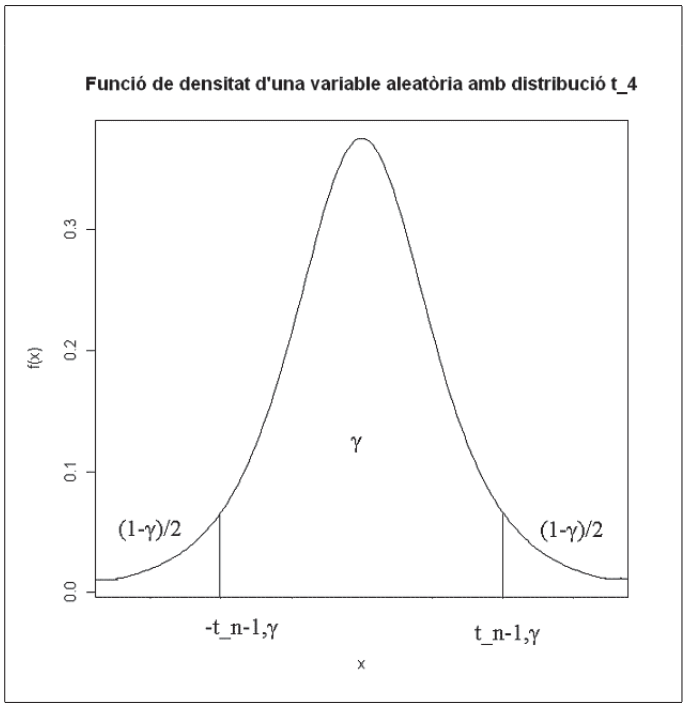
1. Utilizarem la desviació estàndard mostral enlloc de la σ, en aquest cas ens queda que

$$T_{n-1} = \frac{\overline{X}_n - \mu}{\frac{S_n}{\sqrt{n-1}}} \sim t_{n-1}.$$

2. Calculem els valors crítics $t_{n-1,\gamma}$ tal que

$$P\left(-t_{n-1,\gamma} \leq T_{n-1} \leq t_{n-1,\gamma}\right) = \gamma \Rightarrow P(T_{n-1} < t_{n-1,\gamma}) = 1 - \frac{1-\gamma}{2}$$

en aquest cas $t_{n-1,\gamma} = \texttt{qt}(1 - \frac{1-\gamma}{2}, n-1)$.



3. L'interval serà

$$\overline{X}_n \pm t_{n-1,\gamma}\frac{S_n}{\sqrt{n-1}}$$

o el que és el mateix

$$\left[\overline{X}_n - t_{n-1,\gamma}\frac{S_n}{\sqrt{n-1}}, \overline{X}_n + t_{n-1,\gamma}\frac{S_n}{\sqrt{n-1}}\right].$$

Observació 18. La t de Student només la farem servir en els casos en què tinguem mostres de mida $n < 30$, en la resta de casos podem aproxima la t de Student per una $N(0,1)$.

4.2.3 Interval de confiança per a la variància σ^2

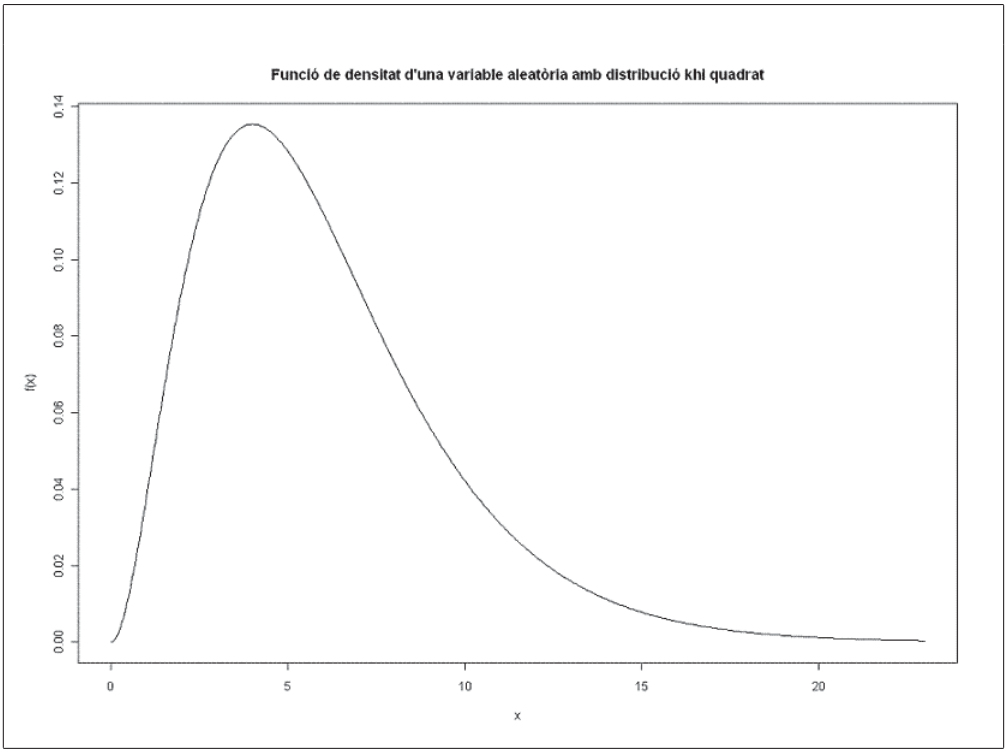
Volem construir un interval de confiança per a σ^2 amb un nivell de confiança γ.

Per construir un interval de confiança per a σ^2, utilitzarem un resultat que ens diu que si $Z_1, \ldots, Z_n$ són variables aleatòries independents idènticament distribuïdes amb distribució normal estàndard, aleshores

$$Z_1^2 + \ldots + Z_n^2 \sim \chi_n^2.$$

on χ_n^2 és una distribució **khi quadrat amb n graus de llibertat**. La distribució χ_n^2, **khi quadrat amb n graus de llibertat**, és una distribució:

- No simètrica,
- que només pren valors positius,
- $E(\chi_n^2) = n$ i $Var(\chi_n^2) = 2 \cdot n$.



Fent servir aquesta nova distribució, podem veure que

Proposició 12. *Si* $X_1, X_2, ..., X_n$ *és una mostra aleatòria simple d'una distribució normal* $N(\mu, \sigma^2)$, *(això vol dir que* $X_i \sim N(\mu, \sigma^2)$*), aleshores*

$$\frac{nS_n^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}{\sigma^2} \sim \chi^2_{n-1}$$

Ara doncs, els passos que seguim per calcular un interval de confiança per a σ^2 són:

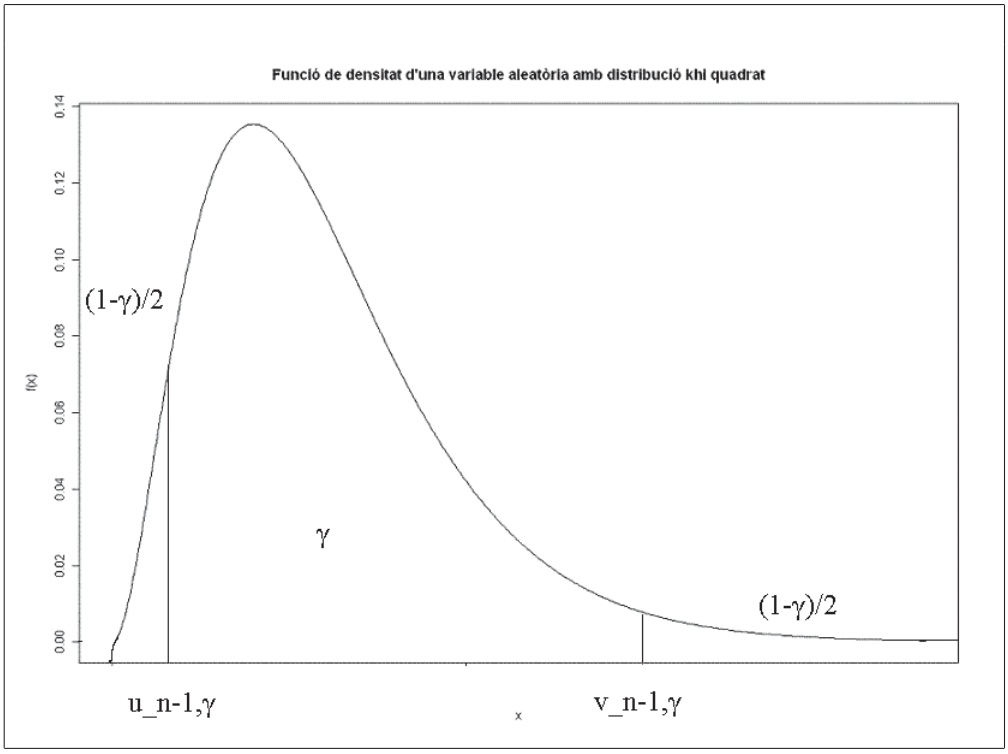
1. Utilitzant el resultat de la proposició anterior tenim

$$\frac{nS_n^2}{\sigma^2} \sim \chi^2_{n-1}$$

2. Calculem els valors crítics $u_{n-1,\gamma}$ i $v_{n-1,\gamma}$ tals que

$$P\left(u_{n-1,\gamma} \leq \chi^2_{n-1} \leq v_{n-1,\gamma}\right) = \gamma,$$

en aquest cas $u_{n-1,\gamma}$ =`qchisq` $(\frac{1-\gamma}{2}, n-1)$ i $v_{n-1,\gamma}$ =`qchisq` $(1\text{-}\frac{1-\gamma}{2}, n-1)$.



3. L'interval serà

$$\left[\frac{nS_n^2}{v_{n-1,\gamma}}, \frac{nS_n^2}{u_{n-1,\gamma}}\right].$$

Exemple 45. *Una mostra aleatòria de les hores que dormen 10 treballadors d'una empresa ens dóna els següents resultats*

$$8.3,\ 8.7,\ 10.1,\ 4.6,\ 7.7,\ 5.4,\ 5.8,\ 6.6,\ 13.1,\ 11.$$

Suposem que les dades són normals.

Volem trobar un interval de confiança al 95% per:

1. *la mitjana hores que dormen, si es coneix que* $\sigma = 2$
   Com que σ és coneguda el nostre interval de confiança serà

$$\left[\overline{x} - u_\gamma \frac{\sigma}{\sqrt{n}}, \overline{x} + u_\gamma \frac{\sigma}{\sqrt{n}}\right].$$

   Hem de calcular: $\overline{x} = 8.13$ i $u_{0.95}$=`qnorm(1-0.05/2)`=1.96, per tant

$$\left[8.13 - 1.96\frac{2}{\sqrt{10}}, 8.13 + 1.96\frac{2}{\sqrt{10}}\right] = [6.89, 9.37].$$

2. *La mitjana hores que dormen, si no coneixem el valor de* σ
   Com que σ és desconeguda el nostre interval de confiança serà

$$\left[\overline{x} - t_{n-1,\gamma} \frac{s_n}{\sqrt{n-1}}, \overline{x} + t_{n-1,\gamma} \frac{s_n}{\sqrt{n-1}}\right].$$

   Hem de calcular: $\overline{x} = 8.13$, $s_n = 2.55$ i $t_{9,0.95}$ =`qt(1-0.05/2,9)`=2.26, per tant

$$\left[8.13 - 2.26\frac{2.55}{\sqrt{9}}, 8.13 + 2.26\frac{2.55}{\sqrt{9}}\right] = [6.209, 10.051].$$

3. *La variabilitat de les hores que dormen els treballadors d'aquesta empresa* Com que les dades són normal l'interval de confiança que volem calcular serà

$$\left[\frac{ns_n^2}{v_{n-1,\gamma}}, \frac{ns_n^2}{u_{n-1,\gamma}}\right].$$

Tenim calculada $s_n = 2.55$, i ens falta calcular $u_{9,0.95} =$ `qchisq(0.05/2,9)`=2.7 i $v_{9,0.95} =$ `qchisq(1-0.05/2,9)`=19.02, per tant

$$\left[\frac{10 \cdot 2.55^2}{19.02}, \frac{10 \cdot 2.55^2}{2.7}\right] = [3.42, 24.08]$$

4.3 Intervals de confiança per la proporció

Fins ara hem treballat per trobar intervals de confiança per paràmetres (μ i σ) d'una distribució normal. Ara canvíem una mica de situació, voldrem trobar un interval de confiança per una proporció.

La situació serà la següent: tenim una mostra aleatòria simple $X_1, \ldots, X_n$ que vé d'una distribució Bernouilli $Ber(p)$, el nostre objectiu serà construir un interval de confiança per a la p amb un nivell de confiança γ.

Els passos que seguim per calcular-lo són:

1. Utilitzant el teorema del límit central tenim que quan la n és gran

$$\frac{\sqrt{n}\left(\overline{X}_n - p\right)}{\sqrt{p(1-p)}}$$

es comporta (no és ben bé igual però s'aproxima prou bé) com una $N(0,1)$.

2. Calculem els valors crítics u_γ tal que

$$P\left(-u_\gamma \leq Z \leq u_\gamma\right) = \gamma \Rightarrow P(Z < u_\gamma) = 1 - \frac{1-\gamma}{2}$$

en aquest cas $u_\gamma =$ `qnorm`$(1 - \frac{1-\gamma}{2})$.

3. L'interval serà

$$\overline{X}_n \pm u_\gamma \sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}$$

o el que és el mateix

$$\left[\overline{X}_n - u_\gamma \sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}, \overline{X}_n + u_\gamma \sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\right].$$

Observem que $\overline{X}_n$ serà la proporció observada a partir de la mostra. Notem també que hem utilitzat $S_n = \overline{X}_n(1-\overline{X}_n)$ enlloc de $\sigma = p(1-p)$ ja que p és el paràmetre que volem estimar.

Exemple 46. *Un radar fix situat en una carretera detecta que 56 cotxes dels 489 que han passat per aquell punt un dia determinat anaven a una velocitat superior a la permesa. Volem trobar un interval de confiança al 90% per a la proporció de cotxes que sobrepassen el límit de velocitat en aquest punt.*

Com que volem calcular un interval de confiança d'una proporció, utilitzarem que serà de la forma

$$\left[\overline{x} - u_\gamma \sqrt{\frac{\overline{x}(1-\overline{x})}{n}}, \overline{x} + u_\gamma \sqrt{\frac{\overline{x}(1-\overline{x})}{n}}\right].$$

on $\overline{x} = \frac{56}{489} = 0.1145$ i $u_{0.9} = \texttt{qnorm(0.9+0.1/2)}$=1.645, per tant

$$\left[0.1145 - 1.645\sqrt{\frac{0.1145 \cdot (1-0.1145)}{489}}, 0.1145 + 1.645\sqrt{\frac{0.1145 \cdot (1-0.1145)}{489}}\right] = [0.0908, 0.1382].$$

4.4 Intervals de confiança per a mostres grans

Hem vist fins ara com calcular intervals de confiança per a paràmetres en el cas que la distribució sigui normal i també en el cas d'una proporció. Però ens preguntem què passa si no tenim cap d'aquests casos, si no que les nostres dades segueixen una distribució qualsevol. Aleshores podrem trobar intervals de confiança per l'esperança sempre que treballem amb mostres grans ($n > 30$).

Així doncs, la situació ara és: tenim $X_1, \ldots, X_n$ una mostra aleatòria simple tal que $n > 30$ d'una variable X amb distribució desconeguda i $E(X) = \mu$, aleshores en aquest cas també podem trobar un interval de confiança per a la μ de la següent manera:

1. Utilizarem pel teorema del límit central que

$$Z = \frac{\overline{X}_n - \mu}{\frac{S_n}{\sqrt{n-1}}}$$

   es comporta com una $N(0,1)$

2. Calculem els valors crítics u_γ tal que

$$P(-u_\gamma \leq Z \leq u_\gamma) = \gamma \Rightarrow P(Z < u_\gamma) = 1 - \frac{1-\gamma}{2}$$

   en aquest cas $u_\gamma = \texttt{qnorm}(1 - \frac{1-\gamma}{2})$.

3. L'interval serà

$$\overline{X}_n \pm u_\gamma \frac{S_n}{\sqrt{n-1}}$$

   o el que és el mateix

$$\left[\overline{X}_n - u_\gamma \frac{S_n}{\sqrt{n-1}}, \overline{X}_n + u_\gamma \frac{S_n}{\sqrt{n-1}}\right].$$

Exemple 47. *Volem estudiar el temps de vida –el temps que poden funcionar fins a espatllar-se– d'una certa classe de bombetes. Hem observat els temps de vida de 1.000 bombetes i hem obtingut unes observacions $x_1, ..., x_{1000}$ tals que $\overline{x} = 1.840$ i $s^2 = 96.5$. A partir d'aquestes dades volem obtenir un interval de confiança al 90% per al temps de vida mitjà de les bombetes.*

En principi no sabem la distribució del temps de vida d'una bombeta. De totes maneres, com que la mostra és molt gran, podem fer un interval de confiança per a l'esperança del temps de vida, que serà

$$\left[\overline{x} - u_\gamma \frac{s_n}{\sqrt{n-1}}, \overline{x} + u_\gamma \frac{s_n}{\sqrt{n-1}}\right].$$

Només necessitem calcular $u_{0.9}$ = `qnorm(1-0.1/2)`=1.64, per tant l'interval de confiança per a la mitjana serà:

$$\left[1.840 - 1.64\left(\frac{\sqrt{96.5}}{\sqrt{999}}\right), 1.840 + 1.64\left(\frac{\sqrt{96.5}}{\sqrt{999}}\right)\right] = [1839.490, 1840.509].$$

Fixeu-vos que com que la mostra és molt gran, l'interval de confiança és bastant petit.

4.5 Intervals de confiança per a dues mostres

Tots els intervals que hem calculat fins ara depenien d'una sola mostra. El problema que se'ns planteja ara és calcular intervals de confiança per a dues mostres. Aquestes mostres són observacions de la mateixa variable en condicions diferents. Ens interessarà estudiar la variabilitat de la variable.

Quan treballem amb dues mostres, la primera característica en la que ens hem de fixar és si les dades són aparellades o no.

Definició 19. *Una **mostra de dades aparellades** és un conjunt d'observacions de dues variables observades en un únic conjunt d'individus, de manera que per a cada individu tenim una observació de cada una de les dues variables.*

Aquest factor és important tenir-lo en compte ja que els càlculs que hem de fer en cada cas són bastant diferents.

4.5.1 Intervals de confiança per a dues mostres de dades aparellades

Per calcular intervals de confiança en aquest cas, no necessitarem saber res de nou. La manera de procedir és: crear una nova variable que serà la variació entre les dades de la primera mostra i les dades de la segona mostra i que s'obté restant les dues variables. Així doncs retornem al cas d'una sola mostra i tot el que hem fet a l'apartat anterior ho podem aplicar per aquesta nova variable que hem creat.

Exemple 48. *Després d'un règim de 3 mesos, el nivell de glucosa de 6 malalts va presentar la variació següent:*

abans règim (mg/dL)	130	121	100	110	110	99
després règim (mg/dL)	105	110	101	82	80	97

Volem estudiar com ha variat el nivell de glucosa. O sigui, volem calcular un interval de confiança al 90% per a la diferència de mitjanes.

Observeu que les dades són aparellades, de manera que si les restem obtenim una nova variable que serà la *variació en el nivell de glucosa:*

abans règim (mg/dL)	130	121	100	110	110	99
després règim (mg/dL)	105	110	101	82	80	97
variació nivell glucosa	-25	-11	1	-28	-30	-2

Podem considerar, per tant, que tenim una única variable. Suposant que tenim normalitat per a les dades inicials, tenim també normalitat per a la nova variable. Així podem buscar un interval de confiança per a la mitjana d'aquesta variable –estem en el cas que hem d'obtenir un interval de confiança per a la mitjana quan la variància és desconeguda i tenim una població normal. Per tant l'interval serà

$$\left[\overline{x} - t_{n-1,\gamma}\frac{s_n}{\sqrt{n-1}}, \overline{x} + t_{n-1,\gamma}\frac{S_n}{\sqrt{n-1}}\right].$$

Necessitem calcular:

$$n = 6 \qquad \overline{x} = -15.833 \quad s_n = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}} = 12.4554.$$

i $t_{5,0.9}$ = `qt(0.9+0.1/2,5)` =2.015. Substituint trobem finalment l'interval per a la mitjana:

$$\begin{aligned} & \left[-15.833 - 2.015\left(\frac{12.4554}{\sqrt{5}}\right), -15.833 + 2.015\left(\frac{12.4554}{\sqrt{5}}\right)\right] \\ = \; & [-27.057,\ -4.6089]. \end{aligned}$$

Per tant, la mitjana de la variació en el nivell de glucosa està en aquest interval amb un nivell de confiança del 90%, és a dir, hi ha hagut una reducció del nivell de glucosa d'entre 4.6089 i 27.057. Fixeu-vos que es tracta d'un interval bastant gran, ja que tenim una mostra molt petita.

4.5.2 Intervals de confiança per a dues mostres de dades no aparellades

Suposarem a partir d'ara que tenim dues mostres aleatòries simples $X_1, \ldots, X_n$ i $Y_1, \ldots, Y_m$ independents entre elles que provenen de d'unes distribucions $N(\mu_x, \sigma_x^2)$ i $N(\mu_y, \sigma_y^2)$ respectivament.

Intervals de confiança per a $\mu_x - \mu_y$, on σ_x^2 i σ_y^2 són conegudes

Volem construir un interval de confiança per a $\mu_x - \mu_y$ suposant que σ_x^2 i σ_y^2 són conegudes amb un nivell de confiança γ.

Els passos que seguim per calcular-lo són:

1. Sabem que

$$\overline{X}_n \sim N\left(\mu_x, \frac{\sigma_x^2}{n}\right) \quad \text{i} \quad \overline{Y}_m \sim N\left(\mu_y, \frac{\sigma_y^2}{m}\right)$$

   i que $\overline{X}_n$ i $\overline{Y}_m$ són independents, per tant

$$\overline{X}_n - \overline{Y}_m \sim N\left(\mu_x - \mu_y, \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}\right).$$

   així doncs,

$$\frac{\overline{X}_n - \overline{Y}_m - (\mu_x - \mu_y)}{\left(\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}\right)^{\frac{1}{2}}} \sim N(0,1).$$

2. Calculem els valors crítics u_γ tal que

$$P\left(-u_\gamma \leq Z \leq u_\gamma\right) = \gamma \Rightarrow P(Z < u_\gamma) = 1 - \frac{1-\gamma}{2}$$

en aquest cas $u_\gamma = \texttt{qnorm}(1 - \frac{1-\gamma}{2})$.

3. L'interval serà

$$\overline{X}_n - \overline{Y}_m \pm u_\gamma \left(\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}\right)^{\frac{1}{2}}$$

o el que és el mateix

$$\left[\overline{X}_n - \overline{Y}_m - u_\gamma \left(\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}\right)^{\frac{1}{2}}, \overline{X}_n - \overline{Y}_m + u_\gamma \left(\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}\right)^{\frac{1}{2}}\right].$$

Intervals de confiança per a $\mu_x - \mu_y$, on σ_x^2 i σ_y^2 desconegudes però iguals

Volem construir un interval de confiança per a $\mu_x - \mu_y$ suposant que $\sigma^2 = \sigma_x^2 = \sigma_y^2$ són desconegudes amb un nivell de confiança γ.

Els passos que seguim per calcular-lo són:

1. Sabem com en el cas anterior que

$$\frac{\overline{X}_n - \overline{Y}_m - (\mu_x - \mu_y)}{\sigma\left(\frac{1}{n} + \frac{1}{m}\right)^{\frac{1}{2}}} \sim N(0,1).$$

Però no podem utilitzar aquest resultat ja que no coneixem el valor de σ. El que fem és substituir el valor de σ per les desviacions típiques respectives S_X^2 i S_Y^2, aleshores si considerem

$$T_{n+m-2} = \frac{\overline{X}_n - \overline{Y}_m - (\mu_x - \mu_y)}{\left(nS_X^2 + mS_Y^2\right)^{\frac{1}{2}}} \sqrt{\frac{nm(n+m-2)}{n+m}}$$

T_{n+m-2} té una distribució t de student amb $n + m - 2$ graus de llibertat.

2. Calculem els valors crítics $t_{n+m-2,\gamma}$ tals que

$$P\left(-t_{n+m-2,\gamma} \leq T_{n+m-2} \leq t_{n+m-2,\gamma}\right) = \gamma \Rightarrow P(T_{n+m-2} < t_{n+m-2,\gamma}) = 1 - \frac{1-\gamma}{2}$$

en aquest cas $t_{n+m-2,\gamma} = \texttt{qt}(1 - \frac{1-\gamma}{2})$.

3. L'interval serà

$$\overline{X}_n - \overline{Y}_m \pm t_{n+m-2,\gamma} \sqrt{\frac{(n+m)\left(nS_X^2 + mS_Y^2\right)}{nm(n+m-2)}}$$

o el que és el mateix

$$\left[\overline{X}_n - \overline{Y}_m - t_{n+m-2,\gamma} \sqrt{\frac{(n+m)\left(nS_X^2 + mS_Y^2\right)}{nm(n+m-2)}}, \overline{X}_n - \overline{Y}_m + t_{n+m-2,\gamma} \sqrt{\frac{(n+m)\left(nS_X^2 + mS_Y^2\right)}{nm(n+m-2)}}\right].$$

Exemple 49. *És conegut que les notes de l'examen d'Estadística segueixen una distribució normal i les notes de l'examen de Probabilitats també segueixen una distribució normal. Volem estudiar si hi ha diferència entre la mitjana de les notes.*
Per fer-ho disposem de les notes de dos grups diferents d'estudiants:

Estadística	4.17	6	4.67	4.83	5	6.5	4	5.5	5.17	5.33	2.33	7.50	7
Probabilitats	5.1	5.2	4.1	4.9	2.3	4.3	3	3	5.1	4.9	4.1		

Ho farem determinant un interval de confiança al 90% per a la diferència de mitjanes. Observem en primer lloc que les dades no són aparellades, corresponen a diferents estudiants. Calcularem aquest interval en dues situacions diferents.

1. *Suposem que les notes de l'examen d'Estadística tenen una variància coneguda* 1.21 *i les notes de l'examen de Probabilitats també tenen una variància coneguda en aquest cas* 1*.*

   Com que estem en el cas que les desviacions són conegudes l'interval que hem de calcular serà de la forma

   $$\left[\overline{x}-\overline{y}-u_\gamma\left(\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}\right)^{\frac{1}{2}},\overline{x}-\overline{y}+u_\gamma\left(\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}\right)^{\frac{1}{2}}\right].$$

   On

   $$n=13,\ \sigma_x^2=1.21,\ m=11 \text{ i } \sigma_y^2=1$$

   i també hem de calcular

   $$\overline{x}=5.23 \quad \text{i} \quad \overline{y}=4.18.$$

   i finalment $u_{0.9}=$ `qnorm(0.95)`=1.64.
   Així, obtenim l'interval de confiança per a $\mu_x-\mu_y$

   $$\left[5.23-4.18-1.64\left(\frac{1.21}{13}+\frac{1}{11}\right)^{\frac{1}{2}},5.23-4.18+1.64\left(\frac{1.21}{13}+\frac{1}{11}\right)^{\frac{1}{2}}\right]$$
   $$=[0.3466,\ 1.7534].$$

2. *Suposem que les notes de l'examen d'Estadística i les notes de l'examen de Probabilitats tenen variàncies desconegudes però iguals*

   Com que estem en el cas de desviacions desconegudes però iguals l'interval serà de la forma

   $$\left[\overline{x}-\overline{y}-t_{n+m-2,\gamma}\sqrt{\frac{(n+m)\,(ns_x^2+ms_y^2)}{nm(n+m-2)}},\overline{x}-\overline{y}+t_{n+m-2,\gamma}\sqrt{\frac{(n+m)\,(ns_x^2+ms_y^2)}{nm(n+m-2)}}\right].$$

   Ens queda calcular $s_x^2=1.70$, $s_y^2=0.92$ i $t_{22,0.9}=$ `qt(0.95,22)`=1.72.

   Així, obtenim l'interval de confiança per a $\mu_x-\mu_y$

   $$\left[5.23-4.18-1.72\sqrt{\frac{24(13\cdot 1.70+11\cdot 0.92)}{13\cdot 11\cdot 22}},5.23-4.18+1.72\sqrt{\frac{24(13\cdot 1.70+11\cdot 0.92)}{13\cdot 11\cdot 22}}\right]$$
   $$=[0.1980,1.9019].$$

Observació 19. Fixeu-vos que el zero –és a dir, quan les mitjanes són iguals– no està en cap d'aquests intervals, així que no podem dir que la mitjana sigui la mateixa en les dues assignatures.

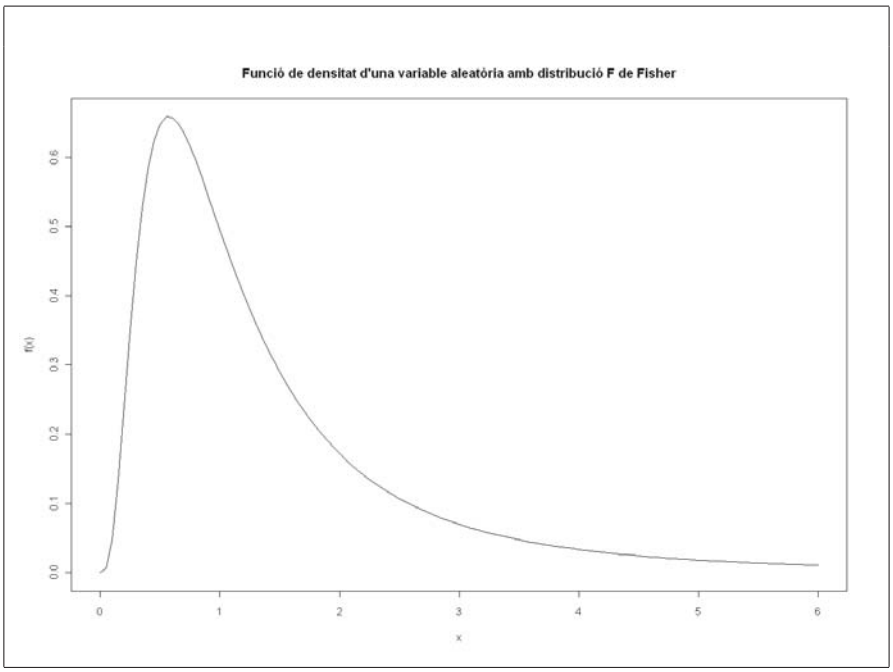
Interval de confiança per a la raó de variàncies

Volem construir un interval de confiança per a $\frac{\sigma_x^2}{\sigma_y^2}$ amb un nivell de confiança γ. Per construir aquest interval utilitzarem que si $U \sim \chi_n^2$ i $V \sim \chi_m^2$ són dues variables aleatòries independents, aleshores

$$\frac{\frac{U}{n}}{\frac{V}{m}} \sim F_{n,m}$$

on $F_{n,m}$ té una té una distribució F **de Fisher amb** n **i** m **graus de llibertat**. La distribució $F_{n,m}$**, de Fisher amb** n **i** m **graus de llibertat**, és una distribució:

- No simètrica,
- que només pren valors positius.



Fent servir aquesta nova distribució, podem veure que:

Proposició 13. *Si* $X_1, X_2, ..., X_n$ *una mostra aleatòria d'una distribució normal* $N(\mu_x, \sigma^2)$ *i* $Y_1, Y_2, ..., Y_m$ *una mostra aleatòria d'una distribució normal* $N(\mu_y, \sigma^2)$*, i suposem a més que les dues mostres són independents i que no coneixem els valors poblacionals de* μ_x, μ_y *i* σ^2*. Aleshores sabem que,*

$$\begin{aligned} \frac{nS_X^2}{\sigma^2} &= \frac{\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2}{\sigma^2} \sim \chi_{n-1}^2, \\ \frac{mS_Y^2}{\sigma^2} &= \frac{\sum_{i=1}^m \left(Y_i - \overline{Y}_m\right)^2}{\sigma^2} \sim \chi_{m-1}^2. \end{aligned}$$

I es compleix que,

$$\frac{\frac{\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2}{n-1}}{\frac{\sum_{i=1}^m \left(Y_i - \overline{Y}_m\right)^2}{m-1}} = \frac{(m-1)\left(\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2\right)}{(n-1)\left(\sum_{i=1}^m \left(Y_i - \overline{Y}_m\right)\right)} \sim F_{n-1,m-1}.$$

Ara doncs, els passos que seguim per calcular l'interval de confiança que volíem són:

1. Utilitzant el resultat de la proposició anterior tenim

$$\frac{1}{\sigma_x^2}\sum_{i=1}^{n}\left(X_i-\overline{X}_n\right)^2=\frac{nS_X^2}{\sigma_x^2}\sim\chi^2_{n-1}\quad\text{i}\quad\frac{1}{\sigma_y^2}\sum_{i=1}^{m}\left(Y_i-\overline{Y}_m\right)^2=\frac{mS_Y^2}{\sigma_y^2}\sim\chi^2_{m-1}$$

i a més també sabem que són independents, aleshores tenim que

$$F_{n-1,m-1}=\frac{\frac{nS_X^2}{(n-1)\sigma_x^2}}{\frac{mS_Y^2}{(m-1)\sigma_y^2}}=\frac{\sigma_y^2\widetilde{S}_X^2}{\sigma_x^2\widetilde{S}_Y^2},$$

on

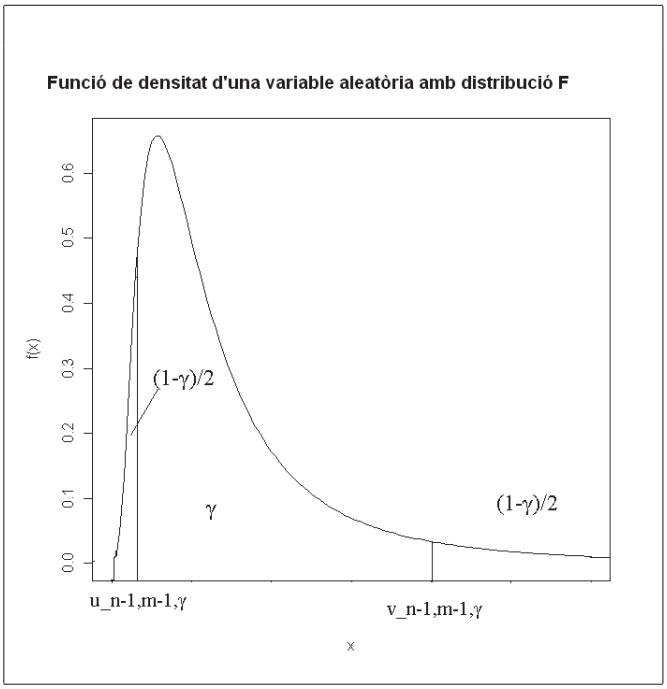
$$\widetilde{S}_X^2=\frac{nS_X^2}{(n-1)}=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}_n\right)^2\quad\text{i}\quad\widetilde{S}_Y^2=\frac{mS_Y^2}{(m-1)}=\frac{1}{m-1}\sum_{i=1}^{m}\left(Y_i-\overline{Y}_m\right)^2.$$

i $F_{n-1,m-1}$ té una té una distribució F **de Fisher amb** $n-1$ **i** $m-1$ **graus de llibertat**.

2. Calculem els valors crítics $u_{n-1,m-1,\gamma}$ i $v_{n-1,m-1,\gamma}$ tals que

$$P\left(u_{n-1,m-1,\gamma}\leq F_{n-1,m-1}\leq v_{n-1,m-1,\gamma}\right)=\gamma.$$

en aquest cas com que la distribució de Fisher no és simètrica agafarem $u_{n-1,m-1,\gamma}$=`qf(`$1-\frac{1-\gamma}{2}$`, n-1, m-1)` i $v_{n-1,m-1,\gamma}$=`qf(`$\frac{1-\gamma}{2}$`, n-1, m-1)`.



3. L'interval serà

$$\left[\frac{\widetilde{S}_X^2}{v_{n-1,m-1,\gamma}\widetilde{S}_Y^2},\frac{\widetilde{S}_X^2}{u_{n-1,m-1,\gamma}\widetilde{S}_Y^2}\right].$$

Exemple 50. *Volem estudiar si la longitud de les ales d'una determinada espècie d'insectes depèn de l'altitud. Per fer-ho agafem 10 individus al nivell del mar i 11 individus a una altitud de 1.000 metres. Obtenim els resultats següents:*

Nivell del mar:	*2.5*	*2.8*	*3.1*	*2.4*	*3.3*	*3.6*	*3*	*2.9*	*2.7*	*3*	
1.000 metres:	*2.4*	*3.2*	*2.6*	*2.3*	*2.9*	*2.8*	*2.2*	*3*	*2.5*	*2.4*	*2.8*

Suposarem normalitat.

Volem comparar per tant les dues mitjanes de les poblacions. És clar que no són dades aparellades.

Com que les variàncies són desconegudes abans de calcular l'interval de confiança per a la diferència de mitjanes hem de comprovar que les variàncies són iguals.

El primer pas serà buscar un interval de confiança al 95% per a la raó de variàncies $\frac{\sigma_x^2}{\sigma_y^2}$, on σ_x^2 és la variància al nivell del mar i σ_y^2 és la variància a 1.000 metres. Per tant serà un interval de la forma

$$\left[\frac{\widetilde{s}_x^2}{v_{n-1,m-1,\gamma}\widetilde{s}_y^2}, \frac{\widetilde{s}_x^2}{u_{n-1,m-1,\gamma}\widetilde{s}_y^2}\right].$$

Hem de calcular

$$n = 10;\ \overline{x} = 2.93,\ \widetilde{s}_x^2 = 0.129,\ s_x^2 = 0.116$$
$$m = 11,\ \overline{y} = 2.64,\ \widetilde{s}_y^2 = 0.100,\ s_y^2 = 0.090.$$

Estem treballant amb una $F_{9,10}$, una F de Fisher amb 9 i 10 graus de llibertat, trobem que $u_{9,10,0.95}$ =`qf(0.025,9,10)`=0.252 i $v_{9,10,0.95}$ =`qf(0.975,9,10)`=3.778.
i això ens dóna l'interval per a $\frac{\sigma_x^2}{\sigma_y^2}$

$$\left[\frac{0.129}{3.778 \cdot 0.1}, \frac{0.129}{0.252 \cdot 0.1}\right] = [0.338,\ 5.076].$$

Com que el valor 1 és a l'interval de confiança per a la raó de variàncies, podrem suposar que $\sigma_x^2 = \sigma_y^2$ –variàncies desconegudes però iguals.

L'interval de confiança per a $\mu_x - \mu_y$ que busquem serà de la forma

$$\left[\overline{x} - \overline{y} - t_{n+m-2,\gamma}\sqrt{\frac{(n+m)\,(ns_x^2 + ms_y^2)}{nm(n+m-2)}}, \overline{x} - \overline{y} + t_{n+m-2,\gamma}\sqrt{\frac{(n+m)\,(ns_x^2 + ms_y^2)}{nm(n+m-2)}}\right].$$

Utilitzant ara una t de Student amb $10 + 11 - 2 = 19$ graus de llibertat, podem determinar $t_{19,0.95}$=`qt(0.975,19)`=2.093 i, substituint, trobem l'interval de confiança per a $\mu_x - \mu_y$. Si anomenem

$$\rho = \sqrt{\frac{21 \cdot (10 \cdot 0,116 + 11 \cdot 0,90)}{10 \cdot 11 \cdot 19}}$$

podem escriure

$$[2.93 - 2.64 - 2.093\rho,\ 2.93 - 2.64 + 2.093\rho] = [-0.024,\ 0.593].$$

Tenim així que els intervals amb un nivell de confiança al 95% :

$$\frac{\sigma_x^2}{\sigma_y^2} \in [0.338,\ 5.076] \quad \text{i} \quad \mu_x - \mu_y \in [-0.024,\ 0.593].$$

Per tant no podem deduir que hi hagi diferències entre la longitud de les ales a nivell del mar o a 1.000 metres.