Universitat Autònoma de Barcelona
Facultat de Ciències
Estadística
M. Saucedo

Seminari 3: Intervals de confiança

En aquesta pràctica veurem com trobar intervals de confiança pels estadístics més coneguts. A part de les estimacions per la mitjana, la variància i la proporció que hem vist (o veureu) a classe de teoria, determinarem intervals de confiança per dues mostres.

Denotem amb γ el nivell de confiança d'un interval i amb α el nivell de significació que compleixen la relació $\gamma = 1 - \alpha$.

Anomenem *regió de confiança* C per a un paràmetre θ a una regió aleatòria de l'espai de paràmetres Θ. La *confiança* de la regió C, com estimador de θ és $\mathbf{P}\{\theta \in \mathbf{C}\}$.

Suposem que $\Theta \subset \mathbb{R}$. A continuació presentarem un mètode heurístic que dóna bons resultats. Suposem que volem construir una regió de confiança $1-\alpha$ per a un paràmetre θ i tenim un estadístic T del que coneixem la seva distribució. Aleshores, podem seguir el següent:

1. Busquem una regió que contingui T amb probabilitat $1 - \alpha$ i que escollim, per comoditat, com un interval
$$a'(\theta) < T < b'(\theta).$$
2. Trobem el conjunt $C(T)$ dels valors de θ que compleixen les desigualtat anteriors:
$$C(T) = \{\theta : a'(\theta) \leq T \leq b'(\theta)\}.$$
Aleshores, $\mathbf{P}\{\theta \in C(T)\} = \mathbf{P}\{a'(\theta) \leq T \leq b'(\theta)\} = 1 - \alpha$ i solucionem per θ.

Quan la distribució de l'estadístic T no depèn del paràmetre θ, aleshores T s'anomena *pivot*.

1 Interval de confiança per la mitjana μ

Suposem que tenim una mostra d'una $N(\mu, \sigma)$ i considerem els diferents casos per obtenir una estimació de la mitjana.

1.1 Amb σ coneguda.

A teoria hem vist que, quan la variància és coneguda,

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

A partir d'aquest resultat i utilitzant els estadístics de la mostra, podem deduir el següent interval de confiança per la mitjana:

$$\mu \in \left(\overline{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right).$$

Recordeu que $z_{1-\frac{\alpha}{2}}$ és el valor tal que $P(Z \leq z_{1-\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$, on $Z \sim N(0, 1)$. Com hem vist a la pràctica anterior, amb `R` el podem calcular utilitzant la funció `qnorm`. Per exemple, per $\gamma = 0.95$,

```
z <- qnorm(0.975)
```

R no conté cap funció que calculi directament l'interval de confiança per la mitjana quan σ és coneguda, per tant, haurem d'utilitzar la fórmula.

Què és el coeficient de confiança?

Per entendre el significat del coeficient de confiança com una probabilitat podem repetir l'experiment moltes vegades. Obtindrem que la freqüència relativa de l'esdeveniment *{L'interval conté μ}* serà pròxima al valor teòric γ. Demostrem-ho amb un exemple.

Exemple 1. Generem $N = 1000$ mostres aleatòries normals amb

$$n = 10, \quad \mu = 4, \quad \sigma = 2.$$

Organitzant la llista de 10000 nombres aleatoris en 1000 files i 10 columnes, de forma que obtenim una matriu on cada fila és una mostra de mida 10:

```
N <- 1000
n <- 10
mu <- 4
sigma <- 2
x <- rnorm(N*n, mu, sigma)
dim(x) <- c(N, n)
```

Calculem el vector M de les 1000 mitjanes empíriques i, a partir d'ell, els vectors A i B que contenen les vores inferior i superior, respectivament, de l'interval de confiança per a cadascuna de les $B = 1000$ mostres.

```
M <- apply(x, 1, mean)
err <- qnorm(0.975) * sigma / sqrt(n)
A <- M - err
B <- M + err
```

Verifiquem, per a cada interval, si el valor teòric $\mu = 4$ hi pertany i comptem quantes vegades passa aquest esdeveniment:

```
u <- (A<4) & (B>4)
sum(u) / length(u)
```

El resultat de u és un vector Booleà, que pren el valor TRUE, és a dir, 1, si l'interval conté 4 i FALSE, és a dir, 0, si no el conté. La proporció de valors TRUE obtinguda (en una determinada realització de la sèrie d'experiments) s'aproxima al valor $\gamma = 0.95$.

Exercici 1. Feu $N = 1000$ simulacions de $n = 20$ valors d'una variable aleatòria X amb llei $N(m, 1)$, amb $m = 3$. Amb cadascuna de les simulacions calculeu un interval de confiança del 90% per a la mitjana de X. Calculeu la proporció dels N intervals obtinguts que contenen el valor m.

1.2 Amb σ desconeguda

Quan la variància és desconeguda, podem obtenir resultats satisfactoris utilitzant la variància mostral corregida. A teoria hem vist que

$$\frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

on t_{n-1} és una t-Student amb $n-1$ graus de llibertat.

A partir d'aquest resultat i utilitzant els estadístics de la mostra, podem deduir el següent interval de confiança per la mitjana:

$$\mu \in \left(\overline{x} - t_{n-1,1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \overline{x} + t_{n-1,1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \right).$$

En aquest cas, hi ha una manera de fer aquest càlcul directament amb `R`. La funció:

```
t.test(x, conf.level = 0.90)$conf.int
```

dóna l'interval de confiança que busquem. Si no s'especifica el nivell de confiança, la funció `t.test` retorna per defecte un interval del 95%. La instrucció

```
t.test(x, conf.level=0.90)
```

dóna més informació.

Exercici 2. Es fan 12 determinacions de la quantitat d'argent d'un mineral (en mg) i s'obté:

5.2 4.8 5.3 5.7 5.0 4.7 4.3 5.5 5.4 5.1 4.9 5.8

Suposeu normalitat i determineu l'interval de confiança amb nivell de confiança 0.95 per la mitjana suposant variància desconeguda. Si augmenteu el nivell de confiança a 0.99, obteniu un interval més o menys ample?

2 Interval de confiança per a una proporció

Pels resultats vists a teoria sabem que l'interval de confiança per una proporció és

$$p \in \left(\widehat{p} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{n}}, \widehat{p} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{n}} \right)$$

on $\widehat{p}$ és la proporció mostral.

Amb `R` es pot calcular directament aquest interval amb la funció `prop.test`:

```
prop.test(x, n, conf.level = 0.90)$conf.int
```

on `x` és el nombre d'elements que presenten la característica d'interès i `n` és el nombre total d'elements.

Si no s'especifica el nivell de confiança, la funció `t.test` retorna per defecte un interval del 95%. Com pel cas de `t.test`, la instrucció

```
prop.test(x, n, conf.level = 0.90)
```

dóna més informació.

Exercici 3. Es vol estimar la proporció p d'animals de granja d'una certa espècie amb una malaltia congènita a una determinada zona. Per això s'agafa una mostra de 200 animals i s'observa que el percentatge de malalts és del 8%. Calculeu l'interval de confiança per a p amb nivell de confiança 0.92.

3 Interval de confiança per a σ

Suposem que tenim una mostra normal. Se sap que

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

A partir d'aquest resultat, la fórmula per l'interval de confiança per a la variància és:

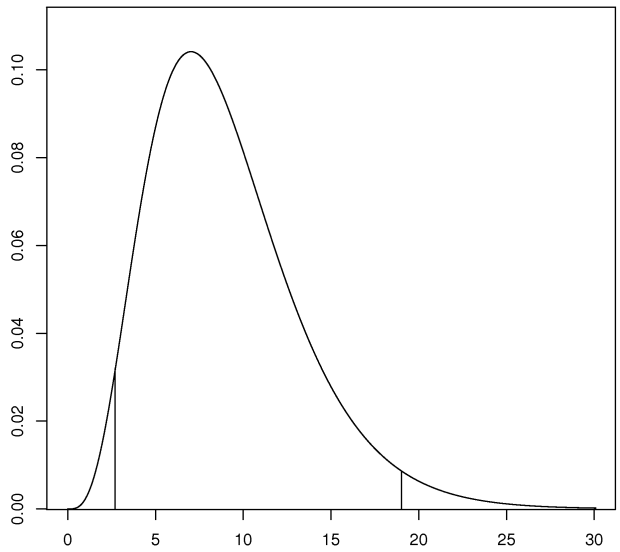
$$\sigma^2 \in \left(\frac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}, \frac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}} \right),$$

on $\chi^2_{n-1,\frac{\alpha}{2}}$ és el valor d'una distribució chi-quadrat amb $n-1$ graus de llibertat que deixa a la seva dreta una probabilitat de $\frac{\alpha}{2}$. La definició de $\chi^2_{n-1,1-\frac{\alpha}{2}}$ és equivalent substituint $\frac{\alpha}{2}$ per $1-\frac{\alpha}{2}$.

Exemple 2. Per $n = 10$ i $\gamma = 0.95$, amb `R` els valors de $\chi^2_{n-1,\frac{\alpha}{2}}$ i $\chi^2_{n-1,1-\frac{\alpha}{2}}$ s'obtenen de la següent manera:

```
n <- 10
gm <- 0.95
alp <- 1 - gm
a <- qchisq(alp/2, 9)
b <- qchisq(1-alp/2, 9)
```

Els podem representar a la gràfica de la funció de densitat de χ^2_{n-1}:



Les instruccions per a dibuixar la figura anterior són:

```
z <- seq(0, 30, by=0.1)
y <- dchisq(z,9)
plot(z, y, type = 'l', xlab = "", ylab = "", ylim = c(0.0041, 0.11))
lines(c(a,a), c(0,dchisq(a,n-1)))
lines(c(b,b), c(0,dchisq(b,n-1)))
```

`R` no conté cap funció que calculi directament l'interval de confiança per a la variància, per tant, haurem de calcular la fórmula.

Exercici 4. Tenim les següents notes d'examen de 12 alumnes:

7.9 8.3 4.8 8.4 7.9 5.2 5.6 3.2 9.1 7.7 6.5 4.4

Calculeu un interval de confiança del 93% per a la variància. Suposeu normalitat.

4 Intervals de confiança per a dues mostres

Suposem que tenim dues mostres que provenen de la distribució normal:

$$X_1, \ldots, X_{n_x} \quad \text{iid} \sim N(\mu_x, \sigma_x^2), \qquad Y_1, \ldots, Y_{n_y} \quad \text{iid} \sim N(\mu_y, \sigma_y^2).$$

4.1 Dades aparellades

X_i són independents (entre elles), les Y_j són independents (entre elles). Hi ha el mateix nombre d'observacions, $n_x = n_y \equiv n$, que venen per parelles, $(X_1, Y_1), \ldots, (X_n, Y_n)$ i, en general, cada X_i és independent de les *altres* Y_j, amb $j \neq i$, però no de la seva parella Y_i. Simètricament, cada Y_i és independent de les *altres* X_j, amb $j \neq i$, però no de la seva parella X_i. Es consideren les n diferències,

$$D_i = X_i - Y_i, \quad 1 \leq i \leq n.$$

Aquestes noves variables aleatòries són normals donat que ho són les variables X_i i Y_i. Es procedeix amb D com al cas d'una sola variable normal.

Si volem calcular l'interval de confiança per la diferència de les mitjanes, en aquest cas, tenim una funció de `R` que ens ho calcula automàticament emprant

```
t.test(x, y, paired = TRUE, conf.level = 0.9)$conf.int
```

Exercici 5. Les dades següents es varen obtenir d'un experiment dissenyat per estimar la reducció en la pressió arterial després de seguir una dieta sense sal durant dues setmanes:

Abans	93	106	87	92	102	95	88	110
Després	92	102	89	92	101	96	88	105

Calculeu un interval de confiança del 97% per la reducció mitjana. Podem dir que la pressió arterial ha disminuït després de la dieta?
Observació: Veureu que l'interval conté tan valors positius com negatius, per tant no podem afirmar que la pressió arterial hagi disminuït.

4.2 Mostres independents

Les X_i són independents (entre elles), les Y_j són independents (entre elles). Cada X_i és independent de cada Y_j. Distingirem entre els casos quan les variàncies són conegudes i quan no ho són.

4.2.1 Interval de confiança per a $\mu_x - \mu_y$, amb σ_x i σ_y conegudes

L'interval que volem calcular amb nivell de confiança γ és:

$$\left((\overline{x} - \overline{y}) - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}, \quad (\overline{x} - \overline{y}) + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}} \right).$$

En aquest cas, no tenim cap funció de l'`R` que ens el calculi, per tant, haurem d'utilitzar la fórmula.

Exercici 6. En l'estudi d'un tipus d'insectes s'han mesurat les longituds (en mm) de les ales de 15 individus nascuts en el laboratori (L) i 15 individus nascuts salvatges (S). Els resultats són els següents:

L	6.7	1.9	6.4	4.8	2.6	4.9	6.7	3.6	1.5	1.2	2.4	2.4	4.6	4.9	4.8
S	6.2	3.7	4.5	6.2	6.0	5.3	3.5	3.6	3.1	0.3	5.3	4.5	4.5	3.6	4.5

Suposem normalitat. Sabem que la variància de les longituds de les ales dels individus nascuts en el laboratori és 3.5, mentre la dels individus nascuts salvatge és 2.2. Calculeu un interval de confiança del 95% per la diferència de les mitjanes de les dues poblacions.

4.2.2 Interval de confiança per a $\mu_x - \mu_y$, amb σ_x i σ_y desconegudes

Si les variàncies són desconegudes però iguals, l'interval que busquem s'obté amb la següent fórmula:

$$\left((\overline{x} - \overline{y}) - t_{\nu,1-\frac{\alpha}{2}}\, S_c \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}, \quad (\overline{x} - \overline{y}) + t_{\nu,1-\frac{\alpha}{2}}\, S_c \sqrt{\frac{1}{n_x} + \frac{1}{n_y}} \right),$$

on $t_{\nu,1-\frac{\alpha}{2}}$ és el valor d'una distribució t-Student amb $\nu = n_x + n_y - 2$ graus de llibertat que deixa a la seva dreta una probabilitat de $1 - \frac{\alpha}{2}$ i S_c^2 és la variància conjunta que es defineix com

$$S_c^2 = \frac{(n_x - 1)S_x^2 + (n_y - 1)S_y^2}{n_x + n_y - 2}.$$

Si les variàncies són desconegudes i no podem assumir que siguin iguals, l'interval que busquem s'obté amb la següent fórmula:

$$\left((\overline{x} - \overline{y}) - t_{k,1-\frac{\alpha}{2}} \sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}, \quad (\overline{x} - \overline{y}) + t_{k,1-\frac{\alpha}{2}} \sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}} \right),$$

on $t_{k,1-\frac{\alpha}{2}}$ és el valor d'una distribució t-Student amb k graus de llibertat que deixa a la seva dreta una probabilitat de $1 - \frac{\alpha}{2}$. El graus de llibertat es calculen com

$$k = \frac{\left(\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}\right)^2}{\frac{\left(\frac{S_x^2}{n_x}\right)^2}{n_x - 1} + \frac{\left(\frac{S_y^2}{n_y}\right)^2}{n_y - 1}}.$$

La funció `t.test` de l'`R` ens permet calcular directament aquests dos interval de confiança utilitzant l'opció `var.equal`. Quan les variàncies són iguals, utilitzem

```
t.test(x, y, var.equal = TRUE, conf.level = 0.9)$conf.int
```

En cas contrari,

```
t.test(x, y, var.equal = FALSE, conf.level = 0.9)$conf.int
```

Exercici 7. Considerem les dades sobre les longitud d'ales d'un tipus d'insectes de l'estudi anterior. Suposem que no coneixem les variàncies. Calculeu els intervals de confiança del 95% per la diferència de les mitjanes de les dues poblacions en el cas que les variàncies es puguin assumir iguals i en cas contrari.

Problemes

1. Obriu el fitxer de dades d'R anomenat `malaria.RData` que trobeu a

   `https://mat.uab.cat/~mbarcelona/est/malaria.RData`

   Aquest document conté les dades de 120 persones que han contret la malària. Les variables són:

   Edat en el moment de l'estudi.
   Sexe (D=Dona, H=Home).

   a) Trobeu un interval de confiança del 95% per la mitjana d'edat. Suposeu normalitat.
   b) Calculeu un interval de confiança del 92% per la proporció de dones i un per la proporció d'homes.
   c) Determineu un interval de confiança del 93% per la diferencia de mitjanes entre l'edat dels homes i l'edat de les dones. Suposeu que les dues poblacions tenen la mateixa variància.
   d) Trobeu un interval de confiança del 90% per la variància de la variable *Edat*. Suposeu normalitat.

2. El fitxer `Nadons.RData` conté les dades de 22 nadons.

   `https://mat.uab.cat/~mbarcelona/est/Nadons.RData`

   Les variables són:

   pH pH del cordó umbilical.
   Mare Indica si la Mare és fumadora (F=Fumadora, NF=No Fumadora).

   Suposeu normalitat.

   a) Calculeu un interval de confiança del 90% per la mitjana del pH.
   b) Calculeu un interval de confiança del 90% per la mitjana del pH dels nadons amb mares fumadores i un altre per la la mitjana del pH dels nadons amb mares no fumadores.
   c) Calculeu l'interval de confiança del 95% per a la diferència de les mitjanes del pH de les dones fumadores i de les no fumadores. Amb aquest interval podeu dir quina mitjana és més elevada?

3. Les notes de 13 alumnes d'una classe en dos exàmens parcials van ser:

Alumnes	1	2	3	4	5	6	7	8	9	10	11	12	13
1er Parcial	7.9	5.4	8.3	6.2	8.2	8.3	7.8	4.9	6.2	8.9	7.8	9.7	7.2
2on Parcial	8.2	5.7	6.0	4.2	7.5	4.6	6.2	5.2	5.3	9.2	6.5	8.1	4.5

   a) Calculeu dos intervals de confiança del 90%: un per la mitjana de les notes del primer parcial i un per la mitjana de les notes del segon parcial. Només observant els dos intervals, podeu dir que la nota mitjana del primer parcial és més alta que la del segon? Canvia la resposta si agafeu un nivell de confiança del 80%?
   b) Calculeu l'interval de confiança del 92% per a la diferència de mitjanes de les notes del primer i del segon examen. Observant aquest interval, podeu dir quina és la nota mitjana més alta?

c) El primer parcial val un 40% de la nota final i el segon parcial un 60%. Calculeu l'interval de confiança del 95% per a la mitjana de les notes finals del curs.

d) Calculeu dos intervals de confiança del 93%: un per la variància de les notes del primer parcial i un per la variància de les notes del segon parcial. Només observant els dos intervals, podeu dir quines dades tenen més variabilitat?