Pràctica 7

Intervals de confiança

7.1 Interval de confiança per la mitjana μ

Suposem que tenim una mostra d'una Normal (μ, σ).

7.1.1 Amb σ coneguda.

Donat el nivell de confiança γ, calculem a, per Z una Normal(0,1),

$$\text{Prob}\,(|Z| < a) = \gamma.$$

Per exemple, per $\gamma = 0.95$,

```
a<-qnorm(0.975)
```

Per n i σ donats, aïllem μ de

$$0.95 = \text{Prob}\left(-a < \frac{\overline{X}_n - \mu}{\sigma}\sqrt{n} < a\right) = \text{Prob}\left(\overline{X}_n - a\,\frac{\sigma}{\sqrt{n}} < \mu < \overline{X}_n + a\,\frac{\sigma}{\sqrt{n}}\right)$$

Per exemple, per $\sigma = 2$, $n = 10$, calculem

```
d<-qnorm(0.975)*2/sqrt(10)
```

Resulta $d = 1.23959$, i l'interval és $\overline{X}_n \pm 1.23959$.

Per exemple, si del resultat de generar una mostra aleatòria normal amb $n = 10$, $\mu = 4$, $\sigma = 2$, obtenim la mitjana empírica:$\overline{x_{10}} = 3.963238$, l'interval de confiança per a μ, amb coeficient $\gamma = 0.95$, és:

$$(\overline{x_{10}} - d,\, \overline{x_{10}} + d) = (3.963238 - 1.23959, 3.963238 + 1.23959) = (2.723647, 5.202828)$$

7.1.2 Què és el coeficient de confiança?

Per entendre el significat del coeficient de confiança com una probabilitat podem repetir l'experiment moltes vegades.

La freqüència relativa de l'esdeveniment *l'interval conté μ* serà pròxima al valor teòric γ.

Per simular-ho, generem $B = 1000$ mostres aleatòries normals amb

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

organitzant una 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:

```
X<-rnorm(10000,4,2)
dim(X)<-c(1000,10)
```

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)
d<-qnorm(0.975)*2/sqrt(10)
A<-M-d
B<-M+d
```

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

```
u<-(A<4) & (4<B)
sum(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) és 0.964, aproximat al valor $\gamma = 0.95$.

7.1.3 Amb σ desconeguda

Donat el nivell de confiança γ, calculem a, per a T_n un t-Student de paràmetre n (mida de la mostra),

$$\text{Prob}\,(|T_n| < a) = \gamma.$$

Per exemple, per $\gamma = 0.90$, si $n = 10$ com abans, amb la distribució $t(n-1)$,

```
a<-qt(0.95,9)
```

Com abans aïllem μ,

$$0.90 = \text{Prob}\left(-a < \frac{\overline{X}_n - \mu}{\widetilde{S}_n}\sqrt{n} < a\right) = \text{Prob}\left(\overline{X}_n - a\,\frac{\widetilde{S}_n}{\sqrt{n}} < \mu < \overline{X}_n + a\,\frac{\widetilde{S}_n}{\sqrt{n}}\right).$$

Per exemple, per $n = 10$, calculem

```
d<-qt(0.95,9)*sd(x)/sqrt(10)
```

i l'interval de confiança per la mostra anterior que tenia $\overline{x_{10}} = 3.963238$, és

$$(\overline{x_{10}} - d, \overline{x_{10}} + d) = (3.963238 - 1.057015, 3.963238 + 1.057015) = (2.906223, 5.0252).$$

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
```

produeix el mateix resultat que acabem d'obtenir. Per defecte calcula un interval amb coeficient 0.95. Si només posem la instrucció `t.test(x, conf.level=0.90)` obtenim més informació.

7.2 Interval de confiança per a una proporció

Tenim n repeticions independents d'un experiment aleatori on en cada realització observem si es produeix o no un cert esdeveniment A. Les variables indicatrius

$$X_1, \ldots, X_n \quad \text{iid} \sim Ber(p),$$

on $p = \text{Prob}(A)$.

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

$$\left[f - a\frac{\sqrt{f\,(1-f)}}{\sqrt{n}},\ f + a\frac{\sqrt{f\,(1-f)}}{\sqrt{n}} \right]$$

on $f = N/n$ (N és el número de vegades que es produeix l'esdeveniment A en la nostra mostra) i `a<-qt(`γ`,n-1)`. En aquest cas, no tenim cap funció de l'`R` que ens el calculi per tant per calcular-lo suposem que tenim les dades volem calcular un interval de confiança amb coeficient de confiança 0.9 i coneixem les dades anteriors, aleshores:

```
a<-qt(0.95, n-1)
d<-a*sqrt((f(1-f))/n)
d1<-f-d
d2<-f+d
```

I l'interval que buscàvem és exactament $[d_1,\, d_2]$.

7.3 Intervals de confiança per a dues mostres

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

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

7.3.1 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,

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

Les diferències 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.
No cal que σ_x i σ_y siguin iguals. Suposarem desconegudes.
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.int
```

7.3.2 Dues 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.

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:

$$\left[\overline{X} - \overline{Y} - \mathtt{qnorm}(\gamma)\left(\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}\right)^{\frac{1}{2}}, \overline{X} - \overline{Y} + \mathtt{qnorm}(\gamma)\left(\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}\right)^{\frac{1}{2}}\right].$$

En aquest cas, no tenim cap funció de l'`R` que ens el calculi per tant per calcular-lo suposem que tenim les dades volem calcular un interval de confiança amb coeficient de confiança 0.9 i coneixem `x` i `y` i els valors `sigma2x` i `sigma2y` (les variàncies):

```
xm<-mean(x)
yn<-mean(y)
a<-qnorm(0.95)
d<-a*sqrt((sigma2x/length(x))+(sigma2y/length(y)))
d1<-xm-yn-d
d2<-xm-yn+d
```

I l'interval que buscàvem és exactament $[d_1, d_2]$.

Interval de confiança per a $\mu_x - \mu_y$, amb σ comú desconeguda

Aquest resultat s'obté amb:

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

Per defecte `t.test` calcula un interval amb coeficient de confiança 0.95.

Interval de confiança per a σ_x/σ_y

Podem calcular directament l'interval amb:

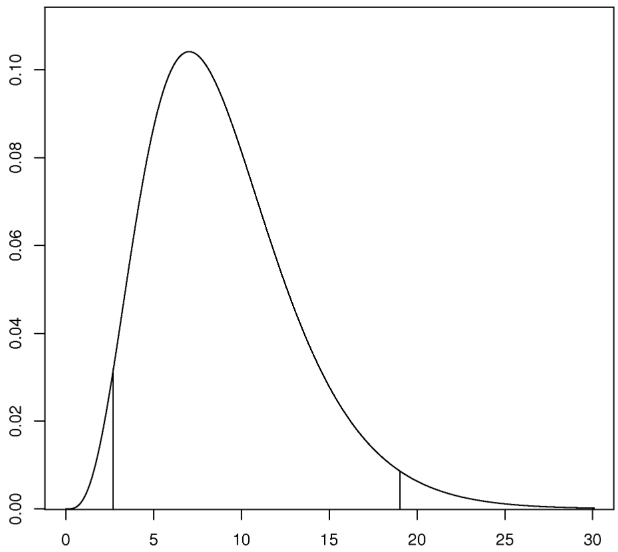
```
var.test(x, y)$conf.int
```

Igual que abans el `var.test` calcula per defecte un interval amb coeficient de confiança 0.95, es pot canviar afegint a la instrucció `conf.level=0.9`.

7.4 Interval de confiança per a σ

Suposem que tenim una mostra normal.
Donat el coeficient de confiança $\gamma \in (0, 1)$, calculem a i b tals que Prob $[a < Q < b] = \gamma$, on $Q \sim \chi^2(n-1)$.
Una manera de fer-ho és repartir $1 - \gamma$ en dues parts iguals, a l'esquerra de a i a la dreta de b.
Per exemple, per $n = 10$ i $\gamma = 0.95$,

```
a<-qchisq(0.025,9)
b<-qchisq(0.975,9)
```



Les instruccions per a dibuixar-lo són:

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

L'interval s'obté aïllant el paràmetre:

$$\frac{n-1}{b}\tilde{S}^2,\ \frac{n-1}{a}\tilde{S}^2.$$

Amb R tenim:

```
d1<-(length(x)-1)/b  · var(x)
d2<-(length(x)-1)/a  · var(x)
```

I l'interval que buscàvem és exactament $[d_1,\, d_2]$.

7.5 Resum-formulari de propietats i distribucions

$$X_1,\ldots,X_n \quad \text{iid} \ \sim\ N(\mu,\sigma^2).$$

Mitjana empírica i mitjana empírica estandarditzada:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim\ N(\mu, \frac{\sigma^2}{n}), \qquad Z = \frac{\overline{X}-\mu}{\sigma}\sqrt{n} \sim\ N(0,1)$$

Sumes de quadrats estandarditzades

$$Q(\mu) = \frac{n\,S^2(\mu)}{\sigma^2} \sim \chi^2(n)$$

$$Q \equiv Q(\bar{X}) = \frac{n\,S^2}{\sigma^2} \sim \chi^2(n-1)$$

El quocient

$$T = \frac{Z}{\sqrt{Q/(n-1)}} = \frac{\overline{X}-\mu}{S}\sqrt{n-1} = \frac{\overline{X}-\mu}{\widetilde{S}}\sqrt{n} \sim t(n-1).$$

7.6 Problemes

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.

   Repetiu amb diferents valors de n (com més gran és n, més estret és l'interval) i amb diferents valors de N (com més gran és N, més s'acosta la proporció al seu valor teòric 0.90).

2. Es fan 5 determinacions de la quantitat d'argent d'un mineral i s'obté:

   5.2,4.8,5.3,5.7 i 5.0 mg d'argent.

   Determineu l'interval de confiança amb coeficient de confiança 0.95 per la mitjana suposant variància desconeguda. Suposeu normalitat

3. Després d'un tractament contra l'obesitat, els pesos en Kg de vuit dones eren:

   58,50,60,65,64,62,56,57

   Si suposem normalitat, calculeu un interval de confiança amb coeficient de confiança 0.95 per la mitjana teòrica si la variància és desconeguda. Què passa si el calculeu amb un coeficient de confiança del 0.9?

4. S'ha mesurat el pH del cordó umbilical de 22 nadons de dones no fumadores i de dones fumadores, obtenint:

   Nadons de dones no fumadores:
   7.28 7.31 7.34 7.34 7.32 7.23 7.31 7.32 7.29 7.35 7.32
   7.34 7.35 7.26 7.18 7.34 7.27 7.34 7.32 7.29 7.26 7.26

   Nadons de dones fumadores:
   7.26 7.27 7.27 7.35 7.29 7.28 7.31 7.34 7.29 7.39 7.21
   7.28 7.30 7.24 7.20 7.28 7.30 7.35 7.32 7.31 7.37 7.26

   (a) Calculeu un interval de confiança del 90% per la mitjana de cada una de les mostres.
   (b) 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.

5. Les notes d'11 almnes d'una classe en dos exàmens consecutius van ser:

Alumnes	1	2	3	4	5	6	7	8	9	10	11
Examen1	7.9	8.3	6.2	8.2	8	7.8	4.9	6.2	8.9	7.8	9.4
Examen2	8.2	7.1	4.8	8.4	7.9	7.4	5.2	5.6	9.2	6.5	8.5

   (a) Calculeu l'interval de confiança del 95% per a la diferència de mitjanes de les notes del primer i del segon examen.
   (b) Calculeu l'interval de confiança del 80% per a la diferència de mitjanes de les notes del primer i del segon examen.Compara els resultats amb l'apartat anterior

6. S'ha dut a terme un estudi per veure l'efecte de l'exercici físic sobre el nivell de colesterol en sang. Es va analitzar la sang de 11 individus abans i després de fer exercici obtenint els següents resultats:

Individus	1	2	3	4	5	6	7	8	9	10	11
Abans	182	232	191	200	148	249	276	213	241	480	262
Després	198	210	194	220	138	220	219	161	210	313	226

(a) Quin és el nivell mig de colesterol en sang abans i després de la prova? Construir un interval de confiança al 95% per cada un d'ells.

(b) Construir un interval de confiança al 95% per a la diferència mitjana entre abans i després del tractament.

7. 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

Calculeu un interval de confiança del 95% per la diferència de les mitjanes de les dues poblacions.

8. 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.