Contrast de dues mostres

Josep Gibergans Bàguena

P08/05057/02309

Universitat Oberta de Catalunya
www.uoc.edu

Índex

Contrastos sobre la diferència de mitjanes

1. Introducció

En aquesta sessió veurem com hem de fer un contrast d'hipòtesis sobre diferències de mitjanes poblacionals. Considerarem dues mostres d'observacions i en compararem les mitjanes contrastant hipòtesis sobre la seva diferència i construint intervals de confiança per a aquesta diferència.

Atenció!

És molt important no confondre aquests tipus de problemes amb els de "dades aparellades" en els quals tenim una mostra d'observacions de dues variables.

Per exemple, podem estar interessats a conèixer si hi ha una diferència significativa entre els següents aspectes:

a) Els sous dels homes i de les dones que treballen en una mateixa empresa.

b) El consum de combustible de dos vehicles de marques diferents.

c) El temps de vida de bombetes de dues marques diferents, etc.

2. Contrastos sobre la diferència de mitjanes

Suposem que tenim una mostra aleatòria de grandària n_1 obtinguda d'una població 1 de mitjana μ_1 i una mostra aleatòria independent de l'anterior de grandària n_2 obtinguda d'una població 2 de mitjana μ_2. Volem contrastar la hipòtesi nul·la (H_0) que afirma que els valors de les mitjanes de totes dues poblacions són iguals:

1) Hipòtesi nul·la: H_0: $\mu_1 = \mu_2$

Notació

A vegades en lloc de:

H_0: $\mu_1 = \mu_2$

escriurem:

H_0: $\mu_1 - \mu_2 = 0$

2) Hipòtesi alternativa. La hipòtesi alternativa (H_1) pot ser bilateral o unilateral:

a) Bilateral. La mitjana d'una població és superior o inferior a la de l'altra població.

$$H_1: \quad \mu_1 \neq \mu_2$$

b) Unilateral, en què es consideren dues situacions:

- La mitjana de la població 1 és superior a la mitjana de la població 2:

$$H_1: \quad \mu_1 > \mu_2$$

- La mitjana de la població 1 és inferior a la mitjana de la població 2:

$$H_1: \quad \mu_1 < \mu_2$$

De la mateixa manera que en el contrast de la mitjana, suposarem poblacions normals i considerarem el cas de variàncies poblacionals conegudes i el de variàncies poblacionals desconegudes.

2.1. Cas de variàncies poblacionals conegudes

A partir de dues poblacions normals $N(\mu_1, \sigma_1)$ i $N(\mu_2, \sigma_2)$ es prenen dues mostres de grandàries n_1 i n_2. Les mitjanes mostrals $\bar{X}_1$ i $\bar{X}_2$ estan distribuïdes segons les distribucions normals:

$$\overline{X_1} \sim N\left(\mu_1, \frac{\sigma_1}{\sqrt{n_1}}\right) \quad \text{i} \quad \overline{X_2} \sim N\left(\mu_2, \frac{\sigma_2}{\sqrt{n_2}}\right)$$

A partir de les propietats de la distribució normal, la variable aleatòria $\overline{X_1} - \overline{X_2}$ també es distribueix normalment:

$$(\bar{X}_1 - \bar{X}_2) \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

Suma de distribucions

Recordem que donades dues distribucions normals: $Z_1 \sim N(\mu_1, \sigma_1)$ i $Z_2 \sim N(\mu_2, \sigma_2)$ la variable $Z_1 + Z_2$ es distribueix segons una normal:

$$N\left(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}\right)$$

Tipificant aquesta variable aleatòria, obtenim una nova variable que es distribueix segons una $N(0,1)$:

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

Si suposem que la hipòtesi nul·la és certa, aleshores $\mu_1 - \mu_2 = 0$; per tant:

$$\frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

En resum, sota el supòsit de la hipòtesi nul·la certa (H_0: $\mu_1 = \mu_2$) tenim que l'estadístic de contrast:

$$Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sigma_{\bar{X}_1 - \bar{X}_2}} \text{ , on } \sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$ és l'**error estàndard,**

correspon a una observació d'una llei $N(0,1)$.

Una vegada hem calculat el valor de l'estadístic de contrast, hem de determinar el p-valor. El p-valor depèn de la hipòtesi alternativa plantejada:

El p-valor

El p-valor és la probabilitat que un resultat sigui "al menys tan extrem" com l'estadístic de contrast obtingut.

- Si H_1: $\mu_1 - \mu_2 \neq 0$, aleshores $p = 2P(Z > |z|)$
- Si H_1: $\mu_1 - \mu_2 < 0$, aleshores $p = P(Z < z)$
- Si H_1: $\mu_1 - \mu_2 > 0$, aleshores $p = P(Z > z)$

Direm que el p-valor és significatiu i rebutjarem la hipòtesi nul·la si és més petit que el nivell de significació α fixat.

Exemple de contrast sobre la diferència de mitjanes en el cas de variàncies poblacionals conegudes

Un fabricant de vidres vol comparar la resistència mitjana dels vidres que fabrica amb la resistència dels que fabrica la competència. S'agafa una mostra de vidres de cada fabricant i es mesuren les resistències. S'obtenen els resultats següents:

Fabricant:	$n_1 = 150$,	$\bar{x}_1 = 111{,}2$
Competència:	$n_2 = 125$,	$\bar{x}_2 = 109{,}6$

Suposant que les mostres són independents i s'han obtingut de dues poblacions normals amb desviacions típiques conegudes $\sigma_1 = 10{,}4$ i $\sigma_2 = 12{,}5$, a quina conclusió podem arribar al 5% de significació?

1. Expressem les hipòtesis:

- Hipòtesi nul·la: H_0: $\mu_1 - \mu_2 = 0$
- Hipòtesi alternativa: H_1: $\mu_1 - \mu_2 \neq 0$

2. Determinem el nivell de significació: $\alpha = 0{,}05$

3. Calculem l'estadístic de contrast, que segueix una distribució $N(0,1)$:

$$z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{111{,}2 - 109{,}6}{1{,}40} = 1{,}14$$

Abans, però, hem hagut de calcular l'error estàndard:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{10{,}4^2}{150} + \frac{12{,}5^2}{125}} = 1{,}40$$

4. Ara ja estem en condicions de calcular el p-valor:

$$p = 2P(Z > |z|) = 2\,P(Z > 1{,}14) = 2 \cdot 0{,}13 = 0{,}26$$

5. Com que $0{,}26 > 0{,}05$, no podem rebutjar la hipòtesi nul·la, és a dir, no podem dir que les resistències dels vidres dels diferents fabricants siguin diferents.

2.2. Cas de variàncies poblacionals desconegudes però iguals

En la vida real, pràcticament mai no es coneixen les variàncies poblacionals, per això és important aquest cas en què considerem que les variàncies poblacionals són desconegudes però iguals a un cert valor σ^2, és a dir, $\sigma_1^2 = \sigma_2^2 = \sigma^2$, amb σ desconeguda. Aquesta desviació típica comuna σ es pot estimar per mitjà de la fórmula:

$$s = \sqrt{\frac{1}{n_1 + n_2 - 2}\left(\sum_{i=1}^{n_1}(x_{i1} - \bar{x}_1)^2 + \sum_{i=1}^{n_2}(x_{i2} - \bar{x}_2)^2\right)}$$

Fórmula

Aquesta fórmula per a estimar σ és semblant a la de la desviació típica per a una única mostra. La diferencia està en el fet que se sumen els totals dels termes al quadrat per separat i després és divideixen per la grandària mostral total menys dos.

on x_{i1} és la *i*-èsima observació de la mostra 1 i x_{i2} és *i*-èsima observació de la mostra 2. També podem escriure aquesta expressió de la manera següent:

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

on s_1^2 i s_2^2 són les variàncies mostrals obtingudes a partir de les mostres.

Si la hipòtesi nul·la és certa ($\mu_1 = \mu_2$), obtenim l'**estadístic de contrast** següent:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(\bar{x}_1 - \bar{x}_2)}{s_{(\bar{X}_1 - \bar{X}_2)}} \text{ amb un } \textbf{error estàndard } s_{(\bar{x}_1 - \bar{x}_2)} = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

En aquest cas, l'estadístic de contrast correspon a una observació d'una distribució t de Student amb $n_1 + n_2 - 2$ graus de llibertat.

Com sempre, a continuació calcularem el *p*-valor corresponent a l'estadístic de contrast calculat. Depenent de la hipòtesi alternativa, tenim:

- Si H_1: $\mu_1 - \mu_2 \neq 0$, aleshores $p = 2P(t_{n1+n2-2} > |t|)$
- Si H_1: $\mu_1 - \mu_2 < 0$, aleshores $p = P(t_{n1+n2-2} < t)$
- Si H_1: $\mu_1 - \mu_2 > 0$, aleshores $p = P(t_{n1+n2-2} > t)$

Segons quin sigui el *p*–valor obtingut en comparació amb el nivell de significació α escollit, rebutjarem o no la hipòtesi nul·la d'igualtat de mitjanes.

Exemple de contrast sobre la diferència de mitjanes en el cas de variàncies poblacionals desconegudes però iguals

Un fabricant de bombetes assegura que les seves bombetes tenen més durada que les d'una nova marca coreana. A partir de la durada (en hores) de $n_1 = 25$ bombetes del fabricant i de $n_2 = 15$ bombetes de la nova marca, triades aleatòriament, hem obtingut:

Per al fabricant:	$\bar{x}_1 = 827$,	$s_1^2 = 9.005$
Per a la nova marca:	$\bar{x}_2 = 812$,	$s_2^2 = 7.984$

Suposarem que totes dues poblacions es distribueixen normalment amb variàncies iguals i desconegudes. Farem un contrast d'hipòtesis a un nivell del 0,05 per determinar si tal com sembla, el fabricant té raó.

Hem de contrastar la diferència de mitjanes per a saber si hi ha una diferència significativa o podem considerar que aquestes són iguals. Ens donen les mitjanes i les variàncies mostrals i desconeixem les variàncies poblacionals, que suposarem iguals. Amb tots aquests supòsits, fem el següent:

1. Expressem les hipòtesis:

- Hipòtesi nul·la: H_0: $\mu_1 - \mu_2 = 0$
- Hipòtesi alternativa: H_1: $\mu_1 - \mu_2 > 0$

2. Determinem el nivell de significació: $\alpha = 0{,}05$

3. Calculem l'estadístic de contrast:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s_{\bar{x}_1 - \bar{x}_2}} = \frac{827 - 812}{0{,}959} = 15{,}64$$

que segueix una distribució *t* de Student amb $n_1 + n_2 - 2 = 38$ graus de llibertat i en què:

$$s_{\bar{x}_1 - \bar{x}_2} = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 2.937\sqrt{\frac{1}{25} + \frac{1}{15}} = 0{,}959$$

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(25-1)9,005 + (15-1)7,984}{25 + 15 - 2}} = \sqrt{\frac{24 \cdot 9,005 + 14 \cdot 7,984}{38}} =$$

$$= \sqrt{\frac{216{,}120 + 111{,}776}{38}} = \sqrt{\frac{327{,}896}{38}} = \sqrt{8,628842} = 2,937$$

4. Calculem el *p*-valor: $p = P(t_{38} > t) \cong 0$

5. Com que $0 < 0{,}05$, aleshores rebutgem la hipòtesi nul·la en favor de la hipòtesi alternativa. Efectivament, el fabricant té raó, és a dir, la mitjana de vida de les seves bombetes és superior a la de la nova marca.

2.3. Cas amb mostres grans no normals

En el cas que no es pugui assegurar que les mostres provenen de poblacions normals, només podrem contrastar la diferència de mitjanes si les grandàries de les mostres són superiors a trenta.

El teorema del límit central ens diu que si les grandàries de les mostres són superiors a 30, l'**estadístic de contrast**:

$$z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(\bar{x}_1 - \bar{x}_2)}{s_{\bar{x}_1 - \bar{x}_2}}$$

és una observació d'una variable aleatòria que és distribueix aproximadament com una $N(0,1)$.

En aquest mòdul no considerarem el cas de mostres petites ($n < 30$) de distribucions no normals.

Exemple de contrast sobre la diferència de mitjanes en el cas de mostres grans no normals

Volem saber si hi ha una diferència significativa entre el consum de benzina de dos motors de cotxe diferents. Recollim les dades sobre el consum a partir d'una mostra aleatòria i independent de cada tipus de motor i obtenim els resultats següents:

Motor 1: $n_1 = 80$, $\bar{x}_1 = 11{,}2$, $s_1 = 2{,}2$

Motor 2: $n_2 = 75$, $\bar{x}_2 = 11{,}8$, $s_2 = 3{,}7$

No tenim informació sobre el tipus de distribució que tenen els consums d'aquests motors. Amb un nivell de significació de l'1%, podem assegurar que el consum és el mateix?

Seguint el mateix plantejament de sempre, farem un contrast d'hipòtesi de la diferència de mitjanes.

1. Expressem les hipòtesis:

- Hipòtesi nul·la: H_0: $\mu_1 - \mu_2 = 0$
- Hipòtesi alternativa: H_1: $\mu_1 - \mu_2 \neq 0$

2. Determinem el nivell de significació: $\alpha = 0{,}01$

3. Calculem l'estadístic de contrast: en aquest cas, no coneixem la distribució del consum, però com que les mostres tenen una grandària superior a 30, l'estadístic de contrast ve donat per:

$$z = \frac{(\bar{x}_1 - \bar{x}_2)}{s_{\bar{x}_1 - \bar{x}_2}} = \frac{11{,}2 - 11{,}8}{0{,}493} = -1{,}217 \approx -1{,}22$$

que segueix una distribució $N(0,1)$ i on:

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_3}} = \sqrt{\frac{2{,}2^2}{80} + \frac{3{,}7^2}{75}} = 0{,}493$$

4. Calculem el p–valor: $p = 2P(Z > |z|) = 2P(Z > |-1{,}22|) = 2 \cdot 0{,}1112 = 0{,}2224$

5. Com que $0{,}2224 > 0{,}01$, aleshores no rebutgem la hipòtesi nul·la en favor de la hipòtesi alternativa, és a dir, el consum no és significativament diferent per a cada tipus de motor.

3. Intervals de confiança per a la diferència de mitjanes

El procediment que cal seguir per a construir un interval de confiança al voltant de la diferència de mitjanes és el següent:

1) Fixem el nivell de confiança, que escrivim com a $(1 - \alpha)$.

2) Calculem l'error estàndard de la diferència de mitjanes.

3) Calculem el valor crític tenint en compte el tipus de distribució del nostre estadístic de contrast.

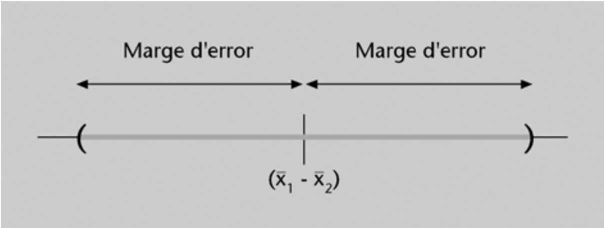
4) Calculem el marge d'error que és el producte entre el valor crític i l'error estàndard.

L'**interval de confiança per a la diferència de mitjanes** ve donat per:

$$(\bar{x}_1 - \bar{x}_2) \pm \text{(valor crític) (error estàndard)}$$

Aquest interval de confiança conté la diferencia de mitjanes amb un nivell de certesa igual a $(1 - \alpha)\%$.

Gràficament, el podem representar així:



A continuació, buscarem per a cada cas l'interval de confiança corresponent.

3.1. Cas de variàncies poblacionals conegudes

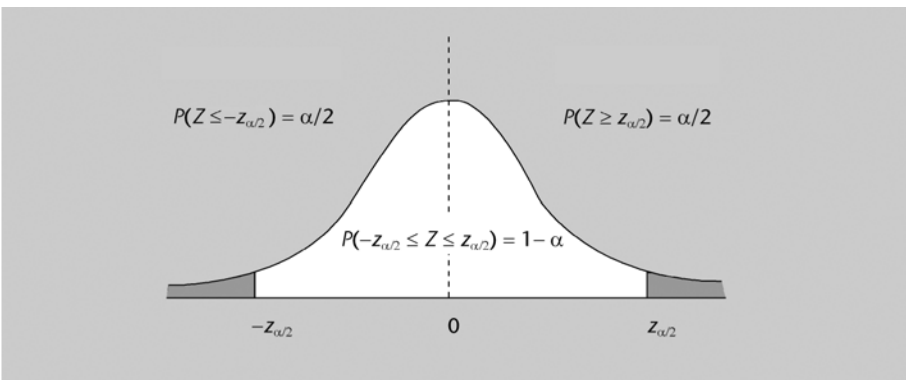
Ja hem dit abans que, donades dues mostres de grandàries n_1 i n_2 obtingudes de dues poblacions normals $N(\mu_1, \sigma_1)$ i $N(\mu_2, \sigma_2)$, la diferència de les seves mitjanes es distribueix segons una llei normal:

$$(\overline{X}_1 - \overline{X}_2) \sim N\left(\mu_1 - \mu_2,\ \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

Suposem que volem un nivell de confiança de l'$(1 - \alpha)$%. En primer lloc, considerarem la variable tipificada:

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

i a continuació construirem un interval centrat en $Z = 0$, de manera que la probabilitat que la variable aleatòria Z prengui un valor en aquest interval sigui d'$1 - \alpha$:



$$P\left(-z_{\alpha/2} \leq \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

on $z_{\alpha/2}$ i $-z_{\alpha/2}$ són els valors crítics. Són aquells que fan que $P(Z \geq z_{\alpha/2}) = \alpha/2$ i $P(Z \leq -z_{\alpha/2} = \alpha/2$. Treballant una mica amb aquesta expressió, tenim:

$$((\overline{X}_1 - \overline{X}_2) - z_{\alpha/2}\sigma_{\overline{X}_1 - \overline{X}_2} \leq (u_1 - u_2) \leq (\overline{X}_1 - \overline{X}_2) + z_{\alpha/2}\sigma_{\overline{X}_1 - \overline{X}_2})$$

Error estàndard

Recordeu que l'error estàndard, o error estàndard, de la diferència de mitjanes és:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Finalment, substituint els valors mostrals obtindrem el corresponent **interval de confiança** per a la diferència de mitjanes en el cas de variàncies conegudes:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sigma_{\bar{x}_1 - \bar{x}_2}$$

Exemple d'intervals de confiança per a la diferència de mitjanes en el cas de variàncies poblacionals conegudes

Si considerem l'exemple del fabricant de vidres, podem calcular un interval de confiança del 95% per a la diferencia de les mitjanes de les resistències de la manera que expliquem a continuació.

Ja tenim dels càlculs anteriors la diferència de mitjanes i l'error estàndard:

$$(\bar{x}_1 - \bar{x}_2) = 111{,}2 - 109{,}6 = 1{,}6 \quad \text{i} \quad \sigma_{(\bar{x}_1 - \bar{x}_2)} = 1{,}40$$

Els valors crítics per a un $\alpha/2 = 0{,}025$ són $\pm z_{\alpha/2} = \pm 1{,}96$. L'interval de confiança és:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sigma_{\bar{x}_1 - \bar{x}_2}$$

$$1{,}6 \pm 1{,}96 \cdot 1{,}40, \text{ és a dir, } 1{,}6 \pm 2{,}74$$

Per tant, l'interval de confiança és (–1,14; 4,34). És important observar que l'interval conté el valor zero, de manera que tenim un 95% de confiança en pensar que la resistència dels vidres és la mateixa. Aquest resultat l'hem obtingut abans fent el contrast d'hipòtesis.

L'interval de confiança també ens serveix per a fer contrastos d'hipòtesis en cas que la hipòtesi alternativa sigui bilateral.

3.2. Cas de variàncies poblacionals desconegudes i iguals

En el contrast d'hipòtesis per a diferències de mitjanes en el cas de variàncies desconegudes i iguals, hem vist que la variable

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}}$$

segueix una distribució t de Student amb $n_1 + n_2 - 2$ graus de llibertat. De manera que si fixem un nivell de confiança $1 - \alpha$, podem determinar els valors crítics $-t_{\alpha/2, n_1+n_2-2}$ i $t_{\alpha/2, n_1+n_2-2}$ que asseguren que:

$$P\left(-t_{\alpha/2, n_1+n_2-2} \leq \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}} \leq t_{\alpha/2, n_1+n_2-2}\right) = 1 - \alpha$$

Per tant:

$$(\bar{X}_1 - \bar{X}_2) - t_{\alpha/2, n_1+n_2-2} \cdot s_{\bar{x}_1 - \bar{x}_2} \leq \mu_1 - \mu_2 \leq (\bar{X}_1 - \bar{X}_2) + t_{\alpha/2, n_1+n_2-2} \cdot s_{\bar{x}_1 - \bar{x}_2}$$

I substituint els valors mostrals tindrem l'**interval de confiança** per a la diferència de mitjanes en el cas de variàncies desconegudes i iguals:

$$(\bar{x}_1 - \bar{x}_2) \mp t_{\alpha/2, n_1+n_2-2} \cdot s_{\bar{x}_1 - \bar{x}_2}$$

Exemple d'intervals de confiança per a la diferència de mitjanes en el cas de variàncies poblacionals desconegudes i iguals

Si considerem ara les dades de l'exemple del fabricant de bombetes, calcularem un interval de confiança per la diferència de durades. Ja havíem trobat la diferencia de mitjanes i l'error estàndard:

$$(\bar{x}_1 - \bar{x}_2) = 827 - 812 = 15 \quad \text{i} \quad s_{\bar{x}_1 - \bar{x}_2} = 0{,}959$$

Els valors crítics per a $\alpha/2 = 0{,}025$ són $\pm t_{\alpha/2, n_1+n_2-2} = \pm t_{0,025;38} = \pm 2{,}0244$

L'interval de confiança és: $(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2, n_1+n_2-2} \cdot s_{\bar{x}_1 - \bar{x}_2}$

$$15 \pm 2{,}0244 \cdot 0{,}959, \text{ és a dir, } 15 \pm 1{,}9414$$

Per tant, l'interval de confiança per la diferència de durades és: (13,06;16,94). En aquest cas, podem veure que l'interval no conté el zero i que és positiu, per tant, podem concloure que hi ha evidències que les bombetes del fabricant duren més que les noves bombetes coreanes.

3.3. Cas amb mostres no normals

Si les mostres no són normals però les seves grandàries són superiors a trenta, aleshores, pel teorema del límit central, tenim que la variable

$$\frac{(\bar{X}_1 - \bar{X}_2)}{s_{\bar{x}_1 - \bar{x}_2}}$$

està distribuïda $N(0,1)$. De manera que fem les mateixes consideracions que en el cas de la secció 3.1.

L'**interval de confiança** per a la diferència de mitjanes pel cas de mostres no normals grans (grandària superior a trenta) ve donat per:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} s_{\bar{x}_1 - \bar{x}_2}$$

Exemple d'intervals de confiança per a la diferència de mitjanes en el cas de mostres no normals

Considerem ara l'exemple del consum de benzina de dos tipus de motors per il·lustrar aquest darrer cas. Teníem les dades següents, la diferencia de mitjanes i l'error estàndard:

$$(\bar{x}_1 - \bar{x}_2) = 11{,}2 - 11{,}8 = -0{,}6 \text{ i } s_{\bar{x}_1 - \bar{x}_2} = 0{,}493$$

Els valors crítics per $\alpha/2 = 0{,}025$ són $\pm z_{\alpha/2} = \pm 1{,}96$.

L'interval de confiança és:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \cdot s_{\bar{x}_1 - \bar{x}_2}$$

$$-0{,}6 \pm 1{,}96 \cdot 0{,}493, \text{ és a dir, } -0{,}6 \pm 0{,}96628$$

Per tant, l'interval de confiança és: (–1,57; 0,366). És important observar que l'interval conté el valor zero. Aquest resultat és totalment coherent amb l'obtingut amb el contrast d'hipòtesis segons el qual el consum no és significativament diferent per a cada tipus de motor.

4. Resum

En aquesta sessió hem vist com fer contrastos d'hipòtesis per a la diferència de mitjanes de dues mostres aleatòries i independents. Hem distingit tres casos:

1) Mostres normals amb variàncies poblacionals conegudes
2) Mostres normals amb variàncies poblacionals desconegudes i iguals
3) Mostres grans no normals

També hem après a construir intervals de confiança per a la diferència de mitjanes considerant aquests tres mateixos casos.

Exercicis

1. Una botiga d'ordinadors portàtils equipa els seus models amb bateries de la marca Durames. Aquestes bateries són de bona qualitat, però passat un cert temps comencen a donar problemes. Per a intentar de donar una millor qualitat a les vendes, el responsable del negoci es planteja la possibilitat de canviar la marca de bateries per la d'Enerplus. Atès que el preu d'aquestes noves bateries és superior al de les de la marca Durames, abans de prendre una decisió vol tenir la seguretat que amb aquest canvi es guanya qualitat en el producte final. Es proven cinquanta bateries Durames i cinquanta-cinc Enerplus i s'obtenen unes durades mitjanes de trenta-set i quaranta-tres mesos, respectivament. Suposant que les desviacions típiques de totes dues marques són conegudes i iguals a cinc mesos, penseu que es guanyarà qualitat amb el canvi de marca?

2. Es vol provar que, a l'hora de carregar-se en posar en funcionament un ordinador, el sistema operatiu A és més ràpid que el sistema operatiu B. S'han mesurat els temps d'arrancada en sis ordinadors equipats amb els sistema A i en altres sis ordinadors amb el sistema B. Els temps (en segons) han estat els següents:

Sistema A	10,7	14,8	12,3	16,5	10,2	11,9
Sistema B	13,4	11,5	11,2	15,1	13,3	12,9

a) Es por acceptar l'afirmació amb un nivell de significació del 5%?
b) Calculeu un interval de confiança del 90% per a la diferència de mitjanes.

Suposarem que els temps estan normalment distribuïts i que les variàncies del temps són les mateixes per a totes dues marques.

Solucionari

1.
Dades del problema:

Bateries Durames: $n_1 = 50$; $\bar{x}_1 = 37$; $\sigma_1 = 5$
Bateries Enerplus: $n_2 = 55$; $\bar{x}_2 = 43$; $\sigma_2 = 5$

En aquest problema haurem de fer un contrast d'hipòtesi sobre la diferència de mitjanes de les durades de les bateries. En aquests cas, coneixem les desviacions típiques de les poblacions. Com que el que ens demanen és saber si "es guanyarà en qualitat", plantejarem un contrast d'hipòtesis amb les hipòtesis següents:

Tipus de contrast

Contrast de diferència de mitjanes en el cas de variàncies conegudes.

1) Expressem les hipòtesis:

- Hipòtesi nul·la. Les mitjanes són iguals:

$$H_0: \quad \mu_1 = \mu_2 \qquad (\mu_1 - \mu_2 = 0)$$

- Hipòtesi alternativa. Les mitjanes no són iguals, ja que la mitjana de durada de la bateria Enerplus és més gran que la mitjana de la durada de la bateria Durames, de manera que "es guanya qualitat amb el canvi":

$$H_1: \quad \mu_1 < \mu_2 \qquad (\mu_1 - \mu_2 < 0)$$

2) Fixem el nivell de significació $\alpha = 0{,}01$.

3) Calculem l'estadístic de contrast. Abans, però, haurem de calcular l'error estàndard

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{25}{50} + \frac{25}{55}} = 0{,}977$$

L'estadístic de contrast: $z = \dfrac{(\bar{X}_1 - \bar{X}_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \dfrac{37 - 43}{0{,}977} = -6{,}14$

4) Mitjançant les taules de la llei normal (0,1), calculem el p-valor corresponent a aquest estadístic de contrast. Tenim que: $p = P(Z < z) = P(Z < -6{,}14) = 0{,}0$.

5) Com que $0{,}0 < 0{,}05$, aleshores rebutgem la hipòtesi nul·la i arribem a la conclusió que amb el canvi de bateries guanyarem qualitat.

2. És un problema de diferència de mitjanes.

Tipus de contrast

Contrast de diferències de mitjanes en el cas de variàncies desconegudes.

a) Hem de contrastar la diferència de mitjanes per a saber si hi ha una diferència significativa o podem considerar que aquestes són iguals. A partir de les dades calculem:

Mitjana de la mostra A: $\bar{x}_A = \dfrac{1}{n_A} \sum_{i=1}^{n} x_{Ai} = 12{,}73$

Desviació típica de la mostra A: $s_A = \sqrt{\dfrac{1}{n_A - 1} \sum_{i=1}^{n} (\bar{x}_A - x_{Ai})^2} = 2{,}44$

Mitjana de la mostra B: $\bar{x}_B = \dfrac{1}{n_B} \sum_{i=1}^{n} x_{Bi} = 12{,}9$

Desviació típica de la mostra B: $s_B = \sqrt{\dfrac{1}{n_B - 1} \sum_{i=1}^{n} (\bar{x}_B - x_{Bi})^2} = 1{,}42$

A continuació, seguirem l'esquema general per fer un contrast d'hipòtesis:

1) Expressem les hipòtesis:

- Hipòtesi nul·la: H_0: $\mu_A - \mu_B = 0$
- Hipòtesi alternativa: H_1: $\mu_A - \mu_B < 0$

2) Fixem un nivell de significació: $\alpha = 0{,}05$

3) Estadístic de contrast: $t = \dfrac{(\bar{x}_A - \bar{x}_B)}{s_{\bar{x}_A - \bar{x}_B}} = -0{,}144$

Aquest estadístic segueix una distribució t de Student, amb $n_A + n_B - 2 = 10$ graus de llibertat i on l'error estàndard $s_{\bar{x}_A - \bar{x}_B}$, ve donat per l'expressió:

$$s_{\bar{x}_A - \bar{x}_B} = s\sqrt{\frac{1}{n_A} + \frac{1}{n_B}} = 1{,}155\text{ , on} \qquad s = \sqrt{\frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}} = 1{,}999$$

4) Finalment calculem el p-valor: $p = P(t_{nA + nB - 2} < t) = P(t_{10} < -0{,}144) = 0{,}444$.

5) Com que 0,444 > 0,05, no rebutgem la hipòtesi nul·la en favor de la hipòtesi alternativa, de manera que aquestes dades no confirmen que el sistema A sigui més ràpid que el sistema B.

b) Calcularem un interval de confiança del 90% per a la diferència de mitjanes. Ara tenim que $\alpha = 0{,}1$, de manera que $\alpha/2 = 0{,}05$ i el valor crític corresponent és $t_{0{,}05;10} = 1{,}8125$. Per tant, l'interval vindrà donat per:

$$(\bar{x}_A - \bar{x}_B) \pm t_{\alpha/2,\, n_A + n_B - 2} \cdot s_{\bar{x}_A - \bar{x}_B}$$

amb: $(\bar{x}_A - \bar{x}_B) = -0{,}17$; $\pm t_{0{,}05;10} = \pm 1{,}8125$; $s_{\bar{x}_A - \bar{x}_B} = 1{,}11$

De manera que obtenim $-0{,}17 \pm 1{,}8125 \cdot 1{,}155$, i fent les operacions tenim l'interval de confiança següent per a la diferència de mitjanes:

$$(-2{,}259;\ 1{,}926).$$

En aquest cas, podem observar que l'interval de confiança conté el zero, per tant, podem concloure amb una confiança del 95% que les mitjanes dels temps de càrrega són iguals. Aquest resultat està totalment d'acord amb l'obtingut en el contrast d'hipòtesis de l'apartat anterior.

Contrastos sobre la diferència de proporcions

1. Introducció

En aquesta sessió veurem com hem de fer un contrast d'hipòtesis sobre la diferència entre dues proporcions i com n'hem de determinar un interval de confiança amb un nivell de significació determinat.

Això pot ser d'interès en alguns casos; veiem alguns exemples:

- Per a saber si hi ha diferència entre la proporció d'alumnes de la UOC que es connecten al matí o a la nit.
- Per a saber si hi ha diferència entre la proporció de persones que estan a favor d'una proposta i de les que hi estan en contra.
- Per a saber si hi ha diferència entre la proporció de consumidors que prefereixen un producte d'un fabricant determinat i dels que el prefereixen de la competència, etc.

Estudiarem la diferència de proporcions per saber com es distribueixen, determinarem l'error estàndard i l'estadístic de contrast. Amb això podrem fer el contrast d'hipòtesis i també trobar intervals de confiança per a la diferència de proporcions.

2. Contrastos sobre la diferència de proporcions

Suposem que tenim: Una mostra de grandària n_1 que prové d'una distribució de Bernoulli de paràmetre p_1. La proporció mostral d'èxits és $\hat{p}_1$ i una mostra independent de l'anterior de grandària n_2 i que prové d'una distribució de Bernoulli de paràmetre p_2. La proporció mostral d'èxits és $\hat{p}_2$. Volem comparar els paràmetres poblacionals p_1 i p_2 a partir de les mostres per poder dir si aquests són iguals. Això ho farem mitjançant el contrast d'hipòtesis:

- Hipòtesis nul·la: H_0: $p_1 = p_2$
- Hipòtesi alternativa:
  - Bilateral: H_1: $p_1 \neq p_2$
  - Unilateral: H_1: $p_1 > p_2$
  - Unilateral: H_1: $p_1 < p_2$

Ja sabem que si la grandària de les mostres és gran (superior a 30), aleshores $\hat{P}_1$ i $\hat{P}_2$ presenten distribucions aproximadament normals, és a dir:

$$\hat{P}_1 \sim N\left(p_1, \sqrt{\frac{p_1(1-p_1)}{n_1}}\right) \text{ i } \hat{P}_2 \sim N\left(p_2, \sqrt{\frac{p_2(1-p_2)}{n_2}}\right)$$

la diferència $\hat{P}_1 - \hat{P}_2$ segueix també una distribució normal:

$$\hat{P}_1 - \hat{P}_2 \sim N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

I estandartitzant-la, obtenim una nova variable que segueix una distribució normal tipificada:

$$Z = \frac{(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim N(0,1)$$

Si es verifica la hipòtesi nul·la, aleshores $p_1 = p_2 = p$ i l'anterior expressió ens queda així:

$$Z = \frac{(\hat{P}_1 - \hat{P}_2)}{\sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}}} = \frac{(\hat{P}_1 - \hat{P}_2)}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

El valor p desconegut que apareix en l'expressió cal substituir-lo per una estimació $\hat{p}$, proporció poblacional comuna que podem estimar a partir de la informació proporcionada per totes dues mostres:

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

Nota

Cal tenir present que com més grans siguin les mostres, més precisa serà l'aproximació. S'obtenen resultats molt bons amb mostres de grandària superior a 100.

Dues maneres de calcular la $\hat{p}$

D'una enquesta a 1.500 habitants de Girona s'obté que 725 estan a favor que la velocitat sigui il·limitada a les autopistes. D'una altra enquesta realitzada a 2.000 habitants de Barcelona, resulta que 1.050 hi estan a favor. En aquests cas, tenim, com a resultat de cada mostra, les proporcions següents:

$$\hat{p}_1 = \frac{725}{1.500} = 0{,}483 \; ; \; \hat{p}_2 = \frac{1.050}{2.000} = 0{,}525$$

Si suposem que la hipòtesi nul·la és certa, aleshores les proporcions poblacionals de Girona i Barcelona són iguals, i són iguals a un valor p. Aquest valor el podem estimar a partir del quocient entre el nombre total d'enquestats que hi estan a favor i el nombre total d'enquestats:

$$\hat{p} = \frac{725 + 1.050}{1.500 + 2.000} = \frac{1.775}{3.500} = 0{,}507$$

Aquest mateix resultat el podem obtenir a partir de les proporcions mostrals, ja que:

- Nombre d'enquestats de Girona que hi estan a favor:

$$\hat{p}_1 \cdot n_1 = 0{,}483 \cdot 1.500 = 724{,}5 \sim 725$$

- Nombre d'enquestats de Barcelona que hi estan a favor:

$$\hat{p}_2 \cdot n_2 = 0,507 \cdot 2.000 = 1.050$$

Per tant: $\hat{p} = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = 0{,}507$

En resum, si la hipòtesi nul·la (H_0: $p_1 = p_2$) és certa, l'**estadístic de contrast** que obtenim i l'error estàndard són:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{\hat{p}_1 - \hat{p}_2}{s_{\hat{p}}}; \quad s_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

i $\hat{p} = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$ és l'**estimació de la proporció poblacional comuna.**

Aquest estadístic de contrast és una observació d'una llei $N(0,1)$.

Com sempre, una vegada calculat el valor de l'estadístic de contrast, determinarem el p-valor. Aquest valor depèn de la hipòtesi alternativa plantejada:

Atenció!

És molt important no confondre la p de la proporció amb la p del p-valor.

- Si H_1: $p_1 - p_2 \neq 0$, aleshores $p = 2\,P(Z > |z|)$
- Si H_1: $p_1 - p_2 < 0$, aleshores $p = P(Z < z)$
- Si H_1: $p_1 - p_2 > 0$, aleshores $p = P(Z > z)$

Exemple de contrast sobre la diferència de proporcions

Es vol construir una central nuclear a prop d'un poble. D'una banda, la central pot donar llocs de treball tant a gent del poble com de la resta de la comarca; de l'altra, algunes persones del poble pensen que pot ser perillosa per a la salut. Es fa una enquesta als habitants del poble i als de la resta de la comarca. Els resultats són els següents: 120 de 200 enquestats del poble i 240 de 500 enquestats de la resta de la comarca estan d'acord amb la seva construcció. Farem un contrast d'hipòtesi a un nivell del 0,05 per determinar si la proporció d'enquestats del poble que estan a favor de la proposta és major que la proporció d'enquestats de la resta de la comarca.

Sigui p_1 la proporció real de votants del poble i p_2 la de la comarca que estan a favor de la proposta. Ara hem de fer una prova de la diferència entre dues proporcions:

1. Expressem les hipòtesis:

Hipòtesi nul·la:	H_0:	$p_1 - p_2 = 0$, és a dir, $p_1 = p_2$
Hipòtesi alternativa:	H_1 :	$p_1 - p_2 > 0$, és a dir, $p_1 > p_2$

2. Seleccionem un nivell de significació: $\alpha = 0{,}05$

3. Calculem el valor de l'estadístic de contrast:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{s_{\hat{p}}} = \frac{0{,}60 - 0{,}48}{0{,}0418} = 2{,}87$$

Aquest valor és una observació d'una variable $N(0,1)$. Abans, però, hem hagut de calcular l'error estàndard:

$$s_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = 0{,}0418 \quad \text{on:} \quad \hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = 0{,}514$$

$$\text{amb } \hat{p}_1 = \frac{x_1}{n_1} = \frac{120}{200} = \frac{3}{5} = 0{,}60 \qquad \text{i} \qquad \hat{p}_2 = \frac{x_2}{n_2} = \frac{240}{500} = \frac{12}{25} = 0{,}48$$

4. Calculem el p-valor: $p = P(Z > 2{,}87) = 0{,}0021$.

5. Conclusió: com que $0{,}0021 < 0{,}05$, aleshores rebutgem H_0 i estem d'acord amb el fet que la proporció de votants del poble que afavoreix la proposta és superior que la proporció de votants de la comarca.

3. Intervals de confiança per a la diferència de proporcions

El procediment que cal seguir per a construir un interval de confiança per a la diferència de proporcions és anàleg al que se segueix per a la diferència de mitjanes i, en general, per a qualsevol tipus d'interval de confiança:

1) Fixem el nivell de confiança, que escrivim com a $(1 - \alpha)$.

2) Calculem l'error estàndard de la diferència de proporcions.

3) Calculem el valor crític tenint en compte el tipus de distribució del nostre estadístic de contrast. En aquest cas, una normal (0,1).

4) Calculem el marge d'error a partir del valor crític i l'error estàndard.

L'**interval de confiança** per a la diferència de proporcions ve donat per:

$$(\hat{p}_1 - \hat{p}_2) \pm \text{(valor crític) (error estàndard)},$$

és a dir, $(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} s_{\hat{p}}$, on:

$$s_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \text{ i } \hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

Exemple de càlcul dels intervals de confiança per a la diferència de proporcions

Considerant de nou l'exemple de la central nuclear, calcularem l'interval de confiança al 95%.

$$\hat{p}_1 - \hat{p}_2 = 0{,}60 - 0{,}48 = 0{,}12; \pm z_{\alpha/2} = \pm 1{,}96; s_{\hat{p}} = 0{,}0418$$

Per tant, tenim: $0{,}12 \pm 1{,}96 \cdot 0{,}0418$

Així, doncs, l'interval de confiança és (0,0380; 0,2019).

Com que estem tractant amb proporcions que es poden expressar en tant per cent, també és possible expressar l'interval de confiança en tant per cent: el valor de la proporció està entre el 3,80% i el 20,19%.

Podem observar que el zero no està dintre d'aquest interval. Aquest resultat està d'acord amb l'obtingut en el contrast d'hipòtesis efectuat anteriorment en què s'ha rebutjat la hipòtesi nul·la, segons la qual la diferència de proporcions és igual a zero.

4. Resum

En aquesta sessió, hem après a fer contrastos d'hipòtesis per a la diferència de dues proporcions en el cas de mostres grans. Després hem vist quin és el procediment per a construir intervals de confiança per a la diferència de proporcions.

Exercicis

1. Una firma manufacturera de cigarrets distribueix dues marques. D'una mostra de 150 fumadors, vint-i-nou prefereixen la marca A i d'una altra mostra de 200 fumadors, cinquanta-sis prefereixen la marca B. A partir d'aquestes dades, podem concloure que els fumadors prefereixen més una marca que l'altre? Feu servir un nivell de significació $\alpha = 0{,}1$.

2. Es realitza un estudi per a determinar l'efectivitat d'una nova vacuna contra la grip. S'administra la vacuna a una mostra aleatòria de 2.000 persones i d'aquest grup, vint-i-tres pateixen la malaltia. Com a grup de control se seleccionen a l'atzar 2.500 persones que no han estat vacunades. D'aquest grup, noranta-vuit pateixen la grip. Construïu un interval de confiança del 95% per a la diferència de proporcions. Què podeu dir de l'efectivitat de la nova vacuna?

Solucionari

1.

Siguin p_1 i p_2 les proporcions reals de consumidors de la marca A i B, respectivament. Ara hem de fer una prova de la diferència entre dues proporcions:

1) Expressem les hipòtesis:

- Hipòtesi nul·la: H_0: $p_1 - p_2 = 0$
- Hipòtesi alternativa: H_1: $p_1 - p_2 \neq 0$

2) Determinem un nivell de significació: $\alpha = 0{,}1$

3) Estadístic de contrast:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})[(1/n_1)+(1/n_2)]}} = -1{,}871 \text{ on: } \hat{p} = \frac{29+56}{150+200} = 0{,}243$$

$$\hat{p}_1 = \frac{29}{150} = 0{,}193 \text{ i } \hat{p}_2 = \frac{56}{200} = \frac{7}{25} = 0{,}280$$

és una observació d'una variable $N(0,1)$.

4) Calculem el p-valor: $p = 2P(Z > |z|) = 2P(Z > 1{,}871) = 2 \cdot 0{,}031 = 0{,}062$.

5) Conclusió: com que $0{,}061 < 0{,}1$, rebutgem H_0 i, per tant, sí que hi ha diferència en les preferències dels consumidors.

2.

Dades:

- Persones vacunades: $n_1 = 2.000$, $x_1 = 23$
- Persones no vacunades: $n_2 = 2.500$, $x_2 = 98$

$$\hat{p}_1 = \frac{23}{2.000} = 0{,}0115, \ \hat{p}_2 = \frac{98}{2.500} = 0{,}0392$$

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{2.000\ \ 0{,}0115 + 2.500\ \ 0{,}0392}{2.000 + 2.500} = 0{,}0269$$

$$s_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{0{,}0269(1-0{,}0296)\left(\frac{1}{2.000} + \frac{1}{2.500}\right)} = 0{,}0049$$

Ja podem calcular l'interval de confiança: $(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}s_{\hat{p}}$

$$(0{,}0115 - 0{,}0392) \pm 1{,}96 \cdot 0{,}0049; \ -0{,}0277 \pm 0{,}0096$$

Per tant, l'interval de confiança és: (–0,0373; –0,0181).

Com que el valor zero no és troba dintre l'interval, arribem a la conclusió que les proporcions són diferents i, atès que l'interval és negatiu, la vacuna té realment algun efecte beneficiós.