Contrastos d'hipòtesis: Dues mostres

Contrastos de dues mostres

Ara volem contrastar el valor d'un mateix paràmetre a dues poblacions

Dos tipus:

- Mostres independents: Les dues mostres s'han obtingut de manera independent

  Exemple: Provam un medicament sobre dues mostres de malalts de característiques diferents

- Mostres aparellades: Les dues mostres corresponen als mateixos individus, o a individus aparellats d'alguna manera

  Exemple: Provam dos medicaments sobre els mateixos malalts

Teniu una Taula de contrastos de dues mostres exhaustiva a Campus Extens

Mostres independents

Tenim dues variables aleatòries (que representen dues poblacions)

Exemple: Homes i Dones

Volem comparar el valor d'un paràmetre a les dues poblacions

Exemple: Són, de mitjana, els homes més alts que les dones?

Ho farem a partir d'una m.a.s. de cada v.a., escollides a més de manera independent

Contrastos per a μ

Tenim dues v.a. X_1 i X_2, de mitjanes μ_1 i μ_2

Prenem m.a.s.

$$\begin{array}{ll} X_{1,1}, X_{1,2}, \ldots, X_{1,n_1} & \text{de } X_1 \\ X_{2,1}, X_{2,2}, \ldots, X_{2,n_2} & \text{de } X_2 \end{array}$$

Siguin $\overline{X}_1$ i $\overline{X}_2$ les seves mitjanes, respectivament

Les hipòtesis

$$H : \mu_1 \left\{ \begin{matrix} > \\ < \\ \neq \\ = \end{matrix} \right\} \mu_2 \quad \text{equivalen a} \quad H : \mu_1 - \mu_2 \left\{ \begin{matrix} > \\ < \\ \neq \\ = \end{matrix} \right\} 0$$

Empram un estadístic de contrast per a $\mu_1 - \mu_2$

C.H. per a μ de poblacions normals o n grans: σ conegudes

Suposem una de les dues situacions següents:

- X_1 i X_2 són normals, o
- n_1 i n_2 són grans ($n_1, n_2 \geqslant 30$ o **40**)

Suposem que coneixem a més les desv. típ. σ_1 i σ_2 de X_1 i X_2

En aquest cas l'estadístic de contrast és

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

i els contrastos són com en el cas quan $H_0 : \mu = 0$

Exemple

Volem comparar els temps de realització d'un test entre estudiants de dos graus G_1 i G_2, i contrastar si és veritat que els estudiants de G_1 empren menys temps que els de G_2

Suposem que les desviacions típiques són conegudes: $\sigma_1 = 1$ i $\sigma_2 = 2$

Disposam de dues mostres independents de tests realitzats per estudiants de cada grau, $n_1 = n_2 = 40$. Calculam les mitjanes dels temps emprats a cada mostra (en minuts):

$$\overline{X}_1 = 9.789, \quad \overline{X}_2 = 11.385$$

Exemple

Contrast:

$$\left\{\begin{array}{l} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 < \mu_2 \end{array}\right. \Longleftrightarrow \left\{\begin{array}{l} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 < 0 \end{array}\right.$$

Estadístic de contrast: $Z = \dfrac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$

Valor que pren: $z_0 = \dfrac{9.789 - 11.385}{\sqrt{\frac{1}{40} + \frac{4}{40}}} = -4.514$

p-valor: $P(Z \leqslant -4.514) \approx 3 \cdot 10^{-6}$ molt petit

Decisió: Rebutjam la hipòtesi que són iguals, en favor que a G_1 tarden menys que a G_2

Exemple

També podem calcular un interval de confiança del 95% per a la diferència de mitjanes $\mu_1 - \mu_2$ al contrast

$$\left\{ \begin{array}{l} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 < 0 \end{array} \right.$$

Segons la Taula de Contrastos, aquest interval és

$$\left]-\infty, \overline{X}_1 - \overline{X}_2 - z_\alpha \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right[$$

Al nostre cas, per a $\alpha = 0.05$, dóna

$$] - \infty, -1.0145[$$

0 no hi pertany: Rebutjam que $\mu_1 - \mu_2 = 0$

C.H. per a μ de normals o n grans: σ desconegudes

Suposem un altre cop una de les dues situacions següents, i que ara no coneixem σ_1 i σ_2:

- X_1 i X_2 són normals, o
- n_1 i n_2 són grans $(n_1, n_2 \geqslant 40)$

En aquest cas, hem de distingir dos subcasos:

(1) Suposam que $\sigma_1 = \sigma_2$

(2) Suposam que $\sigma_1 \neq \sigma_2$

C.H. per a μ de normals o n grans: σ desconegudes

Suposem un altre cop una de les dues situacions següents, i que ara no coneixem σ_1 i σ_2:

- X_1 i X_2 són normals, o
- n_1 i n_2 són grans ($n_1, n_2 \geqslant 40$)

En aquest cas, hem de distingir dos subcasos:

(1) Suposam que $\sigma_1 = \sigma_2$

(2) Suposam que $\sigma_1 \neq \sigma_2$

Com decidim en quin cas estam? Dues possibilitats:

- Fem els dos casos, i si donen el mateix, és el que contestam
- (Si són normals) Fem un contrast d'igualtat de variàncies per decidir quin és el cas

C.H. per a μ de normals o n grans: σ desconegudes

Si suposam que $\sigma_1 = \sigma_2$, l'estadístic de contrast és

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \cdot \frac{(n_1-1)\widetilde{S}_1^2 + (n_2-1)\widetilde{S}_2^2}{n_1+n_2-2}}}$$

que, quan $\mu_1 = \mu_2$, té distribució (aproximadament, en cas de mostres grans) $t_{n_1+n_2-2}$

C.H. per a μ de normals o n grans: σ desconegudes

Si suposam que $\sigma_1 \neq \sigma_2$, l'estadístic de contrast és

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\widetilde{S}_1^2}{n_1} + \frac{\widetilde{S}_2^2}{n_2}}} \sim t_f$$

que, quan $\mu_1 = \mu_2$, té distribució (aproximadament, en cas de mostres grans) t_f amb

$$f = \left\lfloor \frac{\left(\frac{\widetilde{S}_1^2}{n_1} + \frac{\widetilde{S}_2^2}{n_2} \right)^2}{\frac{1}{n_1 - 1} \left(\frac{\widetilde{S}_1^2}{n_1} \right)^2 + \frac{1}{n_2 - 1} \left(\frac{\widetilde{S}_2^2}{n_2} \right)^2} \right\rfloor - 2$$

En els dos casos, els contrastos són com en el cas de $H_0 : \mu = 0$

Exemple

Volem comparar els temps de realització d'un test entre estudiants de dos graus G_1 i G_2, i determinar si és veritat que els estudiants de G_1 empren menys temps que els de G_2

No coneixem σ_1 i σ_2

Disposam de dues mostres independents de tests realitzats per estudiants de cada grau, $n_1 = n_2 = 40$. Calculam les mitjanes i les desviacions típiques mostrals dels temps emprats a cada mostra:

$$\overline{X}_1 = 9.789, \qquad \overline{X}_2 = 11.385,$$
$$\widetilde{S}_1 = 1.201, \qquad \widetilde{S}_2 = 1.579$$

Exemple

Cas 1: Suposam $\sigma_1 = \sigma_2$

Contrast:

$$\left\{\begin{array}{l} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 < \mu_2 \end{array}\right. \iff \left\{\begin{array}{l} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 < 0 \end{array}\right.$$

Estadístic de contrast:

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \cdot \frac{(n_1-1)\widetilde{S}_1^2 + (n_2-1)\widetilde{S}_2^2}{n_1+n_2-2}}} \sim t_{40+40-2}$$

Valor que pren:

$$t_0 = \frac{9.789 - 11.385}{\sqrt{\left(\frac{1}{40} + \frac{1}{40}\right)\frac{39\cdot 1.201^2 + 39\cdot 1.579^2}{78}}} = -5.0881$$

Exemple

Cas 1: Suposam $\sigma_1 = \sigma_2$

p-valor:

$$P(t_{78} < -5.0881) = 1.2 \cdot 10^{-6},$$

molt petit

Decisió: Rebutjam la hipòtesi que són iguals, en favor que a G_1 tarden menys que a G_2

Exemple

Cas 2: Suposam $\sigma_1 \neq \sigma_2$

Contrast:

$$\left\{\begin{array}{l} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 < \mu_2 \end{array}\right. \iff \left\{\begin{array}{l} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 < 0 \end{array}\right.$$

Estadístic de contrast:

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\widetilde{S}_1^2}{n_1} + \frac{\widetilde{S}_2^2}{n_2}}} \sim t_f$$

on

$$f = \left\lfloor \frac{\left(\frac{1.201^2}{40} + \frac{1.579^2}{40}\right)^2}{\frac{1}{39}\left(\frac{1.201^2}{40}\right)^2 + \frac{1}{39}\left(\frac{1.579^2}{40}\right)^2} \right\rfloor - 2 = \lfloor 72.81 \rfloor - 2 = 70$$

Exemple

Cas 2: Suposam $\sigma_1 \neq \sigma_2$

Valor que pren:

$$t_0 = \frac{9.789 - 11.385}{\sqrt{\frac{1.201^2}{40} + \frac{1.579^2}{40}}} = -5.0881$$

p-valor: $P(t_{70} \leqslant -5.0881) = 1.5 \cdot 10^{-6}$ molt petit

Decisió: Rebutjam la hipòtesi que són iguals, en favor que a G_1 tarden menys que a G_2

Exemple

Cas 2: Suposam $\sigma_1 \neq \sigma_2$

Valor que pren:

$$t_0 = \frac{9.789 - 11.385}{\sqrt{\frac{1.201^2}{40} + \frac{1.579^2}{40}}} = -5.0881$$

p-valor: $P(t_{70} \leqslant -5.0881) = 1.5 \cdot 10^{-6}$ molt petit

Decisió: Rebutjam la hipòtesi que són iguals, en favor que a G_1 tarden menys que a G_2

Decisió final: Els dos casos han donat el mateix, així que concloem que a G_1 tarden menys que a G_2

Contrast per a dues proporcions: Test de Fisher

Tenim dues variables aleatòries X_1 i X_2 Bernoulli de proporcions p_1 i p_2

Prenem m.a.s. de cada una i obtenim la taula següent

	X_1	X_2	Total
Èxits	n_{11}	n_{12}	$n_{1\bullet}$
Fracassos	n_{21}	n_{22}	$n_{2\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet\bullet}$

La hipòtesi nul·la que contrastam és $H_0 : p_1 = p_2$

Contrast per a dues proporcions: Test de Fisher

	X_1	X_2	Total
Èxits	n_{11}	n_{12}	$n_{1\bullet}$
Fracassos	n_{21}	n_{22}	$n_{2\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet\bullet}$

Si $p_1 = p_2$, la probabilitat d'obtenir n_{11} èxits dins X_1 és la de:

> *En una bossa hi tenim $n_{1\bullet}$ bolles E i $n_{2\bullet}$ bolles F. Probabilitat d'obtenir n_{11} bolles E si en triam $n_{\bullet 1}$ de cop.*

És una hipergeomètrica $H(n_{1\bullet}, n_{2\bullet}, n_{\bullet 1})$. L'empram per calcular els p-valors.

Contrast per a dues proporcions: Test de Fisher

Contrast: $\left\{\begin{array}{l} H_0 : p_1 = p_2 \\ H_1 : p_1 > p_2 \end{array}\right.$

p-valor: $P(H(n_{1\bullet}, n_{2\bullet}, n_{\bullet 1}) \geqslant n_{11})$

Contrast: $\left\{\begin{array}{l} H_0 : p_1 = p_2 \\ H_1 : p_1 < p_2 \end{array}\right.$

p-valor: $P(H(n_{1\bullet}, n_{2\bullet}, n_{\bullet 1}) \leqslant n_{11})$

Contrast: $\left\{\begin{array}{l} H_0 : p_1 = p_2 \\ H_1 : p_1 \neq p_2 \end{array}\right.$

p-valor: $\min\{2\min\{P(H \leqslant n_{11}), P(H \geqslant n_{11})\}, 1\}$.

(Controvertit; per exemple, R ho fa d'una altra manera)

Exemple

Per determinar si la Síndrome de Mort Sobtada del Nadó (SIDS) té component genètic, es consideren els casos de SIDS en parelles de bessons monozigòtics i dizigòtics. Diguem:

- p_1: proporció de parelles de bessons monozigòtics amb algun cas de SIDS on només un germà la sofrí
- p_2: proporció de parelles de bessons dizigòtics amb algun cas de SIDS on només un germà la sofrí

Si la SIDS té component genètic, és d'esperar que $p_1 < p_2$

Hem de realitzar el contrast

$$\left\{ \begin{array}{l} H_0 : p_1 = p_2 \\ H_1 : p_1 < p_2 \end{array} \right.$$

Exemple

En un estudi (Peterson et al, 1980), s'obtingueren les dades següents:

	Tipus de bessons		
Casos de SIDS	Monozigòtics	Dizigòtics	Total
Un	23	35	58
Dos	1	2	3
Total	24	37	61

p-valor:

$$P(H(58,3,24) \leqslant 23) = \texttt{phyper(23,58,3,24)} = 0.7841$$

No podem rebutjar la hipòtesi nul·la

Contrast per a dues proporcions: mostres grans

Tenim dues variables aleatòries X_1 i X_2 Bernoulli de proporcions p_1 i p_2

Prenem m.a.s. grans ($n_1, n_2 \geqslant 50$ o **100**)

$$\begin{array}{ll} X_{1,1}, X_{1,2}, \ldots, X_{1,n_1} & \text{de } X_1 \\ X_{2,1}, X_{2,2}, \ldots, X_{2,n_2} & \text{de } X_2 \end{array}$$

Siguin $\widehat{p}_1$ i $\widehat{p}_2$ les seves proporcions mostrals

Suposem que els nombres d'èxits i de fracasos a cada mostra són $\geqslant 5$ o **10**)

La hipòtesi nul·la que contrastam és $H_0 : p_1 = p_2$, que hem d'entendre $H_0 : p_1 - p_2 = 0$

Contrast per a dues proporcions: mostres grans

L'estadístic de contrast és

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

i té distribució aproximadament $N(0,1)$ si $p_1 = p_2$.

Els contrastos són com en el cas de $H_0 : p = 0$

Exemple

Es prenen una mostra d'ADN de 100 individus amb almenys tres generacions familiars a l'illa de Mallorca, i una altra de 50 individus amb almenys tres generacions familiars a l'illa de Menorca

Es vol saber si un determinat al·lel d'un gen és present amb la mateixa proporció a les dues poblacions

A la mostra mallorquina, 20 individus el tenen, i a la mostra menorquina, 12

Contrastau la hipòtesi d'igualtat de proporcions al nivell de significació 0.05, i calculau l'interval de confiança per a la diferència de proporcions per a aquest α

Suposau que 50 és prou gran

Exemple

Contrast:

$$\begin{cases} H_0 : p_1 = p_2 \\ H_1 : p_1 \neq p_2 \end{cases}$$

Estadístic de contrast:

$$Z = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\left(\frac{n_1\widehat{p}_1 + n_2\widehat{p}_2}{n_1+n_2}\right)\left(1 - \frac{n_1\widehat{p}_1 + n_2\widehat{p}_2}{n_1+n_2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

Valor que pren: $\widehat{p}_1 = 0.2,\ \widehat{p}_2 = 0.24,\ n_1 = 100,\ n_2 = 50$

$$z_0 = \frac{0.2 - 0.24}{\sqrt{\left(\frac{20+12}{100+50}\right)\left(1 - \frac{20+12}{100+50}\right)\left(\frac{1}{100} + \frac{1}{50}\right)}} = -0.5637$$

Exemple

p-valor: $2 \cdot P(Z \geqslant 0.5637) = 0.573$

Decisió: Com que el *p*-valor és més gran que $\alpha = 0.05$, acceptam la hipòtesi que les dues proporcions són la mateixa

Exemple

L'interval de confiança per a $p_1 - p_2$ al nivell de confiança $(1-\alpha)\cdot 100\%$ en un contrast bilateral és

$$\Bigg]\widehat{p}_1 - \widehat{p}_2 - z_{1-\frac{\alpha}{2}}\sqrt{\left(\frac{n_1\widehat{p}_1+n_2\widehat{p}_2}{n_1+n_2}\right)\left(1-\frac{n_1\widehat{p}_1+n_2\widehat{p}_2}{n_1+n_2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)},$$

$$\widehat{p}_1 - \widehat{p}_2 + z_{1-\frac{\alpha}{2}}\sqrt{\left(\frac{n_1\widehat{p}_1+n_2\widehat{p}_2}{n_1+n_2}\right)\left(1-\frac{n_1\widehat{p}_1+n_2\widehat{p}_2}{n_1+n_2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\Bigg[$$

Exemple

L'interval de confiança per a $p_1 - p_2$ al nivell de confiança $(1-\alpha)\cdot 100\%$ en un contrast bilateral és

$$\Bigg]\widehat{p}_1 - \widehat{p}_2 - z_{1-\frac{\alpha}{2}}\sqrt{\left(\frac{n_1\widehat{p}_1+n_2\widehat{p}_2}{n_1+n_2}\right)\left(1-\frac{n_1\widehat{p}_1+n_2\widehat{p}_2}{n_1+n_2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)},$$
$$\widehat{p}_1 - \widehat{p}_2 + z_{1-\frac{\alpha}{2}}\sqrt{\left(\frac{n_1\widehat{p}_1+n_2\widehat{p}_2}{n_1+n_2}\right)\left(1-\frac{n_1\widehat{p}_1+n_2\widehat{p}_2}{n_1+n_2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\Bigg[$$

$$](0.2-0.24)-1.96\cdot 0.071,(0.2-0.24)+1.96\cdot 0.071[$$
$$=]-0.179, 0.099[$$

Conté el 0, per tant no podem rebutjar que $p_1 - p_2 = 0$

Contrastos una mica més generals

Fins ara hem pres $H_0 : \mu_1 = \mu_2$. Un tipus de contrastos lleugerament més generals serien

$$\left\{\begin{array}{l} H_0 : \mu_1 - \mu_2 = \Delta \\ H_1 : \mu_1 - \mu_2 < \Delta \text{ o } \mu_1 - \mu_2 > \Delta \text{ o } \mu_1 - \mu_2 \neq \Delta \end{array}\right.$$

amb $\Delta \in \mathbb{R}$.

Es fan igual, modificant lleugerament l'estadístic: substituïm als numeradors

$$\overline{X}_1 - \overline{X}_2 \text{ per } \overline{X}_1 - \overline{X}_2 - \Delta$$

Exemple

Tenim dos tractaments, A i B, d'una malaltia. Tractam 50 malalts amb A i 100 amb B. 20 malalts tractats amb A i 25 tractats amb B manifesten haver sentit malestar general durant els 7 dies posteriors a iniciar el tractament.

Podem concloure, a un nivell de significació del 5%, que A produeix malestar general en una proporció dels malalts que és 5 punts percentuals superior a la proporció dels malalts en què el produeix B?

Exemple

p_1: Fracció de malalts en què A produeix malestar general
p_2: Fracció de malalts en què B produeix malestar general

Contrast:

$$\left\{ \begin{array}{l} H_0 : p_1 \leqslant p_2 + 0.05 \\ H_1 : p_1 > p_2 + 0.05 \end{array} \right.$$

Estadístic de contrast:

$$Z = \frac{\widehat{p}_1 - \widehat{p}_2 - \Delta}{\sqrt{\left(\frac{n_1\widehat{p}_1 + n_2\widehat{p}_2}{n_1+n_2}\right)\left(1 - \frac{n_1\widehat{p}_1 + n_2\widehat{p}_2}{n_1+n_2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

Exemple

Valor que pren: $\widehat{p}_1 = 0.4$, $\widehat{p}_2 = 0.25$, $n_1 = 50$, $n_2 = 100$, $\Delta = 0.05$

$$z_0 = \frac{0.4 - 0.25 - 0.05}{\sqrt{\left(\frac{20+25}{50+100}\right)\left(1 - \frac{20+25}{50+100}\right)\left(\frac{1}{50} + \frac{1}{100}\right)}} = 1.26$$

p-valor: $P(Z \geqslant 1.26) = 0.104$

Decisió: Com que el p-valor és més gran que $\alpha = 0.05$, no podem rebutjar la hipòtesi que $p_1 - p_2$ és inferior a un 5%

Exemple

L'interval de confiança per a $p_1 - p_2$ al nivell de confiança $(1-\alpha)\cdot 100\%$ en aquest contrast és

$$\left[\widehat{p}_1 - \widehat{p}_2 + z_\alpha \sqrt{\left(\frac{n_1\widehat{p}_1 + n_2\widehat{p}_2}{n_1+n_2}\right)\left(1 - \frac{n_1\widehat{p}_1 + n_2\widehat{p}_2}{n_1+n_2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}, \infty\right[$$

Operant:

$$[(0.4 - 0.25) - 1.645 \cdot 0.0794, \infty[= [0.0194, \infty[$$

Conté 0.05, per tant no podem rebutjar que $p_1 \leqslant p_2 + 0.05$ (però en canvi, no conté 0 i per tant podríem rebutjar que $p_1 = p_2$)

Contrast per a dues variàncies

Necessitam decidir si les variàncies de les dues poblacions són iguals o diferents, per exemple en el marc d'una comparació de mitjanes de mostres independents

Tenim dues variables aleatòries X_1 i X_2 normals de desviacions típiques σ_1, σ_2 desconegudes

Prenem m.a.s.

$$\begin{array}{ll} X_{1,1}, X_{1,2}, \ldots, X_{1,n_1} & \text{de } X_1 \\ X_{2,1}, X_{2,2}, \ldots, X_{2,n_2} & \text{de } X_2 \end{array}$$

Siguin $\widetilde{S}_1^2$ i $\widetilde{S}_2^2$ les seves variàncies mostrals

Contrast per a dues variàncies

El contrast té hipòtesi nul·la $H_0 : \sigma_1 = \sigma_2$, que correspon a $H_0 : \dfrac{\sigma_1^2}{\sigma_2^2} = 1$

S'empra l'estadístic de contrast

$$F = \frac{\widetilde{S}_1^2}{\widetilde{S}_2^2}$$

que, si les dues poblacions són normals, té distribució *F* de Fisher amb graus de llibertat $n_1 - 1$ i $n_2 - 1$

La distribució F de Fisher

La distribució $F_{n,m}$, on n, m són els graus de llibertat, és la d'una variable aleatòria

$$\chi_n^2/\chi_m^2$$

Té densitat

$$f_{F_{n,m}}(x) = \frac{\Gamma(\frac{n+m}{2}) \cdot (\frac{m}{n})^{m/2} x^{(m-2)/2}}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})(1 + \frac{m}{n}x)^{(m+n)/2}} \quad \text{si } x \geqslant 0$$

on $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\, dt$ si $x > 0$

La distribució està tabulada (Teniu les taules a Campus Extens), i amb R és `f`

No és simètrica. Els p-valors es calculen com en el cas de la χ^2. Alerta en el cas bilateral!

Exemple

Recordau l'exemple on volíem comparar els temps de realització d'un test entre estudiants de dos graus G_1 i G_2. Suposem que aquests temps segueixen distribucions normals.

Disposam de dues mostres independents de tests realitzats per estudiants de cada grau, $n_1 = n_2 = 40$. Calculam les desviacions típiques mostrals dels temps emprats a cada mostra:

$$\widetilde{S}_1 = 1.201, \quad \widetilde{S}_2 = 1.579$$

Contrastau la hipòtesi d'igualtat de variàncies al nivell de significació 0.05

Exemple

Contrast:

$$\begin{cases} H_0 : \sigma_1 = \sigma_2 \\ H_1 : \sigma_1 \neq \sigma_2 \end{cases}$$

Estadístic de contrast:

$$F = \frac{\widetilde{S}_1^2}{\widetilde{S}_2^2} \sim F_{39,39}$$

Valor que pren: $\widetilde{S}_1 = 1.201$, $\widetilde{S}_2 = 1.579$

$$f_0 = \frac{1.201^2}{1.579^2} = 0.5785$$

Exemple

p-valor: No és simètrica

$$2 \cdot P(F_{39,39} \geqslant 0.5785) = 1.909$$
$$2 \cdot P(F_{39,39} \leqslant 0.5785) = 0.091$$

El *p*-valor és 0.091

Decisió: Com que el *p*-valor és més gran que $\alpha = 0.05$, no podem rebutjar la hipòtesi que les dues variàncies són iguals.

Concloem que $\sigma_1 = \sigma_2$. Aquesta seria l'assumpció que hauríem de fer en el test de les μ.

Exemple

Es desitja comparar l'activitat motora espontània d'un grup de 25 rates control i un altre de 36 rates desnodrides. Es va mesurar el nombre de vegades que passaven davant d'una cèl·lula fotoelèctrica durant 24 hores. Les dades obtingudes van ser les següents

	n	$\overline{X}$	$\widetilde{S}$
1. Control	25	869.8	106.7
2. Desnodrides	36	665	133.7

S'observen diferències significatives entre el grup de control i el grup desnodrit?

Exemple

Es desitja comparar l'activitat motora espontània d'un grup de 25 rates control i un altre de 36 rates desnodrides. Es va mesurar el nombre de vegades que passaven davant d'una cèl·lula fotoelèctrica durant 24 hores. Les dades obtingudes van ser les següents

	n	$\overline{X}$	$\widetilde{S}$
1. Control	25	869.8	106.7
2. Desnodrides	36	665	133.7

S'observen diferències significatives entre el grup de control i el grup desnodrit?

Suposarem que aquests nombres de vegades segueixen distribucions normals

Exemple

Contrast:

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$$

Per poder-lo efectuar, efectuarem primer el contrast

$$\begin{cases} H_0 : \sigma_1 = \sigma_2 \\ H_1 : \sigma_1 \neq \sigma_2 \end{cases}$$

per decidir quin test fer

Exemple

Contrast:

$$\begin{cases} H_0 : \sigma_1 = \sigma_2 \\ H_1 : \sigma_1 \neq \sigma_2 \end{cases}$$

Estadístic de contrast:

$$F = \frac{\widetilde{S}_1^2}{\widetilde{S}_2^2} \sim F_{24,35}$$

Valor que pren: $\widetilde{S}_1 = 106.7$, $\widetilde{S}_2 = 133.7$

$$f_0 = \frac{106.7^2}{133.7^2} = 0.637$$

Exemple

p-valor:

$$2 \cdot P(F_{24,35} \leqslant 0.637) = 0.25$$
$$2 \cdot P(F_{24,35} \geqslant 0.637) = 1.75$$

El p-valor és 0.25, gran

Decisió: $\sigma_1 = \sigma_2$

Exemple

Contrast:

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$$

Estadístic de contrast: Com que suposam que $\sigma_1 = \sigma_2$

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \cdot \frac{(n_1-1)\widetilde{S}_1^2 + (n_2-1)\widetilde{S}_2^2}{n_1+n_2-2}}} \sim t_{25+36-2}$$

Valor que pren:

$$t_0 = \frac{869.8 - 665}{\sqrt{\left(\frac{1}{25} + \frac{1}{36}\right) \cdot \frac{24 \cdot 106.7^2 + 35 \cdot 133.7^2}{25+36-2}}} = 6.37$$

Exemple

p-valor: $2 \cdot P(t_{59} \geqslant 6.37) \approx 0$

Decisió: Hi ha diferència (i com que $\overline{x}_2 < \overline{x}_1$, concloem que les desnodrides es mouen menys)

Mostres aparellades

Fins ara hem considerat que les mostres eren independents

Un cas completament diferent és quan les dues mostres corresponen als mateixos individus o a individus aparellats per algun factor

Exemples:

- Mesuram l'estat d'una malaltia als mateixos individus abans i després d'un tractament
- Mesuram la incidència de càncer en parelles de germans bessons

Es parla en aquest cas de mostres aparellades (*paired*)

Mostres aparellades

Per decidir si hi ha diferències, el contrast més comú consisteix a calcular les diferències dels valors de cadascuna de les parelles de mostres i realitzar un contrast per esbrinar si la mitjana de les diferències és 0

Mostres aparellades

Per decidir si hi ha diferències, el contrast més comú consisteix a calcular les diferències dels valors de cadascuna de les parelles de mostres i realitzar un contrast per esbrinar si la mitjana de les diferències és 0

És important observar aquí que hi ha diferents maneres de realitzar un disseny experimental per contrastar una hipòtesi, i que el disseny s'ha de fixar abans de la recollida de dades

Exemple: Contrast de dues mitjanes de mostres aparellades

Disposam de dos algoritmes de plegament de proteïnes. Tots dos produeixen resultats de la mateixa qualitat

Estam interessats a saber quin dels dos és més eficient, en el sentit que té la mitjana de temps d'execució més petita. Suposam que aquests temps d'execució segueixen lleis normals.

Prenem una mostra de proteïnes, li aplicam els dos algoritmes, i anotam els temps d'execució sobre cada proteïna

Exemple: Contrast de dues mitjanes de mostres aparellades

Els resultats obtinguts són:

	1	2	3	4	5	6	7	8	9	10
alg. 1	8.1	11.9	11.4	12.9	9.0	7.2	12.4	6.9	8.9	8.3
alg. 2	6.9	6.7	8.3	8.6	18.9	7.9	7.4	8.7	7.9	12.4
d	1.2	5.2	3.1	4.3	-9.9	-0.7	5.0	-1.8	1.0	-4.1

(La filera d conté les diferències de temps entre el primer i el segon algoritme)

$$\overline{d} = 0.33, \ \widetilde{S}_d = 4.72$$

Volem contrastar la igualtat de mitjanes amb el test que correspongui. I si són diferents, decidir quin té major temps d'execució.

Exemple: Contrast de dues mitjanes de mostres aparellades

Contrast:

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$$

Consultam la taula. L'estadístic de contrast és

$$T = \frac{\overline{d}}{\widetilde{S}_d/\sqrt{n}}$$

que té distribució $t_{n-1} = t_9$. Pren el valor

$$t_0 = \frac{0.33}{4.72/\sqrt{10}} = 0.22$$

Exemple: Contrast de dues mitjanes de mostres aparellades

El p-valor és

$$2P(t_9 \geqslant 0.22) = 0.83$$

molt gran, no podem rebutjar la hipòtesi nul·la que els temps mitjans són iguals. Per tant, no té sentit cercar quin té el temps d'execució més gran

Exemple: Contrast de dues proporcions de mostres aparellades

Es pren una mostra de 1000 persones afectades per migranya. Se'ls facilita un fàrmac perquè n'alleugereixi els símptomes.

Després de l'administració se'ls pregunta si han notat alleujament en el dolor

Al cap d'un temps es subministra als mateixos individus un placebo i se'ls torna a preguntar si han notat o no milloria

Exemple: Contrast de dues proporcions de mostres aparellades

Els resultats són:

		Placebo	
		Sí	No
Fàrmac	Sí	300	62
	No	38	590

És més efectiu el fàrmac que el placebo?

Exemple: Contrast de dues proporcions de mostres aparellades

p_1: Proporció que troba milloria amb el fàrmac

p_2: Proporció que troba milloria amb el placebo

Contrast:

$$\begin{cases} H_0 : p_1 = p_2 \\ H_1 : p_1 > p_2 \end{cases}$$

Exemple: Contrast de dues proporcions de mostres aparellades

Consultam la taula. L'estadístic de contrast és

$$Z = \frac{\frac{b}{n} - \frac{d}{n}}{\sqrt{\frac{b+d}{n^2}}} \sim N(0,1)$$

on

		Placebo	
		Sí	No
Fàrmac	Sí	a	b
	No	d	c

Aquest contrast només és vàlid quan la mostra és gran i el nombre de casos discordants $b + d$ és "bastant gran", posem $\geqslant 20$

Exemple: Contrast de dues proporcions de mostres aparellades

L'estadístic de contrast té el valor $z_0 = 2.4$

El p-valor és

$$P(Z > 2.4) = 0.0082,$$

petit. Per tant, concloem que el fàrmac és més efectiu que el placebo.