Inferència d'informació per a dues poblacions o més

Contrastos d'hipòtesi per a dues poblacions i comparació de grups mitjançant ANOVA

Blanca de la Fuente, Ángel A. Juan i Patricia Carracedo

PID_00242443

Temps de lectura i comprensió: **4 hores**

Universitat Oberta de Catalunya

Índex

Introducció

En els mòduls anteriors s'han introduït els conceptes bàsics d'estimació i de contrast d'hipòtesi relacionats amb una població. En la pràctica quotidiana, tanmateix, és fàcil trobar-se amb situacions en les quals es disposa de dos o més grups d'individus o poblacions i, en aquest cas, l'interès resideix sovint en el fet de ser capaç de discernir si aquests grups o poblacions es poden considerar com a semblants –des d'un punt de vista estadístic– o si, per contra, són grups o poblacions que mostren diferències significatives entre ells. Així, per exemple, potser cal comparar les qualificacions mitjanes de dos grups d'estudiants en funció de si han utilitzat o no una metodologia docent innovadora, comparar els percentatges de recuperació de dos o més grups de malalts segons el tractament rebut, comparar les qualitats mitjanes de diferents accessos a Internet segons l'empresa proveïdora, comparar els preus mitjans dels serveis d'obtenció de documents segons la institució que els ofereixi, etc.

Quan es consideren dos grups o poblacions, les tècniques que es fan servir per a comparar les respectives mitjanes o proporcions són molt semblants a les que s'utilitzen en el cas d'una població: contrastos d'hipòtesi basats en l'ús de la distribució normal (quan es comparen dues proporcions) o de la t de Student (quan es comparen dues mitjanes). En el cas de la comparació entre dues mitjanes de grups diferents, cal distingir si es tracta de dos grups independents (per exemple, quan es comparen els resultats d'un test fet a dos grups diferents d'individus) o bé si es tracta de dos grups dependents (per exemple, quan s'estan considerant els resultats d'un test previ amb els resultats d'un test posterior, tots dos fets en el mateix grup d'individus).

Finalment, en cas que es vulguin comparar més de dos grups o poblacions, els contrastos anteriors ja no serveixen i resulta necessari recórrer a les tècniques ANOVA basades en la distribució F de Snedecor. L'ús d'aquestes tècniques possibilita discernir si les mitjanes corresponents a un conjunt de tres grups o més són totes aproximadament iguals o si, per contra, es pot establir que hi ha diferències significatives entre algunes d'elles (i, conseqüentment, entre els grups associats).

Objectius

Els objectius acadèmics d'aquest mòdul es descriuen a continuació:

1. Comparar dues poblacions fent servir procediments similars als que hem vist per a una única població.

2. Aprendre a formular una hipòtesi sobre la naturalesa de les dues poblacions i la diferència entre les seves mitjanes o proporcions.

3. Conèixer el mètode per a comparar les variàncies de dues poblacions. Per a fer aquests contrastos s'introdueix la distribució *F*.

4. Entendre la importància pràctica de les tècniques ANOVA a l'hora de discernir si hi ha diferències significatives entre més de dos grups o poblacions.

5. Aprendre a usar els tests *F* d'ANOVA i saber interpretar adequadament els resultats que ofereixen.

6. Comprendre la lògica que és subjacent a la metodologia ANOVA.

7. Conèixer les hipòtesis que s'han de satisfer per a poder aplicar les tècniques ANOVA amb garanties.

8. Aprendre a usar maquinari estadístic i d'anàlisi de dades com a instrument bàsic en l'aplicació pràctica dels conceptes i de les tècniques estadístiques.

1. Contrastos d'hipòtesi per a dues poblacions

En aquest mòdul es presenten mètodes per a contrastar les diferències entre les mitjanes o proporcions de dues poblacions i per a contrastar variàncies.

Per a comparar les mitjanes o les proporcions poblacionals, s'extreu una mostra aleatòria de les dues poblacions. La inferència sobre la diferència entre ambdues mitjanes o proporcions es basa en els resultats mostrals. El mètode apropiat per a analitzar la informació depèn del procediment emprat en seleccionar les mostres. Considerem les dues possibilitats següents:

a) Mostres dependents (dades apariades): en aquest procediment les mostres s'elegeixen per parells, una de cada població. La idea és que a part de la característica objecte de l'estudi, els elements de cada un d'aquests parells han d'estar relacionats, de manera que la comparació pot ser establerta directament. Per exemple, suposem que volem mesurar l'eficàcia d'un curs de lectura ràpida. Una manera d'abordar el problema seria prendre nota de les paraules llegides per minut en una mostra d'alumnes abans de prendre el curs i comparar-les amb els resultats obtinguts *pels mateixos alumnes* una vegada completat el curs. En aquest cas cada parell consistiria en mesures de velocitat d'un mateix alumne fetes abans i després del curs, es podria esbrinar si hi ha proves contundents de l'eficàcia del curs de lectura ràpida.

b) Mostres independents: en aquest mètode s'extreuen mostres independents de cada una de les dues poblacions, de manera que els membres d'una mostra no tenen necessàriament relació amb els membres de l'altra. Per exemple, es duu a terme un estudi per a avaluar les diferències en els nivells educatius entre dos centres de capacitació. Per a això, s'aplica un examen comú a persones que assisteixin a cada centre. Les qualificacions de l'examen són un dels factors principals per a avaluar diferències de qualitat entre els centres.

1.1. Contrastos d'hipòtesi per a la diferència de mitjanes

Contrast d'hipòtesi per a la diferència entre les mitjanes de dues poblacions: mostres independents.

En aquest apartat presentarem els procediments per a contrastar les hipòtesis sobre la diferència de mitjanes de dues poblacions.

Se suposa que es disposa de mostres aleatòries independents de n_1 i n_2, observacions procedents de dues poblacions normals amb mitjanes μ_1 u μ_2, i variàncies conegudes σ_1^2 i σ_2^2, respectivament. Volem contrastar la hipòtesi nul·la (H_0) que afirma que els valors de les mitjanes de les dues poblacions són iguals: $\boldsymbol{H_0}$: $\boldsymbol{\mu_1 - \mu_2 = 0}$ davant de qualsevol de les hipòtesis alternatives:

Nota

A vegades en lloc de:
H_0: $\mu_1 - \mu_2 = 0$
escriurem:
H_0: $\mu_1 = \mu_2$

$\boldsymbol{H_1}$: $\boldsymbol{\mu_1 - \mu_2 \neq 0}$, $\boldsymbol{H_1}$: $\boldsymbol{\mu_1 - \mu_2 < 0}$, $\boldsymbol{H_1}$: $\boldsymbol{\mu_1 - \mu_2 > 0}$. Es fixa un nivell de significació α per a fer el contrast.

L'estadístic de contrast serà:

$$z^* = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}},$$

on $\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}} = \sigma_{\bar{x}_1 - \bar{x}_2}$ és l'error estàndard.

És una observació d'una distribució $N(0,1)$.

En el cas que no puguem assegurar que les mostres provenen de poblacions normals, només podrem contrastar la diferència de mitjanes si les grandàries de les mostres són superiors a 30.

El teorema central del límit diu que si tenim un grup nombrós de variables independents i totes segueixen el mateix model de distribució (sigui quin sigui el model), la suma d'aquestes es distribueix segons una distribució normal estàndard.

Per tant, l'estadístic de contrast:

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

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

Recordeu

$\bar{X}_1 \to N(\mu_1, \sigma_1)$ i
$\bar{X}_2 \to N(\mu_2, \sigma_2)$

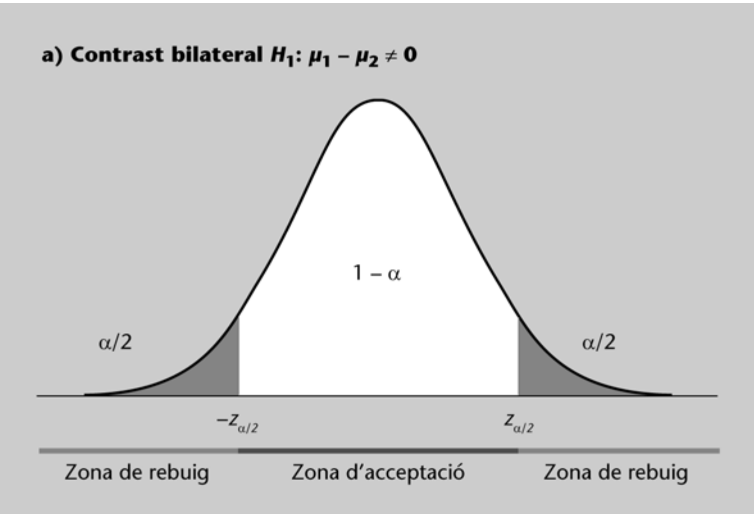
La variable diferencia de grandàries mostrals:

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

Regla de decisió del contrast d'hipòtesi

Les regions de rebuig de la hipòtesi nul·la $\boldsymbol{H_0}$: $\boldsymbol{\mu_1 - \mu_2 = 0}$ són:

Figura 1. Regions de rebuig per a contrastos de les diferències de mitjanes



Variàncies poblacionals conegudes

Podem actuar de dues maneres:

1) a partir del ***p*-valor** segons sigui H_1:

- p-valor = $P(|Z| > |z^*|)$
- p-valor = $P(Z < z^*)$
- p-valor = $P(Z > z^*)$

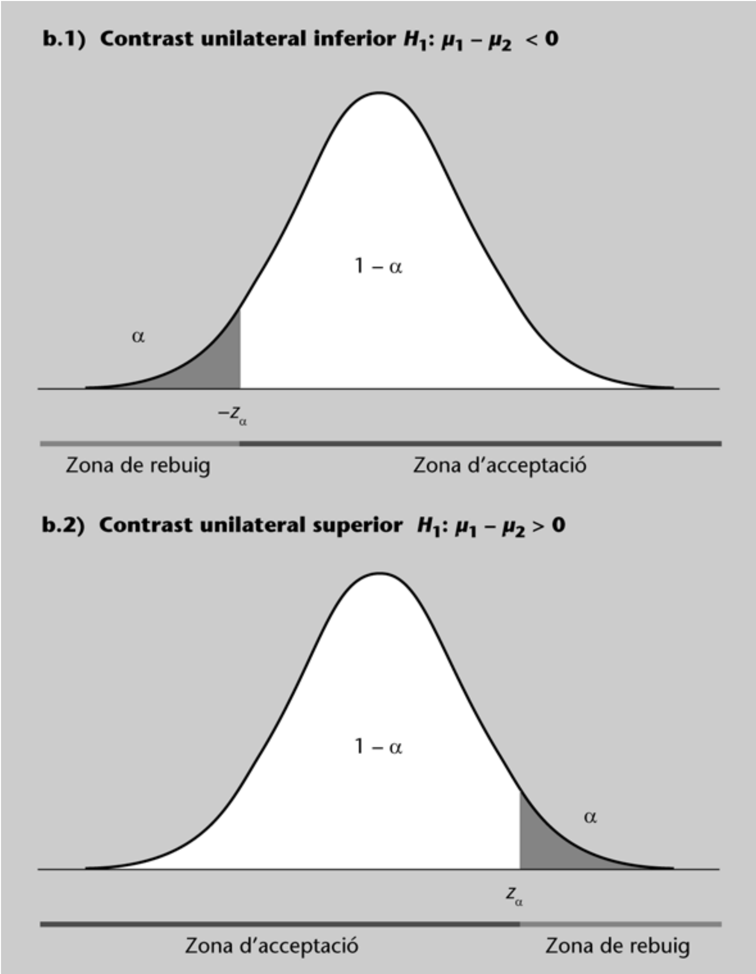
2) si p-valor $\leq \alpha$ **es rebutja $\boldsymbol{H_0}$** a partir dels valors crítics segons sigui H_1:

- Si $|z^*| > z_{\alpha/2}$ es rebutja H_0
- Si $z^* < -z_\alpha$ es rebutja H_0
- Si $z^* > -z_\alpha$ es rebutja H_0

on:

z_α és tal que $P(Z > z_\alpha) = \alpha$ i

$z_{\alpha/2}$ és tal que $P(Z > z_{\alpha/2}) = \alpha/2$



Un cop calculat el valor de l'estadístic de contrast, cal determinar el *p*-valor. El *p*-valor depèn de la hipòtesi alternativa plantejada.

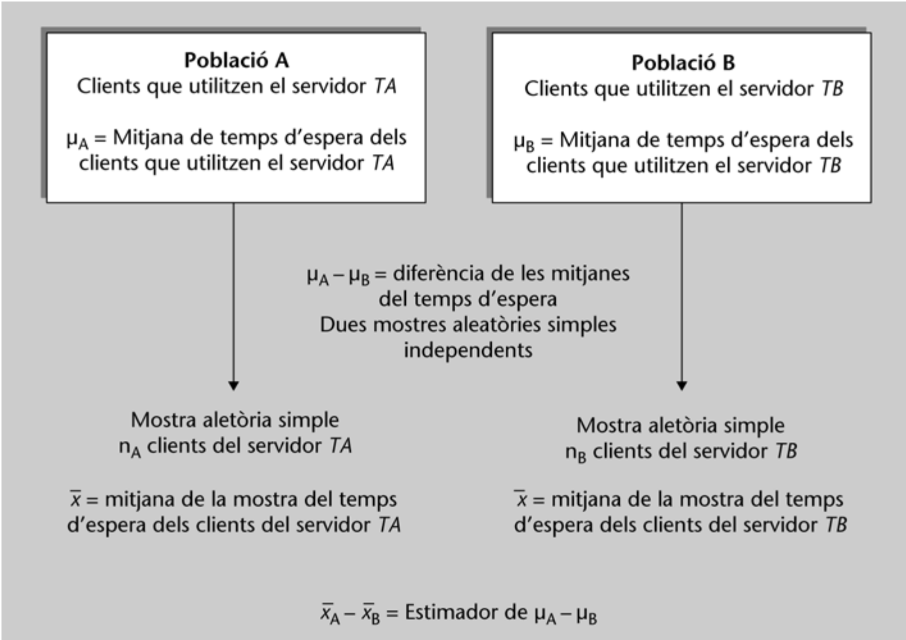
- Si $H_1 : \mu_1 - \mu_2 \neq 0$, aleshores $p = 2P(Z < |z|)$
- Si $H_1 : \mu_1 - \mu_2 < 0$, aleshores $p = P(Z < z)$
- Si $H_1 : \mu_1 - \mu_2 > 0$, aleshores $p = P(Z > z)$

Els *p*-valors d'aquests contrastos són la probabilitat d'obtenir un valor almenys tan extrem com l'estadístic de contrast obtingut.

Si el *p*-valor és significatiu es rebutja la hipòtesi nul·la si és menor que el nivell de significació α fixat.

Exemple 1. Comparació de les mitjanes del temps de resposta de dos servidors

Figura 2. Estimació de la diferència entre les mitjanes de dues poblacions



En una empresa informàtica volem mesurar l'eficiència de dos servidors web. Per a això, mesuren el temps d'espera del client entre la petició que aquest fa i la resposta que li dóna el servidor. A la taula 1 veiem que els temps d'espera (en mil·lisegons) d'ambdós servidors (*TA* i *TB*) per a 50 peticions són:

Taula 1. Dades exemple 1. Comparació de les mitjanes del temps de resposta de dos servidors

Temps d'espera per al servidor *A*				**Temps d'espera per al servidor *B***			
9,67	10,01	8,08	10,01	6,45	6,94	12,11	10,31
9,62	10,55	9,98	9,96	9,64	10,47	12,55	10,83
9,50	11,26	10,30	9,28	8,53	8,47	7,98	8,41
10,88	10,64	7,05	10,30	9,20	7,42	10,20	9,15
8,94	10,23	11,79	11,08	4,55	7,48	11,28	7,06
10,59	11,63	9,59	10,05	8,51	11,01	6,53	8,04
9,81	8,91	10,88	9,74	12,11	9,56	8,14	11,70
9,46	10,27	9,83	11,14	7,65	6,80	8,99	10,56
9,26	9,49	10,92	9,44	8,85	8,99	10,01	7,82
9,02	8,99	10,98	9,17	8,45	7,48	8,14	6,01
8,61	10,09	9,54	10,86	8,80	12,57	9,69	8,82
9,42	9,11	10,17		8,82	7,97	7,03	
10,86	9,47	10,32		9,85	8,62	8,59	

Suposem que les mostres aleatòries dels temps d'espera són independents. L'empresa vol contrastar si el servidor *A* és menys eficient (més lent) que el servidor *B* amb un nivell de confiança del 99%.

Per a contestar aquestes preguntes farem un contrast per a comparar dues mitjanes. Com que l'enunciat ens pregunta si el servidor *A* és menys eficient que el servidor *B*, fonament de dret que un servidor és menys eficient si és més lent, llavors hem de contrastar si la mitjana del temps d'espera del servidor *A* és més gran que la mitjana del temps d'espera del servidor *B*. Així, doncs, hem de plantejar una **hipòtesi alternativa unilateral**.

- Les hipòtesis nul·la i alternativa són: $H_0 : \mu_A - \mu_B = 0$, $H_1 : \mu_A - \mu_B > 0$
- Fixem $\alpha = 0{,}01$.
- No podem assegurar que les poblacions són normals, però com hem esmentat anteriorment, en tractar-se de mostres grans (superiors a 30 observacions) l'estadístic de contrast serà:

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

És una observació d'una variable aleatòria que es **distribueix aproximadament com una *N*(0,1)**.

Per a resoldre'l manualment calcularem primer els valors mostrals com exposem en els mòduls anteriors:

Temps d'espera per al servidor *A*	Temps d'espera per al servidor *B*
$n_A = 50$	$n_B = 50$
$\bar{x}_A = 9{,}94$	$\bar{x}_B = 8{,}90$
$s_A = 0{,}90$	$s_B = 1{,}75$

Les variàncies mostrals s_A^2 i s_B^2 per a estimar les variàncies poblacionals i calcular l'estadístic z^*:

$$z^* = \frac{(9{,}94 - 8{,}90)}{\sqrt{\frac{0{,}90^2}{50} + \frac{1{,}75^2}{50}}} = 3{,}75$$

Ara es pot calcular el *p*-valor $p = P(Z > 3{,}75) = 0{,}00$

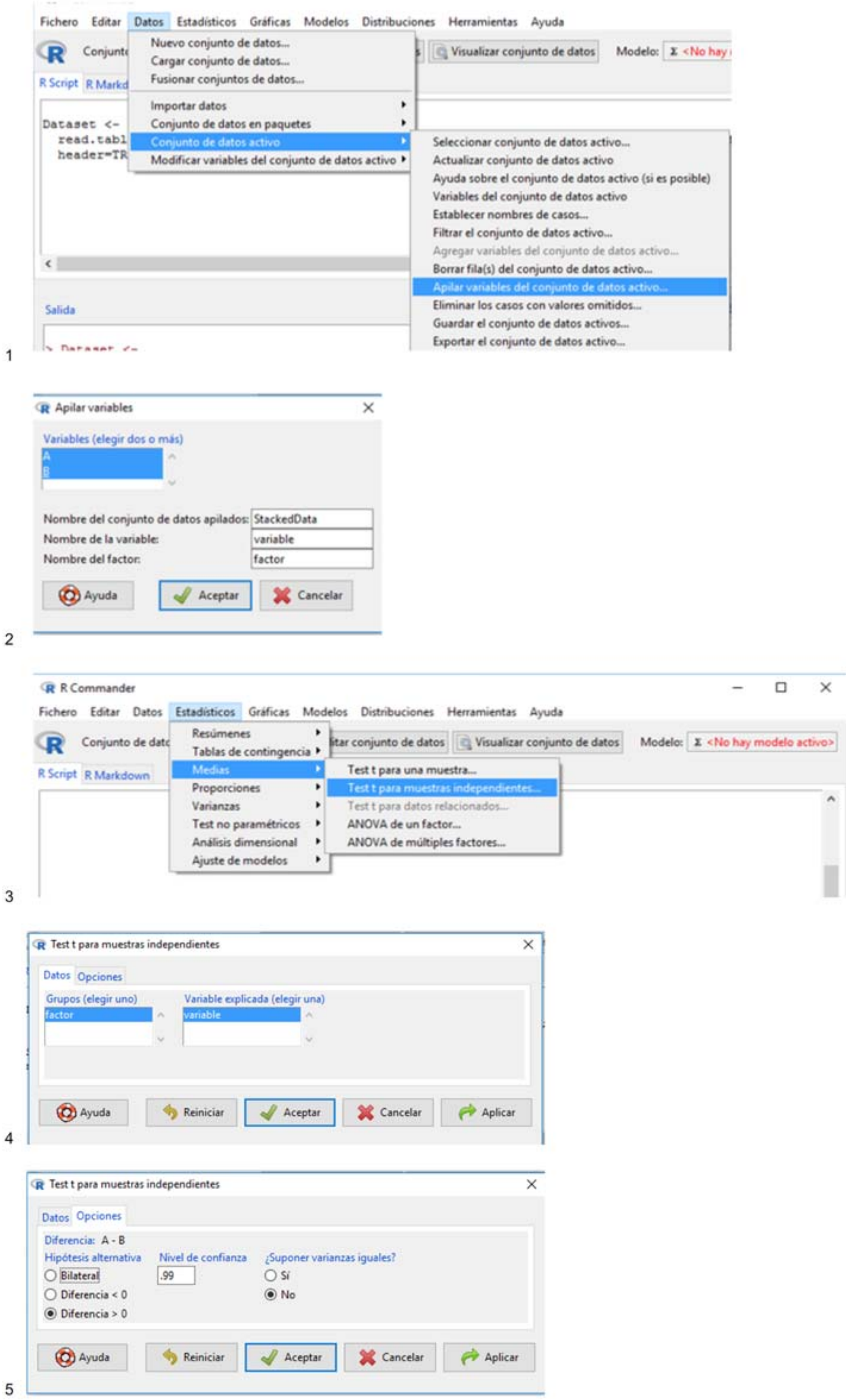
Atès que el *p*-valor és menor que $\alpha = 0{,}01$, es rebutja la hipòtesi nul·la a favor de l'alternativa. Així, el temps mitjà d'espera del servidor *A* és més gran que el del *B*. Per tant, el servidor *A* és menys eficient que el *B*.

Exemple amb el R Commander: si l'exemple anterior es resol amb el R Commander s'observa que el programa no dóna l'opció d'utilitzar la distribució normal. De tota manera, ja que les mostres són molt grans, sabem que la distribució *t* de Student s'apropa a la normal a mesura que augmenta el nombre de graus de llibertat. Per tant, els resultats que dóna R Commander seran similars per l'aproximació a la normal.

Els resultats de la figura 3 mostren el p-valor = 0,000 < 0,001 indica que podem rebutjar la hipòtesi nul·la i concloem que les mitjanes de temps d'espera del servidor A és més gran que les del B. Per tant, el servidor A és menys eficient que el B.

Els graus de llibertat (DF) de l'estadístic t augmenten si les poblacions tenen distribució aproximadament normal, però les variàncies poblacionals no són iguals.

Figura 3. Passos per a fer un contrast d'hipòtesi per a la diferència de mitjanes per a mostres independents



Passos a seguir

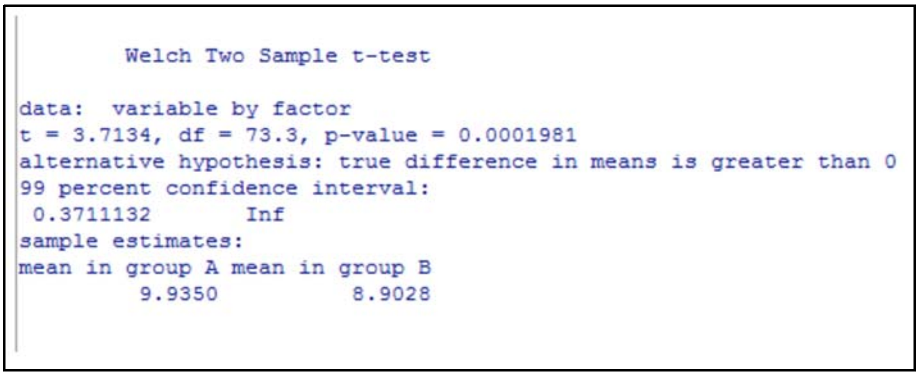
Una vegada introduïdes les dades en el programa, primer les variables han de convertir-se en factors. Per fer-ho se segueix la ruta *Datos* > *Conjunto de datos de activo* > *Apilar variables ...* **(1)**, i se seleccionen les variables a la finestra corresponent **(2)**. A continuació es fa el contrast d'hipòtesi seleccionant *Estadísticos* > *Medias* > *Test t para muestras independientes* **(3)**. En el quadre de diàleg *Datos* i *Opciones* es completen els camps mostrats a **(4)** i **(5)**.

Seleccioneu *OK* per a obtenir el contrast.

Observeu

En el pas **(5)** no assumim que les variàncies siguin iguals.

Figura 4. Resultats del contrast d'hipòtesi



```
        Welch Two Sample t-test

data:  variable by factor
t = 3.7134, df = 73.3, p-value = 0.0001981
alternative hypothesis: true difference in means is greater than 0
99 percent confidence interval:
 0.3711132       Inf
sample estimates:
mean in group A mean in group B
         9.9350          8.9028
```

Contrastos per a mostres amb variàncies poblacionals desconegudes, però iguals.

El procediment que utilitzarem es basa en la distribució t amb $n_1 + n_2 - 2$ graus de llibertat.

L'estadístic de contrast serà:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

on s és la desviació típica comuna.

Exemple 2. Estudi sobre la producció científica

El director d'una escola universitària vol comparar dos departaments, A i B, de grandària similar, pel que fa al nombre total de publicacions o ponències de qualitat que puguin aportar millores a l'activitat docent de l'escola. Es considerarà que una publicació és de qualitat quan hagi estat publicada en una revista indexada o per una editorial de prestigi internacional; es considerarà que una ponència és de qualitat quan aquesta s'hagi desenvolupat en un congrés internacional amb procés de selecció; per a determinar si la publicació o ponència pot aportar millores a l'activitat docent s'ha constituït un tribunal d'experts independents.

Ha pres una mostra aleatòria formada per sis professors del departament A i s'ha trobat el valor de la variable nombre total de publicacions o ponències de qualitat per a cada un dels esmentats professors. S'ha fet el mateix amb una altra mostra aleatòria formada per vuit professors del departament B. Els resultats es presenten a continuació:

Taula 2. Dades exemple 2. Estudi sobre la difusió científica

Dept. A	5	8	7	6	9	7		
Dept. B	8	10	7	11	9	12	14	9

Per a un nivell de significació alpha = 0,05, podem afirmar que la producció mitjana d'ambdós departaments (segons els criteris establerts) és significativament diferent?

Per a fer l'estudi començarem amb el supòsit que no hi ha diferències en el nombre total de publicacions o ponències de qualitat d'ambdós departaments. Per tant, en termes de la mitjana del nombre total de publicacions o ponències de qualitat, la hipòtesi nul·la és que la diferència de mitjanes és zero. Si l'evidència de la mostra condueix al rebuig d'aquesta hipòtesi, arribarem a la conclusió que les mitjanes de qualitat són diferents per a les dues poblacions, la qual cosa indica que hi ha diferència a les publicacions de qualitat dels dos departaments, i això induiria a trobar les raons d'aquesta diferència.

En aquest estudi hi ha dues poblacions: una dels professors del departament A, i una altra dels professors del departament B. Suposem que ambdues poblacions són normals i que les seves variàncies són iguals però desconegudes.

Considerant el nombre de publicacions i ponències, les mitjanes de població són μ_A i μ_B. Es plantegen les hipòtesis de treball següents:

$$H_0\text{: } \mu_1 - \mu_2 = 0 \text{ (ambdues mitjanes són iguals)}$$
$$H_1\text{: } \mu_1 - \mu_2 \neq 0 \text{ (ambdues mitjanes són diferents)}$$

Es tracta d'un contrast d'**hipòtesi bilateral** sobre la mitjana de dues poblacions independents.

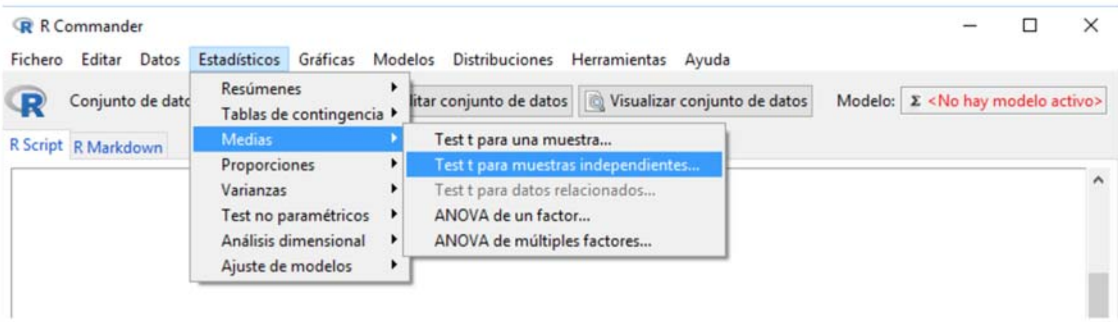
Es farà servir el R Commander per a provar les hipòtesis entorn de la diferència entre les mitjanes de dues poblacions (figura 5).

Figura 5. Passos per a fer un contrast d'hipòtesi per a la diferència de mitjanes per a mostres amb variàncies poblacionals desconegudes

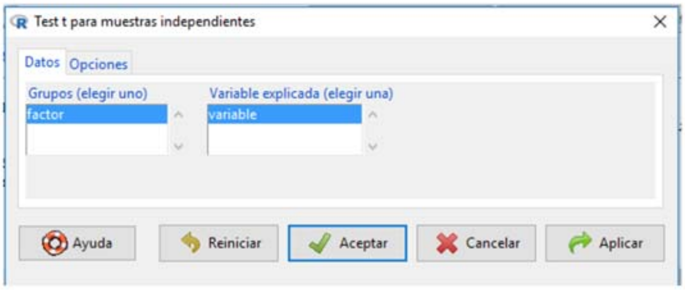
Passos a seguir

Una vegada introduïdes les dades en el programa, aquestes són apilades com en la figura 3 de la forma (1) i (2). A continuació se segueix la ruta *Estadísticos > Medias > Test t para muestras independientes* **(3)**. En el quadre de diàleg *Datos* i *Opciones* es completen els camps mostrats a **(4)** i **(5)**.

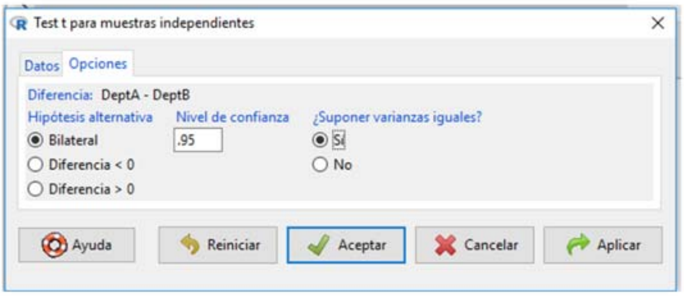
Seleccioneu *OK* per a obtenir el contrast.



3



4



5

Observeu

En el pas **(5)** assumim que les variàncies són desconegudes però iguals i marquem la casella corresponent.

Vam obtenir els resultats de la figura 6. Apareixen els valors mostrals de tots dos departaments. L'estadístic de contrast és un valor t = –2,84 amb 12 graus de llibertat (*DF*) i el p-valor *P-Value* = 0,015.

Figura 6. Resultats del contrast d'hipòtesi

```
        Two Sample t-test

data:  variable by factor
t = -2.8372, df = 12, p-value = 0.01497
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.3038363 -0.6961637
sample estimates:
mean in group DeptA mean in group DeptB
                  7                  10
```

Com que p-valor = 0,015 < 0,05, podem rebutjar la hipòtesi nul·la amb α = 0,05. Així la producció mitjana d'ambdós departaments (segons els criteris establerts) és significativament diferent en els departament A i B. Observeu que la informació del R Commander per a l'interval de confiança del 95% a la figura 5 té com a extrems els valors –5,30 i –0,70 (observeu que el 0 no està inclòs en l'esmentat interval). Això també ens indica que hem de rebutjar la hipòtesi nul·la i acceptar l'alternativa (les mitjanes són diferents).

Així, els resultats permeten al director de l'escola universitària concloure que hi ha diferències significatives entre tots dos departaments en el nombre total de publicacions o ponències de qualitat.

Aplicant **Microsoft Excel** a l'exemple 2. Estudi sobre la difusió científica.

Podeu executar una prova *t* de dues mostres independents per a dades no acoblades. Feu clic a (*t-Test: Two Simple > Assuming Equal Variantes*) *prova t: dues mostres suposant variàncies iguals* i especifiqueu les dues columnes que continguin les dades.

La figura 7 mostra la sortida corresponent que ofereix el **Microsoft Excel**.

Figura 7. Resultats exemple 2. Estudi sobre la difusió científica. Excel

	A	B	C
1	Prova *t* per a dues mostres suposant variàncies iguals		
2			
3		*DepA*	*DepB*
4	Mitjana	7	10
5	Variància	2	5,14285714
6	Observacions	6	8
7	Variància agrupada	3,83333333	
8	Diferència hipotètica de les mitjanes	0	
9	Graus de llibertat	12	
10	Estadístic *t*	-2,8371975	
11	*P*(*T* <= *t*) una cua	0,0074872	
12	Valor crític de *t* (una cua)	1,78228755	
13	*P*(*T* <= *t*) dues cues	0,01497439	
14	Valor crític de *t* (dues cues)	2,17881283	

Com observem, el *p*-valor = 0,0149, en ser menor que el valor de α, pot rebutjar la hipòtesi nul·la amb $\alpha = 0{,}05$.

Contrast d'hipòtesi per a la diferència entre les mitjanes de dues poblacions: Mostres dependents (dades aparellades)

Disposem d'una mostra aleatòria de n parelles d'observacions de distribucions amb mitjanes μ_A i μ_B. Denotem per $\bar{d}$ i s_d la mitjana mostral i la desviació típica observades per a les n diferències $(x_A - x_B)$ i sigui $\mu_d = \mu_A - \mu_B$ mitjana de les diferències per a la població.

Si la distribució poblacional és normal podem fer els contrastos següents per a un nivell de significació α:

la hipòtesi nul·la: $H_0 = \mu_d = 0$

la hipòtesi alternativa (H_1) pot ser bilateral: $H_1 : \mu_d \neq 0$

o unilateral $H_1 : \mu_d > 0$ o $H_1 : \mu_d < 0$

Anàlisi de dades

Per a fer contrastos d'hipòtesi amb l'**MS Excel** és necessari instal·lar un complement anomenat *Anàlisi de dades*. Per a instal·lar les eines d'anàlisis de dades feu clic a *Herramientas > complementos*, en el quadre de diàleg activeu *Herramientas para análisis*.

Mostres dependents

Mostres dependents significa que tenim una mostra d'observacions de dues variables.

La grandària de la mostra és:

$$\bar{d} = \frac{\sum_i^n d_i}{n}$$

La desviació estàndard és:

$$s_d = \sqrt{\frac{\sum_i^n (d_i - \bar{d})^2}{n-1}}$$

la notació *d* és per a recordar que la mostra aparellada proueix dades de *diferència*.

En aquest tipus de contrast fem servir la mateixa metodologia utilitzada per al contrast de la mitjana per a una sola població que vam veure al mòdul anterior.

Per il·lustrar el disseny amb mostres aparellades utilitzarem l'exemple següent:

Exemple 3. Puntuacions d'un test d'actitud

A un grup de persones se'ls va proposar un test d'actitud sobre un tema polèmic i vam obtenir uns resultats. Després el grup va assistir a la projecció d'una pel·lícula favorable al tema i tot seguit se'ls va proposar de nou el test d'actitud, del qual es van obtenir altres resultats. A la taula 3 apareixen les dades sobre les puntuacions del test fet a 11 persones. Cada persona dóna un parell de valors, un per a abans d'assistir a la projecció de la pel·lícula i un altre després d'assistir-hi. Es vol verificar la hipòtesi que la projecció d'una pel·lícula favorable fa que canviï l'actitud desfavorable cap al tema.

Taula 3. Dades exemple 3. Puntuacions d'un test d'actitud

Persona	Puntuació del test abans de veure la pel·lícula	Puntuació del test després de veure la pel·lícula	Diferència de puntuacions del test (d_i)
1	24	16	8
2	20	18	2
3	24	20	4
4	28	24	4
5	30	24	6
6	20	22	–2
7	24	20	4
8	22	18	4
9	18	10	8
10	18	8	10
11	24	20	4

$$\sum_{i=1}^{11} d_i = 52$$

Observeu que l'última columna de la taula 3 conté la diferència entre les puntuacions abans i després de veure la pel·lícula. La clau per a analitzar el disseny amb mostres acoblades és tenir en compte que només es considera la columna de les diferències. Verificarem la hipòtesi d'investigació a un nivell de significació de l'1% ($\alpha = 0{,}01$). Sigui $\mu_d =$ la mitjana de les diferències per a la població de persones.

Les hipòtesis seran:

$$H_0\text{: } \mu_d = 0$$
$$H_1\text{: } \mu_d \neq 0$$

Es tracta d'un contrast bilateral. Si rebutgem H_0 arribem a la conclusió que les mitjanes de les puntuacions del test són diferents al nivell de significació de l'1%. En el mòdul 2 vam veure que si podem suposar que la població té una distribució normal, l'estadístic de contrast és una ***t* de Student** amb $n - 1$ graus de llibertat, per a provar la hipòtesi nul·la sobre la mitjana poblacional, si no coneixem la variància de la població com en aquest exemple.

Amb dades de diferència es calcula l'estadístic de prova per a la hipòtesi nul·la

$$H_0\text{: } \mu d = 0 \text{ és:}$$

$$\text{Com } \bar{d} = \frac{52}{11} = 4{,}72 \text{ i } s_d = \sqrt{\frac{106{,}18}{10}} = 3{,}26$$

$$t^* = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}} = \frac{4{,}72 - 0}{3{,}26 / \sqrt{11}} = 4{,}80$$

Amb $\alpha = 0{,}01$ i $n - 1 = 10$ graus de llibertat ($t_{0,01/2} = t_{0,005} = 3{,}169$), la regla de rebuig per a la prova bilateral és:

$$\text{Rebutjar } H_0 \text{ si } t^* < 3{,}169 \text{ o } t^* > 3{,}169$$

En vista que $t^* = 4{,}80$ és a la regió de rebuig, rebutgem H_0 i acceptem H_1 i podem afirmar al 99% de confiança que la pel·lícula va influir en l'actitud de les persones.

Amb els resultats de la mostra podem definir un interval de confiança de diferència entre les dues mitjanes de la població, amb la metodologia per a població única del mòdul 2 els càlculs són:

$$\bar{d} \pm t_{\alpha/2} \frac{s_d}{\sqrt{n}} = 4{,}72 \pm 3{,}169\left(\frac{3{,}26}{\sqrt{11}}\right) = 4{,}72 \pm 3{,}12 = [1{,}60;\ 7{,}84]$$

En conseqüència, l'interval de confiança de 99% de la diferència de mitjanes entre les mitjanes de les dues puntuacions del test és d'1,6 fins a 7,84. Observem que l'interval no inclou el valor zero. Després com hem vist en el contrast podem rebutjar H_0.

Emprarem el R Commander per a aquest exemple 3. Puntuacions d'un test d'actitud.

La figura 8 mostra els passos bàsics necessaris per a fer el contrast d'hipòtesis.

En primer lloc comprovarem el supòsit que les poblacions segueixen una distribució aproximadament normal:

Figura 8. Passos per a fer un test de normalitat amb R Commander

1

```
> with(Dataset, shapiro.test(Antes))

        Shapiro-Wilk normality test

data:  Antes
W = 0.92417, p-value = 0.355

> with(Dataset, shapiro.test(Después))

        Shapiro-Wilk normality test

data:  Después
W = 0.88424, p-value = 0.1177
```

2

Passos a seguir

Una vegada introduïdes les dades en el programa se segueix la ruta *Estadísticos* > *Resúmenes* > *Test de normalidad de Shapiro-Wilk* i s'emplenen els camps a la finestra corresponent per a cada variable.

En els gràfics resultants (figures 9 i 10) observem que no hi ha indicis per a dubtar que es compleix el supòsit de normalitat, ja que els punts estan molt pròxims a les respectives rectes. Els gràfics ens proporcionen també el p-valor associat al **test de normalitat d'Anderson-Darling**, on l'esmentat p-valor és prou gran en ambdós casos per a no descartar la hipòtesi nul·la d'aquest contrast: que les dades segueixen una distribució normal.

Figura 9. Test de normalitat. R Commander

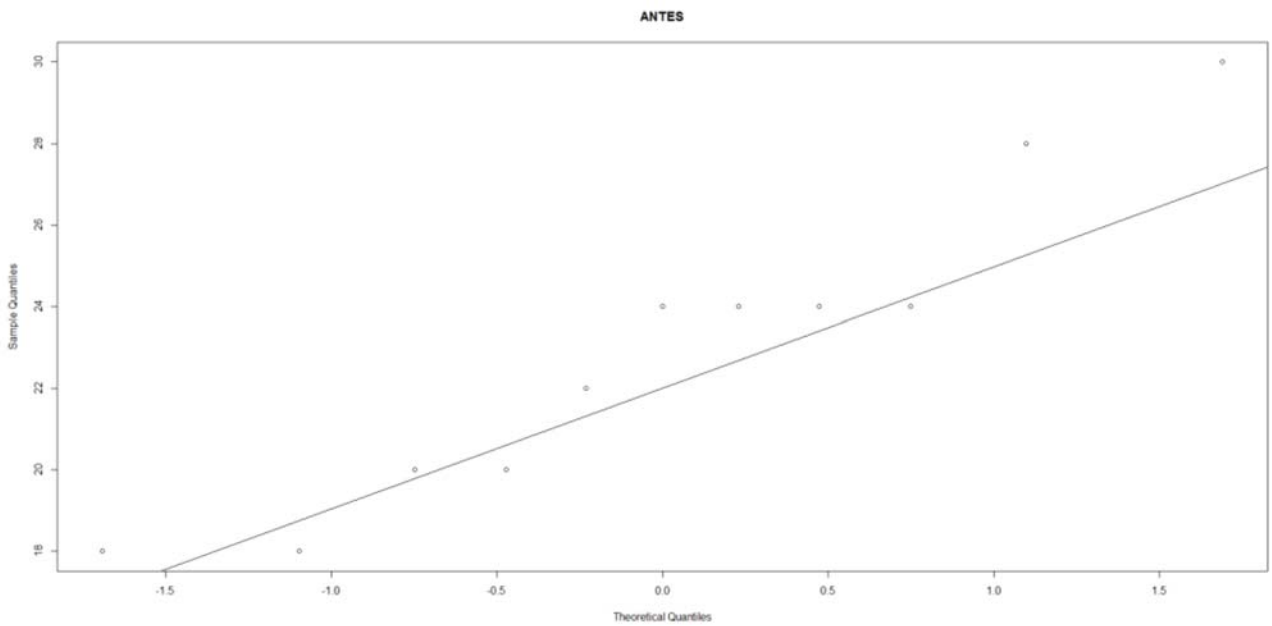
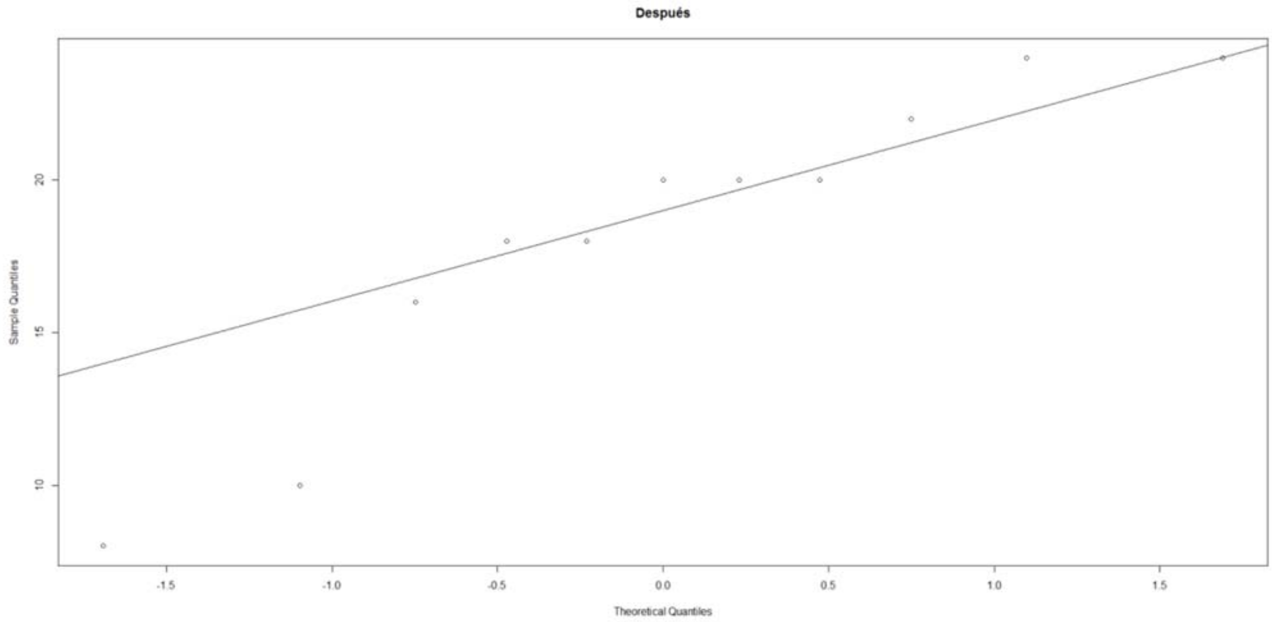
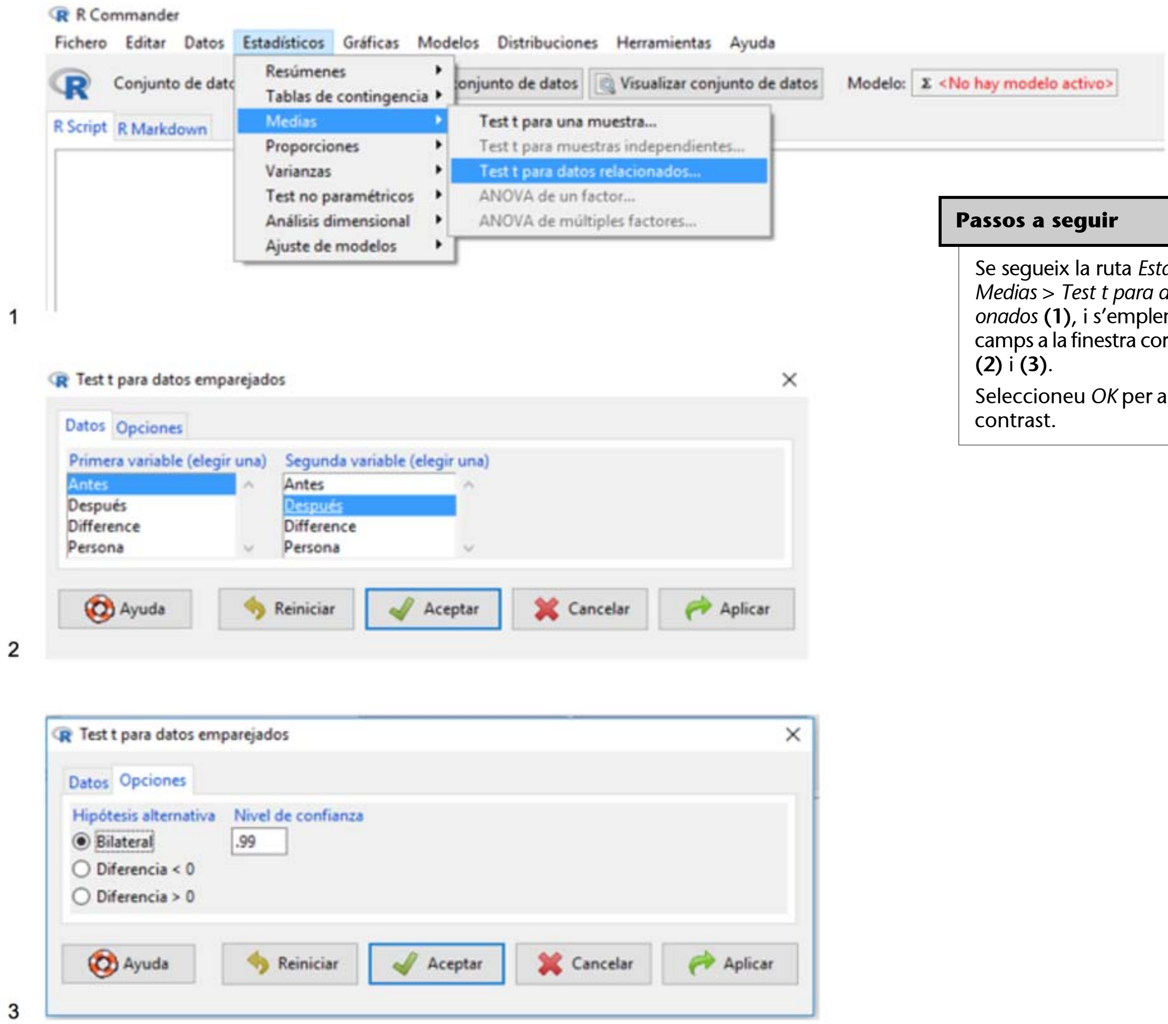


Figura 10. Test de normalitat. R Commander



Passem, doncs, a fer les inferències ja comentades sobre μ_d.

Figura 11. Passos per a fer un contrast d'hipòtesis per a la diferència de mitjanes per a mostres dependents



Passos a seguir

Se segueix la ruta *Estadísticos* > *Medias* > *Test t para datos relacionados* **(1)**, i s'emplenen els camps a la finestra corresponent **(2)** i **(3)**.

Seleccioneu *OK* per a obtenir el contrast.

Els resultats obtinguts en la figura 12 mostren que, partint de les observacions registrades, hi ha una probabilitat de 0,99 que μ_d sigui un valor de l'interval (1,613; 7,841). A més amb un *p*-valor de 0,001 també podem afirmar que hi ha indicis suficients per a rebutjar la hipòtesi nul·la. Per tant, podem concloure que la pel·lícula va influir en l'actitud de les persones.

Figura 12. Resultats del contrast de mitjanes per a dues mostres dependents. R Commander

```
        Paired t-test

data:  Antes and Después
t = 4.8115, df = 10, p-value = 0.0007112
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 1.613490 7.841056
sample estimates:
mean of the differences
               4.727273
```

També podeu executar una prova *t* per parells utilitzant **Excel**.

Des d'*Eines* > *Anàlisi de dades*, feu clic a *Prova t per a mitjanes de dues mostres aparellades* i especifiqueu les dues columnes que contenen les dades per parells. Aquesta ordre no calcula l'interval de confiança, de manera que ha de calcular-lo mitjançant les fórmules que apareixen en aquest mòdul.

La figura 13 mostra la sortida corresponent que ofereix **Microsoft Excel**.

Figura 13. Resultats del contrast de mitjanes per a dues mostres aparellades. Excel

B13 fx 0,000711223615253786

	A	B	C
1	Prova *t* per a dues mostres aparellades		
2			
3		*ABANS*	*DESPRÉS*
4	Mitjana	22,90909091	18,18181818
5	Variància	14,69090909	26,76363636
6	Observacions	11	11
7	Coeficient de correlació de Pearson	0,777564218	
8	Diferència hipotètica de les mitjanes	0	
9	Graus de llibertat	10	
10	Estadístic *t*	4,811515866	
11	*P*(*T* <= *t*) una cua	0,000355612	
12	Valor crític de *t* (una cua)	2,763769458	
13	*P*(*T* <= *t*) dues cues	0,000711224	
14	Valor crític de *t* (dues cues)	3,169272672	

En ser el *p*-valor = 0,0007 < α(0,01), es rebutja H_0.

1.2. Contrastos d'hipòtesi per a la diferència de proporcions

En estudiar la diferència entre dues proporcions poblacionals, l'estimador és $\hat{p}_1 - \hat{p}_2$. Com vam veure en casos anteriors, la distribució de l'estimador de les mostres és un factor clau per a determinar els intervals de confiança i per a provar les hipòtesis dels paràmetres.

Suposem que disposem de dues mostres aleatòries simples i independents de n_1 i n_2 observacions. Les proporcions mostrals d'èxits són respectivament: $\hat{p}_1$ i $\hat{p}_2$.

La distribució de la variable diferència de proporcions mostrals $\hat{p}_1 - \hat{p}_2$ es pot aproximar amb una distribució $N(0,1)$.

Recordeu

Si les grandàries de les mostres són grans:

$$\hat{p}_1 \to N\left(p_1, \sqrt{\frac{p_1(1-p_1)}{n_1}}\right)$$

i

$$\hat{p}_2 \to N\left(p_2, \sqrt{\frac{p_2(1-p_2)}{n_2}}\right)$$

Sota el supòsit de la hipòtesi nul·la certa (H_0: $p_1 - p_2 = 0$), tenim que l'estadístic de contrast és:

$$z^* = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}}$$

On ($\hat{p}_1 - \hat{p}_2$) és la diferència de les proporcions mostrals.

El valor $\hat{p}$ és el valor estimat comú de la proporció poblacional que podem estimar a partir de les dues mostres:

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

Regla de decisió del contrast d'hipòtesis

Nota

A vegades en lloc de:
H_0: $p_1 - p_2 = 0$
escriurem:
H_0: $p_1 = p_2$

Un cop calculat el valor de l'estadístic de contrast, cal determinar el *p*-valor. El *p*-valor depèn de la hipòtesi alternativa plantejada.

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

Si el *p*-valor és significatiu, es rebutja la hipòtesi nul·la si és menor que el nivell de significació α fixat.

Es farà servir l'exemple de l'apartat 1.1, taula 1. Dades de l'exemple 1. "Comparació de les mitjanes del temps de resposta de dos servidors".

En una empresa informàtica es vol mesurar l'eficiència de dos servidors web. Per a això, mesuren el temps d'espera del client entre la petició que aquest fa i la resposta que li dóna el servidor. Els temps d'espera (en mil·lisegons) d'ambdós servidors (*TA* i *TB*) per a 50 peticions són en la taula 2.

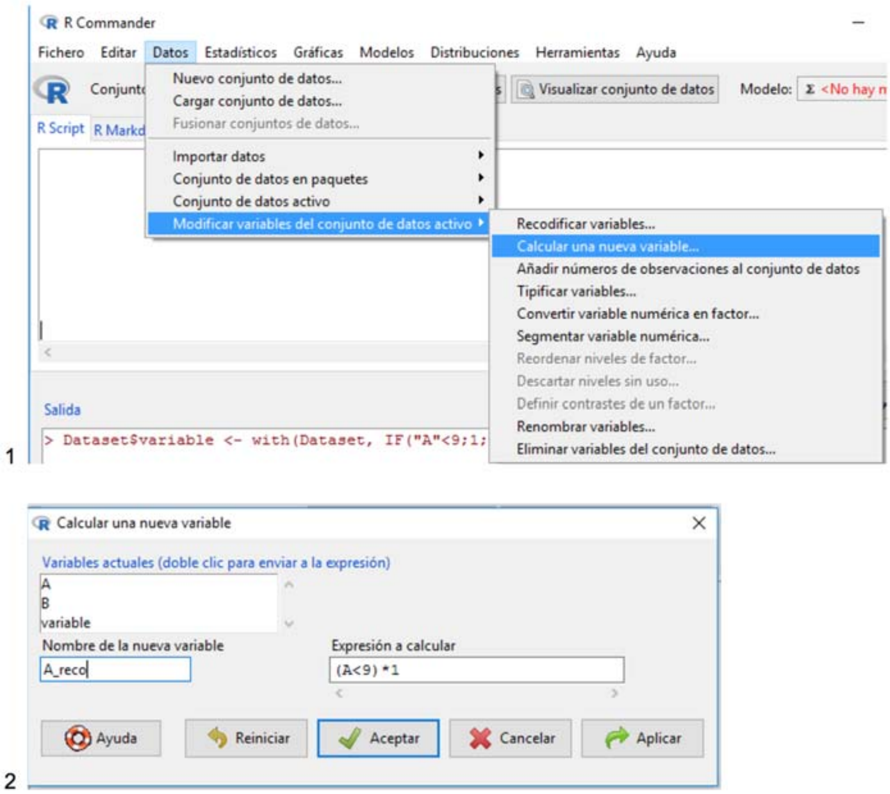
Direm que el temps d'espera és accep.table si és menor que 9 mil·lisegons. Podem dir que la proporció de peticions amb temps d'espera acceptable és diferent per als dos servidors?

Per a contestar aquesta pregunta hem de fer un contrast de diferència de proporcions que resoldrem amb el R Commander.

La primera operació que hem de fer és calcular per a cada tipus de servidor la proporció de temps inferior a 9 mil·lisegons. Per a això, creem una nova columna de nom on posarem un 1 si l'observació de temps d'espera del servidor *A* és inferior a 9 i 0 en cas contrari. Després sumarem els valors de la columna i obtindrem el nombre d'observacions de temps del servidor *A* inferior a 9 mil·lisegons.

En la figura 14 s'indiquen els passos a seguir:

Figura 14. Passos a seguir per a recalcular una variable nova

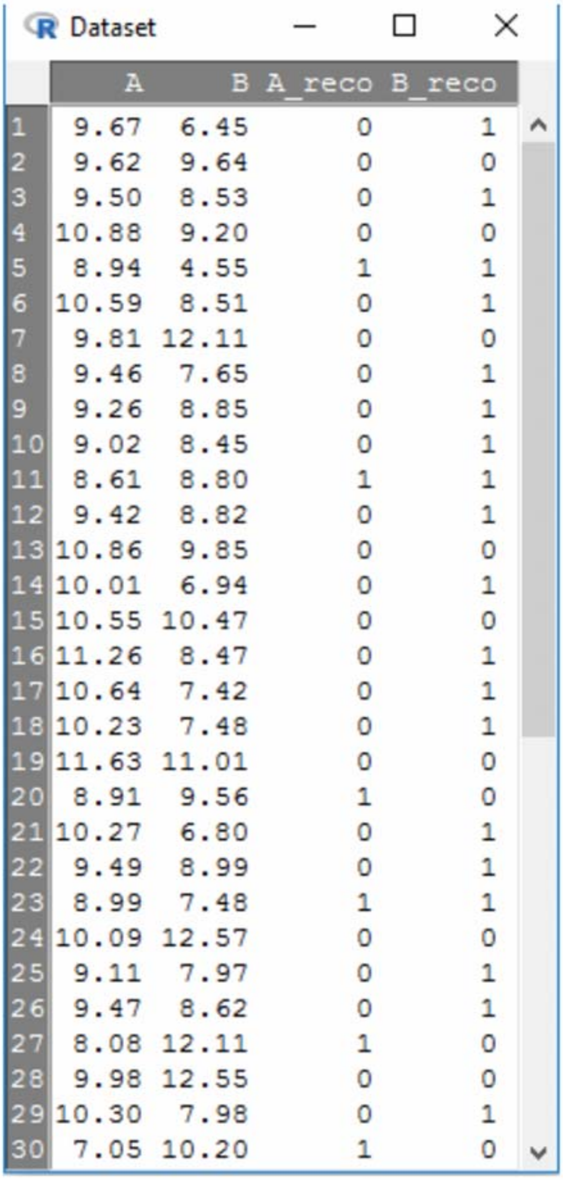


Indicació

Per a fer aquest exercici primer calcularem una nova variable, que valgui 1 si el temps d'espera és menor que 9 mil·lisegons i 0 en cas contrari. Per a calcular aquesta variable, podem utilitzar R Commander seguint la ruta *Datos > Modificar variables del...> Calcular una nueva variable.*

Fem el mateix per al temps del servidor *B* i crearem una columna de nom *B_reco* amb 1 si el temps és inferior a 9 i 0 en cas contrari.

Figura 15. Dades



Dataset

	A	B	A_reco	B_reco
1	9.67	6.45	0	1
2	9.62	9.64	0	0
3	9.50	8.53	0	1
4	10.88	9.20	0	0
5	8.94	4.55	1	1
6	10.59	8.51	0	1
7	9.81	12.11	0	0
8	9.46	7.65	0	1
9	9.26	8.85	0	1
10	9.02	8.45	0	1
11	8.61	8.80	1	1
12	9.42	8.82	0	1
13	10.86	9.85	0	0
14	10.01	6.94	0	1
15	10.55	10.47	0	0
16	11.26	8.47	0	1
17	10.64	7.42	0	1
18	10.23	7.48	0	1
19	11.63	11.01	0	0
20	8.91	9.56	1	0
21	10.27	6.80	0	1
22	9.49	8.99	0	1
23	8.99	7.48	1	1
24	10.09	12.57	0	0
25	9.11	7.97	0	1
26	9.47	8.62	0	1
27	8.08	12.11	1	0
28	9.98	12.55	0	0
29	10.30	7.98	0	1
30	7.05	10.20	1	0

Una vegada tenim aquestes dues noves columnes, calculem la suma de cada una i així tindrem per a cada servidor el nombre d’observacions de temps inferior a 9. Per fer-ho s'ha d'escriure i executar.

Figura 16. Passos a seguir per a obtenir el valor suma

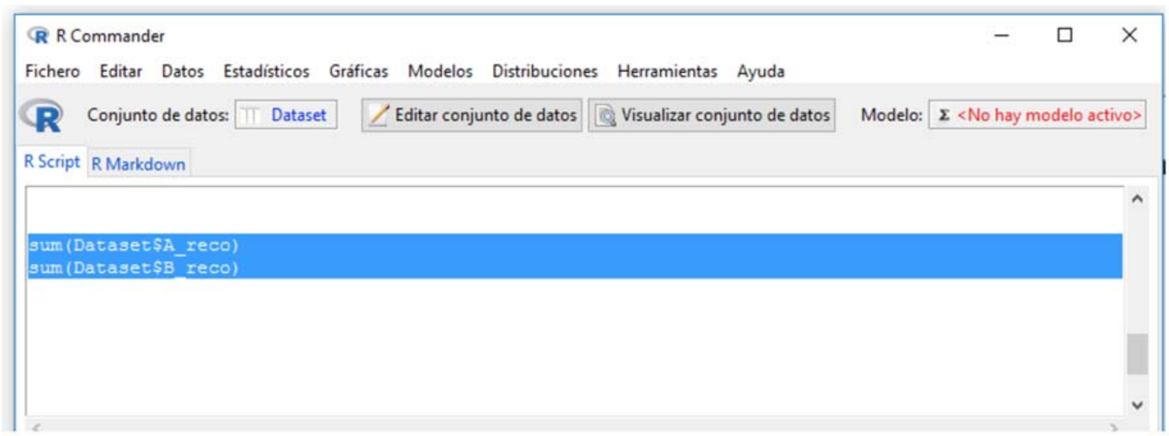


Figura 17. Resultats

```
> sum(Dataset$A_reco)
[1] 6

> sum(Dataset$B_reco)
[1] 31
```

Per al servidor A hi ha 6 observacions amb un temps d'espera inferior a 9 mil·lisegons i per al servidor B el nombre d'observacions menors de 9 mil·lisegons és 31.

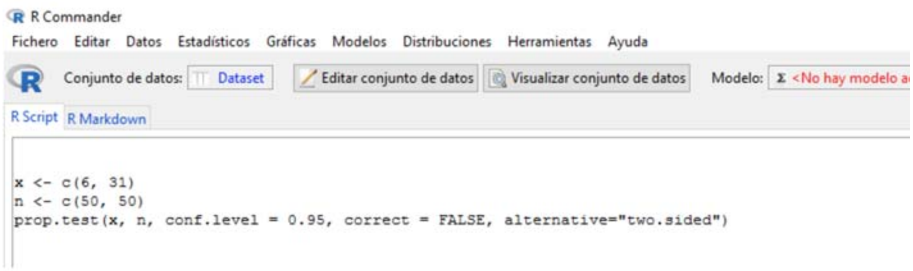
Plantejarem el contrast següent:

$$H_0 : p_A - p_B = 0$$
$$H_1 : p_A - p_B \neq 0$$

Fixem $\alpha = 0{,}05$.

La figura 18 mostra el codi a seguir per a fer el contrast de la diferència de proporcions.

Figura 18. Codi per a fer un contrast d'hipòtesis per a la diferència de proporcions



Els resultats de la figura 19 mostren el p-valor = 0,000 < 0,05 que indica que podem rebutjar la hipòtesi nul·la i concloem que la proporció de peticions amb temps d'espera acceptable és diferent per als dos servidors.

Figura 19. Resultats del contrast de diferència de proporcions. R Commander

```
        2-sample test for equality of proportions without continuity correction

data:  x out of n
X-squared = 26.813, df = 1, p-value = 2.242e-07
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.6619077 -0.3380923
sample estimates:
prop 1 prop 2
  0.12   0.62
```

Passos a seguir

Creem dues variables X i n on:

X són el nombre d'observacions amb un temps d'espera inferior a 9 en els dos servidors, primer A i després B i n és el total d'observacions.

A continuació, apliquem el contrast de diferència de proporcions mitjançant la funció *testprop.test*.

testprop.test és la funció que calcula el contrast d'hipòtesi per a la diferència de proporcions. En aquesta s'han d'indicar les proporcions, nivell de confiança, si la correcció de Yates s'ha d'aplicar, i la hipòtesi alternativa.

1.3. Contrastos d'hipòtesi de comparació de variàncies

Un dels contrastos desenvolupats en l'apartat 1.1 per a la comparació de mitjanes poblacionals depèn del supòsit d'igualtat de les dues variàncies poblacionals. Encara que en moltes aplicacions pràctiques aquest és un supòsit raonable, convé utilitzar les dades disponibles per a contrastar-ne la validesa.

En aquest apartat considerem el cas de dues mostres aleatòries independents de poblacions normals i contrastarem la igualtat de variàncies poblacionals.

Sigui s_1^2 la variància mostral d'una mostra de n_1 observacions d'una població normal amb variància σ_1^2, i s_2^2 la variància mostral d'una mostra independent de n_2 observacions d'una població normal amb variància σ_2^2. Sempre que les dues variàncies poblacionals siguin iguals ($\sigma_1^2 = \sigma_2^2$). La distribució de la relació de les dues variàncies de les mostres s_1^2/s_2^2 està definida per l'estadístic F que segueix una **distribució *F* de Snedecor** amb $n_1 - 1$ graus de llibertat per al numerador i $n_2 - 1$ graus de llibertat per al denominador,

$$F_{n_1-1;n_2-1} = \frac{s_1^2}{s_2^2}$$

Contrast d'igualtat de variàncies de dues poblacions normals

Ara ens interessa contrastar la hipòtesi nul·la que assegura que les variàncies de les poblacions són iguals $\sigma_1^2 = \sigma_2^2$, és a dir, la variància de la població 1 és igual a la variància de la població 2. Primer fixarem el nivell de significació α del contrast. Establirem les hipòtesis nul·la i alternativa:

Hipòtesi alternativa, pot ser:

- Bilateral: $\boldsymbol{H_1}$: $\sigma_1^2 \neq \sigma_2^2$, les variàncies de les dues poblacions són diferents.
- Unilateral: $\boldsymbol{H_1}$: $\sigma_1^2 > \sigma_2^2$, la variància de la població 1 és major que la variància de la població 2.
- Unilateral: $\boldsymbol{H_1}$: $\sigma_1^2 < \sigma_2^2$, la variància de la població 1 és menor que la variància de la població 2.

Nota

A vegades en lloc de:

H_0: $\sigma_1^2 = \sigma_2^2$

escriurem:

H_0: $\sigma_1^2/\sigma_2^2 = 1$

Sota el supòsit de la hipòtesi nul·la certa H_0: $\sigma_1^2 = \sigma_2^2$, l'estadístic de contrast és:

$$F^*_{n_1-1;n_2-1} = \frac{s_1^2}{s_2^2}$$

Regla de decisió del contrast d'hipòtesi

Un cop calculat el valor de l'estadístic de contrast, cal determinar el *p*-valor. El *p*-valor depèn de la hipòtesi alternativa plantejada.

- Si H_1: $\sigma_1^2 \neq \sigma_2^2$, aleshores *p*-valor $= 2P\left(F_{n_1-1,n_2-1} > F^*\right)$
- Si H_1: $\sigma_1^2 < \sigma_2^2$, aleshores *p*-valor $= P\left(F_{n_1-1,n_2-1} < F^*\right)$

- Si H_1: $\sigma_1^2 > \sigma_2^2$, aleshores p-valor $= P\left(F_{n_1-1,n_2 1} > F^*\right)$

- Si p-valor $\leq \alpha$ es rebutja H_0

Exemple 4. Variabilitat de processadors de text

Volem comparar dos tipus de processadors de textos: el LaTeX i l'OpenOffice. Per a fer-ho, considerem textos més o menys de la mateixa longitud i expliquem la variabilitat de l'espai que deixa cada processador entre les paraules. En el cas del LaTeX, considerem 10 textos i obtenim que la desviació estàndard mostral de l'espai que deixa és de 2,5 mm, mentre que per a l'OpenOffice considerem 15 textos i obtenim que la desviació estàndard mostral de l'espai que deixa és de 3,5 mm. Suposant normalitat, podem afirmar que els dos processadors de textos tenen la mateixa variabilitat en l'espai que deixen entre paraules?

Per a contestar la pregunta hem de fer un contrast d'igualtat de variàncies.

El contrast d'hipòtesi és: $\begin{matrix} H_0 : \sigma^2_{LaTeX} = \sigma^2_{OpenOffice} \\ H_1 : \sigma^2_{LaTeX} \neq \sigma^2_{OpenOffice} \end{matrix}$

Fixem el valor de $\alpha = 0{,}05$.

L'estadístic de contrast val: $F^* = \dfrac{s^2_{LaTeX}}{s^2_{OpenOffice}}$. Els valors de s^2_{LaTeX} i $s^2_{OpenOffice}$ són, respectivament, $s^2_{LaTeX} = 6{,}25$ i $s^2_{OpenOffice} = 12{,}25$.

L'estadístic F segueix la distribució F de Fisher-Snedecor amb 9 i 14 graus de llibertat. El valor de l'estadístic de contrast serà:

$$F^* = \frac{6{,}25}{12{,}25} \approx 0{,}51$$

Els valors crítics seran $F_{1-\alpha/2,9,14} = F_{0,975,9,14} \simeq 0{,}265$ i $F_{\alpha/2} = F_{0,025,9,14} \simeq 3{,}21$

Per a calcular els valors crítics utilitzarem la taula F o mitjançant un programari estadístic.

Com que $F_{0,0,25,9,14} < F^* < F_{0,975,9,14}$ acceptem la hipòtesi nul·la i concloem que les variàncies són iguales. Per tant, tots dos processadors tenen la mateixa variabilitat en l'espai que deixen entre paraules. Si volguéssim fer el contrast amb el p-valor, aquest valdria: $p = 2 \cdot p(F_{9,14} < 0{,}51) \approx 0{,}312$. Com que és molt més gran que 0,05, acceptem la hipòtesi nul·la i arribem a la mateixa conclusió.

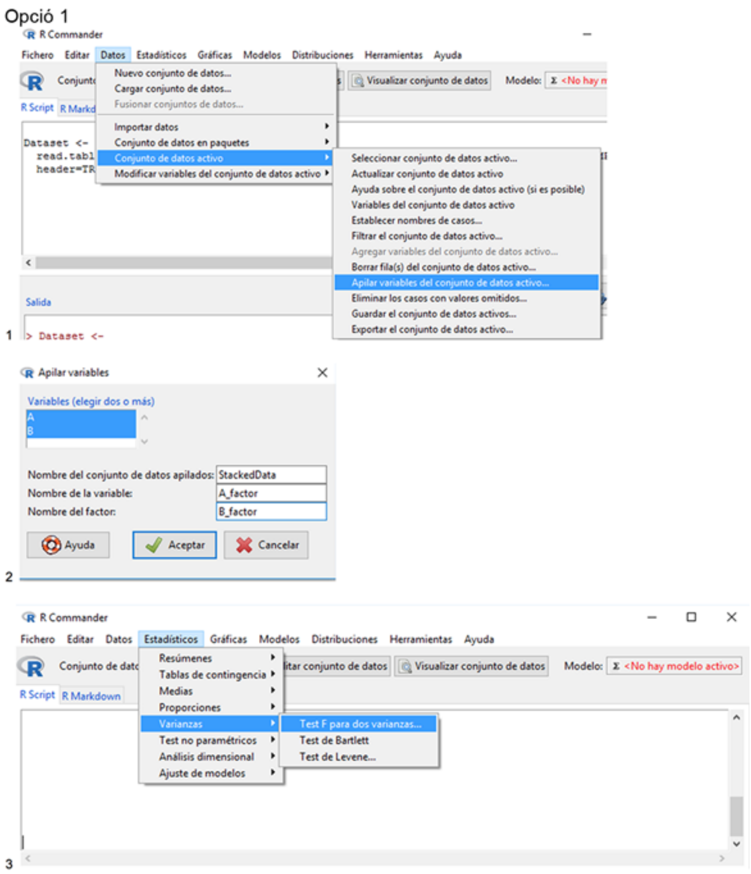
En el exemple 1. "Comparació de les mitjanes del temps de resposta de dos servidors", quan fem el contrast de diferència de mitjanes amb R Commander **no** assumim que les variàncies fossin iguals, ara farem un contrast per a comparar les dues variàncies i veure si són iguals.

Les hipòtesis nul·la i alternativa són:
$$H_0 : \sigma_A^2 = \sigma_B^2$$
$$H_1 : \sigma_A^2 \neq \sigma_B^2$$

Fixem $\alpha = 0{,}1$. L'estadístic de contrast és: $F^* = \frac{s_A^2}{s_B^2}$, on s_A^2 i s_B^2 són respectivament, les variàncies dels temps d'espera dels servidors A i B. La distribució de F és la de la F de Snedecor amb $50 - 1 = 49$ graus de llibertat en el numerador i $50 - 1 = 49$ graus de llibertat en el denominador.

El problema es resoldrà amb el R Commander. Els resultats del R Commander es mostren en la figura 20. Per comparar dues variàncies en R Commander tenim dues opcions. A continuació es mostren.

Figura 20. Passos a seguir en el contrast de variàncies. R Commander

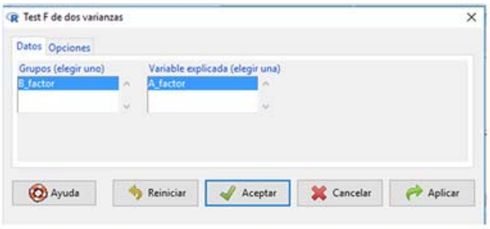


Passos a seguir

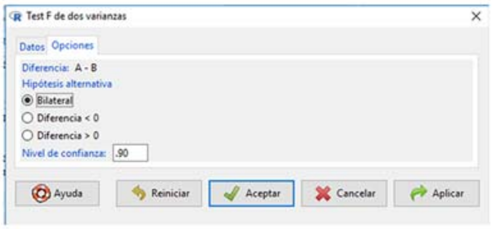
Primer s'apilen les dues variables (1) i (2) i a continuació es mostren les dues opcions disponibles:

Opció 1. Se segueix la ruta *Estadísticos > Varianzas > Test F para dos varianzas* **(3)** i s'emplenen els camps a la finestra corresponent **(4)** i **(5)**. A **(6)** es mostren els resultats del contrast.

Opció 2. Se segueix la ruta *Estadísticos > Varianzas > Test de Levene* **(6)** i s'emplenen els camps a la finestra corresponent **(7)**.

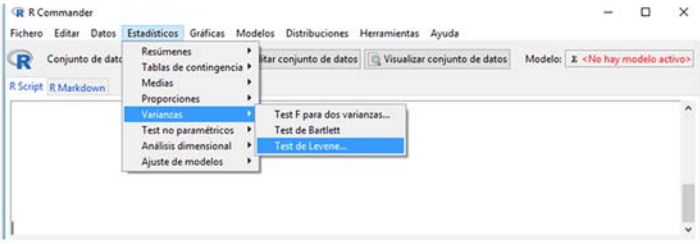


4

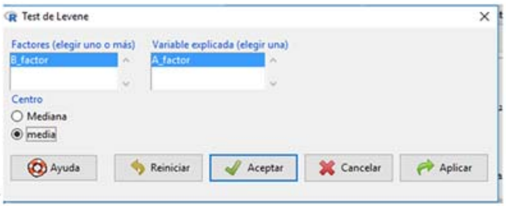


5

Opció 2



6



7

A continuació, es mostren els resultats per al contrast d'igualtat de variàncies:

Figura 21. Resultats del contrast d'igualtat de variàncies. R Commander

Opció 1

```
        F test to compare two variances

data:  A_factor by B_factor
F = 0.26543, num df = 49, denom df = 49, p-value = 7.999e-06
alternative hypothesis: true ratio of variances is not equal to 1
90 percent confidence interval:
 0.1651428 0.4266265
sample estimates:
ratio of variances
         0.2654323
```

Opció 2

```
Levene's Test for Homogeneity of Variance (center = "median")
      Df F value    Pr(>F)
group  1  13.142 0.0004603 ***
      98
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Podem veure que com que el *p*-valor és pràcticament zero, hem de rebutjar la hipòtesi nul·la, és a dir, no podem considerar que les variàncies siguin iguals. Per aquesta raó quan vam fer el contrast de diferència de mitjanes no vam assumir que les variàncies fossin iguals.

2. Comparació de grups mitjançant ANOVA

En l'apartat anterior s'han presentat alguns dels contrastos d'hipòtesi que s'usen habitualment per a determinar si hi ha diferències significatives entre dues poblacions o grups d'individus. Tanmateix, sovint es voldran comparar més de dues poblacions o grups entre ells, per a la qual cosa es faran servir les tècniques d'***analysis of variance*** **(ANOVA)** que s'introdueixen en aquest apartat.

Nota

L'acrònim **ANOVA** prové del terme *analysis of variance* (anàlisi de la variació existent entre les diferents mitjanes considerades, per a veure si hi ha diferències significatives entre aquestes).

Així, per exemple, les tècniques ANOVA es podrien aplicar per a donar respostes a preguntes com ara:

- Hi ha diferències significatives entre la durada mitjana dels judicis segons el tipus de delicte comès (homicidi, abús sexual, robatori, pirateria, frau fiscal, etc.)?

- Hi ha diferències significatives entre la despesa anual mitjana en tecnologia segons la franja d'edat a la qual pertanyi l'individu (nen, jove, adult, gent gran)?

- Hi ha diferències significatives entre el nombre mitjà d'alumnes i ordinadors per centre escolar entre els diferents països de l'eurozona?

- Hi ha diferències significatives entre el nombre mitjà d'autocitacions en revistes científiques segons l'editorial (Elsevier, Inderscience, Taylor & Francis, IGI Global, etc.)?

- Hi ha diferències significatives entre el consum mitjà de combustible segons el model de cotxe utilitzat (esportiu, turisme, tot terreny, monovolum, etc.)?

- Hi ha diferències significatives entre la qualitat mitjana (mesurada a partir d'uns paràmetres definits) dels resultats de cerques en línia segons el tipus de motor utilitzat (Google, Microsoft Bing, Yahoo!, etc.)? (figura 22)

Observeu

Els exemples que es presenten en aquest capítol es caracteritzen perquè la pertinença a una població o a una altra depèn d'un únic factor (tipus de delicte, franja d'edat, país, editorial, model de cotxe, motor de cerca, etc.). En aquests casos, es fa servir ANOVA d'un únic factor (en anglès *one-way ANOVA* o *single-factor ANOVA*). Tot i això, hi ha també tècniques ANOVA per al cas que els grups estiguin determinats per dos factors (per exemple, tipus de delicte i solvència econòmica de l'acusat, franja d'edat i classe social, etc.).

Figura 22. ANOVA permet comparar la qualitat mitjana de diferents serveis

2.1. Comparacions de diverses mitjanes

Quan es volen comparar entre elles les mitjanes corresponents a més de dues poblacions o grups d'individus, es podria pensar a comparar aquestes mitjanes dos a dos mitjançant un contrast d'hipòtesi per a dues poblacions. Així, per exemple, en el cas de tres poblacions es podria pensar a dur a terme una sèrie de tests *t* d'hipòtesi per a comparar les diferents mitjanes entre elles: un primer test *t* per a comparar les mitjanes de les poblacions 1 i 2, un altre per a comparar les mitjanes de les poblacions 1 i 3, i un altre per a comparar les mitjanes de les poblacions 2 i 3. Tot i això, aquesta aproximació té un greu problema: si per a cada test *t* es fa servir un nivell de significació α (generalment es fa servir $\alpha = 0{,}05$), aleshores la probabilitat de cometre un **error de tipus I** és α en cada test; en aquestes condicions, es pot comprovar que la probabilitat de cometre un error de tipus I en el global dels tres tests seria de $1-(1-\alpha)^3$ (si $\alpha = 0{,}05$ aquesta probabilitat seria de 0,14 aproximadament). En altres paraules, comparant les mitjanes dos a dos s'està fent un test global amb una probabilitat d'error de tipus I molt més gran que la que es preveu inicialment per a cada test individual. Per a evitar aquest problema es poden usar les tècniques ANOVA que permeten fer un únic test global amb una probabilitat d'error de tipus I determinada (generalment $\alpha = 0{,}05$).

Recordeu

Un **error de tipus I** consisteix a rebutjar la hipòtesi nul·la quan resulta que aquesta és certa. En aquest cas, la hipòtesi nul·la seria que les mitjanes són coincidents.

El test *F* d'ANOVA

Per a comparar les mitjanes corresponents a *k* poblacions o grups d'individus diferents ($k \geq 3$), es pot plantejar el contrast d'hipòtesi següent, en què el símbol μ_i representa la mitjana de la població *i*-èsima per a $i = 1,2,\ldots,k$:

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \ldots = \mu_k \text{ (totes les mitjanes són iguals)} \\ H_a : \text{no totes les mitjanes són iguals} \end{cases} \tag{1}$$

En altres paraules, la hipòtesi nul·la, H_0, sosté que no hi ha diferències significatives entre les diferents mitjanes poblacionals, mentre que la hipòtesi alternativa, H_a, sosté tot el contrari, com ara que les mitjanes sí que són significativament diferents. És important observar aquí que la hipòtesi nul·la no diu que totes les mitjanes siguin significativament diferents entre elles, sinó simplement que no totes les mitjanes són iguals, tot i que n'hi pot haver algunes que sí que ho siguin (per exemple, podria passar que $\mu_1 \neq \mu_2$ i $\mu_2 \neq \mu_3$ però essent $\mu_1 = \mu_3$). Per tant, si es conclogués que no totes les mitjanes són iguals caldria dur a terme una anàlisi posterior per a determinar quines mitjanes són diferents entre elles.

Programari estadístic

Actualment hi ha una gran varietat de **programes estadístics** o d'anàlisi de dades de gran qualitat, tant comercials (Minitab, SPSS, MS Excel, SAS, S-Plus, etc.) com de codi obert (R, Calc de Open Office, etc.).

El contrast d'hipòtesi (1) s'anomena test *F* de ANOVA, i generalment es recorre a l'ús d'algun **programari estadístic** per a resoldre'l, és a dir, per a obtenir el *p*-valor associat al test. A partir d'aquest *p*-valor correspon a l'investigador determinar si ha estat possible trobar prou evidències per a rebutjar la hipòtesi nul·la o si, per contra, sembla que les dades empíriques no estan en contradicció amb la hipòtesi nul·la i, per tant, s'accepta aquesta com a vàlida. Com en qualsevol altre tipus de contrast estadístic, abans de resoldre el test se sol fixar un valor de significació, α (generalment α = 0,05 o bé α = 0,01). Un cop obtingut el *p*-valor, si $p - valor < \alpha$ es rebutja la hipòtesi nul·la; en cas contrari no hi ha prou indicis per a fer-ho i, per tant, s'acceptarà la hipòtesi nul·la com a vàlida. L'elecció del valor concret per a α dependrà del nivell de confiança, $1 - \alpha$, que es vulgui que tingui la decisió final sobre l'acceptació o no de la hipòtesi nul·la. Així, per exemple, un α = 0,05 implicarà un nivell de confiança en la decisió final del 95%, mentre que un α = 0,01 implicarà un nivell de confiança en la decisió final del 99%. El problema de seleccionar nivells de confiança excessivament elevats (superiors al 99% o, el que és el mateix, valors de α inferiors a 0,01) és que aleshores el contrast d'hipòtesi es torna excessivament "conservador", de manera que només quan les evidències empíriques en contra de la hipòtesi nul·la són totalment aclaparadores (és a dir, només quan les diferències entre algunes de les mitjanes són desproporcionades) és possible obtenir un *p*-valor més petit o igual que α. Per aquest motiu, en la majoria dels casos pràctics se sol fer servir el valor α = 0,05 o bé α = 0,01.

Exemple d'aplicació d'ANOVA: comparant el nombre mitjà d'accessos a continguts en línia segons la posició de l'enllaç en el portal

En un portal web d'accés a publicacions en línia, se sospita que la posició que ocupa l'enllaç a una determinada base de dades afecta el nombre de consultes diàries que rep. Per a comprovar-ho, s'han seleccionat a l'atzar un total de 13 dies laborables d'un mes i, per a cadascun, s'ha comptabilitzat el nombre d'accessos rebuts. La taula 4 mostra els valors obtinguts, els quals han estat agrupats segons la posició diària de l'enllaç (a l'encapçalament de la pàgina, al marge dret o al marge esquerre).

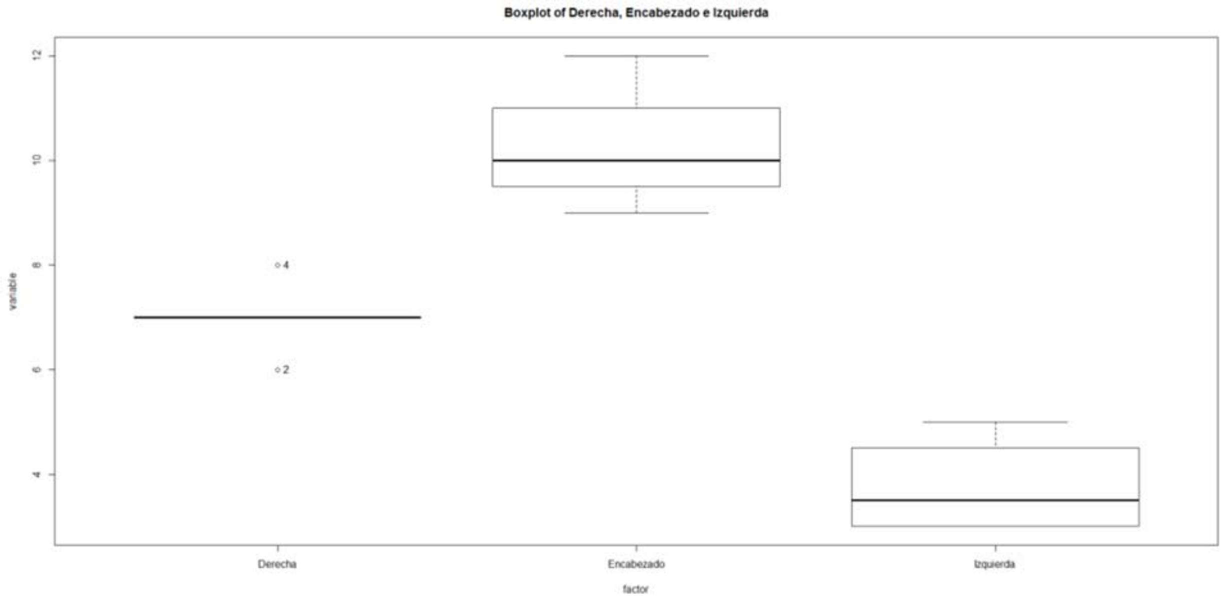
Taula 4. Accessos a una base de dades segons la posició de l'enllaç

	Posició de l'enllaç		
	Encapçalament (1)	**Dreta (2)**	**Esquerra (3)**
	10	7	3
	12	6	3
	10	7	5
	9	8	4
		7	
Total	41	45	15
Mitjana	$\bar{x}_1 = 10{,}25$	$\bar{x}_2 = 7{,}0$	$\bar{x}_3 = 3{,}75$

Es pot afirmar que hi ha diferències significatives entre les diferents mitjanes? És a dir, depèn del nombre mitjà de consultes diàries de la posició que ocupi l'enllaç?

Com a primera aproximació a aquest problema, es pot optar per generar un diagrama de caixa i bigotis (*boxplot*) per a cada un dels grups de dades. En la figura 23 s'aprecien clares diferències en els tres grups considerats, tant pel que fa als *boxplots* com a les respectives mitjanes. Cal destacar que R Commander representa la mediana i no la mitjana als *box plots*. Aquesta és la línia horitzontal dintre de les caixes.

Figura 23. *Boxplot* del nombre de consultes per a cada posició



Tot i això, per a contestar de manera contundent les preguntes anteriors cal fer un test F d'ANOVA. El contrast d'hipòtesi es pot formular de la manera següent:

$$\begin{cases} H_0 : \bar{x}_1 = \bar{x}_2 = \bar{x}_3 \\ H_a : \text{no totes les mitjanes són iguals} \end{cases}$$

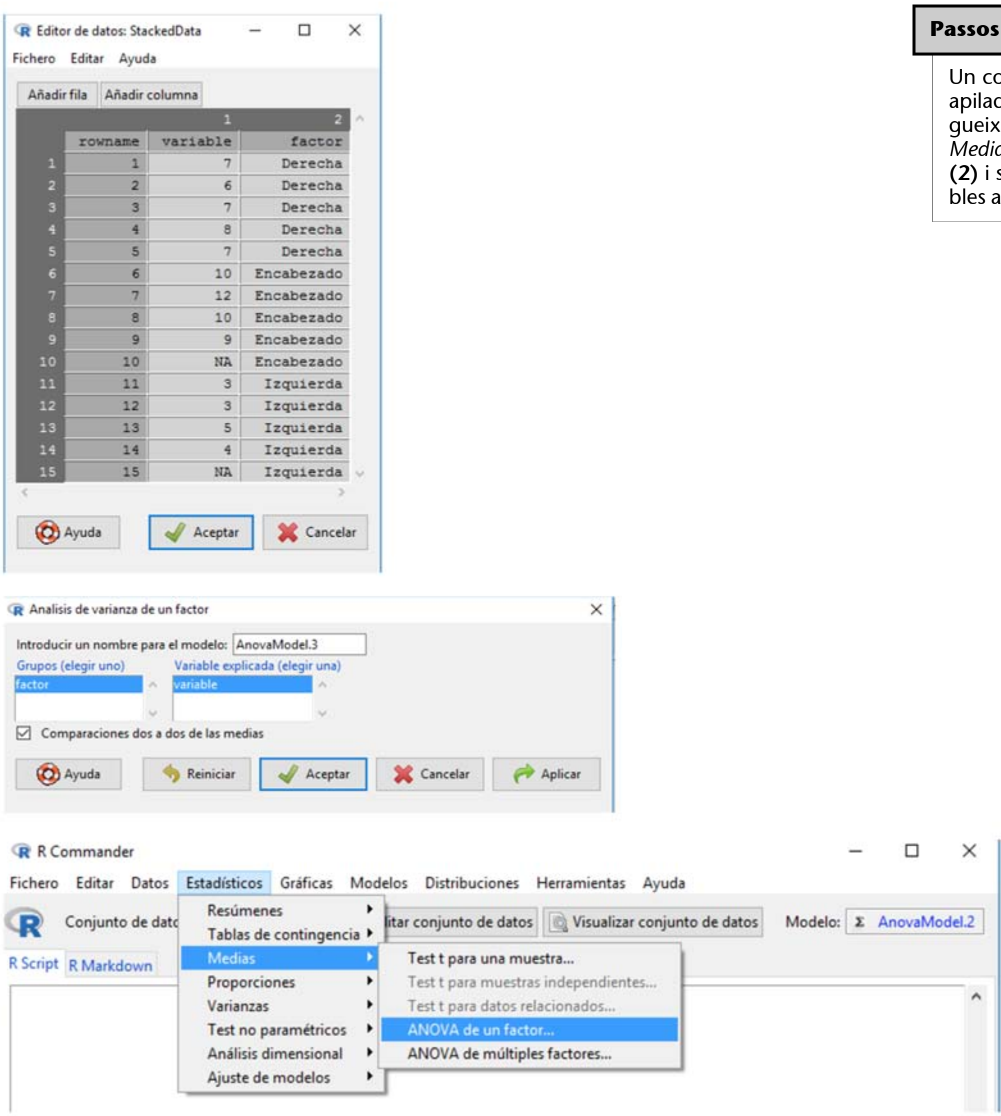
Per a resoldre aquest contrast, es fixarà un valor de significació $\alpha = 0{,}05$ i es recorrerà a l'ús de programari estadístic per a obtenir el p-valor corresponent a les observacions de la taula 4.

La figura 24 mostra els passos bàsics necessaris per a generar una anàlisi ANOVA amb el **programa R Commander**. D'altra banda, la figura 25 mostra la sortida generada per a les dades d'aquest exemple. S'observa que el valor resultant per a l'estadístic del contrast és $F = 44{,}47$. L'estadístic F és una variable aleatòria que es comporta segons una distribució F de Snedecor amb 2 graus de llibertat en el numerador (*Df factor*) i 10 graus de llibertat en el denominador (*Df Residuals*). El p-valor no és res més que la probabilitat que una variable aleatòria amb aquestes característiques superi el valor observat per a l'estadístic de contrast, com ara p-valor = $P(F_{2,10} > 44{,}47)$. Segons s'observa a la sortida, en aquest cas s'obté p-valor = 0,000. Atès que el p-valor és molt menor que el nivell de significació escollit (p-valor = $0{,}000 < 0{,}05 = \alpha$), es conclou que les dades obtingudes semblen contradir la hipòtesi nul·la i, per tant, cal rebutjar-la.

Així, doncs, hi ha indicis clars per a pensar que no totes les mitjanes són iguals, com ara que el nombre mitjà de consultes diàries sí que depèn de la posició que ocupi l'enllaç.

Figura 24. Passos a seguir per a dur a terme una anàlisi ANOVA en el R Commander

1
2
3



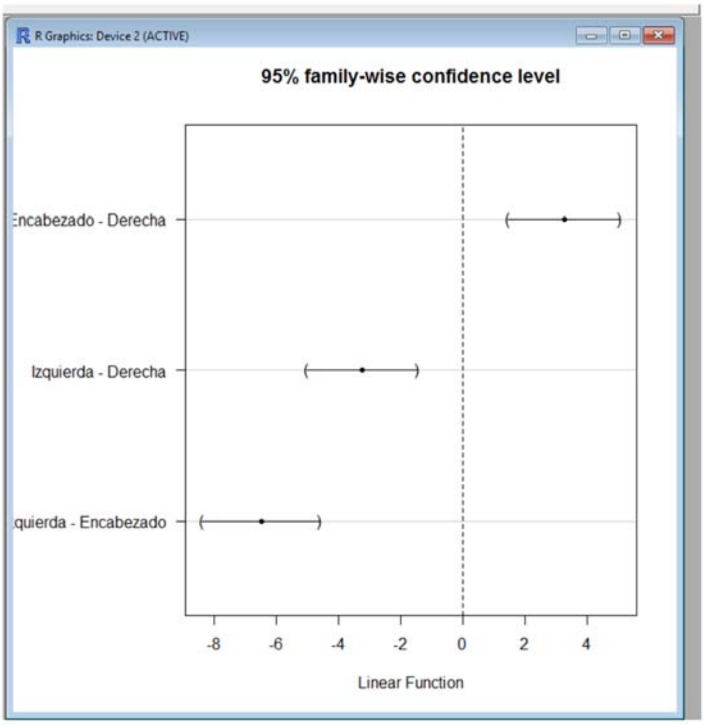
Passos a seguir

Un cop introduïdes les dades apilades al programa **(1)**, se segueix la ruta *Estadísticos* > *Medias* > *ANOVA de un factor* **(2)** i se seleccionen les variables a la finestra d'ANOVA **(3)**.

Figura 25. Sortida ANOVA del R Commander per a la comparació de posicions

```
> summary(AnovaModel.1)
            Df Sum Sq Mean Sq F value   Pr(>F)
factor       2   84.5   42.25   44.47 1.05e-05 ***
Residuals   10    9.5    0.95
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
2 observations deleted due to missingness

> with(StackedData, numSummary(variable, groups=factor, statistics=c("mean", "sd")))
            mean        sd data:n data:NA
Derecha     7.00 0.7071068      5       0
Encabezado 10.25 1.2583057      4       1
Izquierda   3.75 0.9574271      4       1
```



A la segona part de la sortida R Commander es representa cada una de les mitjanes comparades dos a dos amb el seu interval de confiança respectiu per a un nivell de confiança del 95%. Les diferències entre mitjanes en les que l'interval de confiança engloba els límits inferior i superior no contenen el valor 0, són estadísticament significatives amb el mètode de Tukey. Es pot apreciar que per a un nivell de confiança del 95%, cap interval de confiança de diferència entre grups toca 0, per tant, totes les diferències entre mitjanes són estadísticament significatives.

En general, tanmateix, el fet que totes les mitjanes no siguin iguals no implicarà necessàriament que totes siguin diferents (és a dir, hi podria haver intervals que s'encavalquessin i d'altres que no).

La figura 26 mostra la corresponent sortida ANOVA que ofereix el **Microsoft Excel**. S'observa el mateix valor per a l'estadístic $F = 44{,}47$, i també un p-valor = 1,0543E-05 (és a dir, p-valor = 0,00001543 o, arrodonint, p-valor = 0,000).

Nota

Per a poder fer ANOVA amb l'**MS Excel** cal instal·lar prèviament un complement anomenat "Anàlisi de dades". Fent servir el Google o qualsevol altre cercador és fàcil trobar informació detallada sobre el procés d'instal·lació. També hi ha un complement similar per a l'Open Office Calc.

Figura 26. Sortida ANOVA de l'Excel per a la comparació de posicions

	A	B	C	D	E	F	G
1	Anàlisi de variància d'un factor						
2							
3	RESUM						
4	Grups	Compte	Suma	Mitjana	Variància		
5	Encapçalament	4	41	10,25	1,58333333		
6	Dreta	5	35	7	0,5		
7	Esquerra	4	15	3,75	0,91666667		
8							
9							
10	ANÀLISI DE VARIÀNCIA						
11	Origen de les variacions	Suma dels quadrats	Graus de llibertat	Mitjana dels quadrats	*F*	Probabilitat	Valor crític per a *F*
12	Entre grups	84,5	2	42,25	44,4736842	1,05434E-05	4,102821015
13	Dins dels grups	9,5	10	0,95			
14							
15	Total	94	12				
16							

Exemple d'aplicació d'ANOVA: comparant mitjanes de resultats vàlids oferts per un motor de cerca segons l'algoritme utilitzat

Els desenvolupadors d'un nou motor de cerca especialitzat en recursos de recerca estan provant tres algoritmes diferents de recuperació de la informació. Per a comprovar si la mitjana de resultats vàlids que proporciona cada algoritme és el mateix en els tres casos, s'han fet unes proves aleatòries amb cadascun. La taula 5 mostra les observacions que s'han obtingut després de fer les proves.

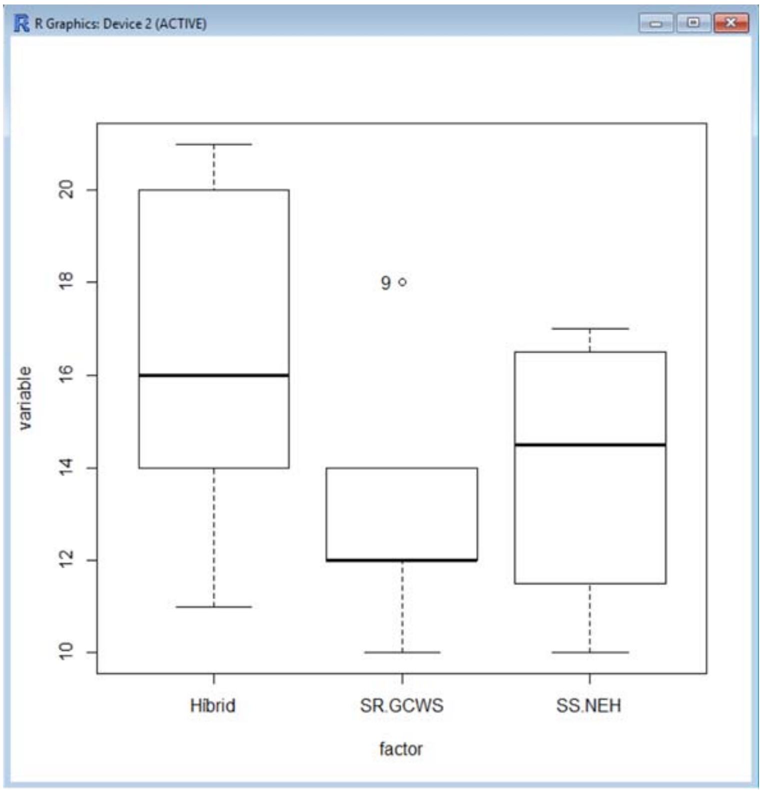
Taula 5. Resultats vàlids obtinguts amb cada algoritme

	Algoritme		
	SR-GCWS	**SS-NEH**	**Híbrid**
	12	10	16
	10	17	14
	18	16	16
	12	13	11
	14		20
			21
Total	66	56	98
Mitjana	$\overline{x}_1 = 13{,}20$	$\overline{x}_2 = 14{,}00$	$\overline{x}_3 = 16{,}33$

Es pot afirmar que hi ha diferències significatives entre les diferents mitjanes? És a dir, la mitjana depèn de resultats vàlids obtinguts de l'algoritme que implementi el motor de cerca?

Novament, per a respondre adequadament aquestes preguntes cal fer un test *F* d'ANOVA. Com a pas previ, tanmateix, podem fer un gràfic dels corresponents *boxplots*. Com s'observa en la figura 27, en aquest cas les diferències entre els diferents grups no semblen excessives, tot i que l'algoritme híbrid sembla haver proporcionat resultats lleugerament superiors a la resta.

Figura 27. *Boxplot* del nombre de resultats vàlids per a cada algoritme



Per a comprovar si les diferències entre les mitjanes són estadísticament significatives o no ho són, es formula el contrast ANOVA següent:

$$\begin{cases} H_0 : \bar{x}_1 = \bar{x}_2 = \bar{x}_3 \\ H_a : \text{no totes les mitjanes són iguals} \end{cases}$$

De nou es farà servir un nivell de significació $\alpha = 0{,}05$ (és a dir, el nivell de confiança usat és del 95%). Les figures 28 i 29 mostren, respectivament, les sortides R Commander i Excel per a aquest exemple.

Figura 28. Sortida ANOVA del R Commander per a la comparació d'algoritmes

```
> summary(AnovaModel.1)
            Df Sum Sq Mean Sq F value Pr(>F)
factor       2   29.2   14.60   1.287  0.312
Residuals   12  136.1   11.34
3 observations deleted due to missingness

> with(StackedData, numSummary(variable, groups=factor, statistics=c("mean", "sd")))
            mean        sd data:n data:NA
Hibrid  16.33333 3.723797      6       0
SR.GCWS 13.20000 3.033150      5       1
SS.NEH  14.00000 3.162278      4       2
```

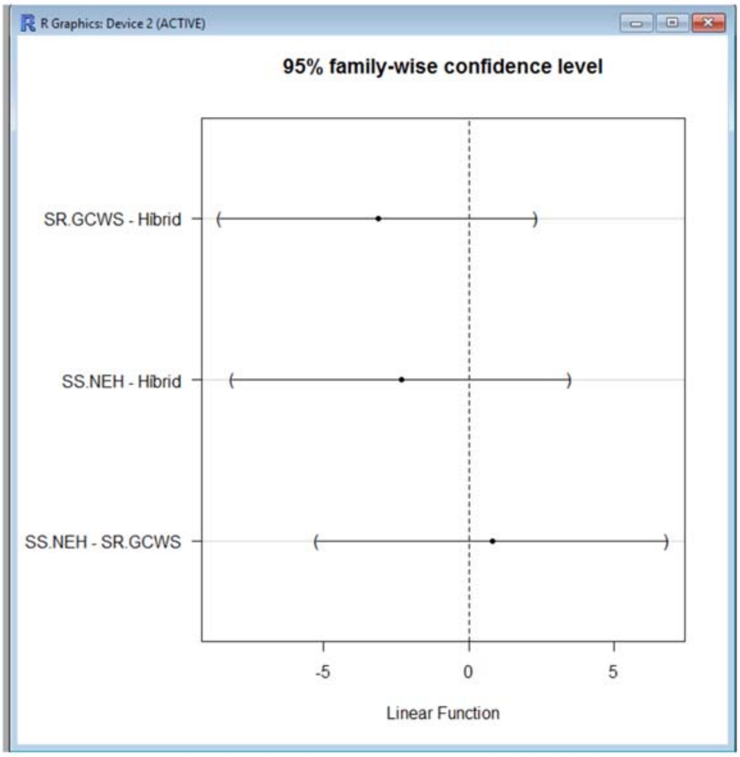


Figura 29. Sortida ANOVA de l'Excel per a la comparació d'algoritmes

	A	B	C	D	E	F	G
1	Anàlisi de variància d'un factor						
2							
3	RESUM						
4	Grups	Compte	Suma	Mitjana	Variància		
5	SR-GCWS	5	66	13.2	9.2		
6	SS-NEH	4	56	14	10		
7	Híbrid	6	98	16,33333333	13,8666667		
8							
9							
10	ANÀLISI DE VARIÀNCIA						
11	Origen de les variacions	Suma dels quadrats	Graus de llibertat	Mitjana dels quadrats	*F*	Probabilitat	Valor crític per a *F*
12	Entre grups	29.2	2	14,6	1,28697356	0,311619296	3,885293835
13	Dins dels grups	136,1333333	12	11,34444444			
14							
15	Total	165,3333333	14				
16							

En totes dues sortides s'observa un valor de l'estadístic $F = 1,29$. En aquesta ocasió, l'estadístic és una variable aleatòria que es distribueix segons una *F* Snedecor amb 2 graus de llibertat en el numerador (*Df factor*) i 12 en el denominador (*Df Residuals*). La probabilitat que una variable com aquesta assoleixi o superi el valor 1,29 obtingut per l'estadístic és de 0,312, i aquest és precisament el *p*-valor que s'observa en totes dues sortides. Atès que *p*-valor = 0,312 > α = 0,05, no sembla que hi hagi prou indicis per a rebutjar la hipòtesi nul·la. Dit d'una altra manera, les dades observades semblen estar en sintonia amb la hipòtesi nul·la, per la qual cosa acceptarem la hipòtesi que les mitjanes de resultats vàlids són equivalents per als tres algoritmes, sense que hi hagi diferències estadísticament significatives entre ells.

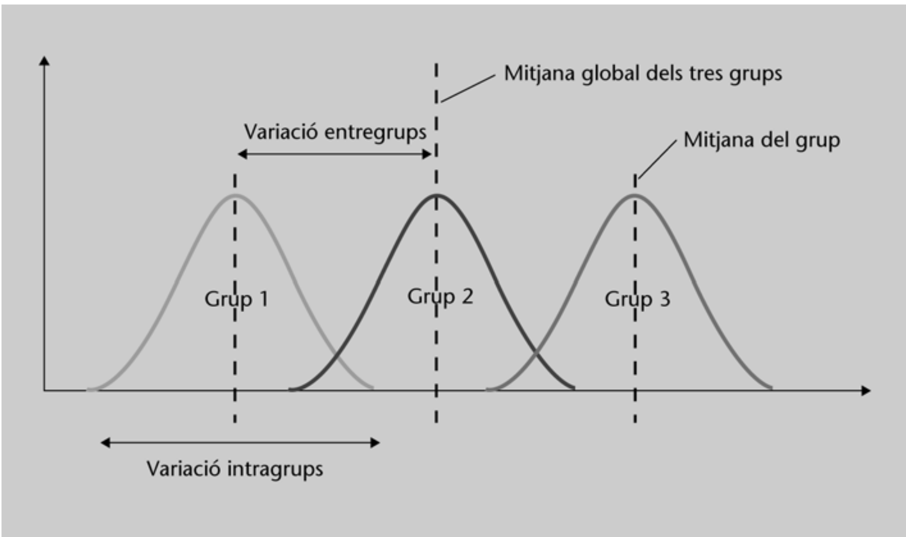
De fet, en la segona part de la sortida R Commander s'observa que tots tres intervals de confiança de diferències entre grups toquen 0, per tant, per a un nivell de confiança del 95% no es pot afirmar que hi hagi diferències significatives entre aquestes mitjanes.

2.2. La lògica del contrast ANOVA

Quan mitjançant un experiment aleatori es recull una sèrie de dades (observacions) que són classificades en grups o nivells diferents segons un factor determinat (franja d'edat, classe social, etc.), es poden analitzar dos tipus diferents de variància en les observacions (figura 30):

- D'una banda, la variació existent entre els diferents grups o nivells (com ara, la variació entre les respectives mitjanes* de cada grup), la qual es coneix com a *variació entregrups* o *Mean factor*.
- De l'altra, la variació existent dins de cada grup o nivell, la qual es coneix com a variació *intragrups* o *Mean Residuals*.

Figura 30. Variació entregrups i variació intragrups



En el fons, el que fa el test ANOVA és comparar totes dues mitjanes de variabilitat, la variació entregrups (*Mean factor*) i la variació intragrups (*Mean Residuals*). Si passa que el *Mean factor* és significativament més gran que el *Mean Residuals* (figura 31), aleshores el test conclourà que les mitjanes dels diferents grups no són iguals en tots els casos (la qual cosa implica que no totes les dades pertanyen a un mateix grup o, dit d'una altra manera, que el valor de les observacions sí que depèn del factor considerat). Si, per contra, el *Mean factor* no és significativament més gran que el *Mean Residuals* (figura

32), aleshores el test conclourà que no s'aprecien diferències significatives entre les mitjanes dels diferents grups (en altres paraules, que totes les observacions semblen procedir d'un únic grup o, dit d'una altra manera, que les observacions no semblen dependre del factor considerat).

Figura 31. La variació entregrups és més gran que la intragrups

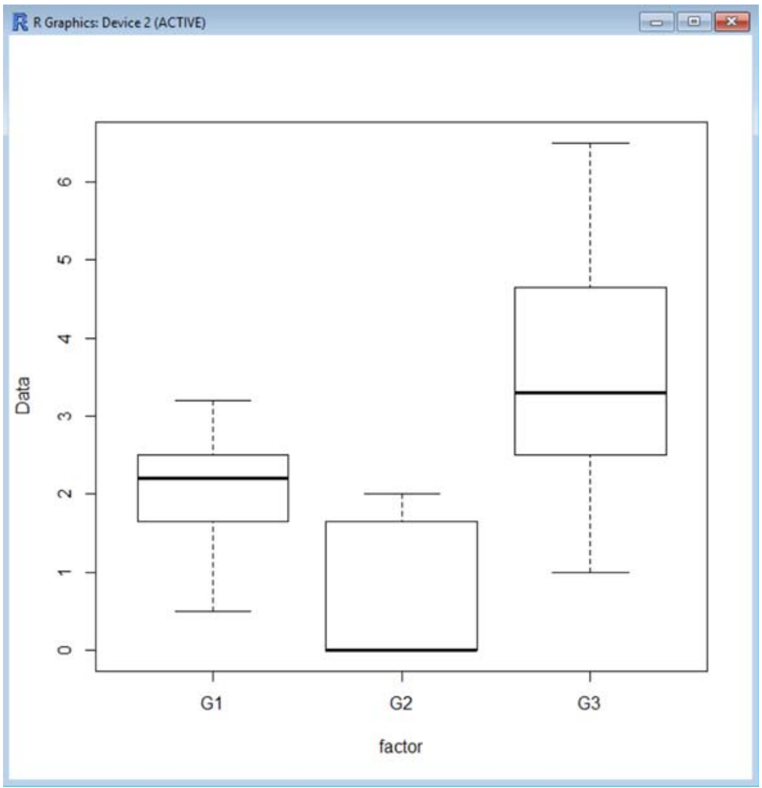
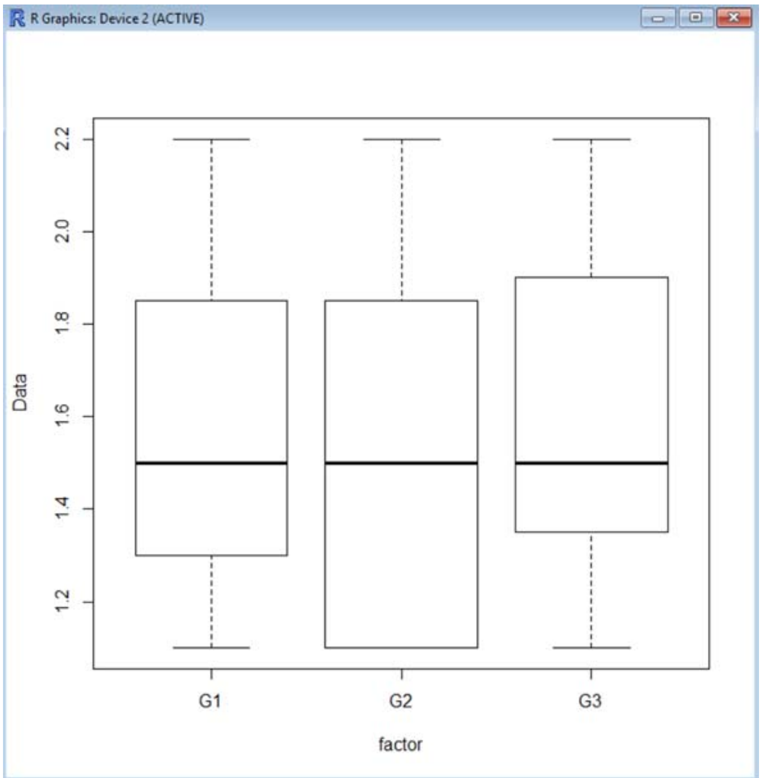


Figura 32. La variació entregrups és més petita que la intragrups



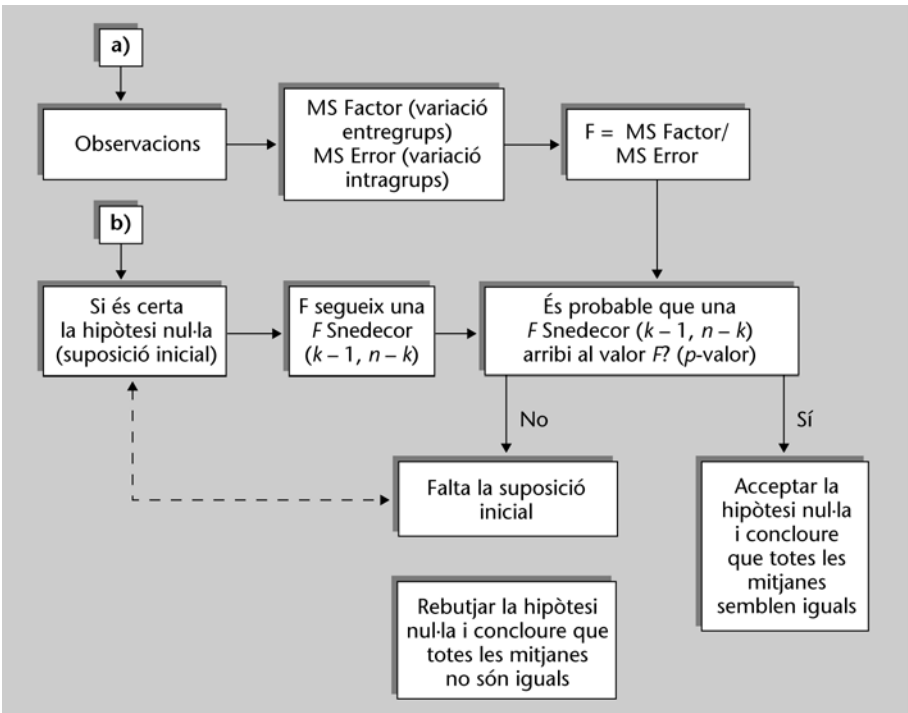
En la figura 25 (sortida ANOVA del R Commander) s'observen els valors *Mean factor* = 42,250 i *Mean Residuals* = 0,950. És a dir, en aquest cas la variació entregrups (MS Factor) és molt més gran que la variació intragrups (MS Error), i això deixa entreveure que, probablement, el test conclogui que no totes les mitjanes són iguals. Però, com arriba el test a la conclusió final? La figura 33 ajuda a entendre millor com funciona el test *F* d'ANOVA:

a) D'una banda, a partir dels valors obtinguts per a MS Factor i MS Error es calcula l'estadístic de contrast $F = \frac{Mean\ factor}{Mean\ residuals}$.
En aquest cas, $F = 44{,}47$.

b) De l'altra, se sap que si la hipòtesi nul·la fos certa (com ara quan totes les mitjanes són iguals), aquest estadístic *F* seria una variable aleatòria que seguiria una distribució *F* de Snedecor amb $k - 1$ graus de llibertat en el numerador (*Df factor*), i $n - k$ graus de llibertat en el denominador (*Df Residuals*), essent k el nombre de grups o nivells i n el nombre total d'observacions.

En l'exemple de la figura 25, *Df factor* = 2 i *Df Residuals* = 10. Ara bé, quina és la probabilitat que una variable aleatòria *F* de Snedecor (2, 10) assoleixi un valor com el que s'obté per l'estadístic de contrast *F*? Dit d'una altra manera, és raonable pensar que una *F* de Snedecor (2,10) hagi assolit un valor de 44,47? La probabilitat que això passi ens la proporciona el *p*-valor. D'aquesta manera, un *p*-valor "petit" (inferior al nivell de significació α) es pot interpretar com una probabilitat massa baixa que una *F* de Snedecor (2, 10) pugui donar el valor obtingut per a *F*, la qual cosa posa en dubte la suposició inicial que la hipòtesi nul·la era certa. D'altra banda, un *p*-valor "gran" (superior al nivell de significació α) es pot interpretar com una probabilitat acceptable que, en efecte, una *F* de Snedecor (2, 10) prengui aquest valor i, per tant, no hi hauria evidències per a dubtar de la hipòtesi nul·la.

Figura 33. Funcionament intern del test *F* d'ANOVA



Observeu

que quan el valor obtingut per a l'estadístic *F* a partir de les observacions no és coherent amb el que caldria esperar d'una *F* Snedecor ($k - 1$, $n - k$), aleshores el que està fallant és la suposició inicial que la hipòtesi nul·la és certa.

2.3. Les hipòtesis del model ANOVA

Com qualsevol altra tècnica d'inferència estadística, el contrast ANOVA es pot fer servir amb garanties per a comparar poblacions o grups, només si es compleixen unes determinades condicions d'entorn o supòsits bàsics:

1) Les observacions són independents entre elles i constitueixen, per a cada població o grup, una mostra aleatòria.

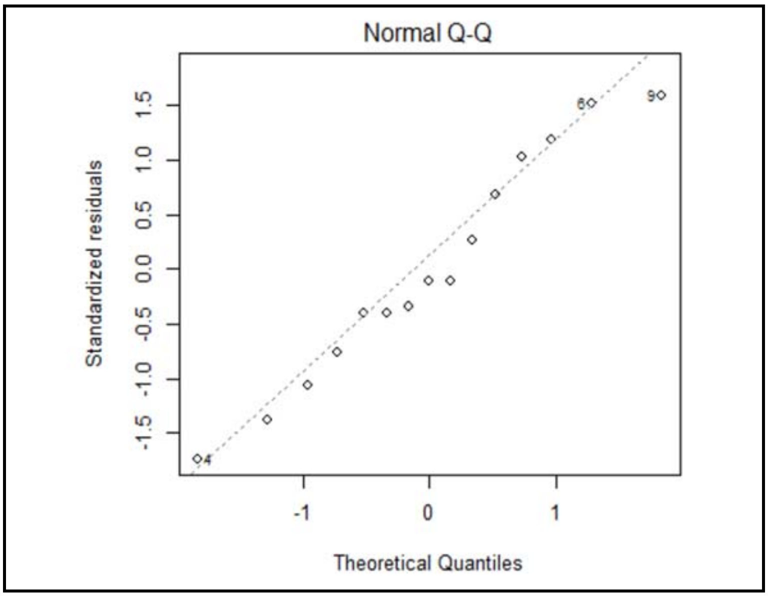
2) Les observacions de cada població o grup segueixen una distribució aproximadament normal.

3) Les observacions de cada població o grup tenen una variància σ^2 que és aproximadament la mateixa per a tots els grups.

El primer supòsit garanteix que les mostres són aleatòries i independents, la qual cosa és un requisit comú en les tècniques d'inferència estadística. Si les mostres no fossin aleatòries o les observacions no fossin independents, la informació que es generaria estaria esbiaixada i, per tant, no seria vàlida. És funció de l'investigador garantir, durant la fase de disseny de l'experiment i recollida de dades posterior, que es compleix aquest supòsit.

Pel que fa al supòsit segon (normalitat de les dades), aquest se sol comprovar mitjançant la realització d'un gràfic de normalitat per al conjunt de les dades. La figura 34 mostra un gràfic per a l'exemple anterior dels algoritmes. Sempre que els punts (que representen les observacions) estiguin raonablement a prop de la línia recta (que representa la distribució normal) i no mostren un patró de comportament estrany, no hi ha motius per a sospitar que falla el supòsit de normalitat. Si s'observa algun patró de comportament anòmal (per exemple, molts punts excessivament allunyats de la línia o bé molts punts consecutius situats al mateix costat de la línia), aleshores el supòsit de normalitat queda en dubte. Per a l'exemple dels algoritmes, no s'observa en el gràfic res estrany i, per tant, es pot validar el supòsit de normalitat de les dades.

Figura 34. Gràfic de normalitat per a les dades de l'exemple d'algoritmes



Passos a seguir

Aquest tipus de gràfic es pot obtenir amb R Commander seguint la següent ruta *Modelos* > *Gráficas* > *Gráficas básicas de diagnostico*.

Finalment, pel que fa al supòsit de variància constant, aquest se sol comprovar o bé calculant les desviacions estàndard de les mostres per a verificar que no hi ha grans diferències entre elles (figura 35), o bé mitjançant un gràfic que permet comparar visualment la dispersió de les dades en cada grup (figura 36). En el cas de l'exemple dels algoritmes no s'observen diferències substancials entre les variàncies dels diferents grups, i això permet validar el supòsit de variància constant.

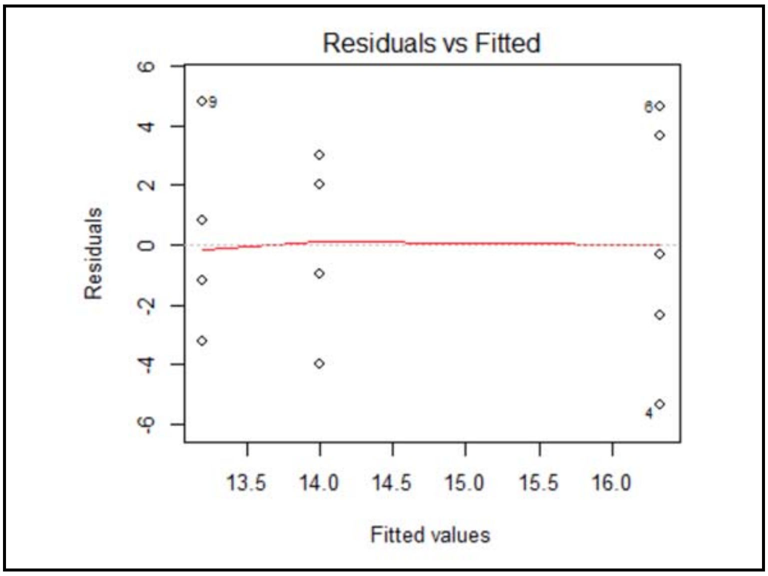
Figura 35. La columna sd permet estimar la variància de cada grup

	mean	sd	IQR	0%	25%	50%	75%	100%	n	NA
Hibrid	16.33333	3.723797	4.5	11	14.50	16.0	19.00	21	6	0
SR.GCWS	13.20000	3.033150	2.0	10	12.00	12.0	14.00	18	5	1
SS.NEH	14.00000	3.162278	4.0	10	12.25	14.5	16.25	17	4	2

Recordeu

La variància, σ^2 és el quadrat de la desviació estàndard o típica, σ. Generalment, el valor exacte de la variància poblacional, σ^2, serà desconegut, però aquest valor es pot estimar mitjançant la variància de la mostra, s^2.

Figura 36. El gràfic mostra la dispersió de cada grup



Passos a seguir

Aquest tipus de gràfic es pot obtenir amb R Commander seguint la següent ruta *Modelos* > *Gráficas* > *Gráficas básicas de diagnostico*.

Exemple d'aplicació d'ANOVA: comparant valoracions mitjanes en un qüestionari d'escala Likert segons el perfil dels enquestats

En una universitat s'ha implementat recentment un nou servei en línia que facilita l'accés a recursos didàctics complementaris. Es vol conèixer l'opinió dels estudiants sobre aquest nou servei i, en particular, si hi ha diferències significatives en la valoració mitjana del servei segons la titulació a la qual pertanyi l'estudiant. Per a fer-ho, un investigador ha seleccionat a l'atzar cinc estudiants de cada un dels principals estudis que s'ofereixen i els ha demanat que emplenin un qüestionari d'avaluació del servei. El qüestionari fa servir una escala Likert entre 1 (mínima valoració) i 7 (màxima valoració). Els resultats obtinguts es mostren en la taula 6.

Taula 6. Valoracions obtingudes segons el perfil de l'estudiant

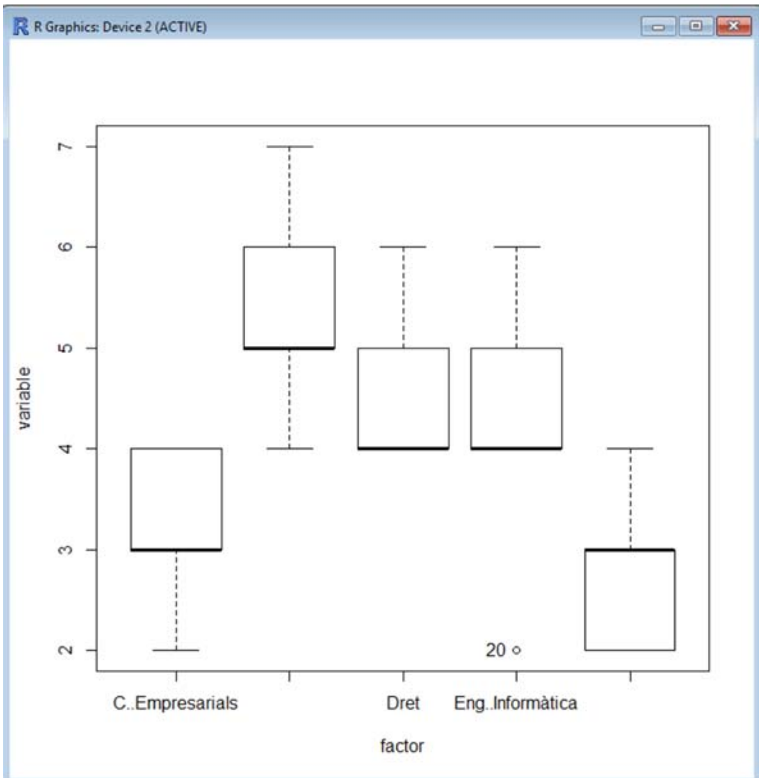
Estudis				
C. Informació	**C. Empresarials**	**Eng. Informàtica**	**Dret**	**Psicologia**
6	4	5	4	3
5	4	4	4	3
5	3	4	5	2
7	3	6	4	4
4	2	2	6	2

La figura 37 mostra la sortida R Commander corresponent als estadístics descriptius per a cada grup o nivell d'observacions. A simple vista sembla que s'aprecien diferències considerables entre la màxima valoració mitjana (C. Informació, amb 5,4) i la mínima (Psicologia, amb 2,8). El *boxplot* de la figura 38 també apunta la possibilitat que les valoracions mitjanes del servei puguin dependre del perfil de l'estudiant, no essent les mateixes per a totes les titulacions.

Figura 37. Estadístics descriptius de les valoracions per grup

```
 C..Informació C..Empresarials Eng..Informàtica      Dret       Psicologia
 Min.    :4.0  Min.    :2.0    Min.    :2.0     Min.    :4.0  Min.    :2.0
 1st Qu.:5.0   1st Qu.:3.0     1st Qu.:4.0      1st Qu.:4.0   1st Qu.:2.0
 Median :5.0   Median :3.0     Median :4.0      Median :4.0   Median :3.0
 Mean   :5.4   Mean   :3.2     Mean   :4.2      Mean   :4.6   Mean   :2.8
 3rd Qu.:6.0   3rd Qu.:4.0     3rd Qu.:5.0      3rd Qu.:5.0   3rd Qu.:3.0
 Max.   :7.0   Max.   :4.0     Max.   :6.0      Max.   :6.0   Max.   :4.0
```

Figura 38. *Boxplot* per a les valoracions del servei per titulació



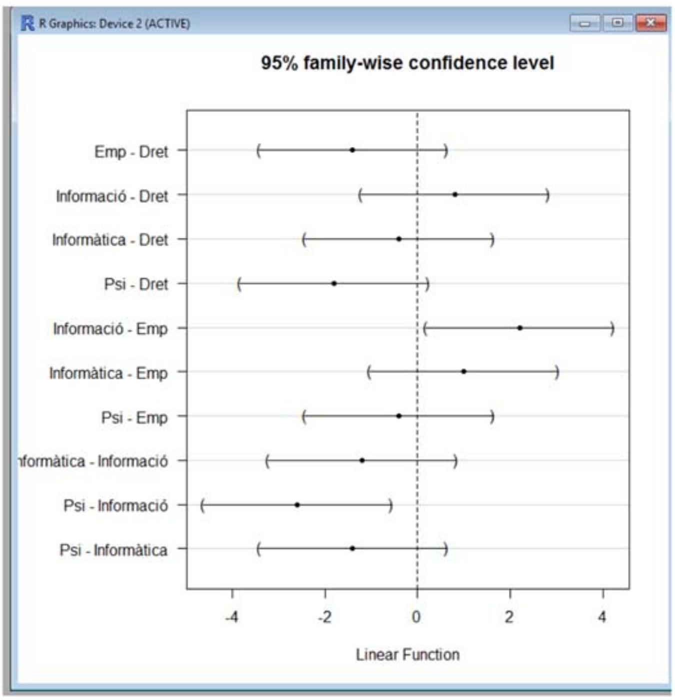
Per a poder corroborar o desmentir aquestes impressions visuals d'una manera més científica, s'opta per fer un test F d'ANOVA amb un nivell de significació $\alpha = 0{,}05$ (és a dir, en aquest cas s'opta per fer servir un nivell de confiança del 95%).

La figura 39 mostra la sortida ANOVA del R Commander, en el qual s'aprecia un *Mean factor* = 5,54 (variació entregrups), un *Mean Residuals* = 1,14 (variació intragrups) i un valor per a l'estadístic de contrast $F = 5{,}54/1{,}14 = 4{,}86$. En el supòsit que la hipòtesi nul·la fos certa, aquest estadístic seguiria una distribució F de Snedecor amb 4 graus de llibertat en el numerador (*Df factor*) i 20 graus de llibertat en el denominador (*Df Residuals*). La probabilitat que una variable aleatòria F de Snedecor (4, 20) prengui un valor igual o superior a 4,86 és 0,007 (p-valor). Aquesta probabilitat és extremadament baixa (més baixa que el valor de significació fixat), la qual cosa fa dubtar sobre el supòsit inicial que la hipòtesi nul·la era certa. En altres paraules: atès que p-valor < α cal rebutjar la hipòtesi nul·la. Així, doncs, segons les evidències empíriques trobades, es pot afirmar amb un 95% de confiança que les valoracions mitjanes dels grups no són totes iguals.

Figura 39. Sortida ANOVA del R Commander per a la comparació de valoracions

```
> summary(AnovaModel.9)
            Df Sum Sq Mean Sq F value  Pr(>F)
factor       4  22.16    5.54    4.86 0.00667 **
Residuals   20  22.80    1.14
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> with(StackedData, numSummary(variable, groups=factor, statistics=c("mean", "sd")))
                 mean        sd data:n
C..Empresarials   3.2 0.8366600      5
C..Informació     5.4 1.1401754      5
Dret              4.6 0.8944272      5
Eng..Informàtica  4.2 1.4832397      5
Psicologia        2.8 0.8366600      5
```



Atenció

Per a evitar confusions en la segona part de la sortida, és important fixar bé el nivell de confiança (1 − α) a la finestra ANOVA (figura 24), de manera que aquest es correspongui amb el nivell de significació α escollit en cada cas.

La segona part de la sortida R Commander ofereix els intervals de confiança, a un nivell de confiança del 95% en aquest cas, comparant dos a dos les mitjanes. S'observa com els intervals més extrems, com ara els corresponents a C. Informació i Psicologia, no s'encavalquen per ben poc. Això és lògic, atès que el p-valor = 0,007 està molt a prop del valor significació escollit α = 0,01. Si el p-valor hagués estat encara més petit, tots dos intervals estarien clarament separats. Si, per contra, el p-valor hagués estat més gran, tots dos intervals s'encavalcarien parcialment com passa en la resta dels casos.

Abans de donar per definitives les conclusions anteriors, convé validar que es compleixen els supòsits bàsics de normalitat i variància constant de les dades. La figura 40 mostra el gràfic de normalitat corresponent a les observacions. No sembla observar-se patrons estranys ni gaire punts excessivament allunyats de la recta, per la qual cosa s'acceptarà com a vàlid el supòsit de normalitat. Per la seva banda, la figura 41 mostra el gràfic de dispersió de cada grup. Tampoc s'hi observen grans diferències entre les dispersions dels diferents nivells, per la qual cosa s'acceptarà com a vàlid el supòsit de variància constant entre els diferents grups.

Figura 40. Gràfic de normalitat de les valoracions registrades

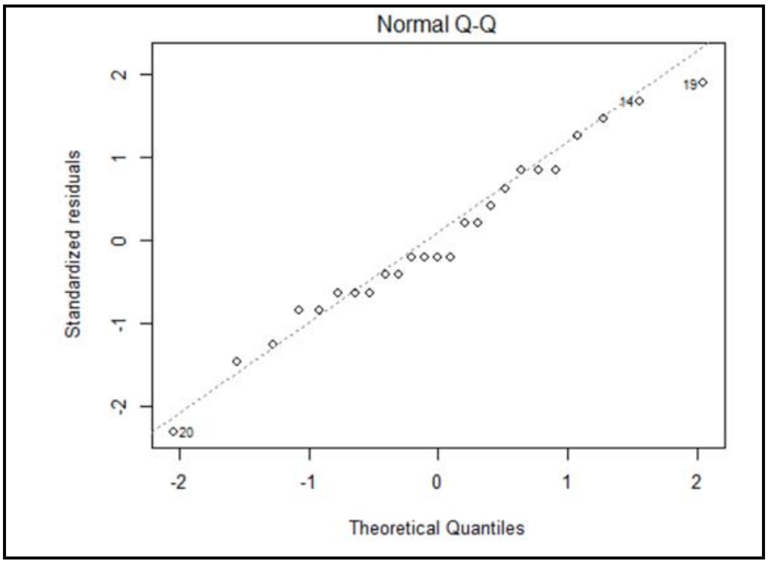
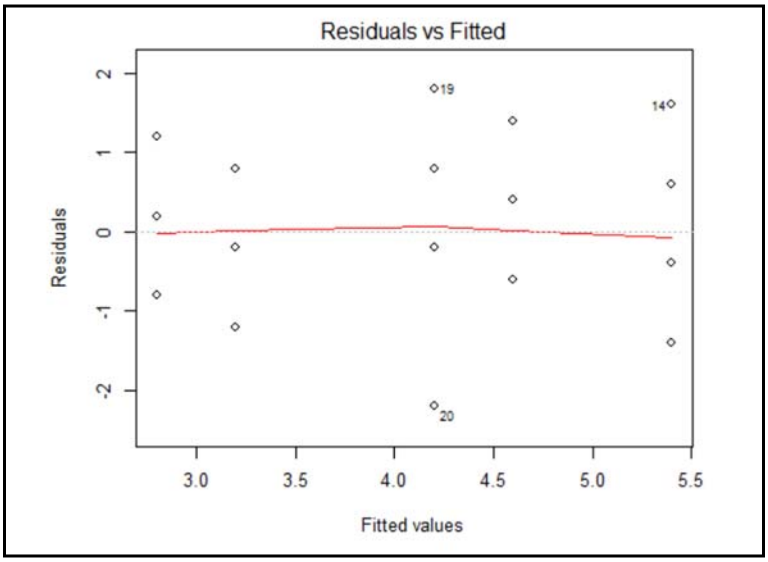


Figura 41. Gràfic de dispersió de cada grup



Resum

En aquest mòdul s'han presentat les principals tècniques estadístiques que permeten comparar estadísticament dos o més grups i discernir si hi ha o no diferències significatives entre ells. En el cas de dos grups, es fa servir un contrast d'hipòtesi basat en la *t* de Student (si s'estan comparant dues mitjanes) o en la normal (si s'estan comparant dues proporcions). Per acabar, s'han estudiat procediments que es poden aplicar per a fer inferències sobre variàncies poblacionals. Es va presentar la distribució *F*, per a usar-la en proves d'hipòtesi sobre les variàncies de dues poblacions normals.

En el cas de tres grups o més, es fa servir un test ANOVA basat en la *F* de Snedecor.

Convé tenir sempre molt present que el que importa més d'un test d'hipòtesi no són els càlculs matemàtics que hi són subjacents (en gran part perquè aquests càlculs es poden automatitzar mitjançant l'ús de programari), sinó la interpretació correcta dels resultats obtinguts i la seva credibilitat, la qual dependrà del fet que es compleixin o no els supòsits necessaris per a poder aplicar cada una de les tècniques d'inferència vistes en aquest mòdul. Si bé l'ordinador pot ser molt útil efectuant els càlculs matemàtics amb precisió i rapidesa, és responsabilitat de l'investigador saber interpretar els resultats i comprovar la validesa dels supòsits.

Exercicis d'autoavaluació

1) Estudia l'impacte que causa la reubicació forçada sobre el bon veïnat. Entrevistem 6 individus tant abans com després que se'ls va obligar a mudar-se. La nostra entrevista produeix les puntuacions següents:

Entrevistat	Abans	Després
1	2	1
2	1	2
3	3	1
4	3	1
5	1	2
6	4	1

Fa un contrast d'hipòtesis al nivell de confiança del 95%.

2) D'una mostra de 85 missatges de correu amb virus que arriba al servidor de la nostra empresa, el nostre programa KILLVIRUS instal·lat al servidor només n'ha detectat 25. Les especificacions del programa deien que el programa en detectava més del 40% del correu amb virus. Estan d'acord els resultats obtinguts amb les especificacions del programa (considereu $\alpha = 0{,}1$)? Trobeu el p-valor del contrast.

3) Volem comparar l'eficiència de dos compiladors de dos sistemes d'indexació diferents A i B. Per a fer-ho, es dissenyen 8 programes en cada un dels dos sistemes i es mesura el temps d'execució que triga cada un dels programes per a resoldre 8 problemes determinats d'optimització. Els resultats es mostren a la taula següent:

Problema d'optimització a resoldre	Temps d'execució usat per l'executable compilat amb el sistema A (en segons)	Temps d'execució usat per l'executable compilat amb el sistema B (en segons)
P1	1,2	1,4
P2	1,3	1,7
P3	1,5	1,5
P4	1,4	1,3
P5	1,7	2,0
P6	1,8	2,1
P7	1,4	1,7
P8	1,3	1,6

Podem assegurar a partir de les dades anteriors que el compilador del sistema A és més eficient que el compilador del sistema B? (considereu $\alpha = 0{,}05$). Trobeu el p-valor del contrast.

4) Dues empreses, A i B, volen comprar un dispositiu d'emmagatzematge amb vista a fer còpies de seguretat. Abans de fer la compra, es fa un estudi de quants gigues necessitarien per a fer la còpia. Aquest estudi consisteix a calcular durant 10 dies tota la informació de l'empresa necessària per a la còpia de seguretat. Els resultats es mostren a la taula següent:

Dia	1	2	3	4	5	6	7	8	9	10
Empresa A (gigues)	34	45	47	49	31	30	24	33	35	40
Empresa B (gigues)	45	47	50	42	40	51	46	59	42	46

Suposant normalitat i a un nivell de significació del 0,05, podem afirmar que l'empresa B necessita més capacitat d'emmagatzematge que l'empresa A? Indicació: abans de res, heu de fer el contrast corresponent per a veure si les variàncies de les dues mostres són iguals a un nivell de significació de 0,05.

5) S'ha dissenyat un experiment aleatori per a analitzar durant quant de temps és efectiva cada una de les quatre drogues que es poden fer servir per a alleugerir el dolor després d'una operació quirúrgica. Les dades obtingudes es mostren en la taula següent:

	Droga			
	A	***B***	***C***	***D***
Temps (hores)	8	6	8	4
	6	6	10	4
	4	4	10	2
	2	4	10	
			12	

Per a un nivell de significació $\alpha = 005$, contrastar la hipòtesi nul·la que les quatre drogues són igualment efectives.

6) A l'hora de baixar programes de codi font obert d'Internet, sol ser habitual poder optar per diversos servidors (rèpliques). Generalment, les velocitats de baixada des de cada servidor depenen de la distància existent entre el servidor i el client que sol·licita la baixada. En aquest cas, es vol estudiar si les velocitats de baixada des de cinc servidors diferents es poden considerar equivalents o no. Per a cada un dels servidors, s'han seleccionat alguns fitxers a l'atzar (tots de la mateixa mida) i s'han baixat en el client, i s'has obtingut els temps de baixada (en segons) que es mostren en la taula següent:

Servidor					
	A	***B***	***C***	***D***	***E***
Temps de baixada (en segons)	3,8	6,8	4,4	6,5	6,2
	4,2	7,1	4,1	6,4	4,5
	4,1	6,7	3,9	6,2	5,3
	4,4		4,5		5,8

Es pot afirmar que la velocitat mitjana de baixada és independent del servidor seleccionat? Feu servir un nivell de significació $\alpha = 0{,}01$.

7) Es volen comparar els ingressos per família (en milers d'euros) corresponents a tres províncies d'una mateixa comunitat autònoma. Per a fer-ho, per a cada província s'han seleccionat 9 famílies a l'atzar i se n'han registrat els ingressos. La taula següent mostra les observacions obtingudes:

Província			
	A	***B***	***C***
Ingressos familiars (milers d'euros)	45	32	40
	39,5	30	42
	42	37	45
	35	35	39,5
	40	28,5	40
	37	37,5	38
	44	31	51
	48,5	37,6	47,5
	50	25	41

Per a un nivell de significació $\alpha = 0{,}05$, es pot afirmar que els ingressos mitjans per família no depenen de la província a la qual aquesta pertany?

8) Una universitat fa servir tres consultories externes que ofereixen serveis d'assessorament tècnic en línia als seus estudiants. Per a cada una d'aquestes consultories, s'han escollit a l'atzar sis serveis prestats durant l'any en curs i s'ha registrat el canvi percentual en el seu preu en relació amb el preu mitjà de l'any anterior. Les dades recollides es mostren en la taula següent:

	Consultoria		
	A	*B*	*C*
Canvi percentual en el preu del servei	3,0	4,5	1,0
	2,5	2,5	–2,5
	–1,5	7,0	–3,5
	4,0	9,0	2,0
	–1,0	1,5	4,6
	5,5	2,0	0,5

Per a un nivell de significació $\alpha = 0{,}01$, es vol contrastar la hipòtesi nul·la que el canvi percentual mitjà en el preu del servei és el mateix per a les tres consultories.

9) Es vol comparar el nivell d'innovació/originalitat de sis revistes diferents, tot i que totes pertanyen a un mateix àmbit temàtic. Per a fer-ho, s'han seleccionat a l'atzar set exemplars de cada una de les revistes i un comitè d'experts ha avaluat el nivell d'innovació/originalitat de cada exemplar, per a la qual cosa s'ha fet servir una escala entre 1 (mínim) i 300 (màxim). Les dades recollides es mostren en la taula següent:

	Revista					
	A	*B*	*C*	*D*	*E*	*F*
Nivell d'innovació/ originalitat	300	190	228	276	162	264
	300	164	300	296	175	168
	300	238	268	62	157	254
	260	200	280	300	262	216
	300	221	300	230	200	257
	261	132	300	175	256	183
	300	156	300	211	92	93

A partir d'aquestes observacions, es pot afirmar que totes las revistes mostren un nivell d'innovació/originalitat equivalent o, per contra, hi ha diferències significatives entre els nivells d'innovació/originalitat de les diferents revistes? Utilitzar un nivell de significació $\alpha = 0{,}05$.

Solucionari

1) Es tracta d'un contrast de diferència de mitjanes per a **mostres dependents o aparellades** (estadístic de contrast t^* =1,49; el valor crític és $t_{\alpha/2=0,05/2,}$ i 5 graus de llibertat = 2,571. $t^* < t_{0,025;\,5}$ no es pot rebutjar H_0. Podem dir al 95% de confiança que el bon veïnat no ha variat quan es produeix la reubicació.

2) Es tracta d'un contrast de **diferència de proporcions**. L'estadístic de contrast segueix aproximadament la distribució normal $N(0,1)$, si la grandària de la mostra és prou gran com és el nostre cas. El valor de l'estadístic de contrast és $z^* = -1{,}993$.

El valor crític serà: $z_{0,1} \approx 1{,}28$. Com que $z < z_{0,1}$, acceptem la hipòtesi nul·la i concloem que les especificacions del servidor són falses.

3) Es tracta d'un contrast de **diferència de mitjanes dependents**. El valor de l'estadístic de contrast val: $t \approx -3{,}481$. El valor crític val $t_{0,05,7} \approx 1{,}895$. Com que $t < -t_{0,05;7}$, rebutgem la hipòtesi.

El *p*-valor és $p = p(t_7 < -3{,}481) \approx 0{,}0051$, valor que és menor que 0,05. Per tant, arribem a la mateixa conclusió: rebutjar la hipòtesi nul·la.

4) El resultat del R Commander per al **contrast de variàncies** és:

```
        F test to compare two variances

data:  variable by factor
F = 2.1915, num df = 9, denom df = 9, p-value = 0.2581
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5443427 8.8230486
sample estimates:
ratio of variances
          2.19152
```

S'accepta la igualtat de variàncies.

El resultat del R Commander per al **contrast de diferència de mitjanes independents** és:

```
        Two Sample t-test

data:  variable by factor
t = -3.2104, df = 18, p-value = 0.00485
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -16.544034  -3.455966
sample estimates:
mean in group A mean in group B
           36.8            46.8
```

Com que *p*-valor < 0,05, rebutgem la hipòtesi nul·la; per tant, acceptem que l'empresa B necessita més capacitat d'emmagatzematge que l'empresa A.

5) Estadístic de contrast $F = 12{,}50$; *p*-valor = 0,001 < α = 0,05 → Rebutjar la hipòtesi nul·la, com ara: no tots els grups tenen el mateix comportament.

6) Estadístic de contrast $F = 31{,}6$; *p*-valor = 0,000 < α = 0,01 → Rebutjar la hipòtesi nul·la, com ara: no tots els grups tenen el mateix comportament.

7) Estadístic de contrast $F = 13{,}83$; *p*-valor = 0,000 < α = 0,05 → Rebutjar la hipòtesi nul·la, com ara: no tots els grups tenen el mateix comportament.

8) Estadístic de contrast $F = 2{,}91$; *p*-valor = 0,085 > α = 0,01 → No rebutjar la hipòtesi nul·la, com ara: tots els grups semblen tenir el mateix comportament.

9) Estadístic de contrast $F = 5{,}30$; *p*-valor = 0,001 < α = 0,05 → Rebutjar la hipòtesi nul·la, com ara: no tots els grups es comporten igual.