PROBLEMES CLAU DE

DISSENY D'EXPERIMENTS

I

ANÀLISI DE DADES

PART I

C. Arenas

Professora del Departament d'Estadística de la UB

Barcelona, 19 de Gener de 2005

Dins de les iniciatives per a la millora de la qualitat de l'ensenyament neix la col·lecció "Problemes Clau de Disseny d'Experiments i Anàlisi de Dades" dirigida als estudiants de Biologia.

Aquesta col·lecció de problemes està formada per una sèrie de manuals que pretenen ajudar l'alumne en el seguiment de l'estudi de l'assignatura de Disseny d'Experiments i Anàlisi de Dades, que s'imparteix actualment a la llicenciatura de Biologia. Els manuals contenen problemes resolts fruit de l'experiència docent en aquesta matèria i estan concebuts perquè es pugui comprovar el nivell actual de l'alumne i superar, si cal, les mancances que s'hi detectin. Els problemes que es presenten són tant teòrics com aplicats i alguns contenen llistats d'ordinador obtinguts amb el paquet estadístic Statgraphics. Enguany apareix el primer títol relatiu al disseny d'un factor.

Vull agrair a tots els meus alumnes de DEAD, en especial als del curs 2003-2004, el seu treball i els seus valuosos comentaris, sense els quals l'elaboració d'aquesta col·lecció de problemes no hauria estat possible.

Espero que aquesta col·lecció sigui útil als estudiants en la seva formació com a persones i futurs professionals.

C. Arenas

Professora del Departament d'Estadística de la UB

Problema

Un zoòleg vol comprovar si l'altura de les caixes niu influeixen en l' èxit reproductor del carboner comú (*Parus major*). Amb aquest objectiu du a terme un experiment al parc natural de Collserola i col·loca caixes niu a sis alçades distintes: 1 (2 m), 2 (3 m), 3 (4 m), 4 (5 m), 5 (6 m) i 6 (7 m). A partir d'aquí es dedica a controlar la supervivència dels pollets a cada caixa niu durant el període reproductor del carboner comú. Les mitjanes i variàncies mostrals són:

Altura	1	2	3	4	5	6
Mitjana	1.92	2.02	2.3	2.13	1.88	1.43
S^2	0.65	0.62	0.69	0.71	0.59	0.64

Comproveu si hi ha diferències significatives en la supervivència dels pollets segons l'alçada del niu si per cada una de les diferents alçades hi ha 5 caixes. Nota: suposem normalitat i igualtat de variàncies i nivell de significació = 0.05.

1·1

Tenim una variable = supervivència pollets, amb un factor = alçada niu, amb 6 nivells (k = 6) i 5 mostres per nivell ($n_1 = 5$, $n_2 = 5$, $n_3 = 5$, $n_4 = 5$, $n_5 = 5$, $n_6 = 5$, $n = 30$); amb variàncies: $S_1^2 = 0.65$, $S_2^2 = 0.62$, $S_3^2 = 0.69$, $S_4^2 = 0.71$, $S_5^2 = 0.59$ i $S_6^2 = 0.64$.

Model ANOVA: supervivència = mitjana general + efecte altura + residus

$$y_{ij} = \mu + \alpha_i + e_{ij} \text{ amb } \sum_i \alpha_i = 0.$$

Hipòtesi:

H_0: $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 = 0$
H_1: Alguna igualtat és falsa.

Com que el Problema ens dóna per suposada la normalitat i l'homogeneïtat de variàncies, calculem directament els valors de la taula ANOVA.

Primer calculem la suma de quadrats entre grups, $Q_e = \sum_i n_i (y_{i\cdot} - y_{\cdot\cdot})^2$; necessitem les mitjanes en cada nivell i la mitjana general que calcularem a partir de les mitjanes de cada nivell

$y_{\cdot\cdot} = (n_1 y_{1\cdot} + n_2 y_{2\cdot} + \cdots + n_6 y_{6\cdot}) / n = (5 \times 1.92 + 5 \times 2.02 + 5 \times 2.3 + 5 \times 2.13 + 5 \times 1.88 + 5 \times 1.43)/30$
$= 1.9467.$

Calculem ara la suma de quadrats dins dels grups:

$$Q_d = \sum_{i,j} \left(y_{ij} - y_{i.}\right)^2 = \sum_j \left(y_{1j} - y_{1.}\right)^2 + \sum_j \left(y_{2j} - y_{2.}\right)^2 + \cdots + \sum_j \left(y_{6j} - y_{6.}\right)^2.$$

Com que tenim les variàncies mostrals i $S_i^2 = \dfrac{\sum_i \left(y_{ij} - y_{i.}\right)^2}{n_i}$,

queda

$$Q_d = n_1 \times S_1^2 + n_2 \times S_2^2 + \cdots + n_6 \times S_6^2 = 19{,}5.$$

Taula ANOVA:

Variable	**SQ**	**Gr.ll.**	**QM**	**F**
Entre grups	2.18	k-1=5	Q_e/k-1=0.436	0.436/0.812=0.537
Dintre de grups	19.5	n-k=24	Q_d/n-k=0.812	
Total	21.68	n-1=29		

Cal comparar el valor de l'estadístic F experimental (el de la taula ANOVA) amb el valor de les taules d'una distribució F amb 5 i 24 graus de llibertat. Com que el valor de $F_{5,24}$ = 2.62 de les taules és més gran que 0.537, rebutgem la hipòtesi alternativa. Per tant, no hi ha diferències significatives en la supervivència dels pollets respecte de les alçades dels nius.

Problema

Se sap que s'ha llençat material tòxic a un riu, que entra en una gran àrea de pesca comercial en aigua salada. Els enginyers civils han estudiat la forma en què l'aigua transporta el material tòxic i han mesurat la quantitat de material (en parts per milió) trobat a les ostres recollides en tres llocs diferents, des de la sortida de l'estuari fins a la badia on es concentra la major part de la pesca comercial. Per un Problema d'organització es van perdre les dades del lloc 3, però per sort es conservaren els valors de la mitjana i la variància.

lloc 1 (estuari	lloc 2 (lluny de la badia)	lloc 3 (prop de la badia)
15	19	
26	15	
20	10	
20	26	
29	11	
28	20	
21	13	
26	15	
	18	

La mitjana per al grup 3 és $y_{3\cdot} = 22$ amb $n_3=7$ i la variància mostral, 2.1.

a) Obtingueu la taula ANOVA i decidiu si hi ha diferències significatives amb un nivell de significació = 0.05.

b) Si hi ha diferències significatives, utilitzeu el mètode de comparacions múltiples LSD per assenyalar-les.

Objectiu: Observar si hi ha diferències significatives en els valors mitjans de parts per milió de material tòxic trobat a les ostres recollides en tots tres llocs.

Variables observables: Y= "ppm de material tòxic en ostres"

1 factor: lloc

k = 3 nivells: lloc 1 (estuari) $\rightarrow n_1 = 8$

lloc 2 (fora badia) $\rightarrow n_2 = 9$

lloc 3 (prop badia) $\rightarrow n_3 = 7$

$n = 24$

Model: $y_{ij} = \mu + \alpha_i + e_{ij}$ amb $\sum_i \alpha_i = 0$.

Hipòtesi que cal testar : H_0: $\alpha_1 = \alpha_2 = \alpha_3 = 0$

H_1: alguna igualtat és falsa

Càlculs:
Calculem les mitjanes mostrals per nivells:

$y_{1.} = (15 + 26 + 20 + 20 + 29 + 28 + 21 + 26) / 8 = 23.125$

$y_{2.} = (19 + 15 + 10 + 26 + 11 + 20 + 13 + 15 + 18) / 9 = 16.33$

$y_{3.} = 22,$

i la mitjana total

$y_{..} = y_{..} = (n_1 y_{1.} + n_2 y_{2.} + n_3 y_{3.}) / n = (8 \text{ x } 23.125 + 9 \text{ x } 16.33 + 7 \text{ x } 22) / 24 = 20.25.$

Calculem ara les sumes de quadrats entre grups i dintre de grups:

$$Q_e = \sum_i n_i (y_{i.} - y_{..})^2 = 8 \text{ x } (23.125 - 20.25)^2 + 9 \text{ x } (16.33 - 20.25)^2 + 7 \text{ x } (22 - 20.25)^2$$

$$= 225.855,$$

$$Q_d = \sum_{i,j} (y_{ij} - y_{i.})^2 = \sum_j (y_{1j} - y_{1.})^2 + \sum_j (y_{2j} - y_{2.})^2 + \sum_j (y_{3j} - y_{3.})^2 =$$

$$= (15 - 23.125)^2 + \cdots + (26 - 23.125)^2 + (19 - 16.33)^2 + \cdots + (18 - 16.33)^2 +$$

$$= (15 - 23.125)^2 + \cdots + (26 - 23.125)^2 + (19 - 16.33)^2 + \cdots + (18 - 16.33)^2 + 7 \times 2.1 = 381.2.$$

Taula ANOVA:

Variabilitat	SQ	Gr. de llibertat	SQM	F
Entre grups	225.855	k –1 = 3 – 1 = 2	225.855 / 2 = 112.93	112.93 / 18.15 = 6.22
Dintre de grups	381.2	n – k = 24 – 3 = 21	381.2 / 21 = 18.15	
Total	607.05	n – 1 = 24 – 1 = 23		

Criteri de decisió:

En no disposar del p-valor, per arribar a la conclusió cal buscar a les taules quin és el valor de l'estadístic F amb 2 i 21 graus de llibertat. Trobem que $F_{2, 21} = 3.47$.

Com que F experimental = 6.22 > $F_{2, 21} = 3.47$, rebutgem H_0.

Conclusió:

Podem acceptar que hi ha diferències significatives entre la quantitat de material tòxic a les ostres segons el lloc de captura.

b) Ara fem les comparacions múltiples següents:

Cas 1	Cas 2	Cas 3
$H_0 : \alpha_1 - \alpha_2 = 0$	$H_0 : \alpha_1 - \alpha_3 = 0$	$H_0 : \alpha_2 - \alpha_3 = 0$
$H_1 : \alpha_1 - \alpha_2 \neq 0$	$H_1 : \alpha_1 - \alpha_3 \neq 0$	$H_1 : \alpha_2 - \alpha_3 \neq 0$

En el mètode LSD, es necessita conèixer el valor de $Q = t_{\alpha/2}\sqrt{Q_d\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$. Observem que com que no és un disseny balancejat, el valor de Q variarà ja que depèn de la mida de les mostres.

Cas 1: el valor de Q és $Q = 2.08\sqrt{18.15}\left(\frac{1}{8}+\frac{1}{9}\right) = 4.34$. L'interval que cerquem serà:

$[y_{1.} - y_{2.} - Q; y_{1.} - y_{2.} + Q] \rightarrow$ [23.125 – 16.33 – 4.34 ; 23.125 – 16.33 + 4.34] =

=[2.45 ; 11.13].

Com que el zero és fora de l'interval, rebutgem H_0, és a dir hi ha diferències significatives entre la quantitat de material tòxic trobat en les ostres recollides a l'estuari i les ostres recollides en el lloc llunya de la badia.

Cas 2: ara $Q = 2.08\sqrt{18.15}\left(\frac{1}{8}+\frac{1}{7}\right) = 4.58$.

$[y_{1.} - y_{3.} - Q; y_{1.} - y_{3.} + Q] \rightarrow$ [23.125 – 22 – 4.58; 23.125 – 22 + 4.58] =

= [-3.45 ; 5.7]

Com el zero és dins de l'interval, rebutgem H_1, és a dir no hi ha diferències significatives entre la quantitat de material tòxic trobat en les ostres recollides a l'estuari i les ostres recollides en el lloc proper a la badia .

Cas 3: ara $Q = 4.46$.

$[y_{2.} - y_{3.} - Q; y_{2.} - y_{3.} + Q] \rightarrow$ [-10.13 ; -1.21].

Com que el zero és dins de l'interval, rebutgem H_1, és a dir, hi ha diferències significatives entre la quantitat de material tòxic trobat en les ostres recollides en el lloc llunyà de la badia i les ostres recollides en el lloc proper a la badia .

Si observem els valors de $y_{1.} = 23.125$; $y_{2.} = 16.33$; $y_{3.} = 22$, es pot veure que el lloc llunyà de la badia és el que presenta menys material tòxic i entre l'estuari i a prop de la badia no detectem diferències significatives.

Problema

Volem estudiar l'efecte de tres nous productes que s'apliquen a la dieta dels animals en una granja que es dedica a la cria d'aviram per fer-los augmentar de pes al cap de mig any i des del naixement. S'aplica el producte A a un grup de pollastres, el B a un segon grup, i el C a un tercer grup. Cada grup d'estudi consta de 5 animals.

Ens donen la següent:

Mitjana producte A (y_a)=3271.6 S_a^2 **=38795.3**

Mitjana producte B (y_b)= 3621.6 S_b^2 **=12650.3**

Mitjana producte C (y_c)= 3323 S_c^2 **=7137**

Mitjana producte Total (y_t)= 3405.4

Taula ANOVA:

Variabilitat	Gr de ll	SQ	QM	F
Entre grups	2	357173.2	178586.6	7.31
Dintre grups	12	292913	24409.41	
Total	14	650086.2		

a) **Hi ha diferències entre tots tres productes?**

b) **Quin producte, amb una probabilitat del 95%, és el més recomanable per fer engreixar els pollastres?**

Assumeix que hi ha normalitat de residus, homogeneïtat de variàncies i nivell de significació del 0.05.

a) Per saber si hi ha diferències entre tots tres productes, com que ja ens donen la taula ANOVA, només cal fer servir el criteri de decisió:
A la taula de la distribució F, per a 2 i 12 graus de llibertat dóna un valor de 3.89.
Com que la F experimental és de 7.31>3.89 rebutgem H_0; podem acceptar que hi ha diferències significatives entre els diferents productes.

b) Per saber quin és el producte més recomanable per fer engreixar l'aviram, calcularem l'interval de confiança de la predicció amb un 95% de confiança.
Sabem que sota A, la predicció segueix una distribució $y_{..} + (y_a - y_{..}) + N(0, \sqrt{QM_d}) =$ $N(y_a, \sqrt{QM_d}) = N(3271.6\ ,\ 156.23) \rightarrow$ els extrems que cal cercar són:

3271.6 - 1.96 x 156.23 i 3271.6 + 1.96 x 156.23. Així doncs amb una confiança del 95%, si fem servir el tractament A, els pollastres engreixaran entre [3580.81 ; 2968.39] g. en un període de mig any.

Sabem que sota B, la predicció segueix una distribució $y_{..} + (y_b - y_{..}) + N(0, \sqrt{QM_d}) =$ $N(y_b, \sqrt{QM_d}) = N(3621.6 , 156.23) \rightarrow$ amb una confiança del 95%, si fem servir el tractament B, els pollastres engreixaran entre
[3621.6 - 1.96 x 156.23 ; 3621.6 + 1.96 x 156.23] = [3927.81 ; 3315.39]g. Al cap de mig any.

Anàlogament, sota C, la predicció segueix una distribució
$N(y_c, \sqrt{QM_d}) = N(3323 , 156.23) \rightarrow$ amb una confiança del 95%, si fem servir el tractament C, els pollastres engreixaran entre [3629.21 ; 3016.79] g. Al cap de mig any.

Així doncs, segons aquestes prediccions, triaríem el tractament B.

Problema

En una piscifactoria es mesura el pes dels peixos segons l'aliment que se'ls dóna. Per això es disposa de 3 tipus diferents de pinso per a peixos: A, B i C i 15 peixos de la mateixa edat, 5 dels quals són alimentats amb A, 5 amb B i 5 amb C.
Es vol saber si hi ha diferències pel que fa al pes d'aquests peixos i, si n'hi ha, quin és l'aliment que els fa tenir un major pes.
Es van obtenir les dades següents:

Variabilitat	gr.ll.	SQ	QM	F
Entre grups	?	?	?	?
Dins els grups	?	143.2	?	
total	?	?		

A més sabem que amb una probabilitat del 95% els intervals de confiança per les prediccions són:
Per a A: [243.83; 257.37]
Per a B: [242.63; 256.17]
Per a C: [253.03; 266.57]

A partir d'aquesta informació completeu la taula ANOVA.

Primerament completem la columna dels graus de llibertat:
Entre grups: 3 – 1 = 2
Dins els grups: 15 – 3 = 12
Total: 15 – 1 = 14
Ara ja podem calcular els QM dintre grups: QM_d=143.2/12=11.938
Per calcular els QM entre grups recordem la seva fórmula:
$Q_e = \sum_i n_i (y_{i.} - y_{..})^2$.
Per tant, a partir de la informació de l'enunciat caldrà determinar els valors de les mitjanes a cada nivell i de la mitjana global. Les grandàries mostrals a cada nivell són conegudes amb valor 5.
Els intervals de confiança per a les prediccions estan determinats per:

Aliment A
predicció$\approx y_{..} + (y_{1.} - y_{..}) + N(0, \sqrt{QM_d})$,

predicció$\approx N(y_{1.}, \sqrt{QM_d})$. Per tant, l'interval de confiança al 95% és:
$\left[y_{1.} - 1.96\sqrt{QM_d}; y_{1.} + 1.96\sqrt{QM_d}\right]$ = [243,83; 257,37]. I, per tant, tindrem que:
$y_{1.}$ =(243,83+ 257,37)/2=250.6

Aliment B
predicció$\approx y_{..} + (y_{2.} - y_{..}) + N(0, \sqrt{QM_d})$
predicció$\approx N(y_{2.}, \sqrt{QM_d})$.
Així doncs,
interval = $\left[y_{2.} - 1.96\sqrt{QM_d}; y_{2.} + 1.96\sqrt{QM_d}\right]$= [242.63;256.17].
I podem deduir que $y_{2.}$ =(242.63+256.17)/2=249.4.

Aliment C, procedint de forma anàloga:
Predicció $\approx N(y_{3.}, \sqrt{QM_d})$; interval = [253.03;266.57] ;
$y_{3.}$ =(253.03+266.57)/2=259.8.

A partir d'aquests valors el valor de $y_{..}$ serà:
$y_{..}$=(5x250.6 + 5x249.4 + 5x259.8)/15=253.26

Ara ja podrem calcular $Q_e = \sum_i n_i (y_{i.} - y_{..})^2$
$= 5(250.6-253.26)^2 + 5(249.4-253.26)^2 + 5(259.8-253.26)^2$
=323.734

Finalment, QM_e =323.734/2=161.867.

Per acabar cal determinar el valor de la F experimental i aplicar el criteri de decisió
Fexp=161.867/11.938=13.56.
El valor teòric segons una distribució $F_{2, 12}$ és 3,89 amb α=0,05
Com que el valor de Fexp=13,56 es troba a la cua (més gran que 3,89), rebutgem la hipòtesi H_0.
Per tant, amb un nivell de significació del 5% podem acceptar que hi ha diferències de pes significatives entre els grups.

Problema

S'estudien 3 marques de piles o bateries. Se sospita que la durada (en setmanes) de totes 3 bateries és diferent. Es proven 5 bateries de cada marca i els resultats que s'obtenen s'indiquen a continuació.

Table of Means for durada by marca with 95,0 percent LSD intervals

marca	Count	Mean
1	5	95.2
2	5	79.4
3	5	100.4
Total	15	91.6667

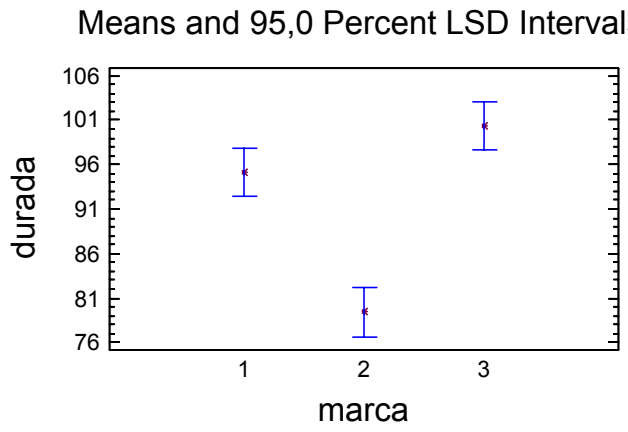
Variance Check

Cochran's C test: 0.444444 P-Value = 0.793832
Bartlett's test: 1.03263 P-Value = 0.840823
Hartley's test: 1.85714
Levene's test: 0.07 P-Value = 0.932772

Computed Chi-Square goodness-of-fit statistic = 12.1333
P-Value = 0.145353
Shapiro-Wilks W statistic = 0.928424
P-Value = 0.255353
Z score for skewness = 0.799461
P-Value = 0.424021
Z score for kurtosis not computed.

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Between groups	?	?	598.067	?	?
Within groups	187.2	12	?		
Total (Corr.)	1383.33	14			



Multiple Range tests (LSD, 95% confiança)

Contrast	Difference	+/- Limits
1-2	15.8	5.44268
1-3	-5.2	5.44268
2-3	-21.0	5.44268

1. Hi ha diferències significatives quant a la durada de les bateries?
2. Quina en triaríeu?

1. Objectiu: Mirar si hi ha diferències significatives quant a la durada de les diferents bateries.
Variable observable: durada de les bateries.
Factor: marca de la bateria.

Nivells k = 3 amb $n_1 = 5$, $n_2 = 5$, $n_3 = 5$ i n = 15.

Model matemàtic: $y_{ij} = \mu + \alpha_i + e_{ij}$ amb $\sum_i \alpha_i = 0$.

Cal que es compleixin les hipòtesi de treball següents per a una ANOVA d'un factor:

- homogeneïtat de variància
- normalitat
- independència

Mirem si hi ha homogeneïtat de variància:
H_0: hi ha homogeneïtat de variància
H_1: no hi ha homogeneïtat de variància

Utilitzem el test de Bartlett. Estadístic: 1.03262; p-valor: 0.8408

Com que el p-valor > 0.05 $\Rightarrow$ Rebutgem H_1, i per tant, podem dir que hi ha homogeneïtat de variància.

Hi ha normalitat?

H_0: hi ha normalitat
H_1: no hi ha normalitat

Per determinar si hi ha normalitat farem servir el test de khi-quadrat

Estadístic: 12.1333; p-valor: 0.145

Com que el p-valor > 0.05 $\Rightarrow$ Rebutgem H_1, i per tant, podem dir que hi ha normalitat.

Fem una ANOVA d'un factor. La taula ANOVA completa es:

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio
Between groups	1196.134	2	598.067	38.34
Within groups	187.2	12	15.6	
Total (Corr.)	1383.33	14		

Hipòtesi de treball:

H_0: $\alpha_1 = \alpha_2 = \alpha_3 = 0$
H_1: alguna igualtat és falsa

Mirem a les taules l'estadístic $F_{2,12}$

$F_{2,12} = 3.89$
$F_{experimental} = 38.34$

Com que $F_{experimental} > F_{2,12}$ $\Rightarrow$ Rebutgem H_0 i, per tant, podem dir que hi ha diferències significatives entre la durada de totes tres bateries.

2. Comparem totes 3 bateries dues a dues. Farem un test de comparació múltiple. (LSD, 95% confiança**).**

Contrast 1-2

H_0: $\alpha_1 - \alpha_2 = 0$
H_1: $\alpha_1 - \alpha_2 \neq 0$

Interval $[y_{1.} - y_{2.} - Q; y_{1.} - y_{2.} + Q]$, Q ens el dóna la taula i val 5.44. També ens dóna la diferència entre les dues mitjanes de cada marca. Per tant,

[15.8 – 5.44; 15.8 + 5.44] = [10.36; 21.24] $\Rightarrow$ 0 està fora de l'interval i per tant rebutgem H_0 . Podem acceptar que hi ha diferències significatives entre la bateria de la marca 1 i la de la 2.

Contrast 1-3

$H_0 : \alpha_1 - \alpha_3 = 0$
$H_1 : \alpha_1 - \alpha_3 \neq 0$

Interval $\left[y_{1.} - y_{3.} - Q; y_{1.} - y_{3.} + Q\right]$= [-5.2 – 5.44; -5.2 + 5.44] = [-10.64; 0.24] $\Rightarrow$ 0 és dins de l'interval i, per tant, rebutgem H_1. Podem acceptar que no hi ha diferències significatives entre la bateria de la marca 1 i la de la 3.

Contrast 2-3

$H_0 : \alpha_2 - \alpha_3 = 0$
$H_1 : \alpha_2 - \alpha_3 \neq 0$

Interval $\left[y_{2.} - y_{3.} - Q; y_{2.} - y_{3.} + Q\right]$= [21 – 5.44; 21 + 5.44] = [15.56; 26.44] $\Rightarrow$ 0 és fora de l'interval i, per tant, rebutgem H_0. Podem acceptar que hi ha diferències significatives entre la bateria de la marca 2 i la de la 3.
Si mirem la gràfica de la mitjana que sol durar cada marca de les diferents bateries estudiades, veiem que la que té més durada sembla que és la 3, seguida de la 1. Però la forma mes acurada és mitjançant els intervals de confiança per a les prediccions.

<u>Predicció marca 1</u>:
predicció$\approx y_{..} + \left(y_{1.} - y_{..}\right) + N(0, \sqrt{QM_d})$,
predicció$\approx N(y_{1.}, \sqrt{QM_d}) = N(95.2, \sqrt{15.6})$
Per tant l'interval de confiança al 95% és:
$\left[y_{1.} - 1.96\sqrt{QM_d}; y_{1.} + 1.96\sqrt{QM_d}\right]$ =[87.45;102.94]

<u>Predicció marca 2</u>:
predicció$\approx y_{..} + \left(y_{2.} - y_{..}\right) + N(0, \sqrt{QM_d})$,
predicció$\approx N(y_{2.}, \sqrt{QM_d}) = N(79.4, \sqrt{15.6})$. Així doncs,
interval = $\left[y_{2.} - 1.96\sqrt{QM_d}; y_{2.} + 1.96\sqrt{QM_d}\right]$= [71.66;87.14]

<u>Predicció marca 3</u>:
predicció$\approx N(y_{3.}, \sqrt{QM_d})$; interval = [92.66;108.14]

Així doncs, amb un 95% de confiança, la marca que dura més és la marca 3.

Problema

Es vol comparar la temperatura de cinc zones de Catalunya. S'han pres 30 mesures a cada indret i s'han obtingut els resultats següents:

ZONA	MITJANA	VARIÀNCIA (S_i^2)
1	25.4	?
2	28	1.3
3	26.3	1
4	?	0.9
5	26.7	1.2

Suposant normalitat i igualtat de variàncies i sabent que la mitjana general val 26.72 i que la suma de quadrats dintre de grups és de 165, calculeu les dades que falten a l'enunciat i obtingueu la taula ANOVA.

Es tracta d'una ANOVA d'un factor (la zona) amb 5 nivells (k = 5):
zona 1: $n_1 = 30$
zona 2: $n_2 = 30$
zona 3: $n_3 = 30$
zona 4: $n_4 = 30$
zona 5: $n_5 = 30$
n = 150
Variable observable: temperatura Y
Model matemàtic: $y_{ij} = \mu + \alpha_i + e_{ij}$ amb $\sum_i \alpha_i = 0.$

Hipòtesi:

H_0: $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0$

H_1: alguna igualtat és falsa

La mitjana de totes les dades:

$$y_{..} = (n_1 y_{1.} + n_2 y_{2.} + \cdots + n_5 y_{5.}) / n =$$

$$= (30 \times 25.4 + 30 \times 28 + 30 \times 26.3 + 30 \times y_{4.} + 30 \times 26.7) / 150 = 26.72$$

aleshores $y_{4.} = 27.2$.

La variabilitat total depèn de la variabilitat que hi ha entre els diferents grups i la variabilitat que hi ha dins els mateixos grups. Aquestes variabilitats les calculem a partir de la suma de quadrats:

Suma de quadrats entre grups: $Q_e = \sum_i n_i \, (y_{i.} - y_{..})^2 =$

$= 30[(25.4-26.72)^2 + (28-26.72)^2 + (26.3-26.72)^2 + (27.2-26.72)^2 + (26.7-26.72)^2]$

$Q_e = 113.64.$

Suma de quadrats dins de grup:

$$Q_d = \sum_{i,j} \left(y_{ij} - y_{i.}\right)^2 = \sum_j \left(y_{1j} - y_{1.}\right)^2 + \sum_j \left(y_{2j} - y_{2.}\right)^2 + \cdots + \sum_j \left(y_{5j} - y_{5.}\right)^2$$

$$Q_d = n_1 \times S_1^2 + n_2 \times S_2^2 + \cdots + n_5 \times S_5^2 = 165.$$

Llavors $S_1^2 = 1.1$

La suma de quadrats totals s'obté fent $Q_d + Q_e = 149$

També cal calcular a la taula ANOVA la suma de quadrats mitjans (QM),

QM entre grups = Q_e / k-1 = 113.64 / (5-1) = 113.64 / 4 = 28.41

QM dins grup = Q_d / n-k = 165 / (150 – 5) = 165 / 145 = 1.138.

Els graus de llibertat seran:

Entre grups: k – 1 = 5 – 1 = 4.

Dins de grup: n – k = 150 – 5 = 145.

L'estadístic F l'obtenim: F = (Q_e / k-1) / (Q_d / n-k) = 28.41 / 1.138 = 24.96

A continuació ens cal un criteri de decisió i per això hem de treballar amb la taula de Fisher amb els graus de llibertat següents: $F_{k-1,\, n-k} = F_{4,145.}$ El valor corresponent a aquests graus de llibertat és 2.37.

Com que la F experimental (24.96) és més gran que la trobada a la taula (2.37) rebutgem la hipòtesi nul·la, és a dir, que amb un nivell de significació del 5% podem dir que hi ha diferències significatives entre les temperatures de les diferents zones.

La taula ANOVA obtinguda és:

Variabilitat	**graus llibertat**	**SQ**	**SQM**	**F**
entre grups	4	113.64	28.41	24.96
dins de grup	145	165	1.138	
Total	149	278.64		

Problema

S'ha analitzat la quantitat de greixos de 3 tipus d'aliments diferents, fent 4 rèpliques de cada anàlisi. Hem obtingut les mitjanes següents i les seves variàncies mostrals:

Aliments	Mitjanes	Variàncies
1	16.5	1.1
2	18.2	0.9
3	15.0	1.2

a) **Construïu la taula ANOVA.**
b) **Per a l'aliment del tipus 2, quina és la probabilitat que la quantitat de greixos sigui superior a 17?**

Objectiu: Observar si hi ha diferències en la quantitat de greixos entre els diferents aliments.

Variable observable: Quantitat de greix.
Factor: aliment → 3 nivells (k=3)
$n_1 = n_2 = n_3 = 4 \quad n = 12$

$y_{1.} = 16.5 \qquad S_1^2 = 1.1$
$y_{2.} = 18.2 \qquad S_2^2 = 0.9$
$y_{3.} = 15 \qquad S_3^2 = 1.2$

$$y_{..} = (n_1 y_{1.} + n_2 y_{2.} + n_3 y_{3.}) / n = 4(16.5) + 4(18.2) + 4(15) / 12 = 16.56$$

a) Taula ANOVA:

$$Q_e = \sum_i n_i (y_{i.} - y_{..})^2 = 4(16.5\text{-}16.56)^2 + 4(18.2\text{-}16.56)^2 + 4(15\text{-}16.56)^2$$
$$= 0.0144 + 10.75 + 9.73 = 20.507.$$

$$Q_d = \sum_{i,j} (y_{ij} - y_{i.})^2 = \sum_j (y_{1j} - y_{1.})^2 + \sum_j (y_{2j} - y_{2.})^2 + \sum_j (y_{3j} - y_{3.})^2 =$$

$Q_d = n_1 \times S_1^2 + n_2 \times S_2^2 + n_3 \times S_3^2 = 12.8$.

QM entre grups = Q_e / k-1 = 20.507/2= 10.253

QM dins grup = Q_d / n-k = 12.8/9= 1.42

F = (Q_e / k-1) / (Q_d / n-k) = 10.253/1.42 = 7.22

Variabilitat	**gr. ll.**	**SQ**	**SQM**	**F**
Entre grups	2 (k-1)	20.507	10.253	7.22
Dintre grups	9 (n-k)	12.8	1.42	
Total	11 (n-1)	33.307		

Atès $F_{2,9}$ = 4.26 és inferior a F= 7.22 rebutgem la hipòtesi nul·la i, per tant, es pot acceptar que hi ha diferències significatives entre aquests 3 aliments.

b) Sabem que sota l'aliment 2

predicció$\approx y_{..} + (y_{2.} - y_{..}) + N(0, \sqrt{QM_d})$,

predicció$\approx N(y_{2.}, \sqrt{QM_d}) = N(18.2, \sqrt{1.42})$.

Així doncs, cal calcular $P\left(N(18.2, \sqrt{1.42}) > 17\right)$.

Reduint a una N(0,1)

$$P\left(N(0,1) > \frac{17 - 18.2}{\sqrt{1.42}}\right) = P\left(N(0,1) > -0.845\right)$$

per la simetria de la normal serà igual a:

$P\left(N(0,1) < 0.845\right)$.

I, finalment, mirant les taules dóna un valor de 0.801. És a dir, la probabilitat que la quantitat de greixos a l'aliment de tipus 2 sigui superior a 17 és del 80%.

Problema

Un grup de paleontòlogues vol demostrar que observant tan sols la mesura basialveolar dels cranis que es troben en un jaciment egipci, poden determinar l'època a la qual pertanyen els cranis. Per això disposen d'una col·lecció de cranis del mateix jaciment ja datats, amb la corresponent mesura basialveolar (expressada en mil·límetres).

a) Determineu si es pot acceptar la hipòtesi de normalitat i homogeneïtat de variàncies.

b) Analitzeu les diferències de la mesura basialveolar entre èpoques.

A : 4000 aC (Early Predynastic)	B : 3300 aC (Late Predynastic)	C: 1850 aC (12th–13th dynasties)	D: 200 aC (Ptolemaic period)	E: 150 dC (Roman period)
101	102	94	87	81
110	98	89	90	81
102	101	90	86	83
114	95	93	85	90
109	99	97	85	82
111	97	88	90	85
100	98	87	83	85
107	96	92	83	81

Objectiu: Determinar si la mesura basialveolar dels cranis depèn de l'època.
Variable observable: mesura basialveolar.

$\Rightarrow$ 1 factor: època

$\Rightarrow$ amb K = 5 nivells

$\rightarrow$ A $\rightarrow n_1 = 8$

$\rightarrow$ B $\rightarrow n_2 = 8$

$\rightarrow$ C $\rightarrow n_3 = 8$

$\rightarrow$ D $\rightarrow n_4 = 8$

$\rightarrow$ E $\rightarrow n_5 = 8$

Model ANOVA d'un factor

$$y_{ij} = \mu + \alpha_i + e_{ij} \text{ amb } \sum_i \alpha_i = 0.$$

Hipòtesi de treball

Els errors són variables aleatòries independents amb distribució normal de mitjana 0 i homogeneïtat de variàncies.

Hipòtesi que s'ha de testar

H_0: $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0$
H_1: alguna = és falsa

Primerament cal de comprovar les hipòtesi de treball:

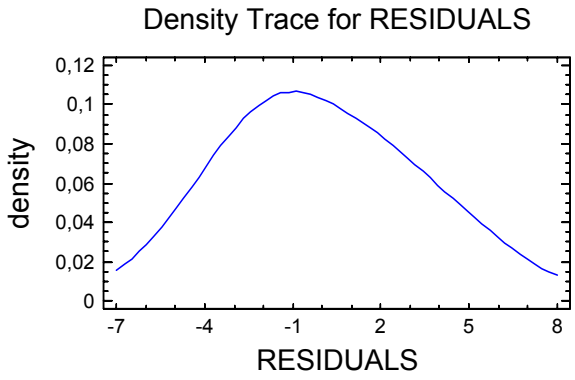
Normalitat:

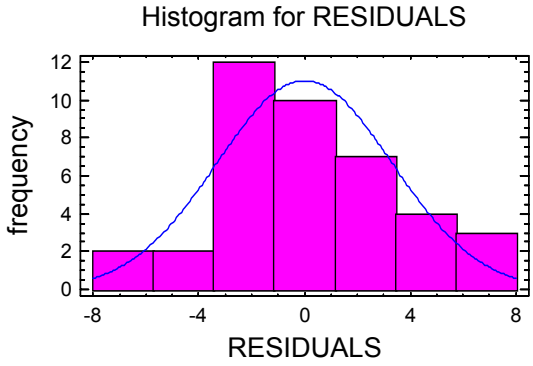
H_o: Residus ≈ Normal
H_1: Residus ≈ Normal

Mirarem si els residus segueixen una distribució normal amb el Chi-Square test:

Valor del estadístic=6.75 p-valor=0.94392

Com p-valor (0,94392) és més gran que el nivell de significació (0,05) rebutgem la hipòtesi alternativa. Podem, doncs, acceptar que els residus segueixen una distribució normal.





Homogeneïtat de variàncies:

H_o: Sí que hi ha homogeneïtat
H_1: No hi ha homogeneïtat

Per decidir si hi ha homogeneïtat de variàncies farem servir la prova de Barlett:

Valor de l'estadístic=1.16599, p-valor=0.278749

Com que el p-valor (0,278749) és més gran que el nivell de significació (0,05) rebutgem la hipòtesi alternativa; per tant, acceptem que hi ha homogeneïtat de variàncies.

2n. Ara continuarem testant:

H_0: $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0$
H_1: alguna = és falsa

Utilitzarem la taula ANOVA com a resultat de l'aplicació del model d'anàlisi de variància d'un factor:

Font	SQ	Gr.ll.	QM	F	P-Valor
Entre grups	2856.4	4	714.1	58.48	0.0000
Dintre grups	427.375	35	12.2107		
Total (Corr.)	3283.78	39			

Com que p-valor (0.0000) és menor que el nivell de significació (0.05) rebutgem la hipòtesi nul·la. Podem acceptar que hi ha diferències significatives entre els cranis de les distintes èpoques.

A continuació realitzarem les comparacions múltiples per determinar les diferències entres èpoques. Ens basarem en el mètode LSD (*Least Significant Difference*):

H_0: $\alpha_i - \alpha_j = 0$
H_1: $\alpha_i - \alpha j \neq 0$

El corresponent interval està determinat per $\left]y_{i.} - y_{j.} - Q; y_{i.} - y_{j.} + Q\right[$ amb $Q = t_{\alpha/2}\sqrt{Q_d\left(\frac{1}{n_i} - \frac{1}{n_j}\right)}$. Com que el disseny és balancejat, el valor de Q sempre serà el mateix. A la taula següent s'adjunten els valors de les diferències $y_{i.} - y_{j.}$ corresponents a cada interval i el valor Q**.**

Contrast	Diferències	+/- Límit
1 - 2	*8.5	3.54699
1 - 3	*15.5	3.54699
1 - 4	*20.625	3.54699
1 - 5	*23.25	3.54699
2 - 3	*7.0	3.54699
2 - 4	*12.125	3.54699
2 - 5	*14.75	3.54699
3 - 4	*5.125	3.54699
3 - 5	*7.75	3.54699
4 - 5	2.625	3.54699

Podem acceptar que hi ha diferències significatives entre totes les èpoques, llevat de la 4 i la 5. Per exemple,

1 - 2	*8.5	3.54699

l'interval és [8.5-3.54699; 8.5+3.54699], com que 0 no és dins l'interval rebutgem la hipòtesi nul·la; per tant hi ha diferències significatives entre els nivells 1 i 2.

Problema

Es vol comprovar si una àrea marina protegida pot influir en la pesca local de la zona. Es fa un estudi en el qual es mesura el pes de les captures que es produeixen durant cinc dies en tres zones: la primera, limítrofa a l'àrea protegida; la segona, a 500 m de distància i la tercera, a 1000 m. Les dades obtingudes en kg són:

Zona A (limítrofa): 113, 105, 111, 113, 105.
Zona B (500 m): 105, 111, 97, 111, 97.
Zona C (1000 m): 81, 85, 80, 89, 87.

- **Analitzeu si hi ha diferències entre totes tres zones estudiades. Si n'hi ha, on són?**

Taula de comparacions múltiples

Contrast	Difference	+/- Limits
1 - 2	5.2	7.15148
1 - 3	25.0	7.15148
2 - 3	19.8	7.15148

Digueu en quina zona tenim més captures.

- **Comproveu si podem acceptar la normalitat dels residus i l'homogeneïtat de variàncies.**

Objectiu: Comprovar si hi ha diferències en el pes de les captures en les diferents zones estudiades.
Variables observades: i = "pes de les captures".
Volem veure si hi ha un factor extern que afecti el pes de les captures. Aquest factor extern és la distància a una zona protegida. El factor distància presenta k = 3 nivells:

- Nivell 1: zona A ($n_1 = 30$)
- Nivell 2: zona B ($n_2 = 30$) k = 3 n = 90
- Nivell 3: zona C ($n_3 = 30$)

Model matemàtic: $y_{ij} = \mu + \alpha_i + e_{ij}$ amb $\sum_i \alpha_i = 0$.

Hipòtesi de treball: Els residus són variables aleatòries independents amb una mitjana igual a 0, que segueixen una distribució Normal i amb homogeneïtat de variàncies. Cal comprovar que aquestes hipòtesi de treball es compleixen:

Normalitat dels residus; fem un contrast d'hipòtesi per veure si els residus segueixen una distribució Normal.

H_0: segueixen distribució Normal
H_1: no segueixen una distribució Normal

Utilitzarem la prova de Khi-quadrat. Valor de l'estadístic: 9.2. P-valor=0.325706
Com que el p-valor és més gran que el nivell de significació (0,05), rebutgem la hipòtesi alternativa, H_1. Amb un nivell de significació del 5% podem acceptar una distribució Normal dels residus.

Homogeneïtat de variàncies; Farem un contrast d'hipòtesis per comprovar l'homogeneïtat de les variàncies:

H_0: hi ha homogeneïtat variàncies.
H_1: no hi ha homogeneïtat variàncies.

Utilitzant la prova de Bartlett: Valor de l'estadístic = 1.1689. P-valor = 0.430525
Com que el p-valor obtingut és 0,430525 i, per tant, és més gran que el nivell de significació (0,05), rebutjarem la hipòtesi alternativa, H_1. Per tant, amb un nivell de significació del 5% podem acceptar l'homogeneïtat de les variàncies.

Un cop comprovades les hipòtesi de treball fem ara el contrast d'hipòtesis:

H_0: $\alpha_1 = \alpha_2 = \alpha_3 = 0$
H_1: alguna igualtat és falsa

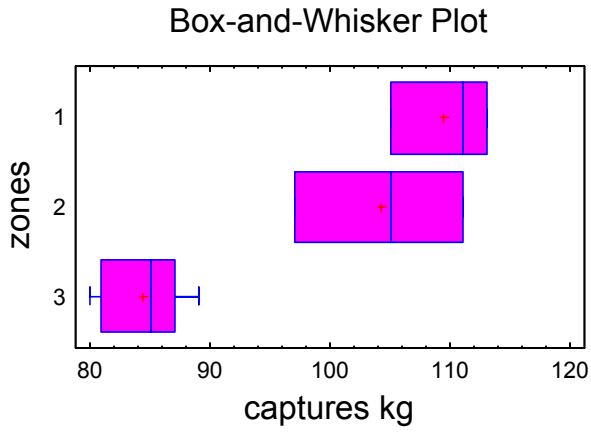
Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Between groups	1740.13	2	870.067	32.30	0.0000
Within groups	323.2	12	26.9333		
Total (Corr.)	2063.33	14			

De la taula ANOVA obtenim un p-valor entre grups igual a 0; per tant, com que és inferior al nivell de significació (0.05), rebutgem la hipòtesi nul·la, H_0.

Conclusió: podem acceptar, amb un nivell de significació del 5%, que hi ha diferències significatives en el pes de les captures a diferents distàncies de l'àrea protegida.

Per comprovar on són les diferències (en kg) farem servir un mètode de comparacions múltiples. La gràfica inclosa (Box-and-Whisker Plot) ens mostra la distribució de les dades.



Anàlisi de la gràfica: Sembla que la zona limítrofa és on es produeixen més captures, seguida de la zona que dista 500 m de l'àrea protegida, on les diferències amb l'anterior són poc acusades. La zona situada a 1000 m és on el pes de les captures és més petit, amb una diferència notable de les anteriors.

Amb les dades obtingudes farem 3 contrasts d'hipòtesi comparant en cada un si hi ha diferències significatives en el pes de les captures de les diferents zones estudiades.

Zones 1 i 2

H_0: $\alpha_1 - \alpha_2 = 0$
H_1: $\alpha_1 - \alpha_2 \neq 0$

Per aplicar el nostre criteri de decisió fem servir un interval mitjançant els valors que ens dóna la taula de comparacions múltiples. L'interval consta dels valors següents:

$$\left[y_{i.} - y_{j.} - Q; y_{i.} - y_{j.} + Q \right].$$

Si els substituïm amb els valors de la taula de comparacions múltiples, obtenim:

$$[\ 5.2 - 7.15148\ ,\ 5.2 + 7.15148\ ] = [\ -1.95148\ ,\ 12.35148\ ].$$

Segons el criteri de decisió, com que el 0 queda dins de l'interval resultant, rebutgem la hipòtesi alternativa, H_1; per tant, podem acceptar que no hi ha diferències significatives entre el pes de les captures de la zona 1 i el pes de les captures de la zona 2.

1) Zones 1 i 3

H_0: $\alpha_1 - \alpha_3 = 0$
H_1: $\alpha_1 - \alpha_3 \neq 0$

Substituint els valors de la taula per a les zones 1 i 3 a l'interval anterior obtenim:

$$[\ 25 - 7.15148\ ,\ 25 + 7.15148\ ] = [\ 17.84852\ ,\ 32.15148\ ].$$

Segons el criteri de decisió, com que el 0 no es troba dins de l'interval, rebutgem la hipòtesi nul·la, H_0. Per tant, podem acceptar que hi ha diferències significatives entre el pes de les captures a la zona 1 i la zona 3.

2) Zones 2 i 3

H_0: $\alpha_2 - \alpha_3 = 0$
H_1: $\alpha_2 - \alpha_3 \neq 0$

Substituint els valors de la taula de comparacions múltiples per a les zones 2 i 3 en l'interval anterior obtenim:

[19.8 – 7.15148 , 19.8 + 7.15148] = [12.64852 , 26.95148].

Segons el criteri de decisió, com que el 0 no es troba dins de l'interval, rebutgem la hipòtesi nul·la, H_0. Per tant, podem acceptar que hi ha diferències significatives entre el pes de les captures entre la zona 2 i la zona 3.

Ara, per saber quina és la zona que produeix més captures (en kg), fem servir el càlcul de prediccions per a tots tres nivells:

1) Nivell 1

predicció$\approx y_{..} + (y_{1.} - y_{..}) + N(0, \sqrt{QM_d})$,

predicció$\approx N(y_{1.}, \sqrt{QM_d}) = N(109.4, \sqrt{26.9333})$

Per tant l'interval de confiança al 95% és:

$\left[y_{1.} - 1.96\sqrt{QM_d}; y_{1.} + 1.96\sqrt{QM_d}\right]$ =[99.2282 , 119.5718].

2) Nivell 2

predicció$\approx y_{..} + (y_{2.} - y_{..}) + N(0, \sqrt{QM_d})$,

predicció$\approx N(y_{2.}, \sqrt{QM_d}) = N(104.2, \sqrt{26.93333})$. Així doncs,

interval = $\left[y_{2.} - 1.96\sqrt{QM_d}; y_{2.} + 1.96\sqrt{QM_d}\right]$=[94.0282 , 114.3718].

3) Nivell 3

predicció$\approx N(y_{3.}, \sqrt{QM_d})$; interval = [74.2282 , 94.5718].

Observant el rang de valors que ens donen els intervals de tots tres nivells i amb un 95% de confiança, podem acceptar que el nivell 1 (zona limítrofa a l'àrea protegida) presenta més abundància en el pes de les captures que les altres zones.

Problema

A un biòleg que treballa al Departament de Qualitat d'una empresa productora de pizzes fresques, se li ha demanat un estudi sobre la maquinària que s'encarrega de dosificar la quantitat de tomàquet que cau a cada pizza (aquesta maquinària consta de tres dosificadors situats en línia). Per fer-ho, el biòleg substitueix les masses de pizza que estan a punt de ser cobertes de tomàquet per safates que prèviament ha tarat, i les recull just abans que passin per la zona del formatge.
El treballador vol agafar i mesurar 10 mostres de cada dosificador, però és una mica despistat i se'n descuida d'una. Obté els resultats següents:

Dosificador 1: (g)	100	105	106	91	97	89	110	100	102	106
Dosificador 2: (g)	96	93	85	98	104	110	109	85	90	100
Dosificador 3: (g)	105	106	112	100	99	97	100	105	98	

a) Interpreteu la taula ANOVA i comproveu si hi ha normalitat de residus i homogeneïtat de variàncies. Digueu, també, si hi ha diferències entre els dosificadors de la maquinària i, per tant, si s'ha de plantejar reparar la maquinària.
b) Segons el gràfic de mitjanes, quina en canviaríeu si a la normativa està estipulat que el pes del tomàquet per pizza ha de ser entre [100-110] g.?
c) És balancejat el disseny?, per què?

a) Objectiu: Esbrinar si tots tres dosificadors funcionen igual.
Variables observables: grams de tomàquets distribuïts per cada dosificador (k). En aquest cas k=3 amb un nombre de mostres $n_1 = 10$, $n_2 = 10$, $n_3 = 9$ i $n = 29$.
Hipòtesi de treball: els errors són variables aleatòries independents, amb distribució normal de mitjana zero i amb homogeneïtat de variàncies.

Hipòtesi que cal testar:

H_0: $\alpha_1 = \alpha_2 = \alpha_3 = 0$
H_1: Alguna igualtat és falsa

Comprovació de les hipòtesis de treball:

- Per saber si els residus segueixen una distribució normal, mirem l'estadístic de la Khi-quadrat o el de Shapiro-Wilks W.

Prova	Valor estadístic	P-valor
Khi-quadrat	5.65	0.93
Shapiro-Wilks W	0.97	0.64

Ho: hi ha normalitat
H_1: no hi ha normalitat

Com que en tots dos casos el p-valor és superior al nivell de significació, rebutgem H_1 i, per tant podem dir que hi ha normalitat.

- Per saber si hi ha homogeneïtat de variàncies, el que fem és mirar la prova de Barlett, que en aquest cas ens dona una F= 1,12 i un p-valor = 0,24.

Ho: sí hi ha homogeneïtat de variància
H_1: no hi ha homogeneïtat de variància

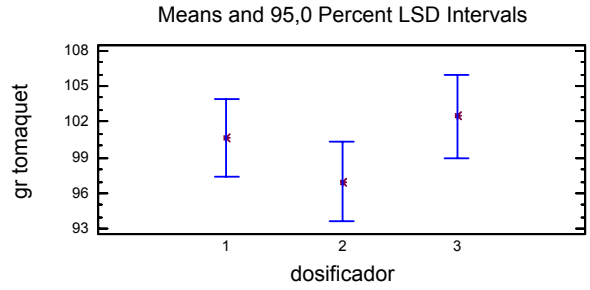
Atès que el p-valor és superior al nivell de significació de 0,05, rebutgem H_1 i, per tant, podem dir que hi ha homogeneïtat de variàncies.

Ara, si mirem l'anàlisi de la variància a la taula ANOVA, ens dóna el resultat següent:

Variabilitat	Gr. de llibertat	Suma de quadrats	Mitjana suma de quadrats	F	p-valor
Entre grups	2	147.24	73.62	1.45	0.2541
Dintre grups	26	1324l.62	50.95		
Total	28	1471.86			

Com que el p-valor = 0.2541 i és > nivell significació=0,05, rebutgem H_1, la qual cosa implica que amb una probabilitat d'error del 5% podem acceptar que no hi ha diferències significatives entre els dosificadors de la màquina i, per tant, en aquest aspecte, no cal plantejar canviar la maquinària.

b) El gràfic de mitjanes segons el mètode LSD és:



Si a la normativa estigués estipulat que la quantitat de tomàquet que hi ha d'haver a cada pizza és de 100-110 g, segons el gràfic de les mitjanes, ens plantejaríem modificar el dosificador 2 i potser també l'1, però abans de fer res, s'haurien de tenir en compte altres factors: com ara el preu reparacions etc.

d) No és un disseny balancejat perquè les grandàries mostrals dels diferents nivells del factor no són iguals. En els dosificadors 1 i 2 la grandària mostral és 10 i en el dosificador 3 és 9.

Problema

En un hivernacle es cultiva una planta ornamental i es vol conèixer quin dels fertilitzants emprats hi proporciona un major creixement. Es mesura l'alçada (en cm) de la part aèria i s'obtenen les dades següents:

Fertilitzant 1 34, 33, 39, 35, 28, 36, 31, 34, 29, 30
Fertilitzant 2 29, 30, 31, 28, 29, 26, 31, 35, 31, 26
Fertilitzant 3 25, 27, 25, 30, 27, 29, 25, 32, 21, 20

Analitzeu les diferències entre tots tres tipus d'adob. Quin en recomanaríeu?. Indiqueu com determinaríeu l'estimació de tots els paràmetres.
Suposeu la hipòtesi de normalitat de residus i la d'homogeneïtat de variàncies.

Objectiu: Comprovar si hi ha diferències en l'alçada segons el fertilitzant emprat
Variables observables: alçada de la part aèria
1 factor: tipus de fertilitzant amb k = 3 nivells i $n_1 = 10$, $n_2 = 10$, $n_3 = 10$

Model: ANOVA 1 factor $y_{ij} = \mu + \alpha_i + e_{ij}$ amb $\sum_i \alpha_i = 0$.

Hipòtesi de treball: Els residus són variables aleatòries independents amb distribució normal de mitjana 0 i amb homogeneïtat de variàncies.

Hipòtesis que cal testar:

H_0: $\alpha_1 = \alpha_2 = \alpha_3 = 0$
H_1: alguna igualtat és falsa

Anàlisi de la variància

Font	SQ	Gr. ll.	QM	F	p-Valor
Entre grups	231.267	2	115.633	10.54	0.0004
Intra grups	296.2	27	10.9704		
Total (Corr.)	527.467	29			

Com que el p-valor < 0.05, rebutgem H_0 . És a dir, amb un error del 5% es pot acceptar que hi ha diferències significatives entre els diferents fertilitzants.

Comparacions múltiples:

Contrast	Diferències	+/- Límits
1 - 2	*3.3	3.03926
1 - 3	*6.8	3.03926
2 - 3	*3.52	3.03926

Contrast :

H_0: $\alpha_1 - \alpha_2 = 0$

H_1: $\alpha_1 - \alpha_2 \neq 0$

Interval $[y_{1.} - y_{2.} - Q; y_{1.} - y_{2.} + Q]$ = [3.3 - 3.03926; 3.3 + 3.03926]
= [0.26074; 6.33926].

Com que zero no és dins de l'interval, rebutgem H_0, es pot acceptar que hi ha diferències significatives entre el fertilitzant 1 i el 2.

Contrast:

H_0: $\alpha_1 - \alpha_3 = 0$

H_1: $\alpha_1 - \alpha_3 \neq 0$

Interval $[y_{1.} - y_{3.} - Q; y_{1.} - y_{3.} + Q]$ = [6.8 - 3.03926; 6.8 + 3.03926]
= [3.76074; 9.83926].

Com que zero no és dins de l'interval, rebutgem H_0, es pot acceptar que hi ha diferències significatives entre el fertilitzant 1 i el 3.

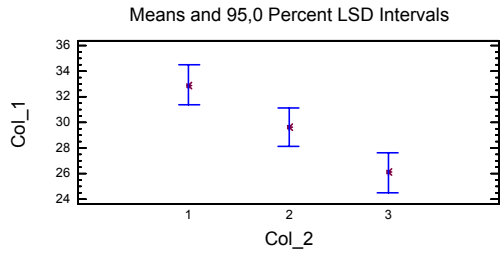
Contrast:

H_0: $\alpha_2 - \alpha_3 = 0$

H_1: $\alpha_2 - \alpha_3 \neq 0$

Interval $[y_{2.} - y_{3.} - Q; y_{2.} - y_{3.} + Q]$= [3.5 - 3.03926; 3.5 + 3.03926]
= [0.46074; 6.53926].

Com que zero no és dins de l'interval, rebutgem H_0, es pot acceptar que hi ha diferències significatives entre el fertilitzant 2 i el 3.
Si observem la gràfica de mitjanes es veu que el fertilitzant 1 és el que dóna un millor resultat.



Estimacions dels paràmetres:

μ = mitjana total = 29.53
$\alpha_1 = 32.9 - 29.53 = 3.37$
$\alpha_2 = 29.6 - 29.53 = 0.07$
$\alpha_3 = 26.1 - 29.53 = -3.43$
$\sigma = \sqrt{QM_d} = \sqrt{10.97} = 3.31$.

Problema

Es vol estudiar l'associació entre diferents al·lels d'un polimorfisme en el receptor d'estròg ens i l'osteoporosi. Per fer-ho s'ha genotipat aquest polimorfisme (er_pvu ; receptor d'estròg ens tallat amb enzim de restricció pvu) en dones postmenopàusiques que no han rebut cap tractament amb estròg ens a les quals també se'ls ha mesurat la densitat mineral òssia (dmo) com a indicador d'osteoporosi. Suposem normalitat i homogeneïtat de variàncies.

Er_pvu	dmo
3	1,022
3	0,619
2	1,297
2	1,021
2	1,011
2	0,946
3	0,721
3	0,93
1	0,937
3	0,912
1	0,731
1	0,764
2	1,076
1	0,935

a) Hi ha diferències, en quant a la densitat mineral òssia segons l'al·lel del polimorfisme?
b) Varien els resultats si considerem un nivell de significació de 0.01?
c) Quin és l'interval de confiança per a la predicció amb un nivell de significació de 0.05 per a cada al·lel?
d) Quin és l'al·lel que està associat a una densitat òssia menor?; I quin és el que està associat a una densitat òssia major?

a) Objectiu: Determinar si hi ha diferències significatives quant a la densitat mineral òssia segons l'al·lel del polimorfisme d'estròg ens.

Variable observable: densitat mineral òssia.

1 factor: al·lel el receptor d'estròg ens. Tres nivells:
R1 amb grandària mostral $n_1 = 4$
R2 amb grandària mostral $n_2 = 5$
R3 amb grandària mostral $n_3 = 5$

Model matemàtic $y_{ij} = y_{ij} = \mu + \alpha_i + e_{ij}$ amb $\sum_{i=1}^{k} \alpha_i = 0$.

Hipòtesi de treball: els residus són variables aleatòries independents amb distribució normal de mitjana 0 i homogeneïtat de variàncies.

Per comprovar la normalitat i l'homogeneïtat de variàncies hem fet el test de normalitat: Khi-quadrat = 9.571 p-valor = 0.296 i el test de Barlett: 1.053 p-valor = 0.778.

Acceptem, doncs, que hi ha normalitat i homogeneïtat de variàncies ja que el p-valor en tots dos casos és superior al nivell de significació 0.05.

Hipòtesi:
H_0: $\alpha_1 = \alpha_2 = \alpha_3 = 0$
H_1: alguna igualtat és falsa

TAULA ANOVA

Variabilitat	graus de llibertat	SQ	SQ mitjana	F
Entre grups	3 - 1 = 2	0.168	0.169/2 = 0.084	0.084/0.020 = 4.25
Dins grups	14 – 3 = 11	0.218	0.218/11 = 0.020	
Total	13	0.367	0.104	

$y_{1\cdot} = \frac{0.937 + 0.731 + 0.764 + 0.935}{4} = 0.842$

$y_{2\cdot} = \frac{1.297 + 1.021 + 1.011 + 0.946 + 1.076}{5} = 1.070$

$y_{3\cdot} = \frac{1.022 + 0.619 + 0.721 + 0.93 + 0.912}{5} = 0.841$

$y_{\cdot\cdot} = 0.923$.

Entre grups:

$Q_e = \sum_i n_i (y_{i\cdot} - y_{\cdot\cdot})^2 = 4 (0.842 – 0.923)^2 + 5 (1.070 – 0.923)^2 + 5 (0.841 – 0.923)^2$

$= 0.168$.

Dins de grups:

$$Q_d = \sum_{i,j} \left(y_{ij} - y_{i.}\right)^2 = (0.937 - 0.842)^2 + (0.731 - 0.842)^2 + (0.764 - 0.842)^2$$
$$+ (0.935 - 0.842)^2 + (1.297 - 1.070)^2 + (1.021 - 1.070)^2 + (1.011 - 1.070)^2$$
$$+ (0.946 - 1.070)^2 + (1.076 - 1.070)^2 + (1.022 - 0.841)^2 + (0.619 - 0.841)^2$$
$$+ (0.721 - 0.841)^2 + (0.93 - 0.841)^2 + (0.912 - 0.841)^2 = 0.218.$$

El valor del nostre estadístic és 4.25.
El valor trobat a les taules de $F_{2,11} = 3.98$.

El nostre estadístic cau dins la cua, per tant, amb una probabilitat d'error del 5% detectem diferències significatives en la densitat mineral òssia segons el tipus d'al·lel.

b) Amb un nivell de significació de 0.01 el valor trobat a les taules és $F_{2,11} = 7.21$.
El valor del nostre estadístic és 4.25.
Llavors amb una probabilitat d'error de l'1 % no podem detectar diferències significatives en la densitat mineral òssia segons el tipus d'al·lel.

c) Prediccions al 95%

ER 1: Predicció = $y_{..} + (y_{1.} - y_{..}) + N(0, \sqrt{QM_{r.}}) = N(y_{1.}, \sqrt{QM_{r.}})$.

Interval de confiança: [0.842 – 1.96 x 0.141 ; 0.842 + 1.96 x 0.141] = [0.566 ; 1.118].

ER 2 : Predicció = $y_{..} + (y_{2.} - y_{..}) + N(0, \sqrt{QM_{r.}}) = N(y_{2.}, \sqrt{QM_{r.}})$.

Interval de confiança: [1.070 – 1.96 x 0.141 ; 1.070 + 1.96 x 0.141] = [0.794 ; 1.346].

ER 3 : predicció = $y_{..} + (y_{3.} - y_{..}) + N(0, \sqrt{QM_{r.}}) = N(y_{3.}, \sqrt{QM_{r.}})$.

Interval de confiança: [0.841 – 1.96 x 0.141 ; 0.841 + 1.96 x 0.141] = [0.565 ; 1.117.]

Amb un 95% de confiança la predicció dels valors de dmo per a:
al·lel ER1 es troba en [0.566 ; 1.118],
al·lel ER2 es troba en [0.794 ; 1.346],
al·lel ER3 es troba en [0.565 ; 1.117].

d) D'acord amb els intervals de confiança calculats anteriorment deduïm amb un 5 % d'error el següent:

L'al·lel associat a una densitat òssia major seria el 2. En canvi, no hi ha diferències significatives entre 1 i 3 per triar quin de tots dos estaria associat a una densitat òssia menor.

Problema

En un estudi per investigar el creixement dels roures vermells americans a tres alçàries diferents (975 metres, 825 metres i 675 metres) se'n van obtenir els centímetres de la medul·la durant un període de 10 anys:

975	825	675
3.8	5.0	1.8
1.3	2.0	2.3
2.6	2.9	2.0
2.2	3.4	2.2
2.8	3.0	2.3
2.0	2	2.4
3.8	1.6	1.1
1.5	1.4	1.1
4.0	3.0	2.6
1.7	1.3	2.1

Es pot deduir que hi ha diferències en el creixement mitjà d'aquests arbres segons l'altura?. Quines són les estimacions dels paràmetres que intervenen en el model utilitzat?. Agafeu $\alpha = 0.05$

Objectiu: Comprovar si hi ha diferències significatives en el creixement dels roures a totes tres alçàries.

Variables observables: *Y= "centímetres de medul·la"*

1 factor: altitud

k = 3 nivells: 975 m $\rightarrow$ 1 $n_1 = 10$

825 m $\rightarrow$ 2 $n_2 = 10$

675 m $\rightarrow$ 3 $n_3 = 10$

$n = 30$

Model matemàtic: $y_{ij} = \mu + \alpha_i + e_{ij}$ amb $\sum_i \alpha_i = 0.$

Hipòtesi de treball: els residus són variables aleatòries independents amb distribució normal de mitjana 0 i amb homogeneïtat de variàncies.

a) Comprovació de normalitat:

H_0: sí que hi ha normalitat
H_1: no hi ha normalitat

Estadístic de $\chi^2 = 13.0$
p-valor = 0.369041

Com que p-valor > α, rebutgem H_1, per tant acceptem la normalitat.

b) Comprovació de l'homogeneïtat de variàncies:

H_0: sí que hi ha homogeneïtat de variàncies.
H_1: no hi ha homogeneïtat de variàncies.

Estadístic de Bartlett = 1.21579
p-valor = 0.08096

Com que p-valor > α, rebutgem H_1, per tant acceptem l'homogeneïtat de variàncies.

Hipòtesi que cal testar:

$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$
H_1 : alguna igualtat és falsa.

Taula ANOVA:

Variabilitat	SQ	Graus de llibertat	SQM	F-ratio	p-valor
Entre grups	2.74867	k-1= 3-1= 2	1.37433	1.61	0.2193
Dintre de grups	23.106	n-k = 30-3= 27	0.855778		
Total	25.8547	n-1= 30-1= 29			

Conclusió: com que el p-valor > α, rebutgem H_1. Per tant, amb un nivell de significació del 5% podem acceptar que no hi ha diferències significatives en el creixement dels roures segon l'alçada, és a dir, cap alçada no afavoreix el creixement dels roures.

Estimació dels paràmetres del model
A continuació s'indiquen els valors mitjans per a cada nivell i les corresponents estimacions dels paràmetres:

$y_{1.} = 2.57$ $\quad \hat{\mu} = y_{...} = 24133$ $\quad \hat{\sigma}^2 = \sqrt{QM}_d = 0.925$
$y_{2.} = 2.68$ $\quad \hat{\alpha}_1 = y_{1.} - y_{..} = 0.1567$
$y_{3.} = 1.99$ $\quad \hat{\alpha}_2 = y_{2.} - y_{..} = 2667$
$y_{..} = 24133$ $\quad \hat{\alpha}_3 = y_{3.} - y_{..} = -0.4233$