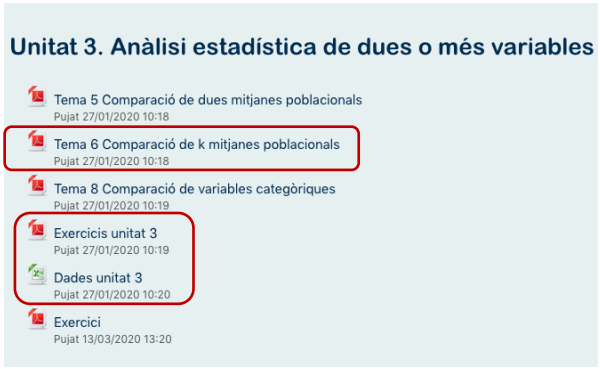
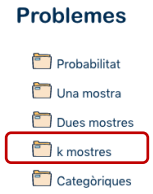
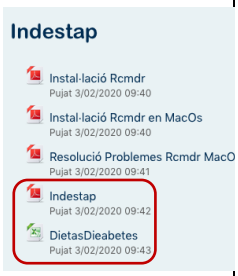
Document de Treball 4

Objectiu

L'objectiu d'aquesta sessió és:

1. Abordar el cas de k mostres independents: WELCH.
2. Abordar el cas de k mostres independents: KRUSKAL-WALLIS.
3. Exemple.
4. Activitat d'entrega.

Material de l'aulavirtual:

Teoria	Activitats amb R	Activitats sense R	Indestap
Secció: Unitat 3. Anàlisi estadística de dues o més variables. Pdf: Tema 6. Comparació de k mitjanes poblacionals. Pàgines: 1 a 27	Document: Exercicis unitat 3. Dades: Dades unitat 3. Exercicis: 10 al 24	Secció: Problemes Carpeta: k mostres.	Document: Indestap Secció del document: 2.5
			

1. K Mostres independents: WELCH

Destresa:

a) Conèixer les condicions sobre el disseny.
-Les mostres de cada grup (i=1,..., k) han de poder considerar-se mostres aleatòries de les seues corresponents poblacions.
-Les mostres han de ser independents entre si.

Utilitzarem l'exemple 4 del document de teoria:
Vol avaluar-se l'eficàcia de distintes dosis d'un fàrmac contra la hipertensió arterial, comparant-la amb la d'una dieta pobra en sal. Per fer-ho es seleccionen a l'atzar 25 hipertensos i es distribueixen aleatòriament en 5 grups. Als del primer grup no se'ls subministra cap tractament, als del segon una dieta amb un contingut pobre en sal, als del tercer una dieta sense sal, als del quart una dosi determinada del fàrmac i als del cinquè un dosi elevada del fàrmac. Les pressions arterials sistòliques dels 25 individus al finalitzar el tractament estan al full **Tensió** del llibre unitat3.xls i són:

Grup 1	Grup 2	Grup 3	Grup 4	Grup 5
180	172	163	158	147
178	152	170	155	152
179	167	158	160	143
182	160	162	161	155
181	180	170	157	160

b) Conèixer les condiciones d'aplicabilitat del test WELCH

Les k distribucions poblacionals han de ser

- **Normals o les mostres haurien de ser suficientment grans (test de shapiro, k vegades)**

X_1=pressió arterial sistòlica en el grup 1.
$\bar{X}_1$segueix una distribució normal?

Si n_1 és la grandària de la mostra de la pressió sistòlica del grup 1, aleshores si $n_1 > 30$ aleshores $\bar{X}_1$segueix una distribució normal pel teorema central de límit (TCL)

Si $n_1 \leq 30$, aleshores haurem de comprovar la normalitat de la població X_1mitjançant el test de Shapiro Wilk:
si el p-valor és més gran que el nivell de significativitat (no rebutgem H_0) aleshores X_1segueix una distribució normal i per tant, $\bar{X}_1$segueix una distribució normal que és el que busquem

Ho repetirem k voltes, per a cadascuna de les variables.

En l'exemple:

```
> tapply(Presion$Presion_arterial,Presion$Grupo,shapiro.test)
$Grupo1

        Shapiro-Wilk normality test

data:  X[[1L]]
W = 0.9868, p-value = 0.9672


$Grupo2

        Shapiro-Wilk normality test

data:  X[[2L]]
W = 0.9951, p-value = 0.9941

$Grupo3

        Shapiro-Wilk normality test

data:  X[[3L]]
W = 0.8826, p-value = 0.321


$Grupo4

        Shapiro-Wilk normality test

data:  X[[4L]]
W = 0.9738, p-value = 0.8989


$Grupo5

        Shapiro-Wilk normality test

data:  X[[5L]]
W = 0.9877, p-value = 0.9712
```

Com que en els tres casos el
p-valor > α, no es rebutja la hipòtesi nul·la H_0 (**Normalitat**)

- **NO totes elles amb la mateixa desviació típica σ (test de Levene)**

Test de Levene d'homogeneïtat de variàncies

H_0: Les variàncies de les dues poblacions són iguals($\sigma_1 = \sigma_2 = \ldots = \sigma_k$)
H_1: No s'acompleix H_0
Si p-valor ≤ α (=0.05) Rebutjar la igualtat de variàncies (H_0) → test WELCH
Si p-valor > α No rebutjat la igualtat de variàncies (H_0)

En l'exemple:
Una vegada contrastada la normalitat en totes les mostres, hem de passar a la segona comprovació, podem assumir que totes les desviacions típiques poblacionals són iguals?
Test de Levene d'homogeneïtat de variàncies

H_0: $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = \sigma_5$

H_1: No s'acompleix H_0

```
> leveneTest(Presion$Presion_arterial, Presion$Grupo, center=mean)
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value  Pr(>F)
group  4  3.7991 0.01866 *
      20
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Com p-valor < α, es rebutja H_0 (Variàncies NO són iguals)

->En aquest cas podem aplicar el test **WELCH**.

c) Saber realitzar el contrast contrast d'hipòtesi per a la igualtat de mitjanes poblacionals amb variàncies distintes (WELCH).
H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$
H_1: No s'acompleix H_0

En l'exemple (pàgina 29):
H_0: les pressions arterials mitjanes no són les mateixes en tots els grups ($\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$)
H_1: les pressions arterials mitjanes no són les mateixes en tots els grups

```
> oneway.test(Presion_arterial~Grupo, data=Presion, var.equal=FALSE)

        One-way analysis of means (not assuming equal variances)

data:  Presion_arterial and Grupo
F = 72.7905, num df = 4.000, denom df = 9.231, p-value = 5.723e-07
```

el p-valor és < 0.05, per tant, rebutgem H_0, és a dir, Existeix suficient evidència per afirmar que les pressions arterials mitjanes no són les mateixes en tots els grups.

Destresa	Pàgina del document on es troba la resposta	Exemples en el document
a)	2	3 a 6
b)	8	9 a 14
c)	28	29

Important
No anem a estudiar les comparacions a posteriori (pàgina 29 part de baix del document), per tant, en els exercicis quan us pregunten pels grups homogenis no heu de fer cas d'eixos apartats.

Activitat
Recordeu que podeu realitzar totes les activitats esmenades en l'apartat de material de l'aula virtual.

2. K Mostres independents: KRUSKAL-WALLIS

Destresa:
a) Conèixer les condicions sobre el disseny.
-Les mostres de cada grup (i=1,..., k) han de poder considerar-se mostres aleatòries de les seues corresponents poblacions.
-Les mostres han de ser independents entre si.

Exemple: La tabla següent mostra el contingut de calci (mg/100 gr. producte) de 17 productes lactis agrupats en tres categories: formatge, llet i iogurt/mousse/petit suisse (veure també en **Calci3** del llibre unitat3.xls)

formatge	llet	iogurt/mousse/petit suisse
295	120	100
740	114	85
623.46	110	96
838.47	183	120
714.77		96
809.01		127
		131

b) Conèixer les condiciones d'aplicabilitat del test KRUSKAL-WALLIS

No totes les k distribucions de les mitjanes són normals.

X_1=contingut de calci en el formatge.
$\bar{X}_1$segueix una distribució normal?

Si n_1 és la grandària de la mostra del contingut del formatge, aleshores si $n_1 > 30$ aleshores $\bar{X}_1$segueix una distribució normal pel teorema central de límit (TCL)

Si $n_1 \leq 30$, aleshores haurem de comprovar la normalitat de la població X_1mitjançant el test de Shapiro Wilk:
si el p-valor és més gran que el nivell de significativitat (no rebutgem H_0) aleshores X_1segueix una distribució normal i per tant, $\bar{X}_1$segueix una distribució normal que és el que busquem

Ho repetirem k voltes, per a cadascuna de les variables.

En l'exemple

```
> shapiro.test(Queso$Calcio)

        Shapiro-Wilk normality test

data:  Queso$Calcio
W = 0.8258, p-value = 0.09899
> shapiro.test(Leche$Calcio)

        Shapiro-Wilk normality test

data:  Leche$Calcio
W = 0.7369, p-value = 0.02893
> shapiro.test(Yogur$Calcio)

        Shapiro-Wilk normality test

data:  Yogur$Calcio
W = 0.897, p-value = 0.313
```

Existeix un p-valor < α i, per tant es rebutja H_0 (Normalitat) per a la població LLET

En aquest cas ja no cal comprovar la igualtat de variàncies. Com que una distribució poblacional no és normal, utilitzarem el test no paramètric de **Kruskal-Wallis**

c) Saber realitzar el contrast contrast d'hipòtesi en el cas no paramètric.
H_0: la mediana és la mateixa en els tres grups
H_1: No s'acompleix H_0

En l'exemple de les dietes (pàgina 30):
H_0: El contingut de calci, la mediana, en els tres grups és la mateixa
H_1: El contingut de calci, la mediana, en els tres grups no és la mateixa

```
> tapply(calci$contingut, calci$producte, median, na.rm=TRUE)
formatge  iogurts     llet
 727.385  100.000  117.000

> kruskal.test(contingut ~ producte, data=calci)

        Kruskal-Wallis rank sum test

data:  contingut by producte
Kruskal-Wallis chi-squared = 11.4942, df = 2, p-value = 0.003192
```

el p-valor és 0.00319 < 0.05, per tant, rebutgem H_0, és a dir, tenim suficient evidència estadística per rebutjar H_0, per la qual cosa existeix evidència estadística a un nivell de significativitat de 0.05 que les tres dietes no tenen el mateix efecte.

Destresa	Pàgina del document on es troba la resposta	Exemples en el document
a)	2	3 a 6
b)	8	9 a 14
c)	30	31 i 32

Important
No anem a estudiar les comparacions a posteriori (pàgina 33 del document de teoria), per tant, en els exercicis quan us pregunten pels grups homogenis no heu de fer cas d'eixos apartats.

Activitat
Recordeu que podeu realitzar totes les activitats esmenades en l'apartat de material de l'aula virtual.

3. Exemples

Exemple:

L'aparició d'hepatitis C recurrent és freqüent en persones que reben un trasplantament de fetge. Les biòpsies del fetge segueixen sent la millor manera de vigilar la progressió de la malaltia. A causa de les limitacions d'una biòpsia hepàtica, existeix un interès en el desenvolupament de marcadors no invasius de fibrosis hepàtica.

El fitxer fibrosis.xlsx conté dades de 323 pacients hepàtics d'un hospital de la ciutat de València als quals se'ls va realitzar una biòpsia hepàtica per determinar la presència de fibrosis i la seua severitat (absència, lleu, moderada i important) i, per a cadascun dels pacients, tenim els valors de tres variables: l'albúmina (*ALB*), la gamma glutamil transpeptidasa (*GGT*) i el sodi (*NA*).

Per a cadascuna de les tres variables *ALB, GGT* i *NA*, segons la severitat de la fibrosis, contesta a les següents preguntes, amb un nivell de significativitat $\alpha = 0.05$:

i) Analitza les condicions del disseny i de les distribucions poblacionals i tria raonadament la tècnica estadística adequada per comparar els nivells mitjans de cada variable segons la severitat de la fibrosis.

ii) Amb la tècnica estadística que hages seleccionat a l'apartat anterior, investiga si hi ha diferències entre els resultats dels diferents nivells de severitat de la fibrosis i, en cas d'haver-la, estableix els grups homogenis.

Primer importem les dades i cridem fibrosis al conjunt de dades.

Contestem als dos subapartats, per a cada variable per separat.

Variable *GGT*:

i) Ja sabem que, dels 4 grups, solament l'important és xicotet ($n_{absència}$ = 141, $n_{important}$ = 20, n_{lleu} = 110, $n_{moderada}$ = 52). Per tant, solament hem de fer la prova de normalitat al grup "important". El contrast de normalitat és:

H_0: la GGT és normal per al grup important

H_A: la GGT no és normal per al grup important

```
> with(fibro_importante, shapiro.test(ggt))

        Shapiro-Wilk normality test

data:  ggt
W = 0.9402, p-value = 0.2418
```

Com el p-valor és 0.2418, major que α = 0.05, hi ha evidència estadística que la GGT és normal en el grup amb fibrosis important. Per tant, podem aplicar mètodes paramètrics a aquestes mostres: ANOVA o Welch.

Per decidir si utilitzem ANOVA o Welch, hem d'utilitzar el test del Levene (Estadístics -> Variàncies -> Test de Levene) per veure si podem assumir que les variàncies de la GGT són iguals o no segons la severitat de la fibrosis:

H_0: $\sigma^2_{absència} = \sigma^2_{lleu} = \sigma^2_{moderada} = \sigma^2_{important}$

H_A : No totes les variàncies són iguals

```
> leveneTest(ggt ~ fibrosis, data=fibrosis, center="mean")
Levene's Test for Homogeneity of Variance (center = "mean")
       Df F value    Pr(>F)
group   3  7.3543 8.834e-05 ***
      319
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

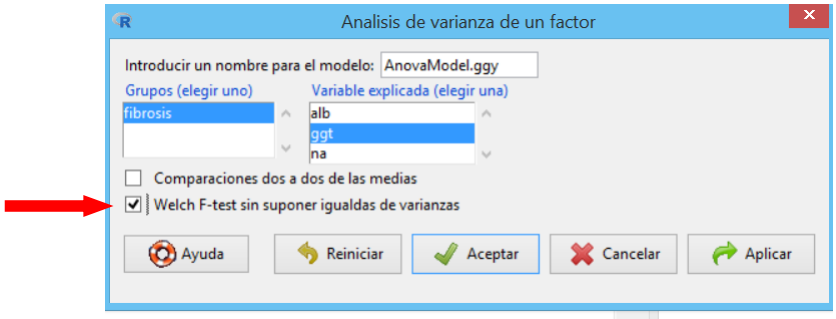
Com el p-valor és 8.834×10^{-5}, menor que α = 0.05, hi ha evidència estadística de que les variàncies no són totes iguals. Per tant, hem d'utilitzar Welch.

i) El contrast d'hipòtesi (que coincideix amb el d'ANOVA) a resoldre és:

H_0: $\mu_{absència} = \mu_{lleu} = \mu_{moderada} = \mu_{important}$

H_A : No totes les mitjanes són iguals

Triem el menú Estadístics -> Mitjanes -> ANOVA d'un factor, marcant l'opció de Welch:



En la sortida, primer trobem la taula d'ANOVA i els estadístics descriptius per grup. Per a Welch, hem de centrar-nos en:

```
> oneway.test(ggt ~ fibrosis, data=fibrosis) # Welch test

        One-way analysis of means (not assuming equal variances)

data:  ggt and fibrosis
F = 4.7903, num df = 3.000, denom df = 99.049, p-value = 0.003695
```

Com el p-valor és 0.003695, menor que α = 0.05, hi ha evidència estadística de que les mitjanes del GGT en funció de la fibrosis no són totes iguals.

Variable *NA*:

i) Com en els casos anteriors, solament hem de fer la prova de normalitat al grup "important". El contrast de normalitat és:

H_0: el sodi és normal per al grup important

H_A : el sodi no és normal per al grup important

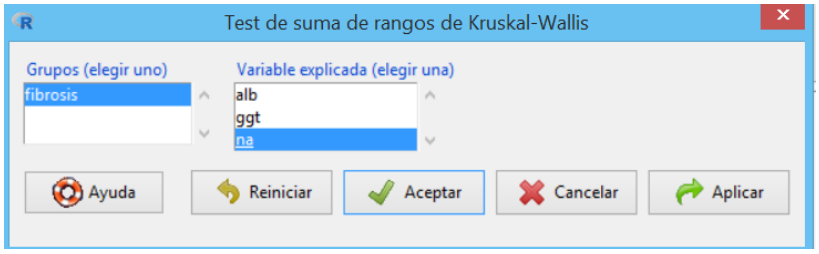
```
> with(fibro_importante, shapiro.test(na))

        Shapiro-Wilk normality test

data:  na
W = 0.857, p-value = 0.006999
```

Com el p-valor és 0.006999, menor que α = 0.05, hi ha evidència estadística que el sodi no és normal en el grup amb fibrosis important. Per tant, hem d'aplicar mètodes no paramètrics a aquestes mostres: Kruskal-Wallis.

ii) El contrast d'hipòtesi a resoldre és:

H_0: El sodi es distribueix igual en tots els nivells de fibrosis
H_A : El sodi no es distribueix igual en tots els nivells de fibrosis

Triem el menú Estadístics -> Tests no paramètrics -> Test de Kruskal-Wallis:



La sortida és:

```
> kruskal.test(na ~ fibrosis, data=fibrosis)

        Kruskal-Wallis rank sum test

data:  na by fibrosis
Kruskal-Wallis chi-squared = 4.7636, df = 3, p-value = 0.19
```

Com el p-valor és 0.19, major que α = 0.05, hi ha evidència estadística de que el sodi es distribueix igual en tots els graus de fibrosis.

4. Activitat d'entrega

Podeu trobar en l'aulavirtual un exercici que s'anomena Activitat2, és un exercici d'un examen de Juliol de 2015. Crearé una tasca perquè entregueu tot l'exercici i els pugeu a l'aulavirtual. S'ha d'entregar en word, i heu d'assenyalar la població, mostra, variables i els p-valors així com les taules utilitzades.