Principales técnicas estadísticas

(ver 2019-2020)

Emilio Letón y Elisa M. Molanes-López

2-Oct-2019

Página legal

PRINCIPALES TÉCNICAS ESTADÍSTICAS

Madrid, versión 2019-2020

ISBN electrónico: xxx.
Edición digital (epub): xxx.

Prefacio

Este material está diseñado con el paradigma del grupo de innovación docente miniXmodular de generar material mini y modular, según se puede ver en la página web www.minixmodular.ia.uned.es, donde se introduce, entre otros, el concepto de mini-libro electrónico modular.

En este mini-libro se contemplan dos variables X e Y donde X es la variable explicativa (también llamada independiente o factor de exposición) e Y es la variable respuesta (también llamada dependiente, resultado u objetivo). De forma esquemática se escribirá $Y \leftarrow X$ para indicar que se está ante un modelo estadístico que establece que se está intentando explicar Y a través de X.

Los mini-capítulos de este mini-libro se estructuran dependiendo de la naturaleza de la variable respuesta, considerando que ésta pueda ser dicotómica, continua, nominal u ordinal.

Antes de comenzar es conveniente, tener presentes algunos conceptos estadísticos básicos (ver, por ejemplo [3] ó [5]) y que básicamente son los siguientes:

- En Estadística (frecuentista) a menudo se está interesado en parámetros poblacionales θ que son constantes desconocidas y que se pretenden conocer.
- Dicho conocimiento se realiza a través de estadísticos muestrales $\widehat{\Theta}$ que son variables aleatorias (v.a.) ya que varían de muestra a muestra.
- Si se conoce la distribución (el modelo teórico de probabilidad) de $\widehat{\Theta}$ se pueden calcular intervalos de confianza para el parámetro desconocido θ y estadísticos de contraste que permitan decidir sobre afirmaciones acerca de θ.
- Para lograr los objetivos anteriores hay que recoger datos experimentales que se disponen en filas (individuos) y columnas (variables).

Por ejemplo, en el caso de que el parámetro poblacional θ de interés sea la proporción poblacional π, dada por $\pi = P(Y = 1)$, siendo Y la v.a. resultado codificada como 1 en el caso de presencia de Y y como 0 en el caso de ausencia de Y, se considera el estadístico $\widehat{\Theta}$ dado por la v.a. proporción muestral $\widehat{\Pi}$ con

$$\widehat{\Theta} = \widehat{\Pi} = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

donde se entiende que Y_i es la v.a. Y asociada al i-ésimo individuo. Esta situación se da, por ejemplo, cuando se quiere estudiar la prevalencia de una enfermedad (probabilidad de estar enfermo), donde Y es "Enfermedad" codificada con $Y = 1$ para "Sí enfermo" y con $Y = 0$ para "No enfermo", y por tanto cada Y_i está reflejando si el individuo i está o no enfermo. En esta situación los datos experimentales tendrán el patrón dado en la tabla 1, donde el número de individuos 1 es r_1, el número de individuos 0 es r_0 y el número total de individuos es $n = r_1 + r_0$.

Y
1
1
...
1
0
0
...
0

Tabla 1: Datos genéricos de una variable dicotómica

El marco teórico poblacional que está generando los datos experimentales de la tabla 1 es el que especifica que Y es una v.a. Bernoulli y por tanto todas las Y_i también lo son, siendo independientes entre sí (y por tanto incorreladas). La v.a. Y verifica que su media poblacional viene dada por

$$E[Y] = 1 \cdot P(Y=1) + 0 \cdot P(Y=0) = 1 \cdot \pi + 0 \cdot (1-\pi) = \pi$$

y que su varianza poblacional viene dada por

$$V[Y] = E[(Y - E[Y])^2] = E[Y^2] - E^2[Y] = 1^2 \cdot P(Y=1) + 0^2 \cdot P(Y=0) - \pi^2 = \pi(1-\pi).$$

Por otra parte, $\sum_{i=1}^{n} Y_i$ es una v.a. Binomial (por ser suma de v.a. Bernoulli independientes) que verifica que su media poblacional viene dada por

$$E\Big[\sum_{i=1}^{n} Y_i\Big] = \sum_{i=1}^{n} E[Y_i] = \sum_{i=1}^{n} \pi = n\pi$$

y que su varianza poblacional viene dada por

$$V\Big[\sum_{i=1}^{n} Y_i\Big] \overset{incorr.}{=} \sum_{i=1}^{n} V[Y_i] = \sum_{i=1}^{n} \pi(1-\pi) = n\pi(1-\pi).$$

En lo que respecta a $\widehat{\Pi} = \overline{Y}$, utilizando el teorema central del límite, se tiene que $\widehat{\Pi} = \overline{Y}$ sigue una distribución normal. Además, es fácil ver que la media teórica de $\widehat{\Pi} = \overline{Y}$ está dada por

$$E[\widehat{\Pi}] = E[\overline{Y}] = E\Big[\frac{1}{n}\sum_{i=1}^{n} Y_i\Big] = \frac{1}{n} E\Big[\sum_{i=1}^{n} Y_i\Big] = \frac{1}{n} n\pi = \pi$$

y la varianza teórica de $\widehat{\Pi} = \overline{Y}$ está dada por

$$V[\widehat{\Pi}] = V[\overline{Y}] = V\Big[\frac{1}{n}\sum_{i=1}^{n} Y_i\Big] = \frac{1}{n^2} V\Big[\sum_{i=1}^{n} Y_i\Big] \overset{incorr.}{=} \frac{1}{n^2} n\pi(1-\pi) = \pi(1-\pi)\frac{1}{n},$$

con lo que el error estándar de $\widehat{\Pi} = \overline{Y}$ es

$$EE[\widehat{\Pi}] = EE[\overline{Y}] = \sqrt{V[\overline{Y}]} = \sqrt{\pi(1-\pi)\frac{1}{n}}.$$

Con la información anterior se define un nuevo estadístico dado por

$$\frac{\widehat{\Pi} - E[\widehat{\Pi}]}{EE[\widehat{\Pi}]} = \frac{\widehat{\Pi} - \pi}{\sqrt{\pi(1-\pi)\frac{1}{n}}}$$

que sigue una $N(0,1)$ y que servirá para construir un intervalo de confianza y un contraste de hipótesis para π.

En el caso del intervalo de confianza hay que estimar $EE[\widehat{\Pi}]$ a través de

$$\widehat{EE}[\widehat{\Pi}] = \sqrt{\widehat{\pi}(1-\widehat{\pi})\frac{1}{n}}$$

con $\widehat{\pi} = \frac{r_1}{n}$, con lo que el intervalo de confianza al $(1-\alpha)\,\%$ está dado por

$$\begin{aligned} IC(1-\alpha)\,\%(\pi) &= \Big(\widehat{\pi} \mp z_{1-\alpha/2}\sqrt{\widehat{\pi}(1-\widehat{\pi})\frac{1}{n}}\Big) \\ &= \Big(\frac{r_1}{n} \mp z_{1-\alpha/2}\sqrt{\frac{r_1}{n}\Big(1-\frac{r_1}{n}\Big)\frac{1}{n}}\Big). \end{aligned}$$

En el caso del contraste de hipótesis

$$\begin{aligned} H_0 &: \pi = \pi_t \\ H_1 &: \pi \neq \pi_t \end{aligned}$$

hay que estimar $EE[\widehat{\Pi}]$ bajo H_0 (con lo que $\pi = \pi_t$) a través de

$$\widehat{EE}_0[\widehat{\Pi}] = \sqrt{\pi_t(1-\pi_t)\frac{1}{n}}$$

y el contraste de hipótesis al nivel de significación α se realiza con

$$z = \frac{\frac{r_1}{n} - \pi_t}{\sqrt{\pi_t(1-\pi_t)\frac{1}{n}}}$$

tomando la decisión de que si $|z| > z_{1-\alpha/2}$ se rechaza la H_0.

Como recomendación práctica, la asunción del test z de una proporción es que el valor $n\widehat{\pi}$ sea mayor que 5 lo que, intuitivamente, corresponde a tamaño muestral grande y proporción no muy extrema.

Por último, conviene mencionar que a lo largo de este mini-libro se presenta código en el lenguaje R [4] para realizar los cálculos estadísticos. Existen numerosas páginas web y libros que introducen este lenguaje; dos referencias clásicas que combinan Estadística y R son [1] y [2]. No obstante, conviene señalar que el código que se utiliza en este mini-libro es muy sencillo y autoexplicativo.

Mini-capítulo 1

Respuesta dicotómica

En este mini-capítulo se contemplan el test z de diferencia de dos proporciones y el test exacto de Fisher, que están enmarcados en el esquema $D \leftarrow D$ que indica que se está intentando explicar la variable respuesta Y dicotómica a través de la variable explicativa X dicotómica.

Esta situación se da, por ejemplo, cuando se quiere estudiar si la curación de una enfermedad está influida por el hecho de dar un tratamiento u otro. Para ello se realiza un estudio en n individuos a los que se les suministra el tratamiento 1 (codificado con 1 en la variable tratamiento) a n_1 individuos y el tratamiento 2 (codificado con 2 en la v.a. tratamiento) a n_2 individuos con $n = n_1 + n_2$ y se observa quién se cura (codificándolo con 1 en la variable respuesta) y quién no (codificándolo con 0 en la variable respuesta. En esta situación los datos experimentales tendrán el patrón dado en la tabla 1.1, donde el número de individuos con perfil $Y = 1$ y $X = 1$ es a_1, con perfil $Y = 0$ y $X = 1$ es $n_1 - a_1$, con perfil $Y = 1$ y $X = 2$ es a_2 y con perfil $Y = 0$ y $X = 2$ es $n_2 - a_2$.

Y	X
1	1
...	...
1	1
0	1
...	...
0	1
1	2
...	...
1	2
0	2
...	...
0	2

Tabla 1.1: Datos genéricos del esquema $D \leftarrow D$

La pregunta de si X influye en Y se traduce en comparar la proporción poblacional de curaciones con el tratamiento 1 (que se denota por π_1) y la propoción poblacional de curaciones con el tratamiento 2 (que se denota por π_2).

1.1. z de diferencia de dos proporciones

El test z de diferencia de dos proporciones está enmarcado en el esquema $D \leftarrow D$ y está indicado en tamaños muestrales grandes. En este caso, el parámetro poblacional θ de interés es la diferencia de proporciones poblacionales, entre π_1 y π_2 (es decir, $\theta = \pi_1 - \pi_2$). Para ello se considera el estadístico $\widehat{\Theta}$ dado por la v.a. $\widehat{\Pi}_1 - \widehat{\Pi}_2$ (es decir, $\widehat{\Theta} = \widehat{\Pi}_1 - \widehat{\Pi}_2$), siendo $\widehat{\Pi}_1$ y $\widehat{\Pi}_2$ las proporciones muestrales entendidas como v.a. ya que variarán de muestra a muestra. Es sabido (utilizando el teorema central del límite) que $\widehat{\Pi}_1$ y $\widehat{\Pi}_2$ siguen sendas distribuciones normales y, por lo tanto, será normal la diferencia $\widehat{\Pi}_1 - \widehat{\Pi}_2$. Además, es fácil ver que la media teórica de $\widehat{\Pi}_1 - \widehat{\Pi}_2$ está dada por

$$E[\widehat{\Pi}_1 - \widehat{\Pi}_2] = E[\widehat{\Pi}_1] - E[\widehat{\Pi}_2] = \pi_1 - \pi_2$$

y la varianza teórica de $\widehat{\Pi}_1 - \widehat{\Pi}_2$ está dada por

$$V[\widehat{\Pi}_1 - \widehat{\Pi}_2] = V[\widehat{\Pi}_1] + V[\widehat{\Pi}_2] - 2Cov(\widehat{\Pi}_1, \widehat{\Pi}_2) = \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2},$$

ya que independencia implica incorrelación que equivale a covarianza nula, con lo que el error estándar de $\widehat{\Pi}_1 - \widehat{\Pi}_2$ es

$$EE[\widehat{\Pi}_1 - \widehat{\Pi}_2] = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}.$$

Con la información anterior se define un nuevo estadístico dado por

$$\frac{\widehat{\Pi}_1 - \widehat{\Pi}_2 - E[\widehat{\Pi}_1 - \widehat{\Pi}_2]}{EE[\widehat{\Pi}_1 - \widehat{\Pi}_2]} = \frac{\widehat{\Pi}_1 - \widehat{\Pi}_2 - (\pi_1 - \pi_2)}{\sqrt{\pi_1(1-\pi_1)\frac{1}{n_1} + \pi_2(1-\pi_2)\frac{1}{n_2}}}$$

que sigue una $N(0,1)$ y que servirá para construir un intervalo de confianza y un contraste de hipótesis para $\pi_1 - \pi_2$.

En el caso del intervalo de confianza hay que estimar $EE[\widehat{\Pi}_1 - \widehat{\Pi}_2]$ a través de

$$\widehat{EE}[\widehat{\Pi}_1 - \widehat{\Pi}_2] = \sqrt{\widehat{\pi}_1(1-\widehat{\pi}_1)\frac{1}{n_1} + \widehat{\pi}_2(1-\widehat{\pi}_2)\frac{1}{n_2}}$$

con $\widehat{\pi}_1 = \frac{a_1}{n_1}$ y $\widehat{\pi}_2 = \frac{a_2}{n_2}$, con lo que el intervalo de confianza al $(1-\alpha)\,\%$ está dado por

$$\begin{aligned} IC(1-\alpha)\,\%(\pi_1 - \pi_2) &= \left(\widehat{\pi}_1 - \widehat{\pi}_2 \mp z_{1-\alpha/2}\sqrt{\widehat{\pi}_1(1-\widehat{\pi}_1)\frac{1}{n_1} + \widehat{\pi}_2(1-\widehat{\pi}_2)\frac{1}{n_2}}\right) \\ &= \left(\frac{a_1}{n_1} - \frac{a_2}{n_2} \mp z_{1-\alpha/2}\sqrt{\frac{a_1}{n_1}\Big(1-\frac{a_1}{n_1}\Big)\frac{1}{n_1} + \frac{a_2}{n_2}\Big(1-\frac{a_2}{n_2}\Big)\frac{1}{n_2}}\right). \end{aligned}$$

En el caso del contraste de hipótesis

$$\begin{aligned} H_0 &: \pi_1 - \pi_2 = 0 \\ H_1 &: \pi_1 - \pi_2 \neq 0 \end{aligned}$$

hay que estimar $EE[\widehat{\Pi}_1 - \widehat{\Pi}_2]$ bajo H_0 (con lo que $\pi_1 = \pi_2$ y se suponen igual a un valor común π_c que se estima por $\widehat{\pi}_c = \frac{n_1\widehat{\pi}_1 + n_1\widehat{\pi}_2}{n_1+n_2} = \frac{a_1+a_2}{n_1+n_2}$) a través de

$$\begin{aligned} \widehat{EE}_0[\widehat{\Pi}_1 - \widehat{\Pi}_2] &= \sqrt{\widehat{\pi}_c(1-\widehat{\pi}_c)\frac{1}{n_1} + \widehat{\pi}_c(1-\widehat{\pi}_c)\frac{1}{n_2}} \\ &= \sqrt{\widehat{\pi}_c(1-\widehat{\pi}_c)\Big(\frac{1}{n_1} + \frac{1}{n_2}\Big)} = \sqrt{\frac{a_1+a_2}{n_1+n_2}\Big(1-\frac{a_1+a_2}{n_1+n_2}\Big)\Big(\frac{1}{n_1} + \frac{1}{n_2}\Big)} \end{aligned}$$

y el contraste de hipótesis al nivel de significación α se realiza con

$$z = \frac{\frac{a_1}{n_1} - \frac{a_2}{n_2}}{\sqrt{\frac{a_1+a_2}{n_1+n_2}\Big(1-\frac{a_1+a_2}{n_1+n_2}\Big)\Big(\frac{1}{n_1} + \frac{1}{n_2}\Big)}}$$

tomando la decisión de que si $|z| > z_{1-\alpha/2}$ se rechaza la H_0.

Como recomendación práctica, las asunciones del test z de diferencia de dos proporciones son que todos los valores $n_1\widehat{\pi}_1$, $n_1(1-\widehat{\pi}_1)$, $n_2\widehat{\pi}_2$ y $n_2(1-\widehat{\pi}_2)$, sean mayores que 5 lo que, intuitivamente, corresponde a tamaños muestrales grandes y proporciones no muy extremas.

Ejemplo

1. Se ha realizado un estudio para ver si influye una variable explicativa dicotómica en una variable respuesta dicotómica. Han participado en el estudio 90 individuos. Los datos experimentales se dan en la tabla 1.2, donde el número de individuos con perfil $rta = 1$ y $exp = 1$ es 8, con perfil $rta = 1$ y $exp = 2$ es 14, con perfil $rta = 0$ y $exp = 1$ es 42 y con perfil $rta = 0$ y $exp = 2$ es 26.

   SOLUCIÓN:

rta	*exp*
1	1
...	...
1	1
1	2
...	...
1	2
0	1
...	...
0	1
0	2
...	...
0	2

Tabla 1.2: Datos observados del esquema $D \leftarrow D$

En primer lugar se muestran los cálculos detallados para el intervalo de confianza al 95 % dado por

$$\begin{aligned} IC95\,\%(\pi_1-\pi_2) &= \Big(\frac{8}{50}-\frac{14}{40} \mp 1.96\sqrt{\frac{8}{50}\Big(1-\frac{8}{50}\Big)\frac{1}{50}+\frac{14}{40}\Big(1-\frac{14}{40}\Big)\frac{1}{40}}\Big) \\ &= (0.16-0.35 \mp 1.96\cdot 0.0915) \\ &= (-0.3694,-0.0106) \end{aligned}$$

y para el contraste de hipótesis dado por

$$z=\frac{\frac{8}{50}-\frac{14}{50}}{\sqrt{\frac{8+14}{50+40}\Big(1-\frac{8+14}{50+40}\Big)\Big(\frac{1}{50}+\frac{1}{40}\Big)}}=-2.0841$$

con lo que al ser $|z| = 2.0841 > 1.96$ se rechaza la hipótesis nula de igualdad de proporciones y se concluye que la variable explicativa influye en la variable respuesta de forma significativa con un p-valor$= 2\cdot(1-pnorm(abs(-2.0841)) = 0.0371$.

En segundo lugar se muestra la sintaxis en R para el cálculo del IC y del p-valor del test z de diferencia de dos proporciones. Si se supone que los datos sin agrupar están en el fichero d_d_1.txt la sintaxis es

```
rm(list=ls())

datos=read.table('d_d_1.txt',header=T)
attach(datos)

tabla=table(-rta,exp); tabla

ind1=which(exp==1);
ind2=which(exp==2);
ind11=which(rta==1 & exp==1);
ind12=which(rta==1 & exp==2);
a1=length(rta[ind11]); a1
a2=length(rta[ind12]); a2
n1=length(rta[ind1]); n1
n2=length(rta[ind2]); n2

alfa=0.05
pi1=a1/n1
pi2=a2/n2

ee=sqrt(pi1*(1-pi1)*(1/n1)+pi2*(1-pi2)*(1/n2)); ee

ic1=pi1-pi2-qnorm(1-alfa/2)*ee; ic1
ic2=pi1-pi2+qnorm(1-alfa/2)*ee; ic2
```

```
pit=(a1+a2)/(n1+n2); pit

ee0=sqrt(pit*(1-pit)*(1/n1+1/n2))
z=(pi1-pi2)/ee0; z
z*z

p_valor= 2*(1-pnorm(abs(z))); p_valor

a=c(a1,a2); a
n=c(n1,n2); n
prop.test(a,n,correct=F)
```

y si se supone que se introducen los datos agrupados directamente hay que sustituir las primeras líneas de la sintaxis anterior por

```
rm(list=ls())

a1=8; n1=50
a2=14; n2=40
```

1.2. Exacta de Fisher

El test exacto de Fisher está enmarcado en el esquema $D \leftarrow D$ y está indicado en tamaños muestrales pequeños donde no se cumplen las asunciones del test z de diferencia de dos proporciones. A partir de los datos experimentales se construye la tabla de datos cruzados genérica dada en la tabla 1.3.

	$X=1$	$X=2$	
$Y=1$	a	b	$r_1 = a+b$
$Y=0$	c	d	$r_2 = c+d$
	n_1	n_2	n

Tabla 1.3: Tabla con datos genéricos del esquema $D \leftarrow D$

A partir de la tabla 1.3, se construyen todas las tablas posibles 2x2 con celdas a', b', c' y d' siendo:

- $0 \leq a' \leq \min\{n_1, r_1\}$.
- $b' = r_1 - a'$.
- $c' = n_1 - a'$.
- $d' = r_2 - c'$.

A continuación, a partir de dichas tablas, se calcula

$$p_{a'} = \frac{r_1! r_2! n_1! n_2!}{n! a'! b'! c'! d'!}$$

donde $x!$ indica el factorial de x que se calcula como $x \cdot (x-1) \cdot (x-2) \cdot \ldots \cdot 1$ (por ejemplo, $5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$). Por último el p-valor viene dado por

$$p\text{-valor} = \sum_{p_{a'} \leq p_a} p_{a'}.$$

Ejemplo

1. Se ha realizado un estudio para ver si influye una variable explicativa dicotómica en una variable respuesta dicotómica. Han participado en el estudio 42 individuos. Los datos experimentales se dan en la tabla 1.4, donde el número de individuos con perfil $rta = 1$ y $exp = 1$ es 4, con perfil $rta = 1$ y $exp = 2$ es 1, con perfil $rta = 0$ y $exp = 1$ es 16 y con perfil $rta = 0$ y $exp = 2$ es 21.

   SOLUCIÓN:

rta	exp
1	1
...	...
1	1
1	2
...	...
1	2
0	1
...	...
0	1
0	2
...	...
0	2

Tabla 1.4: Datos observados del esquema $D \leftarrow D$

	$exp = 1$	$exp = 2$	
$rta = 1$	4	1	5
$rta = 0$	16	21	37
	20	22	42

Tabla 1.5: Tabla con datos observados del esquema $D \leftarrow D$

a'	b'	c'	d'	n	$p_{a'}$
0	5	20	17	42	0.0310
1	4	19	18	42	0.1720
2	3	18	19	42	0.3440
3	2	17	20	42	0.3096
4	1	16	21	42	0.1253
5	0	15	22	42	0.0182

Tabla 1.6: Tablas variando a' en el esquema $D \leftarrow D$

En primer lugar se muestran los cálculos intermedios detallados que toman por base la tabla de datos cruzados dada en la tabla 1.5 y que sirven para construir la tabla 1.6.

Utilizando la tabla 1.6, el cálculo del p-valor es

$$p\text{-valor} = \sum_{p_{a'} \leq p_a} p_{a'} = \sum_{p_{a'} \leq 0.1253} p_{a'} = 0.1253 + 0.0182 + 0.0310 = 0.1745 > 0.05,$$

por lo que no hay evidencia para concluir que la variable explicativa influya en la variable respuesta de forma significativa.

En segundo lugar se muestra la sintaxis en R para el cálculo del p-valor del test exacto de Fisher. Si se supone que los datos sin agrupar están en el fichero d_d_2.txt la sintaxis es

```
rm(list=ls())

datos=read.table('d_d_2.txt',header=T)
attach(datos)

tabla=table(rta,exp)

fisher.test(x=tabla,alternative="two.sided")
```

y si se supone que se introducen los datos agrupados directamente la sintaxis es

```
rm(list=ls())
```

```
tabla <- matrix(c(4,1,16,21),ncol=2,byrow=TRUE)
colnames(tabla) <- c("exp=1","exp=2")
rownames(tabla) <- c("rta=1","rta=0")
tabla <- as.table(tabla); tabla
fisher.test(x=tabla,alternative="two.sided")
```

Mini-capítulo 2

Respuesta continua

En este mini-capítulo se contemplan el test t de Student y el test ANOVA de un factor (ANOVA significa "Analysis of variance"; a veces, por ejemplo en [3], se traduce por ADEVA). El test t de Student está enmarcado en el esquema $C \leftarrow D$ que indica que se está intentando explicar la variable respuesta Y continua a través de la variable explicativa X dicotómica y el test ANOVA de un factor está enmarcado en el esquema $C \leftarrow N$ que indica que se está intentando explicar la variable respuesta Y continua a través de la variable explicativa X nominal.

2.1. t de Student

El test t de Student de diferencia de dos medias está enmarcado en el esquema $C \leftarrow D$ y está indicado para variables Y normales en cada categoría de la variable dicotómica X. Esta situación se da, por ejemplo, cuando se quiere estudiar si una variable bioquímica está influida por el hecho de recibir un tipo u otro de tratamiento. Para ello se realiza un estudio en n individuos a los que se les asigna el tratamiento 1 (codificado con 1 en la variable tratamiento) a n_1 individuos y el tratamiento 2 (codificado con 2) a n_2 individuos con $n = n_1 + n_2$ y se observa qué valor continuo tienen en una determinada variable bioquímica. En esta situación los datos experimentales tendrán el aspecto dado en la tabla 2.1, donde el número de individuos con $X = 1$ es n_1 y con $X = 2$ es n_2.

Y	X
y_{11}	1
$\dots$	$\dots$
y_{1n_1}	1
y_{21}	2
$\dots$	$\dots$
y_{2n_2}	2

Tabla 2.1: Datos genéricos del esquema $C \leftarrow D$

La pregunta de si X influye en Y se traduce en comparar la media poblacional de la variable bioquímica con la dieta 1 (que se denota por μ_1) y la media poblacional de la variable bioquímica con la dieta 2 (que se denota por μ_2). En este caso, el parámetro poblacional θ de interés es la diferencia de medias poblacionales, entre μ_1 y μ_2 (es decir, $\theta = \mu_1 - \mu_2$). Para ello se considera el estadístico $\widehat{\Theta}$ dado por la v.a. $\overline{Y}_1 - \overline{Y}_2$ (es decir, $\widehat{\Theta} = \overline{Y}_1 - \overline{Y}_2$), siendo $\overline{Y}_1$ y $\overline{Y}_2$ las medias muestrales entendidas como v.a. ya que variarán de muestra a muestra. Es sabido (utilizando que combinación lineal de normales es normal) que $\overline{Y}_1$ y $\overline{Y}_2$ siguen sendas distribuciones normales y, por lo tanto, será normal la diferencia $\overline{Y}_1 - \overline{Y}_2$. Además, es fácil ver que la media teórica de $\overline{Y}_1 - \overline{Y}_2$ está dada por

$$E[\overline{Y}_1 - \overline{Y}_2] = E[\overline{Y}_1] - E[\overline{Y}_2] = \mu_1 - \mu_2$$

y la varianza teórica de $\overline{Y}_1 - \overline{Y}_2$ está dada por

$$V[\overline{Y}_1 - \overline{Y}_2] = V[\overline{Y}_1] + V[\overline{Y}_2] - 2Cov(\overline{Y}_1, \overline{Y}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2},$$

ya que independencia implica incorrelación que equivale a covarianza nula, con lo que el error estándar de

$\overline{Y}_1 - \overline{Y}_2$ es

$$EE[\overline{Y}_1 - \overline{Y}_2] = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Con la información anterior se define un nuevo estadístico dado por

$$\frac{\overline{Y}_1 - \overline{Y}_2 - E[\overline{Y}_1 - \overline{Y}_2]}{EE[\overline{Y}_1 - \overline{Y}_2]} = \frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

que sigue una $N(0,1)$ y que servirá para construir un intervalo de confianza y un contraste de hipótesis para $\mu_1 - \mu_2$.

En el caso del intervalo de confianza hay que estimar $EE[\overline{Y}_1 - \overline{Y}_2]$. Hay dos situaciones posibles:

- Que se pueda asumir $\sigma_1^2 = \sigma_2^2$ (homocedasticidad), en cuyo caso se suponen igual a un valor común σ_c^2 y que se estima por

  $$\widehat{s}_c^2 = \frac{(n_1 - 1)\widehat{s}_1^2 + (n_2 - 1)\widehat{s}_2^2}{(n_1 - 1) + (n_2 - 1)},$$

  con $\widehat{s}_1^2$ y $\widehat{s}_1^2$ las cuasivarianzas muestrales en cada grupo. En este supuesto

  $$\widehat{EE}[\overline{Y}_1 - \overline{Y}_2] = \sqrt{\widehat{s}_c^2\Big(\frac{1}{n_1} + \frac{1}{n_2}\Big)}$$

  y el intervalo de confianza al $(1-\alpha)\,\%$ está dado por

  $$IC(1-\alpha)\,\%(\mu_1 - \mu_2) = \Big(\widehat{y}_1 - \widehat{y}_2 \mp t_{1-\alpha/2, n_1+n_2-2}\sqrt{\widehat{s}_c^2\Big(\frac{1}{n_1} + \frac{1}{n_2}\Big)}\Big).$$

- Que no se pueda asumir $\sigma_1^2 = \sigma_2^2$ (heterocedasticidad). En este supuesto

  $$\widehat{EE}[\overline{Y}_1 - \overline{Y}_2] = \sqrt{\frac{\widehat{s}_1^2}{n_1} + \frac{\widehat{s}_2^2}{n_2}}$$

  y el intervalo de confianza al $(1-\alpha)\,\%$ está dado por

  $$IC(1-\alpha)\,\%(\mu_1 - \mu_2) = \Big(\widehat{y}_1 - \widehat{y}_2 \mp t_{1-\alpha/2, gl}\sqrt{\frac{\widehat{s}_1^2}{n_1} + \frac{\widehat{s}_2^2}{n_2}}\Big),$$

  con

  $$gl = \frac{\widehat{EE}^4[\overline{Y_1} - \overline{Y}_2]}{\frac{1}{n_1 - 1}\widehat{EE}^4[\overline{Y_1}] + \frac{1}{n_2 - 1}\widehat{EE}^4[\overline{Y_2}]}.$$

En el caso del contraste de hipótesis

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 \neq 0$$

hay que considerar también los supuestos de homocedasticidad y heterocedasticidad.

- En el caso de homocedasticidad, la estimación de $EE[\overline{Y}_1 - \overline{Y}_2]$ bajo H_0 dada por $\widehat{EE}_0[\overline{Y}_1 - \widehat{\overline{Y}}_2]$ coincide con $\widehat{EE}[\overline{Y}_1 - \widehat{\overline{Y}}_2] = \sqrt{\widehat{s}_c^2\Big(\frac{1}{n_1} + \frac{1}{n_2}\Big)}$ y el contraste de hipótesis al nivel de significación α se realiza con

  $$t = \frac{\overline{y}_1 - \overline{y}_2}{\sqrt{\widehat{s}_c^2\Big(\frac{1}{n_1} + \frac{1}{n_2}\Big)}}$$

  tomando la decisión de que si $|t| > t_{1-\alpha/2, n_1+n_2-2}$ se rechaza la H_0.

- En el caso de heterocedasticidad, la estimación de $EE[\overline{Y}_1 - \overline{Y}_2]$ bajo H_0 dada por $\widehat{EE}_0[\overline{Y}_1 - \widehat{\overline{Y}}_2]$ coincide con $\widehat{EE}[\overline{Y}_1 - \widehat{\overline{Y}}_2] = \sqrt{\frac{\widehat{s}_1^2}{n_1} + \frac{\widehat{s}_2^2}{n_2}}$ y el contraste de hipótesis al nivel de significación α se realiza con

$$t = \frac{\overline{y}_1 - \overline{y}_2}{\sqrt{\frac{\widehat{s}_1^2}{n_1} + \frac{\widehat{s}_2^2}{n_2}}}$$

tomando la decisión de que si $|t| > t_{1-\alpha/2,gl}$ se rechaza la H_0, siendo

$$gl = \frac{\widehat{EE}^4[\overline{Y_1} - \overline{Y}_2]}{\frac{1}{n_1-1}\widehat{EE}^4[\overline{Y_1}] + \frac{1}{n_2-1}\widehat{EE}^4[\overline{Y_2}]}$$

.

Por último, como recomendaciones prácticas, conviene explicitar que las asunciones en la t de Student de normalidad se contrastan con el test de Shapiro-Wilk y la de homocedasticidad con el test de Levene.

Ejemplo

1. Se ha realizado un estudio para ver si influye una variable explicativa dicotómica en una variable respuesta continua. Han participado en el estudio 17 individuos. Los datos experimentales se dan en la tabla 2.2. ¿Influye la variable explicativa en la variable respuesta?

rta	*exp*
15	1
15	1
25	1
25	1
25	1
33	1
43	1
15	2
16	2
16	2
25	2
28	2
28	2
28	2
28	2
35	2
43	2

Tabla 2.2: Datos observados del esquema $C \leftarrow D$

SOLUCIÓN:

En primer lugar se realiza una estadística descriptiva numérica básica de los datos con la sintaxis siguiente

```
rm(list=ls())

datos=read.table('c_d_1.txt',header=T)
attach(datos)

ind1=which(exp==1)
ind2=which(exp==2)
n1=length(rta[ind1]); n1
n2=length(rta[ind2]); n2
tapply(rta,exp,mean)
tapply(rta,exp,sd)
```

que proporcionan los estadísticos:

$$n_1 = 7 \quad ; \quad n_2 = 10$$
$$\overline{y}_1 = \frac{1}{n_1}\sum_{i_1=1}^{n_1} y_{i_1} = \frac{181}{7} = 25.8571 \quad ; \quad \overline{y}_2 = \frac{1}{n_2}\sum_{i_2=1}^{n_2} y_{i_2} = \frac{262}{10} = 26.2000$$
$$\widehat{s}_1 = \sqrt{\frac{1}{n_1-1}\sum_{i_1}^{n_1}(y_{i_1}-\overline{y})^2} = 9.8561 \quad ; \quad \widehat{s}_2 = \sqrt{\frac{1}{n_2-1}\sum_{i_2}^{n_2}(y_{i_2}-\overline{y})^2} = 8.8669$$

En segundo lugar se contrasta la normalidad y la homocedasticidad de la variable respuesta en cada grupo con la sintaxis siguiente

```
shapiro.test(rta[ind1])
shapiro.test(rta[ind2])

library(lawstat)
levene.test(rta,exp,location="mean")
```

que proporciona, en cuanto a la normalidad, para el grupo 1 un p-valor=0.3300 y para el grupo 2 un p-valor=0.2268, por lo que se asume normalidad de la respuesta en cada grupo, y en cuanto a la homocedasticidad, un p-valor=0.8982 por lo que se asume homocedasticidad.

Por último, dado que se cumplen las condiciones de aplicabilidad de la t de Student, se calcula el estadístico de contraste y su p-valor asociado con la sintaxis

```
t.test(rta[ind1],rta[ind2],var.equal=TRUE)
```

que proporciona,

$$t = \frac{\overline{y}_1 - \overline{y}_2}{\sqrt{\widehat{s}_c^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{25.8571 - 26.2000}{\sqrt{86.0305\left(\frac{1}{7}+\frac{1}{10}\right)}} = -0.0750,$$

con $gl = 7 + 10 - 2 = 15$ y un p-valor asociado de $= 2 \cdot (1 - pt(abs(-0.0750), n1 + n2 - 2)) = 0.9412 > 0.05$, por lo que no hay evidencia suficiente para rechazar la igualdad de medias con lo que la variable explicativa no influye en la respuesta (no hay diferencias estadísticamente significativas entre $\overline{y}_1 = 25.8571$ y $\overline{y}_2 = 26.2000$).

2.2. ANOVA de un factor

El test ANOVA de un factor está enmarcado en el esquema $C \leftarrow N$ y está indicado para variables Y normales con igual varianza (homocedasticidad) en cada categoría de la variable nominal X. La situación de $C \leftarrow N$ se da, por ejemplo, cuando se quiere estudiar si la variable pérdida de peso (medidad en g) está influida por el hecho de recibir un tipo u otro de dieta. Para ello se realiza un estudio en n individuos a los que se les asigna la dieta 1 (codificada con 1 en la variable dieta) a n_1 individuos, la dieta 2 (codificada con 2) a n_2 individuos o la dieta 3 (codificada con 3) a n_3 individuos con $n = n_1 + n_2 + n_3$ y se observa qué valor tienen en la variable pérdida de peso. En esta situación, los datos experimentales tendrán el aspecto dado en la tabla 2.3, donde el número de individuos con $X = 1$ es n_1, con $X = 2$ es n_2 y con $X = 3$ es n_3.

La pregunta de si X influye en Y se traduce en realizar el contraste de hipótesis

$$H_0 : \mu_1 = \mu_2 = \mu_3$$
$$H_1 : \text{Alguna } \mu_i \text{ distinta}$$

El ANOVA de un factor se basa en el cáculo de la tabla ANOVA que se muestra en la tabla 2.4, siendo SCE la suma de cuadrados explicada (por la varible explicativa), SCR la suma de cuadrados residual (la que queda sin explicar por la variable explicativa), SCT la suma de cuadrados total, GLE los grados de libertad de la parte explicada, GLR los grados de libertad de la parte residual, GLT los grados de libertad total, CME el cuadrado medio explicado, CMR el cuadrado medio residual y CMT el cuadrado medio total que se calculan

Y	X
y_{11}	1
$\dots$	$\dots$
y_{1n_1}	1
y_{21}	2
$\dots$	$\dots$
y_{2n_2}	2
y_{31}	3
$\dots$	$\dots$
y_{3n_3}	3

Tabla 2.3: Datos genéricos del esquema $C \leftarrow N$

como

$$\begin{aligned}
SCE &= \sum_{m=1}^{r} n_m(\overline{y}_{.m} - \overline{y}_{..})^2 \\
SCR &= \sum_{m=1}^{r}\sum_{i=1}^{n_m}(\overline{y}_{im} - \overline{y}_{.m})^2 \\
SCT &= \sum_{m=1}^{r}\sum_{i=1}^{n_m}(\overline{y}_{im} - \overline{y}_{..})^2 = SCE + SCR \\
GLE &= r - 1 \\
GLR &= n - r \\
GLT &= GLE + GLR \\
CME &= \frac{SCE}{GLE} \\
CMR &= \frac{SCR}{GLR} \\
F &= \frac{CME}{CMRE}
\end{aligned}$$

con $\overline{y}_{.m}$ la media muestral de la variable respuesta en el grupo $m = 1, \dots, r$ e $\overline{y}_{..}$ la media muestral global de la variable respuesta sin tener en cuenta los grupos.

	GL	SC	CM	F	p-valor
Explicada	GLE	SCE	CME	F	asociado a F
Residual	GLR	SCR	CMR		
Total	GLT	SCT			

Tabla 2.4: Tabla ANOVA con datos genéricos del esquema $C \leftarrow N$

Por último, como recomendaciones prácticas, conviene explicitar que las asunciones en el ANOVA de un factor de normalidad se contrastan con el test de Shapiro-Wilk y la de homocedasticidad con el test de Levene.

Ejemplo

1. Se ha realizado un estudio para ver si influye una variable explicativa nominal en una variable respuesta continua. Han participado en el estudio 22 individuos. Los datos experimentales se dan en la tabla 2.5. ¿Influye la variable explicativa en la variable respuesta?

   SOLUCIÓN:

   En primer lugar se realiza una estadística descriptiva numérica básica de los datos con la sintaxis siguiente

```
rm(list=ls())

datos=read.table('c_n_1.txt',header=T)
attach(datos)
```

rta	*exp*
15	1
15	1
25	1
25	1
25	1
33	1
43	1
15	2
16	2
16	2
25	2
28	2
28	2
28	2
28	2
35	2
43	2
13	3
15	3
25	3
25	3
35	3

Tabla 2.5: Datos observados del esquema $C \leftarrow N$

```
ind1=which(exp==1);
ind2=which(exp==2);
ind3=which(exp==3);
n1=length(rta[ind1]); n1
n2=length(rta[ind2]); n2
n3=length(rta[ind3]); n3
tapply(rta,exp,mean); tapply(rta,exp,sd)
```

que proporcionan los estadísticos:

$$n_1 = 7;\ n_2 = 10;\ n_3 = 5$$
$$\overline{y}_1 = 25.8571; \overline{y}_2 = 26.2000; \overline{y}_3 = 22.6000$$
$$\widehat{s}_1 = 9.8561; \widehat{s}_2 = 8.8669; \widehat{s}_2 = 8.8769$$

En segundo lugar se contrasta la normalidad y la homocedasticidad de la variable respuesta en cada grupo con la sintaxis siguiente

```
shapiro.test(rta[ind1]); shapiro.test(rta[ind2]); shapiro.test(rta[ind3])

library(lawstat)
levene.test(rta,exp,location="mean")
```

que proporciona, en cuanto a la normalidad, para el grupo 1 un p-valor=0.3300, para el grupo 2 un p-valor=0.2268 y para el para el grupo 3 un p-valor=0.5347, por lo que se asume normalidad de la respuesta en cada grupo, y en cuanto a la homocedasticidad, un p-valor=0.9891 por lo que se asume homocedasticidad.

Por último, dado que se cumplen las condiciones de aplicabilidad del ANOVA de un factor, se calcula el estadístico de contraste y su p-valor asociado con la sintaxis

```
summary(aov(rta~factor(exp)))
```

que proporciona, la tabla 2.6.

	GL	SC	CM	F	p-valor
Explicada	2	46.7065	23.3532	0.2763	0.7616
Residual	19	1605.6571	84.5083		
Total	21	1652.3636			

Tabla 2.6: Tabla ANOVA con datos observados del esquema $C \leftarrow N$

El cálculo de los valores de la tabla 2.6 es el siguiente:

$$\begin{aligned}
SCE &= \sum_{m=1}^{r} n_m(\overline{y}_{.m} - \overline{y}_{..})^2 \\
&= 7(25.8571 - 25.2727)^2 + 10(26.200 - 25.2727)^2 + 5(22.6 - 25.2727)^2 \\
&= = 46.7065 \\
SCR &= \sum_{m=1}^{r}\sum_{i=1}^{n_m}(\overline{y}_{im} - \overline{y}_{.m})^2 = 1605.6571 \\
SCT &= \sum_{m=1}^{r}\sum_{i=1}^{n_m}(\overline{y}_{im} - \overline{y}_{..})^2 = 1652.3636 = SCE + SCR \\
GLE &= r - 1 = 2 \\
GLR &= n - r = 22 - 3 = 19 \\
GLT &= n - 1 = 21 = GLE + GLR \\
CME &= \frac{SCE}{GLE} = 23.3532 \\
CMR &= \frac{SCR}{GLR} = 84.5083 \\
F &= \frac{CME}{CMRE} = 0.2763
\end{aligned}$$

con un p-valor asociado de $1 - pf(0.2763, 2, 19) = 0.7616 > 0.05$, por lo que no hay evidencia suficiente para rechazar la igualdad de medias con lo que la variable explicativa no influye en la respuesta (no hay diferencias estadísticamente significativas entre $\overline{y}_1 = 25.8571$, $\overline{y}_2 = 26.2000$ y $\overline{y}_3 = 22.6000$).

Mini-capítulo 3

Respuesta nominal

En este mini-capítulo se contempla el test χ^2 de homogeneidad que está enmarcado en el esquema $N \leftarrow N$ que indica que se está intentando explicar la variable respuesta Y nominal a través de la variable explicativa X nominal.

3.1. χ^2 de homogeneidad

El test χ^2 de homogeneidad está enmarcado en el esquema $N \leftarrow N$. La situación de $N \leftarrow N$ se da, por ejemplo, cuando se quiere estudiar si distintos tipos de efectos adversos están influidos por distintos tipos de dieta. Para ello se realiza un estudio en n individuos a los que se les asigna, por ejemplo, la dieta 1 (codificada con 1 en la variable dieta) a n_1 individuos, la dieta 2 (codificada con 2), la dieta 3 (codificada con 3) a n_3 individuos, con $n = n_1 + n_2 + n_3$ y se observa qué tipo de efecto adverso presenta (por ejemplo, 1, 2, 3 ó 4). En esta situación los datos experimentales tendrán el patrón dado en la tabla 3.1, donde el número de individuos con perfil $Y = 1$ y $X = 1$ es n_{11}, con perfil $Y = 1$ y $X = 2$ es n_{12}, con perfil $Y = 1$ y $X = 3$ es n_{13}, con perfil $Y = 2$ y $X = 1$ es n_{21}, con perfil $Y = 2$ y $X = 2$ es n_{22}, con perfil $Y = 2$ y $X = 3$ es n_{23}, con perfil $Y = 3$ y $X = 1$ es n_{31}, con perfil $Y = 3$ y $X = 2$ es n_{32} y con perfil $Y = 3$ y $X = 3$ es n_{13}.

La pregunta de si X influye en Y se traduce en realizar el contraste de hipótesis

$$H_0 : \text{No diferencia en los grupos}$$
$$H_1 : \text{Sí diferencia en los grupos}$$

A partir de los datos experimentales se construye la tabla de datos cruzados genérica dada en la tabla 3.2, donde r_i es la suma por filas y c_j es la suma por columnas.

La expresión para el cálculo del test χ^2 de Pearson, publicada en 1900, viene dada por

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

que sigue una $\chi^2_{(r-1)(c-1)}$ donde e_{ij} son las frecuencias esperadas bajo H_0 y que se calculan como

$$e_{ij} = \frac{r_i}{n}\frac{c_j}{n}n = \frac{r_i c_j}{n}.$$

Por último, como recomendación práctica, conviene explicitar las asunciones del test χ^2 de homogeneidad que son que todos los valores de las frecuencias esperadas e_{ij}, sean mayores que 5, que intuitivamente corresponde a tamaños muestrales grandes. Así mismo, hay que señalar que numéricamente el test χ^2 de homogneidad también se usa en el marco de asociación entre dos variables nominales donde ninguna de ellas hace el papel de variable explicativa, en ese caso se le denomina χ^2 de independencia.

Ejemplo

1. Se ha realizado un estudio para ver si influye una variable explicativa nominal en una variable respuesta nominal. Han participado en el estudio 100 individuos. Los datos experimentales corresponden al esquema dado en la tabla 3.1 y se muestran ya tabulados en la tabla 3.3. ¿Influye la variable explicativa en la variable respuesta?

   SOLUCIÓN:

Y	X
1	1
...	...
1	1
1	2
...	...
1	2
1	3
...	...
1	3
2	1
...	...
2	1
2	2
...	...
2	2
2	3
...	...
2	3
3	1
...	...
3	1
3	2
...	...
3	2
3	3
...	...
3	3

Tabla 3.1: Datos genéricos del esquema $N \leftarrow N$

	$X=1$	$X=2$	$X=3$	
$Y=1$	n_{11}	n_{12}	n_{13}	r_1
$Y=2$	n_{21}	n_{22}	n_{23}	r_2
$Y=3$	n_{31}	n_{32}	n_{33}	r_3
$Y=4$	n_{41}	n_{42}	n_{43}	r_4
	c_1	c_2	c_3	n

Tabla 3.2: Tabla con datos genéricos del esquema $N \leftarrow N$

	$X=1$	$X=2$	$X=3$	
$Y=1$	3	10	14	27
$Y=2$	10	10	14	34
$Y=3$	6	10	3	19
$Y=4$	14	3	3	20
	33	33	34	100

Tabla 3.3: Tabla con datos observados del esquema $N \leftarrow N$

En primer lugar se muestra la tabla 3.4 con los datos esperados. A partir de esta tabla se calcula

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-e_{ij})^2}{e_{ij}} = 23.8325$$

con un p-valor asociado de $1 - pchisq(23.8325, 6) = 0.0006 < 0.05$, por lo que sí hay evidencia suficiente para rechazar la H_0 con lo que se concluye que la variable explicativa sí influye en la respuesta (hay diferencias estadísticamente significativas entre los valores oservados y los esperados bajo homogeneidad de no diferencia entre grupos).

	$X=1$	$X=2$	$X=3$	
$Y=1$	8.91	8.91	9.18	27
$Y=2$	11.22	11.22	11.56	34
$Y=3$	6.27	6.27	6.46	19
$Y=4$	6.6	6.6	6.8	20
	33	33	34	100

Tabla 3.4: Tabla con datos esperados del esquema $N \leftarrow N$

En segundo lugar se muestra la sintaxis en R para el cálculo del χ^2. Si se supone que los datos sin agrupar están en el fichero n_n_1.txt la sintaxis es

```
rm(list=ls())

datos=read.table('n_n_1.txt',header=T)
attach(datos)

tabla=table(rta,exp); tabla
chisq.test(tabla)\$expected

chisq.test(tabla)
```

y si se supone que se introducen los datos agrupados directamente la sintaxis es

```
rm(list=ls())

tabla <- matrix(c(3,10,14,10,10,14,6,10,3,14,3,3),ncol=3,byrow=TRUE)
colnames(tabla) <- c("exp=1","exp=2","exp=3")
rownames(tabla) <- c("rta=1","rta=2","rta=3","rta=4")
tabla <- as.table(tabla); tabla
chisq.test(tabla)\$expected
```

Mini-capítulo 4

Respuesta ordinal

En este mini-capítulo se contemplan el test U de Mann-Whitney, el test W de Wilcoxon y el test H de Kruskal–Wallis. El test U de Mann-Whitney y el test W de Wilcoxon están enmarcados en el esquema $O \leftarrow D$ que indica que se está intentando explicar la variable respuesta Y ordinal a través de la variable explicativa X dicotómica y el test H de Kruskal-Wallis está enmarcado en el esquema $O \leftarrow N$ que indica que se está intentando explicar la variable respuesta Y ordinal a través de la variable explicativa X nominal.

4.1. U de Mann-Whitney

El test U de Mann-Whitney está enmarcado en el esquema $O \leftarrow D$, aunque también se puede usar en el esquema $C \leftarrow D$ donde no se cumpla la asunción de que la variable Y sea normal en cada categoría de la variable dicotómica X. La situación de $O \leftarrow D$ se da, por ejemplo, cuando se quiere estudiar si la variable grado de mejoría (codificada por ejemplo con 1 si algo, 2 bastante, 3 casi total y 4 total) está influida por el hecho de recibir un tratamiento u otro. Para ello se realiza un estudio en n individuos a los que se les asigna el tratamiento 1 (codificado con 1 en la variable tratamiento) a n_1 individuos y el tratamiento 2 (codificado con 2 en la variable tratamiento) a n_2 individuos con $n = n_1 + n_2$ y se observa qué valor ordinal tienen en la variable grado de mejoría. En esta situación los datos experimentales tendrán el aspecto dado en la tabla 4.1, donde el número de individuos con $X = 1$ es n_1 y con $X = 2$ es n_2.

Y	X
y_{11}	1
$\dots$	$\dots$
y_{1n_1}	1
y_{21}	2
$\dots$	$\dots$
y_{2n_2}	2

Tabla 4.1: Datos genéricos del esquema $O \leftarrow D$

La pregunta de si X influye en Y se traduce en realizar el contraste de hipótesis

$$H_0 : F_1 = F_2$$
$$H_1 : F_1 \neq F_2$$

siendo F_1 y F_2 las funciones de distribución de la variable respuesta para el grupo 1 y el grupo 2, respectivamente.

La expresión para el cálculo de la U de Mann-Whitney, publicada en 1947, viene dada por $U_{Y_1Y_2}$ o por $U_{Y_2Y_1}$ donde

$$U_{Y_1Y_2} = \#\{y_{1i} < y_{2i}\}$$
$$U_{Y_2Y_1} = \#\{y_{2i} < y_{1i}\},$$

siendo y_{1i} los valores de la variable respuesta para el grupo 1, y_{2i} los valores de la variable respuesta para el grupo 2 y donde la notación $\#\{\ \}$ representa el número de pares que verifica la condición entre llaves, teniendo en cuenta que cualquier par con $y_{1i} = y_{2i}$ suma $\frac{1}{2}$ en el cálculo de $U_{Y_1Y_2}$ o de $U_{Y_2Y_1}$. Es fácil ver que $U_{Y_1Y_2} + U_{Y_2Y_1} = n_1 n_2$.

Para llevar a cabo el contraste requerido se construye el estadístico de contraste

$$z = \frac{U_{Y_1Y_2} - \widehat{E}[U_{Y_1Y_2}]}{\widehat{EE}[U_{Y_1Y_2}]} = \frac{U_{Y_2Y_1} - \widehat{E}[U_{Y_2Y_1}]}{\widehat{EE}[U_{Y_2Y_1}]}$$

que sigue una $N(0,1)$, siendo

$$\widehat{E}[U_{Y_1Y_2}] = \widehat{E}[U_{Y_2Y_1}] = \frac{1}{2}n_1n_2$$

$$\widehat{EE}[U_{Y_1Y_2}] = \widehat{EE}[U_{Y_2Y_1}] = \sqrt{\frac{1}{12}\frac{n_1n_2}{n(n-1)}\Big(n^3 - n - \sum_{j=1}^{k}(d_j^3 - d_j)\Big)},$$

con d_j el número de empates en $j = 1, \ldots, k$ siendo k el número de valores distintos que se produce para cada valor de la variable respuesta sin distinguir grupos.

Ejemplo

1. Se ha realizado un estudio para ver si influye una variable explicativa dicotómica en una variable respuesta ordinal. Han participado en el estudio 17 individuos. Los datos experimentales se dan en la tabla 4.2. ¿Influye la variable explicativa en la variable respuesta?

rta	*exp*
1	1
1	1
3	1
3	1
3	1
5	1
7	1
1	2
2	2
2	2
3	2
4	2
4	2
4	2
4	2
6	2
7	2

Tabla 4.2: Datos observados del esquema $O \leftarrow D$

SOLUCIÓN:

En primer lugar se realiza una estadística descriptiva numérica y gráfica básica de los datos con la sintaxis siguiente

```
rm(list=ls())

datos=read.table('o_d_1.txt',header=T)
attach(datos)

ind1=which(exp==1)
ind2=which(exp==2)
n1=length(rta[ind1]); n1
n2=length(rta[ind2]); n2

tapply(rta,exp,median); tapply(rta,exp,IQR)

boxplot(rta[ind1],rta[ind2],names=c("1","2"),col=(c("white","gray")),
xlab="Grupo 1 (white), Grupo 2 (gray)", ylab="Rta")
```

que proporcionan los estadísticos:

$$\begin{aligned} n_1 = 7 \quad &; \quad n_2 = 10 \\ \text{mediana } (y)_1 = 3 \quad &; \quad \text{mediana } (y)_2 = 4 \\ \text{rango intercuartílico } (y)_1 = 2.0000 \quad &; \quad \text{rango intercuartílico } (y)_2 = 1.7500 \end{aligned}$$

y el gráfico de Box-Plot dado en la figura 4.1.

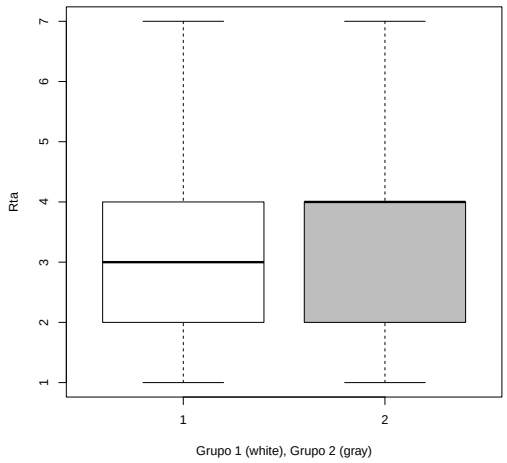

Figura 4.1: Box-Plot para los datos observados del esquema $O \leftarrow D$

A continuación se muestran los cálculos detallados de la U de Mann-Whitney:

$$\begin{aligned}
U_{Y_1Y_2} &= \Big(\frac{1}{2}+9\Big)+\Big(\frac{1}{2}+9\Big)+\Big(\frac{1}{2}+6\Big)+\Big(\frac{1}{2}+6\Big) \\
&\quad +\Big(\frac{1}{2}+6\Big)+(2)+\Big(\frac{1}{2}\Big)=41 \\
U_{Y_2Y_1} &= \Big(\frac{1}{2}+\frac{1}{2}+5\Big)+(5)+(5)+\Big(\frac{3}{2}+2\Big)+(2)+(2) \\
&\quad +(2)+(2)+(1)+\Big(\frac{1}{2}\Big)=29 \\
U_{Y_1Y_2}+U_{Y_2Y_1} &= 70=n_1n_2 \\
\widehat{E}[U_{Y_1Y_2}] &= \widehat{E}[U_{Y_2Y_1}]=\frac{1}{2}n_1n_2=\frac{1}{2}\cdot 70=35 \\
\sum_{j=1}^{k}(d_j^3-d_j) &= (3^3-3)+(2^3-2)+(4^3-4)+(4^3-4)+(2^3-2)=156 \\
\widehat{EE}[U_{Y_1Y_2}] &= \widehat{EE}[U_{Y_2Y_1}]=\sqrt{\frac{1}{12}\frac{n_1n_2}{n(n-1)}\Big(n^3-n-\sum_{j=1}^{k}(d_j^3-d_j)\Big)} \\
&= \sqrt{\frac{1}{12}\frac{70}{17\cdot 16}\Big(17^3-17-156\Big)}=\sqrt{101.6544}=10.0824 \\
z &= \frac{U_{Y_2Y_1}-\widehat{E}[U_{Y_2Y_1}]}{\widehat{EE}[U_{Y_2Y_1}]}=\frac{29-35}{10.0824}=-0.5951
\end{aligned}$$

Dichos cálculos se realizan con la sintaxis

```
wilcox.test(rta[ind1],rta[ind2],correct=FALSE)
```

que proporciona, $U_{Y_2Y_1}=29$ y un p-valor asociado igual a $2\cdot(1-pnorm(abs(-0.5951)))=0.5518>0.05$, por lo que no hay evidencia suficiente para rechazar la igualdad de funciones de distribución con lo que la variable explicativa no influye en la respuesta.

4.2. W de Wilcoxon

El test W de Wilcoxon está enmarcado en el esquema $O \leftarrow D$, al igual que el test U de Mann-Whitney, por lo que los datos experimentales tendrán el aspecto anterior dado en la tabla 4.1. Conviene remarcar que ambos tests son además equivalentes.

Como en el test de U de Mann-Whitney, la pregunta de si X influye en Y se traduce en realizar el contraste de hipótesis

$$H_0 : F_1 = F_2$$
$$H_1 : F_1 \neq F_2$$

siendo F_1 y F_2 las funciones de distribución de la variable respuesta para el grupo 1 y el grupo 2, respectivamente.

La expresión para el cálculo de la W de Mann-Whitney publicada en 1945, viene dada por W_1 o por W_2 con

$$W_1 = \sum_{grupo_1} Rangos$$

$$W_2 = \sum_{grupo_2} Rangos,$$

donde los rangos se calculan en relación a la muestra conjunta y en el caso de empates como promedios de los órdenes de las observaciones empatadas. Es fácil ver que $W_1 + W_2 = \frac{1}{2}n(n+1)$.

Para llevar a cabo el contraste requerido se construye el estadístico de contraste

$$z = \frac{W_1 - \widehat{E}[W_1]}{\widehat{EE}[W_1]} = \frac{W_2 - \widehat{E}[W_2]}{\widehat{EE}[W_2]}$$

que sigue una $N(0,1)$, siendo

$$\widehat{E}[W_1] = \frac{1}{2}n_1(n+1)$$

$$\widehat{E}[W_2] = \frac{1}{2}n_2(n+1)$$

$$\widehat{EE}[W_1] = \widehat{EE}[W_2] = \sqrt{\frac{1}{12}\frac{n_1 n_2}{n(n-1)}\Big(n^3 - n - \sum_{j=1}^{k}(d_j^3 - d_j)\Big)},$$

con d_j el número de empates en $j = 1, \ldots, k$ siendo k el número de valores distintos que se produce para cada valor de la variable respuesta sin distinguir grupos.

Ejemplo

1. Se ha realizado un estudio para ver si influye una variable explicativa dicotómica en una variable respuesta ordinal. Han participado en el estudio 17 individuos. Los datos experimentales son los mismos que se utilizaron en el test U de Mann-Whitney y se dan en la tabla 4.2. ¿Influye la variable explicativa en la variable respuesta?

   SOLUCIÓN:

   La estadística descriptiva numérica y gráfica básica de los datos es la que se mostró en el test U de Mann-Whitney. Los cálculos detallados del test W de Wilcoxon se basan en rangos por lo que como paso previo se asigna a cada observación un rango según el orden que ocupe dicha obervación en el cómputo total de datos, asignando el rango medio en caso de empates, según se muestra en la tabla 4.3.

rta	*exp*	*Rango(rta)*
1	1	$\frac{1+2+3}{3} = 2$
1	1	$\frac{1+2+3}{3} = 2$
3	1	$\frac{6+7+8+9}{4} = 7.5$
3	1	$\frac{6+7+8+9}{4} = 7.5$
3	1	$\frac{6+7+8+9}{4} = 7.5$
5	1	14
7	1	$\frac{16+17}{2} = 16.5$
1	2	$\frac{1+2+3}{3} = 2$
2	2	$\frac{4+5}{2} = 4.5$
2	2	$\frac{4+5}{2} = 4.5$
3	2	$\frac{6+7+8+9}{4} = 7.5$
4	2	$\frac{10+11+12+13}{4} = 11.5$
4	2	$\frac{10+11+12+13}{4} = 11.5$
4	2	$\frac{10+11+12+13}{4} = 11.5$
4	2	$\frac{10+11+12+13}{4} = 11.5$
6	2	15
7	2	$\frac{16+17}{2} = 16.5$

Tabla 4.3: Rangos para los datos observados del esquema $O \leftarrow D$

A partir de estos rangos, se calculan los valores siguientes:

$$
\begin{aligned}
W_1 &= \sum_{grupo_1} Rangos = 2+2+7.5+7.5+7.5+14+16.5 = 57 \\
W_2 &= \sum_{grupo_2} Rangos = 2+4.5+4.5+7.5+11.5+11.5+11.5+11.5+16.5 \\
&= 96 \\
W_1 + W_2 &= 153 = \frac{1}{2}n(n+1) \\
\widehat{E}[W_1] &= \frac{1}{2}n_1(n+1) = \frac{1}{2}\cdot 7 \cdot 18 = 63 \\
\widehat{E}[W_2] &= \frac{1}{2}n_2(n+1) = \frac{1}{2}\cdot 10 \cdot 18 = 90 \\
\sum_{j=1}^{k}(d_j^3 - d_j) &= (3^3-3)+(2^3-2)+(4^3-4)+(4^3-4)+(2^3-2) = 156 \\
\widehat{EE}[W_1] &= \widehat{EE}[W_2] = \sqrt{\frac{1}{12}\frac{n_1 n_2}{n(n-1)}\Big(n^3 - n - \sum_{j=1}^{k}(d_j^3 - d_j)\Big)} \\
&= \sqrt{\frac{1}{12}\frac{70}{17\cdot 16}\Big(17^3 - 17 - 156\Big)} = \sqrt{101.6544} = 10.0824 \\
z &= \frac{W_1 - \widehat{E}[W_1]}{\widehat{EE}[W_1]} = \frac{29-35}{10.0824} = -0.5951
\end{aligned}
$$

Dichos cálculos se realizan con la sintaxis

```
wilcox.test(rta[ind1],rta[ind2],correct=FALSE)
```

que no proporciona el test de Wilcoxon, sino su equivalente de U de Mann-Whitney, con un p-valor igual a $2 \cdot (1 - pnorm(abs(-0.5951))) = 0.5518 > 0.05$, por lo que no hay evidencia suficiente para rechazar la igualdad de funciones de distribución con lo que la variable explicativa no influye en la respuesta.

4.3. H de Kruskal-Wallis

El test H de Kruskal-Wallis está enmarcado en el esquema $O \leftarrow N$, aunque también se puede usar en el esquema $C \leftarrow N$ donde no se cumpla la asunción de que la variable Y sea normal en cada categoría de la variable dicotómica X. La situación de $O \leftarrow N$ se da, por ejemplo, cuando se quiere estudiar si la variable grado de pérdida de peso (codificada por ejemplo con 1 si se ha perdido entre 0-5 kilos, 2 si 5-10, 3 si 10-15 y 4 si > 15) está influida por el hecho de recibir un tipo u otro de dieta. Para ello se realiza un estudio en n individuos a los que se les asigna la dieta 1 (codificada con 1 en la variable dieta) a n_1 individuos, la dieta 2 (codificado con 2) a n_2 individuos o la dieta 3 (codificada con 3) a n_3 individuos con $n = n_1 + n_2 + n_3$ y se observa qué valor ordinal tienen en la variable grado de de pérdida de peso. En esta situación, los datos experimentales tendrán el aspecto dado en la tabla 4.4, donde el número de individuos con $X = 1$ es n_1, con $X = 2$ es n_2 y con $X = 3$ es n_3.

Y	X
y_{11}	1
$\dots$	$\dots$
y_{1n_1}	1
y_{21}	2
$\dots$	$\dots$
y_{2n_2}	2
y_{31}	3
$\dots$	$\dots$
y_{3n_3}	3

Tabla 4.4: Datos genéricos del esquema $O \leftarrow N$

La pregunta de si X influye en Y se traduce en realizar el contraste de hipótesis

$$H_0 : F_1 = F_2 = F_3$$
$$H_1 : \text{Alguna } F_i \text{ distinta}$$

siendo F_1, F_2 y F_3 las funciones de distribución de la variable respuesta para el grupo 1, el grupo 2 y el grupo 3, respectivamente.

La expresión para el cálculo de la H de Kruskal-Wallis, publicada en 1952, viene dada por

$$H = \frac{\frac{12}{n(n+1)} \sum_{m=1}^{r} \frac{1}{n_m} \left(R_m - \widehat{E}[R_m] \right)^2}{1 - \frac{\sum_{j=1}^{k} (d_j^3 - d_j)}{n^3 - n}}$$

que sigue una χ^2_{r-1}, siendo

$$R_m = \sum_{grupo_m} Rangos \ , \ \widehat{E}[R_m] = \frac{n_m(n+1)}{2}$$

con $m = 1, \dots, r$ indicando el grupo y d_j el número de empates en $j = 1, \dots, k$ siendo k el número de valores distintos que se produce para cada valor de la variable respuesta sin distinguir grupos.

Ejemplo

1. Se ha realizado un estudio para ver si influye una variable explicativa nominal en una variable respuesta ordinal. Han participado en el estudio 22 individuos. Los datos experimentales se dan en la tabla 4.5. ¿Influye la variable explicativa en la variable respuesta?

   SOLUCIÓN:

   En primer lugar se realiza una estadística descriptiva numérica y gráfica básica de los datos con la sintaxis siguiente

rta	*exp*
2	1
2	1
4	1
4	1
4	1
6	1
8	1
2	2
3	2
3	2
4	2
5	2
5	2
5	2
5	2
7	2
8	2
1	3
2	3
4	3
4	3
7	3

Tabla 4.5: Datos observados del esquema $O \leftarrow D$

```
rm(list=ls())

datos=read.table('o_n_1.txt',header=T)
attach(datos)

ind1=which(exp==1);
ind2=which(exp==2);
ind3=which(exp==3);
n1=length(rta[ind1]); n1
n2=length(rta[ind2]); n2
n3=length(rta[ind3]); n3
tapply(rta,exp,median); tapply(rta,exp,IQR)

boxplot(rta[ind1],rta[ind2],rta[ind3],names=c("1","2","3"),
col=(c("white","gray80","gray50")),
xlab="Grupo 1 (white), Grupo 2 (gray80), Grupo 3 (gray50)", ylab="Rta")
```

que proporcionan los estadísticos:

$$\begin{aligned} n_1 = 7 \quad &; \quad n_2 = 10 \ ; \ n_3 = 5 \\ \text{mediana } (y)_1 = 4 \quad &; \quad \text{mediana } (y)_2 = 5 \ ; \ \text{mediana } (y)_3 = 4 \\ \text{rango inter. } (y)_1 = 2.0000 \quad &; \quad \text{rango inter. } (y)_2 = 1.7500 \ ; \ \text{rango inter. } (y)_3 = 2.0000 \end{aligned}$$

y el gráfico de Box-Plot dado en la figura 4.2.

Los cálculos detallados del test H de Kruskal-Wallis se basan en rangos por lo que como paso previo se asigna a cada observación un rango según el orden que ocupe dicha obervación en el cómputo total de datos, asignando el rango medio en caso de empates, según se muestra en la tabla 4.6.

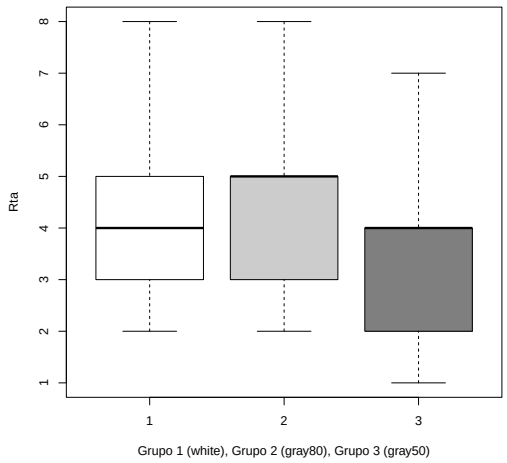

Figura 4.2: Box-Plot para los datos observados del esquema $O \leftarrow N$

A partir de estos rangos, se calculan los valores siguientes:

$$\begin{aligned}
R_1 &= \sum_{grupo_1} Rangos = 78 \\
R_2 &= \sum_{grupo_2} Rangos = 130 \\
R_3 &= \sum_{grupo_3} Rangos = 45 \\
\widehat{E}[R_1] &= \frac{1}{2} n_1(n+1) = \frac{1}{2} \cdot 7 \cdot 23 = 80.50 \\
\widehat{E}[R_2] &= \frac{1}{2} n_2(n+1) = \frac{1}{2} \cdot 10 \cdot 23 = 115.00 \\
\widehat{E}[R_3] &= \frac{1}{2} n_3(n+1) = \frac{1}{2} \cdot 5 \cdot 23 = 57.00 \\
\sum_{j=1}^{k} (d_j^3 - d_j) &= (4^3 - 4) + (2^3 - 2) + (6^3 - 6) + (4^3 - 4) \\
&+ (2^3 - 2) + (2^3 - 2) = 348 \\
H &= \frac{\frac{12}{22 \cdot 23}\left(\frac{1}{7}(78 - 80.50)^2 + \frac{1}{10}(130 - 115.00)^2 + \frac{1}{5}(45 - 57.50)^2\right)}{1 - \frac{348}{22^3 - 22}} \\
&= 1.3398
\end{aligned}$$

Dichos cálculos se realizan con la sintaxis

```
kruskal.test(rta~exp)
```

que proporciona, $H = 1.3398$ con $gl = r-1 = 3-1 = 2$ y un p-valor asociado igual a $1-pchisq(1.3398, 2) = 0.5118 > 0.05$, por lo que no hay evidencia suficiente para rechazar la igualdad de funciones de distribución con lo que la variable explicativa no influye en la respuesta.

rta	*exp*	*Rango*(*rta*)
2	1	$\frac{2+3+4+5}{4} = 3.5$
2	1	$\frac{2+3+4+5}{4} = 3.5$
4	1	$\frac{8+9+10+11+12+13}{6} = 10.5$
4	1	$\frac{8+9+10+11+12+13}{6} = 10.5$
4	1	$\frac{8+9+10+11+12+13}{6} = 10.5$
6	1	18
8	1	$\frac{21+22}{2} = 21.5$
2	2	$\frac{2+3+4+5}{4} = 3.5$
3	2	$\frac{6+7}{2} = 6.5$
3	2	$\frac{6+7}{2} = 6.5$
4	2	$\frac{8+9+10+11+12+13}{6} = 10.5$
5	2	$\frac{14+15+16+17}{4} = 15.5$
5	2	$\frac{14+15+16+17}{4} = 15.5$
5	2	$\frac{14+15+16+17}{4} = 15.5$
5	2	$\frac{14+15+16+17}{4} = 15.5$
7	2	$\frac{19+20}{2} = 19.5$
8	2	$\frac{21+22}{2} = 21.5$
1	3	1
2	3	$\frac{2+3+4+5}{4} = 3.5$
4	3	$\frac{8+9+10+11+12+13}{6} = 10.5$
4	3	$\frac{8+9+10+11+12+13}{6} = 10.5$
7	3	$\frac{19+20}{2} = 19.5$

Tabla 4.6: Rangos para los datos observados del esquema $O \leftarrow N$

Referencias

[1] Davies, T.M., *The Book of R: A first course in programming and statistics*, No Starch Press, 2016.

[2] García-Pérez, A., *Estadística básica con R*, Universidad Nacional de Educación a Distancia (UNED) (Madrid), 2010.

[3] Peña, D., *Fundamentos de Estadística*, Alianza editorial (Madrid), 2014.

[4] R Core Team, *R: A Language and Environment for Statistical Computing. R*, Foundation for Statistical Computing, Vienna, Austria, 2017. URL https: //www.R-project.org/.

[5] Ramos, E., Vélez, R., Hernández, V., *Modelos probabilísticos y optimización*, Sanz y Torres (Madrid), 2019.

Acerca de

Emilio Letón

Figura 4.3: Emilio Letón

Emilio Letón nace en Madrid en 1966. Es Licenciado en Matemáticas por la UCM en 1989 y doctor en Matemáticas en la misma universidad en 2002. En la actualidad es profesor contratado doctor de la UNED en el departamento de Inteligencia Artificial, al que se incorporó en 2009. Anteriormente fue profesor del departamento de Estadística de la UC3M durante 5 años. Asimismo, ha trabajado durante 15 años en departamentos de Planificación y Estadística dentro del sector bancario y de la industria farmacéutica. Sus líneas de investigación incluyen el Análisis de Supervivencia, tests no paramétricos, PLS, Meta-Análisis, Bioestadística y B-Learning. Ha participado en más de 30 proyectos de innovación docente (siendo coordinador en más de 10 de ellos) colaborando con distintas universidades: UNED, UC3M, UCM y UPM. Ha recibido 1 premio en excelencia en publicaciones científicas (UC3M) y 5 premios en excelencia docente (1 en UC3M, 3 en UNED y 1 en OCW Consortium). En @emilioleton se pueden encontrar sus tweets y su página web personal con información ampliada de su curriculum. https://twitter.com/emilioleton

Elisa M. Molanes-López

Figura 4.4: Elisa M. Molanes-López

Elisa M. Molanes López nace en Vigo en 1976. Se licencia en Matemáticas por la Universidad de Santiago de Compostela en 2000 y consigue el grado de doctora, con acreditación europea, por la Universidad de A Coruña en 2007. Entre 2003 y 2007, al amparo de una beca FPI, realiza varias estancias de investigación

en las siguientes universidades extranjeras: Universiteit Hasselt (Bélgica), Université Catholique de Louvain (Bélgica) y The University of Texas (EE.UU.). Durante 8 años, entre 2007 y 2015, es profesora visitante en el departamento de Estadística de la Universidad Carlos III de Madrid. Actualmente, desde octubre de 2015, es profesora ayudante doctora en la Unidad Departamental de Bioestadística del Departamento de Estadística e Investigación Operativa de la Universidad Complutense de Madrid. Durante su etapa docente ha participado en 10 proyectos de innovación docente. En diciembre de 2013, recibe un accésit a la mejor práctica docente en los Premios del Consejo Social de la UNED por su participación en el MOOC "Mini-vídeos docentes modulares: un elemento crítico en el diseño de un MOOC", y en abril de 2014, el OpenCourseWare Consortium le otorga un premio de excelencia por su participación en el curso OCW "Mini-vídeos docentes modulares para diseñar un MOOC". Sus líneas de investigación incluyen la estadística no paramétrica, el análisis de supervivencia, las curvas ROC y las funciones cópula. En la página web personal de Elisa M. Molanes-López, se puede encontrar información más detallada. https://elisamariamolanes.wixsite.com/emolanes

Índice general

1. Respuesta dicotómica **1**
1.1. z de diferencia de dos proporciones 1
1.2. Exacta de Fisher 4

2. Respuesta continua **7**
2.1. t de Student 7
2.2. ANOVA de un factor 10

3. Respuesta nominal **15**
3.1. χ^2 de homogeneidad 15

4. Respuesta ordinal **19**
4.1. U de Mann-Whitney 19
4.2. W de Wilcoxon 21
4.3. H de Kruskal-Wallis 24

Índice de figuras

4.1. Box-Plot para los datos observados del esquema $O \leftarrow D$ 21
4.2. Box-Plot para los datos observados del esquema $O \leftarrow N$ 26
4.3. Emilio Letón 31
4.4. Elisa M. Molanes-López 31

Índice de tablas

1. Datos genéricos de una variable dicotómica III

1.1. Datos genéricos del esquema $D \leftarrow D$ 1
1.2. Datos observados del esquema $D \leftarrow D$ 3
1.3. Tabla con datos genéricos del esquema $D \leftarrow D$ 4
1.4. Datos observados del esquema $D \leftarrow D$ 5
1.5. Tabla con datos observados del esquema $D \leftarrow D$ 5
1.6. Tablas variando a' en el esquema $D \leftarrow D$ 5

2.1. Datos genéricos del esquema $C \leftarrow D$ 7
2.2. Datos observados del esquema $C \leftarrow D$ 9
2.3. Datos genéricos del esquema $C \leftarrow N$ 11
2.4. Tabla ANOVA con datos genéricos del esquema $C \leftarrow N$ 11
2.5. Datos observados del esquema $C \leftarrow N$ 12
2.6. Tabla ANOVA con datos observados del esquema $C \leftarrow N$ 13

3.1. Datos genéricos del esquema $N \leftarrow N$ 16
3.2. Tabla con datos genéricos del esquema $N \leftarrow N$ 16
3.3. Tabla con datos observados del esquema $N \leftarrow N$ 16
3.4. Tabla con datos esperados del esquema $N \leftarrow N$ 17

4.1. Datos genéricos del esquema $O \leftarrow D$ 19
4.2. Datos observados del esquema $O \leftarrow D$ 20
4.3. Rangos para los datos observados del esquema $O \leftarrow D$ 23
4.4. Datos genéricos del esquema $O \leftarrow N$ 24
4.5. Datos observados del esquema $O \leftarrow D$ 25
4.6. Rangos para los datos observados del esquema $O \leftarrow N$ 27