

# Taller 2 entrega problema en grupo. MAT3 (estadística) GIN2 2020-2021 - Estadística inferencial mayo 2021.

## Soluciones

### Contenidos

<b>1 Taller 2 evaluable. Entrega de problemas</b>	<b>1</b>
1.1 Problema 1	1
1.1.1 Solución	2
1.2 Problema 2	6
1.2.1 Solución	7
1.3 Problema 3	11
1.3.1 Solución	11
1.4 Problema 4	14
1.4.1 Solución	15

## 1 Taller 2 evaluable. Entrega de problemas

Taller en grupo entregad las soluciones en .Rmd y .html o .pdf. o escribirlas de forma manual y escanear el resultado, en un solo fichero. Cada apartado vale 1 punto en total hay 15 puntos y se pondera la 10 puntos.

### 1.1 Problema 1

- Consideremos la siguiente muestra aleatoria simple de una v.a. continua  $X$ :  $-3, -2, -1, 0, 0, 1, 2, 3, 4$  de tamaño  $n = 9$ . Calcular, en esta muestra, el error estándar de estadístico media aritmética de la muestra.
- Consideremos la siguiente muestra aleatoria simple de tamaño  $n = 10$  de una v.a.  $X$  con distribución  $Ber(p)$ :  $1, 0, 1, 0, 1, 1, 1, 1, 1, 0$  Calcular, en esta muestra, el estadístico proporción muestral y su error estándar.
- Suponiendo que la población es normal calcular un intervalo de confianza del 95% para  $\mu_X$ .
- Suponiendo que la población es normal calcular un intervalo de confianza del 95% para  $\sigma_X^2$ .

Ayuda de R, acabad vosotros los cálculos

```
muestra1=c(-3,-2,-1,0,0,1,2,3,4)
mean(muestra1)
```

```
## [1] 0.4444444
```

```
sum(muestra1)
```

```
## [1] 4
```

```
sum(muestra1^2)
```

```
## [1] 44
```

```
n=length(muestra1)
n
```

```
## [1] 9
muestra2=c(1,0,1,0,1,1,1,1,0)
table(muestra2)

## muestra2
## 0 1
## 3 7

length(muestra2)

## [1] 10
```

### 1.1.1 Solución

#### Apartado a)

La muestra es  $x_1 = -3, x_2 = -2, x_3 = -1, x_4 = 0, x_5 = 0, x_6 = 1, x_7 = 1, x_8 = 2, x_9 = 3, x_{10} = 4$ , es de tamaño  $n = 10$  media aritmética es

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{-3 - 2 - 1 + 0 + 0 + 1 + 1 + 2 + 3 + 4}{10} = \frac{4}{10} = 0.4444444.$$

La desviación típica de la muestra es

$$\tilde{s}_X = \sqrt{\left(\frac{n}{n-1}\right) \cdot \left(\frac{\sum_{i=1}^n x_i^2}{n} - \bar{x}^2\right)} = \sqrt{\left(\frac{10}{9}\right) \cdot \left(\frac{44}{10} - 0.4444444^2\right)} = 2.2973415$$

Donde 44 es el resultado del código `sum(muestra1^2)`

Por último el error estándar de  $\bar{x}$  es

$$\frac{\tilde{s}_X}{\sqrt{n}} = \frac{2.2973415}{\sqrt{10}} = 0.7657805.$$

Con R

```
muestra1=c(-3,-2,-1,0,0,1,2,3,4)
media=mean(muestra1)
media

## [1] 0.4444444

desv_tip=sd(muestra1)
desv_tip

## [1] 2.297341

error_estandar_media=desv_tip/sqrt(n)
error_estandar_media

## [1] 0.7657805
```

### Apartado b)

La muestra es  $x_1 = 1, x_2 = 0, x_3 = 1, x_4 = 0, x_5 = 0, x_6 = 1, x_7 = 1, x_8 = 1, x_9 = 0$ , es de tamaño  $n = 10$  media aritmética es

$$\hat{p} = \frac{\text{número de 1's}}{n} = \frac{7}{10} = 0.7.$$

Error estándar de  $\hat{p}$  es

$$\sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} = \sqrt{\frac{0.7 \cdot (1 - 0.7)}{10}} = 0.1449138$$

Con R

```
muestra2=c(1,0,1,0,1,1,1,1,1,0)
frecuencias=table(muestra2)
frecuencias
```

```
## muestra2
## 0 1
## 3 7
```

```
n=length(muestra2)
n
```

```
## [1] 10
```

```
 exitos=frecuencias[2]
 exitos
```

```
## 1
## 7
```

```
phat= exitos/n
phat
```

```
## 1
## 0.7
```

```
error_estandar_phat=sqrt(phat*(1-phat)/n)
names(error_estandar_phat)=NULL
error_estandar_phat
```

```
## [1] 0.1449138
```

### Apartado c

Bajo estas condiciones población normal  $\sigma$  desconocida el intervalo para  $\mu$  al nivel de confianza del 95% es el del caso IV de la tabla de contrastes de una muestra

Tipo de contraste y condiciones				
Hipótesis nula	Condiciones	Muestra	Hipótesis alternativa	Caso
$H_0 : \mu = \mu_0$	Población normal o $n$ grande. $\sigma$ conocida.	$n$ observaciones independientes.	$H_1 : \mu \neq \mu_0$	<b>I</b>
			$H_1 : \mu < \mu_0$	<b>II</b>
			$H_1 : \mu > \mu_0$	<b>III</b>
	Población normal. $\sigma$ desconocida.	$n$ observaciones independientes.	$H_1 : \mu \neq \mu_0$	<b>IV</b>
			$H_1 : \mu < \mu_0$	<b>V</b>
			$H_1 : \mu > \mu_0$	<b>VI</b>

Detalles del contraste				
Caso	Estadístico	Región crítica	Intervalo confianza	$p$ -valor
<b>I</b>	$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$ es $N(0, 1)$	$\{Z \leq -z_{1-\frac{\alpha}{2}}\} \cup \{Z \geq z_{1-\frac{\alpha}{2}}\}$	$\left[ \bar{X} - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right]$	$2P(Z \geq  z )$
<b>II</b>		$\{Z \leq z_{\alpha}\}$	$\left[ -\infty, \bar{X} - z_{\alpha} \cdot \frac{\sigma}{\sqrt{n}} \right]$	$P(Z \leq z)$
<b>III</b>		$\{Z \geq z_{1-\alpha}\}$	$\left[ \bar{X} - z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}, \infty \right]$	$P(Z \geq z)$
<b>IV</b>	$T = \frac{\bar{X} - \mu_0}{\frac{\hat{S}}{\sqrt{n}}}$	$\{T \leq -t_{n-1, 1-\frac{\alpha}{2}}\} \cup \{T \geq t_{n-1, 1-\frac{\alpha}{2}}\}$	$\left[ \bar{X} - t_{n-1, 1-\frac{\alpha}{2}} \cdot \frac{\hat{S}}{\sqrt{n}}, \bar{X} + t_{n-1, 1-\frac{\alpha}{2}} \cdot \frac{\hat{S}}{\sqrt{n}} \right]$	$2P(t_{n-1} \geq  T )$
<b>V</b>		$\{T \leq t_{n-1, \alpha}\}$	$\left[ -\infty, \bar{X} - t_{n-1, \alpha} \cdot \frac{\hat{S}}{\sqrt{n}} \right]$	$P(t_{n-1} \leq T)$

```
media=mean(muestra1)
media
```

```
## [1] 0.4444444
```

```
n=length(muestra1)
n
```

```
## [1] 9
```

```
desv_tip=sd(muestra1)
desv_tip
```

```
## [1] 2.297341
```

```
alpha=1-0.95# 1-alpha/2=0.975
```

```
cuantil=qt(1-alpha/2,df=n-1)# cuantil 0.975 de la t de student con n-1 grados de libertad.
cuantil
```

```
## [1] 2.306004
```

El intervalo de confianza para  $\mu$  al nivel de confianza del 95% es

$$= \left[ \bar{x} - t_{n-1, 1-\alpha/2} \cdot \frac{\tilde{s}_X}{\sqrt{n}}, \bar{x} + t_{n-1, 1-\alpha/2} \cdot \frac{\tilde{s}_X}{\sqrt{n}} \right]$$

$$= \left[ 0.4444444 - 2.3060041 \cdot \frac{2.2973415}{3}, 0.4444444 + 2.3060041 \cdot \frac{2.2973415}{3} \right]$$

$$= [-1.3214485, 2.2103374, [$$

Con R se puede calcular así también

```
t.test(muestra1, alternative = "two.sided", conf.level = 0.95) -> solucion
solucion$conf.int
```

```
## [1] -1.321449 2.210337
## attr(,"conf.level")
## [1] 0.95
```

### Apartado d

Bajo estas condiciones población normal el intervalo para  $\sigma_X^2$  al nivel de confianza del 95% es el del caso XIII de la tabla de contrastes de una muestra

$H_0 : \sigma^2 = \sigma_0^2$	Población Normal. $\mu$ desconocida	$n$ observaciones independientes.	$H_1 : \sigma^2 \neq \sigma_0^2$	<b>XIII</b>
			$H_1 : \sigma^2 < \sigma_0^2$	<b>XIV</b>
			$H_1 : \sigma^2 > \sigma_0^2$	<b>XV</b>

```
knitr::include_graphics("casoXIII_2.PNG", dpi=180)
```

<b>XIII</b> <sup>1</sup>	$\chi^2 = \frac{(n-1)\tilde{s}^2}{\sigma_0^2} /$	$\{\chi^2 \leq \chi_{n-1, \frac{\alpha}{2}}^2\} \cup \{\chi^2 \geq \chi_{n-1, 1-\frac{\alpha}{2}}^2\}$	$\left[ \frac{(n-1)\tilde{s}^2}{\chi_{n-1, 1-\frac{\alpha}{2}}^2}, \frac{(n-1)\tilde{s}^2}{\chi_{n-1, \frac{\alpha}{2}}^2} \right]$	$2 \min\{P(\chi_{n-1}^2 \leq \chi^2), P(\chi_{n-1}^2 \geq \chi^2)\}$
--------------------------	--	--	---	--

```
media=mean(muestra1)
media
```

```
## [1] 0.4444444
```

```
n=length(muestra1)
n
```

```
## [1] 9
```

```
desv_tip=sd(muestra1)
desv_tip
```

```
## [1] 2.297341
```

```
alpha=1-0.95# 1-alpha/2=0.975
```

```
cuantil_1=qchisq(1-alpha/2, df=n-1)# cuantil 0.975 de una chi^2 con n-1 grados de libertad.
cuantil_1
```

```
## [1] 17.53455
```

```
cuantil_2=qchisq(alpha/2,df=n-1)# cuantil 0.025 de una chi^2 con n-1 grados de libertad.
cuantil_2
```

```
## [1] 2.179731
```

El intervalo de confianza para  $\sigma^2$  al nivel de confianza del 95% es

$$\left[ \frac{(n-1) \cdot \tilde{s}_X^2}{\chi_{n-1,1-\alpha/2}^2}, \frac{(n-1) \cdot \tilde{s}_X^2}{\chi_{n-1,\alpha/2}^2} \right] = \left[ \frac{8 \cdot 2.2973415^2}{17.5345461}, \frac{8 \cdot 2.2973415^2}{2.1797307} \right] = [2.407945, 19.3703843[.$$

## 1.2 Problema 2

Queremos comparar los rendimientos medidos en consumo de CPU de dos configuraciones (C1 y C2) de un servidor de datos tienen una media similar, de hecho queremos tener evidencia contra que el rendimiento medio del servidor C1 es superior al del servidor C2. No conocemos  $\sigma_1$  y  $\sigma_2$ . Disponemos de dos muestras independientes de consumo por hora realizados para cada configuración C1 y C2, de tamaños  $n_1 = n_2 = 100$ , respectivamente.

Para bajarlos utilizad la dirección de los ficheros `raw` que se muestran en el siguiente código

```
C1=read.csv(
  "https://raw.githubusercontent.com/joanby/estadistica-inferencial/master/datasets/C1.csv",
  header=TRUE)$time
C2=read.csv(
  "https://raw.githubusercontent.com/joanby/estadistica-inferencial/master/datasets/C2.csv",
  header=TRUE)$time

n1=length(na.omit(C1))
n1
```

```
## [1] 100
```

```
n2=length(na.omit(C2))
n2
```

```
## [1] 100
```

```
media.muestra1=mean(C1,na.rm=TRUE)
media.muestra1
```

```
## [1] 38.5841
```

```
media.muestra2=mean(C2,na.rm=TRUE)
media.muestra2
```

```
## [1] 33.7953
```

```
desv.tip.muestra1=sd(C1,na.rm=TRUE)
desv.tip.muestra1
```

```
## [1] 3.014567
```

```
desv.tip.muestra2=sd(C2,na.rm=TRUE)
desv.tip.muestra2
```

```
## [1] 6.727062
```

Calculamos las medias y las desviaciones típicas muestrales de los tiempos empleados para cada muestra. Los datos obtenidos se resumen en la siguiente tabla:

$$\begin{array}{ll} n_1 &= 100, & n_2 &= 100 \\ \bar{x}_1 &= 38.5841, & \bar{x}_2 &= 33.7953 \\ \tilde{s}_1 &= 3.014567, & \tilde{s}_2 &= 6.7270621 \end{array}$$

Se pide:

- Comentad brevemente el código de R explicando que hace cada instrucción.
- Contrastad si hay evidencia de que los rendimientos medios son distintas entre los dos grupos. En dos casos considerando las varianzas desconocidas pero iguales o desconocidas pero distintas. Tenéis que hacer el contraste de forma manual y con funciones de R y resolver el contraste con el  $p$ -valor.
- Calculad e interpretad los intervalos de confianza BILATERALES al nivel de confianza del 95% para la diferencia de medias de los rendimientos en los casos anteriores.
- Comprobad con el test de Fisher y el de Levene si las varianzas de las dos muestras son iguales contra que son distintas. Tenéis que resolver el test de Fisher con R y de forma manual y el test de Levene con R y decidir utilizando el  $p$ -valor.

### 1.2.1 Solución

**Apartado 1.** El código R carga en las variables C1 y C2 las variables `time` de dos data frames de un servidor en github y por lo tanto hemos tenido que pasar la url del fichero original o *raw*.

Luego calcula los estadísticos básicos para realizar las siguientes preguntas. Para los tamaños muestrales  $n_1$  y  $n_2$  se omiten los valores NA antes de asignar la `length` de los arrays. También se calculan las medias y las desviaciones típicas muestrales omitiendo (si es que hay) los valores no disponibles.

**Apartado 2.** Denotemos por  $\mu_1$  y  $\mu_2$  las medias de los tiempos de las configuraciones 1 y 2 respectivamente. El contraste que se pide es

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

Estamos en un diseño de comparación de medias entre dos muestras independientes ambas de tamaño 100 que es grande. Tenemos dos casos varianzas desconocidas pero iguales y varianzas desconocidas pero distintas. Las funciones de R del contraste para estos casos son:

#### Varianzas iguales

```
# test para varianzas iguales
t.test(C1,C2,var.equal = TRUE,alternative = "greater")

##
## Two Sample t-test
##
## data: C1 and C2
## t = 6.4963, df = 198, p-value = 0.0000000003258
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
## 3.570574 Inf
## sample estimates:
## mean of x mean of y
## 38.5841 33.7953
```

#### Varianzas distintas

```
# test para varianzas distintas
t.test(C1,C2,var.equal = FALSE,alternative = "greater")
```

```
##
## Welch Two Sample t-test
##
## data: C1 and C2
## t = 6.4963, df = 137.22, p-value = 0.0000000007014
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
## 3.568032 Inf
## sample estimates:
## mean of x mean of y
## 38.5841 33.7953
```

El  $p$ -valor en ambos casos es muy pequeño así que la muestra no aporta evidencias rechazar la hipótesis nula las medias son iguales contra que son distintas.

Veamos el cálculo manual.

### Varianzas desconocidas pero iguales, $n_1$ y $n_2$ grande

Si suponemos que  $\sigma_1 = \sigma_2$ , el estadístico de contraste es

$$t_0 = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \cdot \frac{((n_1-1)\tilde{S}_1^2 + (n_2-1)\tilde{S}_2^2)}{(n_1+n_2-2)}}} = \frac{38.5841 - 33.7953}{\sqrt{\left(\frac{1}{100} + \frac{1}{100}\right) \cdot \frac{((100-1)3.014567^2 + (100-1)6.727062^{12})}{(100+100-2)}}}$$

```
t0=(media.muestra1-media.muestra2)/
  sqrt((1/n1+1/n2)*
((n1-1)*desv.tip.muestra1^2+(n2-1)*desv.tip.muestra2^2)/(n1+n2-2))
t0
```

```
## [1] 6.496254
```

operando obtenemos que  $t_0 = 6.496254$ . y sabemos que sigue una distribución  $t$ -Student  $t_{n_1+n_2-2} = t_{198}$ . Para este hipótesis alternativa el  $p$ -valor es

$P(t_{198} > 6.4962536)$ , lo calculamos con R

```
t0=(media.muestra1-media.muestra2)/
  sqrt((1/n1+1/n2)*
((n1-1)*desv.tip.muestra1^2+(n2-1)*desv.tip.muestra2^2)/(n1+n2-2))
t0
```

```
## [1] 6.496254
```

```
n1
```

```
## [1] 100
```

```
n2
```

```
## [1] 100
```

```
(1-pt(t0,df=n1+n2-2)) # calculo la probabilidad del complementario
```

```
## [1] 0.0000000003257543
```

```
pt(t0,df=n1+n2-2,lower.tail = FALSE)# calcula el área la cola superior
```

```
## [1] 0.0000000003257544
```

### Varianzas desconocidas pero distintas, $n_1$ y $n_2$ grande



Si suponemos que  $\sigma_1 \neq \sigma_2$ , el estadístico de contraste es  $t_0 = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\tilde{S}_1^2}{n_1} + \frac{\tilde{S}_2^2}{n_2}}} \sim t_f$ , que, cuando  $\mu_1 = \mu_2$ , tiene distribución (aproximadamente, en caso de muestras grandes)  $t_f$  con

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

Calculamos el estadístico y los grados de libertad con R

```
t0=(media.muestra1-media.muestra2)/
  sqrt(desv.tip.muestra1^2/n1+desv.tip.muestra2^2/n2)
t0
```

```
## [1] 6.496254
```

```
f=(desv.tip.muestra1^2/n1+desv.tip.muestra2^2/n2)^2/
  ((1/(n1-1))*(desv.tip.muestra1^2/n1)^2+
   (1/(n2-1))*(desv.tip.muestra2^2/n2)^2)
f
```

```
## [1] 137.2203
```

El  $p$  valor es

```
# el p-valor de la función t.test de R
pt(t0,f,lower.tail = FALSE)
```

```
## [1] 0.0000000007014172
```

### Apartado 3

Los intervalos de confianza BILATERALES al nivel del 95% los podemos obtener así

```
t.test(C1,C2,var.equal=TRUE,
  alternative="two.sided",
  conf.level=0.95)$conf.int
```

```
## [1] 3.335101 6.242499
```

```
## attr("conf.level")
```

```
## [1] 0.95
```

```
t.test(C1,C2,var.equal=FALSE,
  alternative="two.sided",
  conf.level = 0.95)$conf.int
```

```
## [1] 3.331131 6.246469
```

```
## attr("conf.level")
```

```
## [1] 0.95
```

Son similares, podemos asegurar que la diferencia de medias se encuentra  $3.33 < \mu_1 - \mu_2 < 6.24$  al nivel del 95 la CPU del tipo C1 tiene una media de tiempo entre 3.33 y 6.14 mayor que la del y tipo C2. aproximadamente.

**Apartado 4** El test que nos piden es el de igualdad de varianzas

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 \neq \sigma_2^2 \end{cases}.$$

El test de Fisher de igualdad de varianzas

```
var.test(C1,C2,alternative ="two.sided" )

##
## F test to compare two variances
##
## data: C1 and C2
## F = 0.20082, num df = 99, denom df = 99, p-value = 0.0000000000000314
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 0.1351176 0.2984600
## sample estimates:
## ratio of variances
## 0.2008163
```

Obtenemos un  $p$ -valor bajo muy bajo no podemos aceptar la igualdad de varianzas.

De forma manual el estadístico de este test sabemos que es

$$f_0 = \frac{\tilde{S}_1^2}{\tilde{S}_2^2} = \frac{9.0876143}{45.2533646} = 0.2008163.$$

Que sigue una ley de distribución de Fisher y el  $p$ -valor es  $\min\{2 \cdot P(F_{n_1-1, n_2-1} \leq f_0), 2 \cdot P(F_{n_1-1, n_2-1} \geq f_0)\}$ .

que con R es

```
n1
## [1] 100
n2
## [1] 100
f0=desv.tip.muestra1^2/desv.tip.muestra2^2
f0
## [1] 0.2008163
pvalor=min(2*pf(f0,n1-1,n2-2),2*pf(f0,n1-1,n2-2,lower.tail = FALSE))
pvalor
## [1] 0.000000000000033926
```

Obtenemos los mismos resultados que con la función `var.test`.

El test de Levene con R tiene las mismas hipótesis que el anterior

```
library(car,quietly = TRUE)# pongo quietly para que quite avisos
tiempo=c(C1,C2)
grupo=as.factor(c(rep(1,length(C1)),rep(2,length(C1))))
leveneTest(tiempo~grupo)

## Levene's Test for Homogeneity of Variance (center = median)
##      Df F value    Pr(>F)
## group  1  42.221 0.0000000006461 ***
##      198
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El  $p$ -valor obtenido es bajo así que el test de Levene aporta evidencias contra la igualdad de varianzas entre de los tiempos de los dos grupos.

### 1.3 Problema 3

Se prueba la misma implementación de un algoritmo para reconocer caras de la base de datos de una empresa con dos diferentes tipos de cámaras.

Para ello  $n = 100$  trabajadores pasan por cada una de las cámaras 1 vez.

Los resultados se pueden cargar con el siguiente código.

```
caras=read.csv(
  "https://raw.githubusercontent.com/joanby/estadistica-inferencial/master/datasets/caras.csv",
  header=TRUE)
str(caras)

## 'data.frame': 100 obs. of 3 variables:
## $ empleado: int 1 2 3 4 5 6 7 8 9 10 ...
## $ aciertoA: int 0 1 1 1 1 1 1 1 1 1 ...
## $ aciertoB: int 1 1 1 1 1 1 1 1 1 1 ...

table(caras$aciertoA,caras$aciertoB)

##
##      0  1
## 0  0 12
## 1  1 87
```

Donde `empleadop` es la variable el identificador del empleado y `aciertoA` y `aciertoB` valen 1 si se acierta la identidad y 0 si se falla para el mismo empleado en cada una de las cámaras.

Se pide:

- Cargar los datos desde el servidor y calcular el tamaño de las muestras y la proporción de aciertos de cada muestra.
- Contrastar si hay evidencia de que las proporciones de aciertos con la cámara A son iguales que las del algoritmo con la cámara B. Definir bien las hipótesis y las condiciones del contraste. Resolver el contraste de forma manual utilizando R solo como calculadora y resolver el contraste con el  $p$ -valor (calculado con R).
- Resolver el contraste con funciones de R.
- Calcular un intervalo de confianza bilateral para la diferencia de las proporciones al nivel de confianza del 95% con R y de forma manual utilizando R como calculadora y para calcular los cuantiles.

#### 1.3.1 Solución

##### Apartado a

Cargamos los datos y hacemos los cálculos preliminares directamente desde el raw del github y construimos la tabla de contingencia de aciertos y fallos en las cámaras A y B

```
caras=read.csv(
  "https://raw.githubusercontent.com/joanby/estadistica-inferencial/master/datasets/caras.csv",
  header=TRUE)
str(caras)

## 'data.frame': 100 obs. of 3 variables:
## $ empleado: int 1 2 3 4 5 6 7 8 9 10 ...
## $ aciertoA: int 0 1 1 1 1 1 1 1 1 1 ...
## $ aciertoB: int 1 1 1 1 1 1 1 1 1 1 ...
```

```
tabla=table(caras$aciertoA,caras$aciertoB)
tabla
```

```
##
##      0   1
##  0   0  12
##  1   1  87
```

## Apartado b

Lo haremos por la tabla de comparación de dos proporciones para muestras emparejadas es el caso XXII de la tablas de contrastes de dos muestras

Si denotamos por  $p_A$  a la proporción de aciertos en la cámara A y  $p_B$  proporción de aciertos en la cámara B para muestras emparejadas. El contraste es

$$\begin{cases} H_0 : p_A = p_B \\ H_1 : p_A \neq p_B \end{cases}$$

	$H_0 : p_a = p_d$ Casodependiente	Poblaciones Bernoulli, $n_1$ y $n_2$ grandes, muchos casos discordants.	Dos m.a.s. dependientes de tamaño $n$	$H_1 : p_a \neq p_b$	XXII
				$H_1 : p_a < p_b$	XXIII
				$H_1 : p_a > p_b$	XXIV
XXII	$Z = \frac{\hat{p}_{1\bullet} - \hat{p}_{\bullet 1}}{\sqrt{\frac{b+d}{n^2}}}$	$\{Z \leq -z_{1-\frac{\alpha}{2}}\} \cup \{Z \geq z_{1-\frac{\alpha}{2}}\}$	$\left[ \hat{p}_{1\bullet} - \hat{p}_{\bullet 1} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{b+d}{n^2}}, \right.$ $\left. \hat{p}_{1\bullet} - \hat{p}_{\bullet 1} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{b+d}{n^2}} \right]$	$2P(Z \geq  z )$	
XXIII	es $N(0, 1)$ (vegeu (i))	$\{Z \leq z_\alpha\}$	$]-\infty, \hat{p}_{1\bullet} - \hat{p}_{\bullet 1} - z_\alpha \sqrt{\frac{b+d}{n^2}}[$	$P(Z \leq z)$	
XXIV		$\{Z \geq z_{1-\alpha}\}$	$\left[ \hat{p}_{1\bullet} - \hat{p}_{\bullet 1} - z_{1-\alpha} \sqrt{\frac{b+d}{n^2}}, +\infty \right]$	$P(Z \geq z)$	

(i) Para hacer el contraste, hemos de construir la tabla siguiente:

		Muestra después			
		éxito	Fracaso	Frecuencia	Proporción
Muestra antes	éxito	$a$	$b$	$a + b$	$\hat{p}_{1\bullet} = \frac{a+b}{n}$
	Fracaso	$d$	$c$	$c + d$	$\hat{p}_{2\bullet} = \frac{c+d}{n}$
	Frecuencia	$a + d$	$b + c$	$n$	
	Proporción	$\hat{p}_{\bullet 1} = \frac{a+d}{n}$	$\hat{p}_{\bullet 2} = \frac{b+c}{n}$		1

Entonces, el estadístico de contraste se puede escribir como:

$$Z = \frac{\frac{b}{n} - \frac{d}{n}}{\sqrt{\frac{b+d}{n^2}}}$$

Así que el estadístico son las discordancias

```
tabla=table(caras$aciertoA,caras$aciertoB)
tabla
```

```
##
```

```
##      0  1
##    0  0 12
##    1  1 87
```

Tenemos que las discordancias son  $b = 1$  es la frecuencia éxito en la A (filas) y fracaso en la B (columna) y  $d = 12$  es la frecuencia fracaso en la A (filas) y éxito en la B (columna) y  $n = 100$

```
b=1
d=12
n=100
z=(b/n-d/n)/sqrt((b+d)/n^2)
z
```

```
## [1] -3.050851
```

```
pvalor=2*pnorm(abs(z),lower.tail=FALSE)
pvalor
```

```
## [1] 0.002281937
```

El estadístico es

$$Z = \frac{\frac{b}{n} - \frac{d}{n}}{\sqrt{\frac{b+d}{n^2}}} = \frac{\frac{1}{100} - \frac{12}{100}}{\sqrt{\frac{1+12}{100^2}}} = -3.0508511.$$

### Apartado c

Es un diseño de muestras emparejadas y podemos por ejemplo con R utilizar el `mcnemar.test` (aunque no es exactamente el mismo que el test anterior):

```
mcnemar.test(tabla)
```

```
##
## McNemar's Chi-squared test with continuity correction
##
## data:  tabla
## McNemar's chi-squared = 7.6923, df = 1, p-value = 0.005546
```

El  $p$ -valor es 0.1356 no podemos rechazar la igualdad de la proporción de aciertos.

### Apartado d

Necesitamos estos cálculos

```
tabla
```

```
##
##      0  1
##    0  0 12
##    1  1 87
```

```
pA=(87+1)/100
pA
```

```
## [1] 0.88
```

```
pB=(12+87)/100
pB
```

```
## [1] 0.99
```

```
b=1
d=12
n=100
alpha=1-0.95
alpha
```

```
## [1] 0.05
```

```
cuantil=qnorm(1-0.05/2)
cuantil
```

```
## [1] 1.959964
```

Viendo las tablas tenemos que calcular

$\hat{p}_{\bullet 1} = \hat{p}_A = 0.88$  proporción aciertos en  $A$ ,  $\hat{p}_{1\bullet} = \hat{p}_B = 0.99$  proporción aciertos en  $B$

El intervalo de confianza al nivel 95% es ( $\alpha = 0.05$ )

$$\begin{aligned} & \left[ \hat{p}_A - \hat{p}_B - z_{1-\frac{\alpha}{2}} \sqrt{\frac{b+d}{n^2}}, \hat{p}_A - \hat{p}_B + z_{1-\frac{\alpha}{2}} \sqrt{\frac{b+d}{n^2}} \right] \\ &= \left[ 0.88 - 0.99 - 1.959964 \cdot \sqrt{\frac{1+12}{100^2}}, 0.88 - 0.99 + 1.959964 \cdot \sqrt{\frac{1+12}{100^2}} \right] \\ &= ]-0.1806675, -0.0393325[ \end{aligned}$$

## 1.4 Problema 4

El encargado de calidad piensa que  $X =$  número de quejas de clientes por día en las oficinas de atención al cliente de una determinada zona de una ciudad sigue una ley  $Po(\lambda = 5)$ . Para comprobarlo toma una muestra de  $n = 100$  días:

```
quejas=read.csv(
  "https://raw.githubusercontent.com/joanby/estadistica-inferencial/master/datasets/quejas.csv",
  header=TRUE)
str(quejas)
```

```
## 'data.frame': 100 obs. of 1 variable:
## $ Num_quejas: int 4 6 4 2 6 2 7 10 7 4 ...
```

```
ni=c(0,table(quejas))
names(ni)[1]="0"
ni
```

```
## 0 1 2 3 4 5 6 7 8 9 10 11
## 0 1 8 11 16 16 14 14 11 4 4 1
```

```
n=sum(ni)
n
```

```
## [1] 100
```

```
pi=c(dpois(0:10,lambda=5),1-sum(dpois(0:10,lambda=5)))
names(pi)=c(paste0("Prob(X=",0:10,")"),"Prob(X>=11)")
pi
```

```
## Prob(X=0) Prob(X=1) Prob(X=2) Prob(X=3) Prob(X=4) Prob(X=5)
## 0.006737947 0.033689735 0.084224337 0.140373896 0.175467370 0.175467370
```

```
## Prob(X=6) Prob(X=7) Prob(X=8) Prob(X=9) Prob(X=10) Prob(X>=11)
## 0.146222808 0.104444863 0.065278039 0.036265577 0.018132789 0.013695269
```

```
sum(pi)
```

```
## [1] 1
```

```
ei=n*pi
ei
```

```
## Prob(X=0) Prob(X=1) Prob(X=2) Prob(X=3) Prob(X=4) Prob(X=5)
## 0.6737947 3.3689735 8.4224337 14.0373896 17.5467370 17.5467370
## Prob(X=6) Prob(X=7) Prob(X=8) Prob(X=9) Prob(X=10) Prob(X>=11)
## 14.6222808 10.4444863 6.5278039 3.6265577 1.8132789 1.3695269
```

```
ei>5
```

```
## Prob(X=0) Prob(X=1) Prob(X=2) Prob(X=3) Prob(X=4) Prob(X=5)
## FALSE FALSE TRUE TRUE TRUE TRUE
## Prob(X=6) Prob(X=7) Prob(X=8) Prob(X=9) Prob(X=10) Prob(X>=11)
## TRUE TRUE TRUE FALSE FALSE FALSE
```

```
# no se cumple la condición para el test chi^2
#hay que agrupar los 3 primeros y los 3 últimos
# test chi^2 sin agrupar...
chi0=sum((ei-ni)^2/ei)
chi0
```

```
## [1] 10.36668
```

```
k=12# clases de 0 a mayor o igual 11
k=12# clases de 0 a 11
pchisq(chi0,df=k-1,lower.tail=FALSE)
```

```
## [1] 0.4977365
```

Se pide:

- Plantead un contraste de bondad de ajuste  $\chi^2$   $H_0$ : los datos siguen una distribución  $Po(\lambda = 5)$ . Calculas las probabilidades y frecuencias esperadas utilizando los datos del código anterior.
- Reagrupar los datos y resolver el test manualmente pero usando R para el cálculo del  $p$ -valor. Resolver el contraste
- Resolver el contraste con la función adecuada de R.

### 1.4.1 Solución

#### Apartado a

El contraste que se pide es

$$\begin{cases} H_0 : & \text{El número de quejas diarias sigue una distribución } Po(\lambda = 5) \\ H_1 : & \text{El número de quejas diarias sigue otra distribución} \end{cases}$$

Con el código que se dan las frecuencias observadas es el array `ni` hay  $k = 12$  clases y las probabilidades de cada clase bajo la hipótesis nula estén en `pi` el valor de  $n = 100$  y las frecuencias esperadas están en `ei`

```
#observadas
ni
```

```
## 0 1 2 3 4 5 6 7 8 9 10 11
## 0 1 8 11 16 16 14 14 11 4 4 1
```

```

k=length(ni)# numero de clases
k

## [1] 12

pi# probabilidad de cada supuesto que H0 es cierta Po(lambda=5)

## Prob(X=0) Prob(X=1) Prob(X=2) Prob(X=3) Prob(X=4) Prob(X=5)
## 0.006737947 0.033689735 0.084224337 0.140373896 0.175467370 0.175467370
## Prob(X=6) Prob(X=7) Prob(X=8) Prob(X=9) Prob(X=10) Prob(X>=11)
## 0.146222808 0.104444863 0.065278039 0.036265577 0.018132789 0.013695269

n# tamaño de la muestra

## [1] 100

ei# frecuencias esperadas de cada supuesto que H0 es cierta Po(lambda=5)

## Prob(X=0) Prob(X=1) Prob(X=2) Prob(X=3) Prob(X=4) Prob(X=5)
## 0.6737947 3.3689735 8.4224337 14.0373896 17.5467370 17.5467370
## Prob(X=6) Prob(X=7) Prob(X=8) Prob(X=9) Prob(X=10) Prob(X>=11)
## 14.6222808 10.4444863 6.5278039 3.6265577 1.8132789 1.3695269

```

### Apartado b

El siguiente código hace los cálculos manualmente para agrupar las clases que obtienen frecuencias absolutas esperadas “ei” inferiores a 5. Agrupamos las tres primeras clases y las tres últimas quedando ahora  $k = 8$  clases/grupos.

```

ni

## 0 1 2 3 4 5 6 7 8 9 10 11
## 0 1 8 11 16 16 14 14 11 4 4 1

pi

## Prob(X=0) Prob(X=1) Prob(X=2) Prob(X=3) Prob(X=4) Prob(X=5)
## 0.006737947 0.033689735 0.084224337 0.140373896 0.175467370 0.175467370
## Prob(X=6) Prob(X=7) Prob(X=8) Prob(X=9) Prob(X=10) Prob(X>=11)
## 0.146222808 0.104444863 0.065278039 0.036265577 0.018132789 0.013695269

chisq.test(ni,p=pi)

## Warning in chisq.test(ni, p = pi): Chi-squared approximation may be incorrect
##
## Chi-squared test for given probabilities
##
## data: ni
## X-squared = 10.367, df = 11, p-value = 0.4977

chisq.test(ni,p=pi,simulate.p.value = TRUE,B=5000)# test simulando 5000 las ni

##
## Chi-squared test for given probabilities with simulated p-value (based
## on 5000 replicates)
##
## data: ni
## X-squared = 10.367, df = NA, p-value = 0.4871

```



```

# de muestra de tamaño 100 con estas pi
ni_agrupado=c(sum(ni[1:3]),ni[4:9],sum(ni[10:12]))
ni_agrupado

##      3  4  5  6  7  8
##  9 11 16 16 14 11  9

pi_agrupado=c(sum(pi[1:3]),pi[4:9],sum(pi[10:12]))
pi_agrupado

##      Prob(X=3) Prob(X=4) Prob(X=5) Prob(X=6) Prob(X=7) Prob(X=8)
## 0.12465202 0.14037390 0.17546737 0.17546737 0.14622281 0.10444486 0.06527804
##
## 0.06809363
sum(pi_agrupado)

## [1] 1
n=sum(ni)
n

## [1] 100
ei_agrupado=n*pi_agrupado
ei_agrupado

##      Prob(X=3) Prob(X=4) Prob(X=5) Prob(X=6) Prob(X=7) Prob(X=8)
## 12.465202 14.037390 17.546737 17.546737 14.622281 10.444486  6.527804  6.809363
ei_agrupado>=5

##      Prob(X=3) Prob(X=4) Prob(X=5) Prob(X=6) Prob(X=7) Prob(X=8)
##      TRUE      TRUE      TRUE      TRUE      TRUE      TRUE      TRUE      TRUE
k=length(ei_agrupado)
k

## [1] 8
El estadístico de contraste calculado manualmente es
chi0=sum((ni_agrupado-ei_agrupado)^2/ei_agrupado)
chi0

## [1] 6.898705
El p-valor es  $P(\chi^2_{8-1} > 6.8987051)$  lo calculamos con R
1-pchisq(chi0,df=8-1,lower.tail=TRUE)

## [1] 0.4395024
pchisq(chi0,df=8-1,lower.tail=FALSE)

## [1] 0.4395024
El p-valor es alto no podemos rechazar que la distribución sea  $Po(\lambda = 5)$ 

```

### Apartado c

Con la función de R es muy sencillo

```
chisq.test(ni_agrupado,p=pi_agrupado)
```

```
##  
## Chi-squared test for given probabilities  
##  
## data: ni_agrupado  
## X-squared = 6.8987, df = 7, p-value = 0.4395
```

Notemos que si no agrupamos

```
chisq.test(ni,p=pi)
```

```
## Warning in chisq.test(ni, p = pi): Chi-squared approximation may be incorrect  
##  
## Chi-squared test for given probabilities  
##  
## data: ni  
## X-squared = 10.367, df = 11, p-value = 0.4977
```

El test nos avisa que la aproximación del  $p$ -valor por una  $\chi^2_{12-1}$  puede ser incorrecta.

Otra opción recurrir a la simulación del test (*Monte Calo*) con el código

```
chisq.test(ni,p=pi,simulate.p.value = TRUE,B =5000)
```

```
##  
## Chi-squared test for given probabilities with simulated p-value (based  
## on 5000 replicates)  
##  
## data: ni  
## X-squared = 10.367, df = NA, p-value = 0.4829
```

En cualquier caso los  $p$ -valores son altos y se acepta la hipótesis nula.