Universidad Rey Juan Carlos

FACULTAD DE CIENCIAS DE LA SALUD

GRADO DE MEDICINA

BIOESTADÍSTICA

Curso 2017-18

Prof. Jesús Esteban Hernández.

Área de Medicina Preventiva y Salud Pública.

Dpto. de Medicina y Cirugía, Psicología,
Medicina Preventiva y Salud Pública,
Inmunología y Microbiología Médicas ,
Enfermería y Estomatología.

ÍNDICE DE CONTENIDOS

SUGERENCIAS PARA LAS CLASES PRÁCTICAS

1. Si vas a asistir a clase, por favor sé puntual.
2. Apaga el móvil cuando entres en clase.
3. Si no estamos en el aula de informática, acude a clase de problemas con una calculadora científica sencilla. Casi seguro que tu móvil incluye o puede instalar una. No es esencial que tenga modo estadístico.
4. El profesor te irá indicando los ejercicios a realizar cada semana.
5. Dada la limitación de tiempo, en clase solo da tiempo a corregir algunos de ellos. Por ello este año hemos incluido en el cuaderno su resolución, aunque te recomendamos que intentes hacerlos por tu cuenta antes de acudir a la clase en la que se comenten las soluciones.
6. Dejaremos en el aula virtual un manual de R en español, aunque puedes encontrar multitud de ellos y foros de consulta en la red. En cualquier caso, todo el código que necesitas está en las diapositivas de las sesiones y por supuesto en la ayuda de R.
7. Si tienes cualquier duda sobre las prácticas, por favor pregunta, pregunta, pregunta...
8. R está instalado en todos los ordenadores de la universidad, pero te recomendamos que te lo instales en tu ordenador personal para practicar en casa. Al ser software libre, no tendrás problemas con las licencias.
9. Recuerda que **el examen 'teórico'** incluye la resolución de varios problemas y cuestiones para los que se te suministrará una hoja con las fórmulas y por supuesto las tablas de las distribuciones de probabilidad.
10. **El examen práctico consistirá en la respuesta a varias preguntas a partir del análisis de una base de datos problema utilizando R**. En la prueba se analizará su competencia para:
    a. Describir datos con los estadísticos descriptivos y los gráficos adecuados a la variable descrita.
    b. Saber contrastar la normalidad de una variable.
    c. Construir intervalos de confianza para media y proporción de manera global y por grupos.
    d. Realizar e interpretar contrastes de hipótesis de conformidad y homogeneidad para medias y proporciones.
    e. Conocer los criterios para saber cuándo es necesario recurrir a métodos no paramétricos y cómo pedirlo en R

Inferencia estadística: Contraste de hipótesis.
Cuaderno de problemas.

OBJETIVOS DE ESTA SECCIÓN.

- Identificar los tipos de contrastes de hipótesis.
- Identificar las hipótesis (nula y alternativa) en cada caso.
- Resolver contrastes de hipótesis mediante el estadístico y cuando sea posible obtener el intervalo de confianza adecuado para:
    - Contrastar una media respecto a un valor teórico de media poblacional.
    - Contrastar una media respecto a un valor teórico de proporción poblacional.
    - Comparar dos medias de grupos independientes.
    - Comparar dos medias de grupos emparejados.
    - Comparar proporciones de k grupos independientes.
    - Comparar proporciones de dos grupos emparejados.

Recuerda que es muy importante <u>obtener e interpretar</u> el intervalo de confianza que corresponda en función del contraste.

1. Estatura

Se sabe que la estatura de los individuos de una ciudad se distribuye según una normal. A partir de una muestra de 25 personas, se obtuvo que por término medio, los individuos medían 170 cm, con una desviación típica de 10 cm.

Cuestiones:

a) ¿Podría haber salido esta muestra de una población con una media de altura de 174 cm? Para contestar, sigue los pasos siguientes:

1. Plantea las hipótesis nula y alternativa.
2. Calcula el valor del estadístico de contraste.
3. Calcula el p-valor.
4. ¿Qué decisión tomarías con $\alpha = 0.05$?

b) ¿Qué dirías si $\alpha = 0.01$?

Solución.

Para resolver el ejercicio es necesario primero identificar el tipo de prueba en la que nos encontramos. Al disponer de una única muestra y con el objetivo de comprobar si nuestra altura media es igual o diferente a una dada, nos encontramos ante una prueba de conformidad de medias. Para resolverlo planteamos nuestra hipótesis nula y alternativa:
Contraste.

$$H_0: \mu = 174$$
$$H_1: \mu \neq 174$$

$t_{exp} = \frac{\bar{x}-\mu}{s/\sqrt{n}}$, utilizamos s porque desconocemos σ

$t_{exp} = \frac{170-174}{10/\sqrt{25}} = -2$

$$p\ valor = 2 \cdot P[t_{n-1} \geq |t_{exp}|]$$
$$p\ valor = 2 \cdot P[t_{25-1} \geq 2] = 2 \cdot 0.02847$$
$$p\ valor\ bilat. = 0.05694$$

Si construimos el IC95%:

	mean	t/z	0.025	se	lower	upper	texp	pvalue
ci usando t	170	2.0639	2		**165.8722**	**174.1278**	-2	0.0569
ci usando Z	170	1.9600	2		166.0801	173.9199	-2	0.0455

"*Lo correcto es usar la *t-Student*"

Con este p-valor no rechazaríamos la hipótesis nula pues la diferencia encontrada no es suficientemente importante como para afirmar que esta muestra no ha salido de una población cuya media de altura es de 174. Como se observa el p-valor es cercano a alfa y si hubiésemos utilizado la z para decidir la conclusión habría sido rechazarla (p-valor=0.0455). Lo correcto es usar la distribución t-student, pues los gl son menos de 30. Este ejemplo abunda en el sinsentido de interpretar los contrastes de hipótesis y en concreto el p-valor como meras dicotomías.

Apartado b) CI90%.

	mean	t/z	0.05	se	lower	upper	texp	pvalue
ci usando t	170	1.7109	2		**166.5782**	**173.4218**	-2	0.0569
ci usando Z	170	1.6449	2		166.7103	173.2897	-2	0.0455

En este caso $(\alpha = 0.01)$ habríamos rechazado H_0 al asumir un mayor riesgo de equivocarnos al rechazar H_0.

2. Accidentes mortales

El número de accidentes mortales en una ciudad es, en promedio, de 12 mensuales. Tras una campaña de señalización y adecentamiento de las vías urbanas se contabilizaron en los siguientes 6 meses 8, 11, 9, 7, 10, 9 accidentes mortales.

Cuestiones:

Suponemos que el número de accidentes mortales sigue una distribución normal. ¿Fue efectiva la campaña? Para contestar, sigue los pasos siguientes:

1. Plantea las hipótesis nula y alternativa.
2. Calcula el valor del estadístico de contraste.
3. ¿Qué decisión tomarías?
4. Si la campaña fue efectiva, ¿con qué nivel de significación podemos afirmarlo? Razona la respuesta.

Solución.

Nos dicen que la distribución es normal. Este dato es importante dado que el número de meses en los que se ha recogido el dato es 6 y por tanto no podríamos apoyarnos en el Teorema del Límite Central para construir el IC y el contraste.

$$H_0: \mu = 12$$
$$H_1: \mu \neq 12$$
$$t_{exp} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

utilizamos s porque desconocemos σ

$$t_{exp} = \frac{9 - 12}{\frac{1.414}{\sqrt{6}}} = \frac{9 - 12}{0.577} = -5.196$$

En las tablas podemos comprobar que el p-valor es menor que 0.005 (bilateral). Según la tabla de la distribución t-student el punto que deja un área de 0.0025 (0.005 considerando ambas colas) a derecha es 4.773. El estadístico de contraste está más lejos por lo que el p-valor<0.005.

Con la ayuda de R,

```
alfa<-c(.05,.01,.005,.001)
qt(alfa/2,5)
[1] -2.570582 -4.032143 -4.773341 -6.868827
```

Como podemos ver es <0.005 pero >0.001. De hecho el p-valor (bilat.) del contraste es 0.0035).

```
            mean t/z 0.025     se  lower   upper    texp pvalue
ci usando t    9    2.5706 0.5774 7.5159 10.4841 -5.1962 0.0035
ci usando Z*   9    1.9600 0.5774 7.8684 10.1316 -5.1962 0.0010
```

"*Recuerde que este estadístico muestral en realidad sigue una t-student, aunque como se ha visto en clase, conforme aumenta n la t-student se parece más a una z"

Rechazamos la hipótesis nula. El promedio de accidentes mensual (12) se ha reducido significativamente (a 9) pues es poco probable (p-valor<alfa)

haber observado una diferencia como esta o mayor si la hipótesis nula fuese cierta. Esto se observa también al leer el IC(95%). Tenemos una confianza del 95% de que el valor de μ esté entre 7.52 y 10.48, no estándo lo que afirma la H_0 entre estos valores. Dicho de otra manera, es menos probable del 5% que una muestra con una media como la encontrada o mayor, haya salido de una población en la que $\mu = 12$.
Por tanto parece que la campaña ha sido eficaz.
Con R esto mismo se podría haber obtenido utilizando un vector.

```
acc<-c(8, 11, 9, 7, 10, 9)
t.test(acc,mu=12)
```

```
 One Sample t-test

data:  acc
t = -5.1962, df = 5, p-value = 0.003478
alternative hypothesis: true mean is not equal to 12
95 percent confidence interval:
  7.515874 10.484126
sample estimates:
mean of x
        9
```

3. Peso de recién nacidos

En un estudio sobre el retraso del crecimiento intrauterino llevado a cabo en un Hospital madrileño, se registraron los pesos y longitudes cráneo-raquis (CRL) de 400 recién nacidos. 100 eran hijos de fumadoras y los otros 300 eran hijos de no fumadoras. Los resultados fueron los siguientes:

Peso		
	Media ($\bar{x}$)	Desviación típica (s)
Fumadoras	2,0 Kg	0,3 Kg
No fumadoras	2,8 Kg	0,5 Kg

Longitud C-R		
	Media ($\bar{x}$)	Desviación típica (s)
Fumadoras	60	15
No fumadoras	65	10

Cuestiones:

a) ¿Es diferente el peso en hijos de fumadoras que de no fumadoras? Resuelva el contraste asumiendo que las **varianzas no son homogéneas**.
   1. Plantea la hipótesis nula y alternativa para esta comparación.
   2. Calcula el valor del estadístico de contraste.
   3. Para un nivel de significación α = 0,05 ¿Qué decisión tomarías?

b) Si rechaza la hipótesis nula, ¿con qué nivel de significación podemos llegar a hacerlo? Razone la respuesta.

c) ¿Es diferente la longitud C-R en hijos de fumadoras que de no fumadoras? Resuelva el contraste asumiendo que las **varianzas no son homogéneas.**
   1. Plantea las hipótesis nula y alternativa para esta comparación.
   2. Calcula el valor del estadístico de contraste.
   3. Para un nivel de significación α = 0,01 ¿Qué decisión tomaría?

d) Si es diferente ¿para qué nivel de significación podríamos llegar a afirmarlo? Razone la respuesta.

Solución: `Comparación de medias 2 grupos no emparejados.`

`Hipótesis:`

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

luego,

$$EE_{dm} = \sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}$$

por tanto

$$t_{exp} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

pero con gl de Welch

$$\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

PESO.

$$EE_{dm} = \sqrt{\left(\frac{0.3^2}{100} + \frac{0.5^2}{300}\right)}$$

por tanto

$$t_{exp} = \frac{2.8 - 2}{\sqrt{\left(\frac{0.09}{100} + \frac{0.025}{300}\right)}} = \frac{0.8}{0.0416} = 19.22$$

pero con gl de Welch

$$\frac{\left(\frac{0.3^2}{100} + \frac{0.5^2}{300}\right)^2}{\frac{\left(\frac{0.3^2}{100}\right)^2}{99} + \frac{\left(\frac{0.5\text{^}2}{300}\right)^2}{299}} = 286$$

	Mean	**SD**	**N**
Group1	2.0	0.3	100
Group2	2.8	0.5	300

	F	p-value	Var. Equal
Hom.Var	1.667	<0.001	FALSE

	group1	group2	diff.	SError	CI.lower	CI.upper	t	df	p-value
t-Test	2	2.8	-0.8	0.04163	-0.8819	-0.7181	-19.2169	286	< 0.001

En nuestra muestra, la media de Peso es **0.8 kg mayor** en los hijos de no fumadoras. Tenemos una confianza del 95% de que de promedio el Peso en hijos de **no fumadoras** sea **entre 0.71 y 0.8 kg superior** en los hijos de **fumadoras**.
Esto es lo mismo que decir que la probabilidad de haber observado unas diferencias como las observadas o mayores **si la H_0 fuese cierta** es **<0.001,** por lo que rechazaríamos la **H_0**.

LCR

	Mean	SD	N
Group1	60	15	100
Group2	65	10	300

	F	p-value	Var. Equal
Hom.Var	1.5	0.005	FALSE

	group1	group2	diff.	SError	CI.lower	CI.upper	t	df	p-value
t-Test	60	65	-5	1.607	-8.1793	-1.8207	-3.11	129.6	0.0023

En nuestra muestra, la media de LCR es **5 cm mayor** en los hijos de no fumadoras. Tenemos una confianza del 95% de que de promedio el LCR en hijos de **no fumadoras** sea **entre 1.8 y 8.2 cm superior** al de los hijos de **fumadoras**.

Esto es lo mismo que decir que la probabilidad de haber observado unas diferencias como las observadas o mayores **si la H_0 fuese cierta** es de **0.0023,** por lo que rechazaríamos la **H_0.**

Dado que el p-valor es también menor que 0.01 (alfa), habríamos rechazado la hipótesis nula de igualdad también si ese hubiera sido alfa.

4. Vacunación

Se está probando un nuevo sistema de aplicación de una vacuna entre los estudiantes de una Universidad madrileña. Se desea saber si la proporción de acontecimientos adversos es diferente de la proporción del sistema antiguo (30%). En nuestro estudio se vacunaron 150 alumnos y la proporción de efectos adversos fue del 24%.

Cuestiones:

¿Es esta proporción distinta de la del sistema anterior? ¿Con qué nivel de confianza se puede afirmar esto? No olvides:

1. Plantear la hipótesis nula y la hipótesis alternativa.
2. Calcular el valor del estadístico de contraste.
3. ¿Qué decisión tomarías para un nivel de significación α = 0,01?

Solución.

1. Hipótesis.

$$\begin{cases} H_0: \pi = 0.3 \\ H_1: \pi \neq 0.3 \end{cases}$$

2. Estadístico de contraste y p-valor.

$$z_{exp} = \frac{p - \pi_0}{\sqrt{\frac{\pi_0 \cdot (1-\pi_0) q}{n}}},$$

Puesto que se trata de una prueba de conformidad contra valor teórico para una proporción, el estadístico para este tamaño muestral y proporción se construiría así:

$$z_{exp} = \frac{0.24 - 0.3}{\sqrt{\frac{0.3 \cdot 0.7}{150}}} = -1.6$$

El área que deja -1.6 a su izquierda es >0.025, que sumado al que deja +1.6 a su derecha es >0.05. Dicho de otro modo lo observado está solo a -1.6 veces el EE (recordad que esto en el fondo es una 'desviación típica' del estadístico proporción muestral asumiendo que se cumple el TLC) y por tanto no lo suficientemente lejos para rechazar la hipótesis nula.

```
       prop propTeor    EEp    Zexp   pval
Z-test 0.24      0.3 0.0374 -1.6043 0.1087
Exact  0.24      0.3     NA      NA 0.1294
                  Point.Est     EE CI.Lower CI.Upper
Asymp. CI with Pi      0.24 0.0374   0.1716   0.3084
Asymp. CI with p       0.24 0.0349   0.1667   0.3133
Exact CI               0.24          0.1741   0.3165
```

3. Decisión.

El cambio observado (p=24%) no es lo suficientemente importante como para rechazar la hipótesis nula. La probabilidad de haber observado una muestra como esta o más distante si la proporción de efectos adversos en la población fuese del 30% es de 0.11 (**p-valor>alfa**) y por tanto no podemos rechazar la hipótesis nula. Como se ve el IC alrededor de mi estimador no incluye lo que afirma la hipótesis nula (H_0).

5. Infarto en mujeres jóvenes.

En un estudio de casos y controles sobre la relación entre el consumo de cierto fármaco A y el infarto de miocardio en mujeres entre 30 y 45 años (una edad muy temprana para que una mujer padezca un infarto), se reclutaron mujeres con y sin infarto de forma que por cada mujer con infarto se buscaba otra sin infarto de la misma edad y hábito tabáquico.

Se reclutaron 300 mujeres y otras 300 que no lo habían padecido hasta la fecha. En la siguiente tabla se muestra la distribución de la exposición en cada grupo:

		SANAS		TOTAL
		SÍ FÁRMACO A	**NO FÁRMACO A**	
INFARTO	**SÍ FÁRMACO A**	40	82	122
	NO FÁRMACO A	6	172	178
	TOTAL	46	254	300

Cuestiones

1. ¿Qué tipo de estudio han realizado los investigadores? Descríbalo con todo el detalle que pueda y razonando su elección.
2. Teniendo en cuenta el diseño escogido, ¿Se asoció a un mayor riesgo de infarto el consumo del fármaco A? Para responder a esta pregunta:
   a. Formalice las hipótesis estadísticas del contraste.
   b. Elija y calcule el estadístico de contraste adecuado.
   c. Obtenga el estimador de la diferencia adecuado para este estudio.
   d. Calcule el intervalo de confianza del 95% (IC95%) de dicho estimador.
   e. Interprete el resultado en términos de p-valor y de IC95%.

Solución.

Por la descripción sabemos que se trata de un diseño emparejado (por edad) con dos grupos tratadas y no tratadas. En concreto es un estudio de casos y controles emparejado por la variable edad en la que se explorar la exposición previa (consumo) del fármaco A.

Planteamiento de las hipótesis estadísticas:

$$\begin{cases} H_0: \pi_{disc_{SANAS}} = \pi_{disc_{IAM}} \\ H_1: \pi_{disc_{SANAS}} \neq \pi_{disc_{IAM}} \end{cases}$$

$$EE_{dp} = \frac{1}{300}\sqrt{82+6)}$$

$$CI(1-\alpha\%): \left(\frac{82}{300} - \frac{6}{300}\right) \pm z_{\frac{\alpha}{2}} \cdot \frac{1}{300}\sqrt{(82+6)}$$

$$z_{exp} = \frac{27.3\% - 2\%}{\frac{1}{300}\sqrt{(82+6)}} =$$

$$z_{exp} = \frac{\frac{b}{N} - \frac{c}{N}}{\frac{1}{N}\sqrt{b+c)}} = \frac{82-6}{\sqrt{82+6}} = 8.102$$

$$z^2_{exp} = \frac{(b-c)^2}{b+c} \rightsquigarrow \chi^2$$
$$gl = r - 1$$
$$\chi^2_{Mcnemar_{exp}} = \frac{(82-6)^2}{82+6} = 65.6$$

	iam no / fármaco A sí	**fármaco A no**
iam sí / fármaco A sí	40	82
iam sí / fármaco A no	6	172

	prob b	**prob c**	**dif.prop**	**EEdp**	**CIlower**	**CIupper**	**z/chi2**	**df**	**p.value**
Z	0.2733	0.02	0.2533	0.0313	0.192	0.3146	8.1016	NA	<0.001
chi2	NA	NA	NA	NA	NA	NA	65.6364	1	<0.001
chi2Corr	NA	NA	NA	NA	NA	NA	63.9205	1	<0.001

Las diferencias (IC95%) 25.3%(19.2-31.5) son suficientemente importantes (p-valor<0.001) como para rechazar la hipótesis nula.
Vemos que las discrepancias son más frecuentes en el sentido que asocia el fármaco con la aparición de infarto, dicho de otro modo la exposición al fármaco A es más frecuente en las pacientes que lo han sufrido que en las mujeres sanas.

6. Colesterol infantil

En la realización de un estudio sobre los niveles de colesterol en niños se compararon los niveles medios de colesterol total en sangre entre hijos de pacientes que habían padecido un Infarto Agudo de Miocardio (IAM) obteniéndose los siguientes resultados:

Niveles de colesterol	N	Media (mg/dl)	Desviación típica (s)
Hijos de pacientes con IAM	44	200	30
Hijos de pacientes sanos	78	197	40

Cuestiones:

¿Existen diferencias significativas entre los niveles medios de colesterol entre los niños de los dos grupos? Resuelve el contraste siguiendo los pasos que se indican a continuación, asumiendo varianzas homogéneas.

1. Plantea la hipótesis nula y la hipótesis alternativa.
2. Calcula el valor del estadístico de contraste.
3. Para un nivel de significación α = 0,05 ¿Qué decisión tomarías?

Solución.

1. $H_0: \mu_1 = \mu_2$
   $H_1: \mu_1 \neq \mu_2$
2. Estadístico de contraste.

$$S_p^2 = (n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2$$
$$S_p = \sqrt{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}$$
$$S_p = \sqrt{(44 - 1) \cdot 30^2 + (78 - 1) \cdot 40^2}$$
$$S_p = 36.73$$

$$EE_{dm} = \sqrt{S_p^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$
$$EE_{dm} = 36.73 \cdot \sqrt{\left(\frac{1}{44} + \frac{1}{78}\right)}$$
$$EE_{dm} = 36.73 \cdot \sqrt{0.035548}$$
$$EE_{dm} = 6.9253 \; \upsilon = 44 + 72 - 2 = 114.$$

y por tanto…

$$t_{exp} = \frac{200 - 197}{6.9253} = 0.43319$$

En la tabla no aparece una $t_{\upsilon=114}$ y la más similar sería la $t_{\upsilon=120}$ en la que el cuantil que deja un área de 0.025 (0.050 bilateral) a su izquierda es -1.98 (+1.98 deja el mismo área por su derecha)[1]. Nuestro estadístico de contraste ha quedado a su izquierda, el área que deja es menor y por tanto no rechazamos la hipótesis nula.

Si utilizamos el EE podemos obtener el IC de la diferencia de medias.

$$CI(1 - \alpha)\% = (\bar{x}_1 - \bar{x}_2) \pm \sqrt{S_p^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$
$$(200 - 197) \pm 1.98 \cdot 6.9253$$
$$CI95\%: (-10.71; +16.71)$$

[1] En la $t_{\upsilon=114}$ el cuantil para el mismo área es -1.977

La diferencia entre las medias ha sido 3 mg/dL superior en los hijos de pacientes con IAM, pero no es tan poco probable como para rechazar H_0 y afirmar que las poblaciones de las que han salido son diferentes o, traducido a la hipótesis, no tenemos evidencia suficiente como para afirmar que las poblaciones de las que han salido (a las que representan) estos grupos son diferentes en lo que respecta a los niveles de colesterol total.

	Mean	**SD**	**N**
Group1	200	30	44
Group2	197	40	78

	F	**p-value**	**Var. Equal**	**Sp2**
Hom.Var	1.333	0.1354	TRUE	1349.167

	group1	**group2**	**diff.**	**SError**	**CI.lower**	**CI.upper**	**t**	**df**	**p-value**
t-Test	200	197	3	6.925308	-10.7116	16.7116	0.4332	120	**0.6656**

El p-valor es **0.6656** muy superior a alfa. No podemos arriesgarnos a rechazar H_0 con el riesgo alfa propuesto.

7. Osteoporosis y sedentarismo

En un estudio[2] llevado a cabo en el servicio de Reumatología del Hospital Ramón Barros Luco Trudeau de Santiago de Chile en el año 2005, se seleccionaron 860 mujeres en periodo posmenopáusico **con** Osteoporosis. Por cada una de ellas, se escogió una mujer en periodo posmenopáusico **sin** osteoporosis de la misma edad y que acudiese al mismo centro de salud.

A todas ellas se les realizó la misma encuesta en la que se recogieron las siguientes variables: edad, peso, altura, actividad física actual (sedentaria/activa), escolaridad y nivel socioeconómico.

A continuación se muestra la tabla de contingencia en la que se recoge la información sobre la actividad física realizada por las mujeres de cada grupo.

Tabla 1 Osteoporosis y actividad física

		SIN OSTEOPOROSIS		
		SEDENTARIA	**ACTIVA**	**TOTAL**
CON OSTEOPOROSIS	**SEDENTARIA**	420	130	550
	ACTIVA	80	230	310
	TOTAL	500	360	860

Cuestiones.

1. ¿A qué tipo de diseño corresponde este estudio? Especifique todo lo que pueda el tipo de diseño y razone su respuesta.

Teniendo en cuenta el diseño:

2. Plantee las hipótesis estadísticas de este contraste.
3. ¿Es significativamente diferente la proporción de mujeres físicamente activas en uno de los grupos? Si lo es, ¿en cuál es mayor? Razone su respuesta.
4. Construya e interprete el Intervalo de confianza del 95% para el estimador de la diferencia elegido en función del tipo de estudio.
5. Tras el análisis, responda razonadamente a las siguientes preguntas:
   a. Sólo con la información que se desprende de este estudio ¿podría afirmar que el sedentarismo se asocia de alguna manera con la osteoporosis?
   b. Sólo con la información que se desprende de este estudio ¿podría afirmar que ser sedentario aumenta el riesgo (está relacionado causalmente) de padecer osteoporosis? Razone su respuesta.

Solución.

Se trata de un estudio de casos y controles emparejados y por lo tanto:

$$EE_{dp} = \frac{1}{N}\sqrt{b + c)}$$

[2] Ejercicio adaptado a partir del artículo González AL, Espinosa VF, López FA, Fernández LM. Estilo de vida saludable en la prevención de la osteoporosis en la mujer climatérica. Rev Chil Obstet Ginecol 2007;72:383-9.

$$EE_{dp} = \frac{1}{860}\sqrt{80 + 130)} = 0.01685$$

$$\chi^2_{McNemar_{exp}} = \frac{(50)^2}{210} = 11.9049\ ,$$

$$\chi^2_{McNemar_{\exp_{cor}}} = \frac{(49)^2}{210} = 11.43$$

$$gl = r - 1$$

$$CI95\%: \left(\frac{130}{860} - \frac{80}{860}\right) \pm 1.957 \cdot 0.01685$$

$$CI95\%: 0.05814 \pm 0.033026$$

$$CICI95\%: (0.0251; 0.9117)$$

	osteopor. no / sedent sí	sedent no
osteopor. sí / sedent sí	420	130
osteopor. sí / sedent no	80	230

	prob b	prob c	dif.prop	EEdp	CIlower	CIupper	z/chi2	df	p.value
Z	0.1512	0.093	0.0581	0.0169	0.0251	0.0912	3.4503	NA	6e-04
chi2	NA	NA	NA	NA	NA	NA	11.9048	1	6e-04
chi2Corr	NA	NA	NA	NA	NA	NA	11.4333	1	7e-04

>

Como se observa en la tabla resumen anterior, el contraste lleva a rechazar la hipótesis nula de igualdad (p-valor<.001), por lo que podemos afirmar que la diferencia entre la proporción de pares discordantes es significativamente superior en el sentido mujer con osteoporosis sedentaria, mujer sin osteoporosis físicamente activa, es decir el sedentarismo fue superior en las primeras.

Esto también se observa en el IC95% de la diferencia de proporciones. La diferencia favorable a mayor proporción de actividad física en las que no tenían osteoporosis es del 5.8% CI95%(2.5%, 9.12%).

En respuesta a las preguntas 5 y 6, con este estudio solo hemos demostrado que la osteoporosis y actividad física están relacionadas estadísticamente (no son variables independientes), pero no hemos demostrado que esta relación sea causal. Por ejemplo, si la pregunta es sobre la actividad física actual, puede haber ocurrido que las mujeres dejasen de realizar actividad física tras el diagnóstico.

8. Estancia post-quirúrgica

En el servicio de cirugía torácica de un hospital español, los cirujanos, desean conocer si pueden disminuir el tiempo de estancia post-quirúrgica aplicando técnicas de fisioterapia respiratoria poscirugía. Para ello recogen el dato de duración de la estancia tras la cirugía, de las historias clínicas de los pacientes que durante el último año han pasado por quirófano. Por otro lado registran el mismo dato de los pacientes ingresados en el servicio para cirugía en el último mes. Los resultados fueron los siguientes:

Duración estancia post-cirugía	**N**	**Media (día)**	**Desviación típica (s)**
Pacientes con fisioterapia	50	10	2
Pacientes sin fisioterapia	900	13	3

Cuestiones:

a) ¿Son diferentes las duraciones de las estancias entre ambos grupos?
   1. Plantea la hipótesis nula y la hipótesis alternativa.
   2. Calcula el valor del estadístico de contraste.
   3. ¿Qué decisión tomarías para un nivel de significación α = 0,05?
   4. ¿Con qué nivel de seguridad podemos asegurarlo?

b) En el caso de que las diferencias sean estadísticamente significativas, ¿crees que se debe a que las muestras tienen tamaños muy descompensados?

Solución:

```
Comparación de medias grupos independientes.
      Mean SD   N
Group1   10  2  50
Group2   13  3 900
          F p-value Var. Equal Sp2
Hom.Var 1.5  0.0163      FALSE  NA
       group1 group2 diff. SError CI.lower CI.upper   t df p-value
t-Test     10     13    -3    0.3  -3.5997  -2.4003 -10 62 < 0.001
```

Tenemos una confianza del 95% de que la diferencia en los media de días de estancia entre las poblaciones a las que representan estas muestras estará entre 2.4 y 3.6 días, siendo menor en los pacientes con fisioterapia. Dado que la H_0 no está en dicho intervalo podemos rechazarla.
Podríamos llegar a rechazarlo aunque bajásemos alfa hasta valores inferiores al 0.001 o siendo más precisos hasta
2*pt(-10,62)=$1.49 \cdot 10^{-14}$
Por tanto estamos bastante seguros de no equivocarnos cuando afirmemos que la fisioterapia ha sido eficaz reduciendo significativamente el periodo posquirúrgico (reduciendo el tiempo de alta) en estos pacientes. Como siempre, esto no confirma la relación causa efecto entre ambos hechos pues puede haber otros factores no incluidos en este análisis que expliquen esta reducción de la estancia media. Un paso más en la evidencia sería realizar un ensayo clínico aleatorizando la intervención[3].

[3] Estos y otros diseños se estudiarán en las clases de epidemiología.

9. Fumadores

En una universidad se desea conocer si la proporción (prevalencia) de fumadores entre los alumnos es diferente de la de la población general (45%). Para ello se prepara una encuesta y el resultado es que en la muestra entrevistada (n=150) (con representación proporcional de cada curso y carrera) el 40 % fueron fumadores.

Cuestiones:

a) ¿Qué hipótesis plantearías para resolver la cuestión?

b) Con un nivel de confianza del 5% ¿podemos afirmar que en nuestra muestra la proporción de fumadores es diferente a la que existe en la población general?

Solución.

Prueba de conformidad contra valor teórico para proporción.

a) $H_0: \pi = \pi_0$
$H_1: \pi \neq \pi_0$

$$z_{exp} = \frac{\bar{p} - \pi_0}{\sqrt{\frac{\pi_0 \cdot (1-\pi_0)}{n}}}$$ [4]

	prop	propTeor	EEp	Zexp	pval
Z-test	0.4	0.45	0.0406	-1.2315	0.2181
Exact	0.4	0.45	NA	NA	0.2505

	Point.Est	EE	CI.Lower	CI.Upper
Asymp. CI with Pi	0.4	0.0406	0.3216	0.4784
Asymp. CI with p	0.4	0.04	0.3204	0.4796
Exact CI	0.4		0.321	0.4831

Como se observa el p-valor del contraste es >0.05 por lo que no podemos rechazar la H_0. Desde el punto de vista el IC tenemos una confianza de que la proporción en la población de la que ha salido la muestra esté entre 0.32 y 0.47, siendo lo que defiende la H_0 uno de los valores del IC. No es suficientemente improbable que esta muestra haya salido de una población en la que la proporción es del 45% y por tanto no podemos rechazar la H_0.
Como se ve el intervalo obtenido por métodos exactos no es muy diferente del obtenido por métodos asintóticos. Esto es debido a que n es grande y la proporción no es extrema.
NOTA: Recuerde que aunque cuando construimos los IC, utilizamos p y q en el EEp, si contamos con π_0 hemos de utilizarlo para construir el estadístico de contraste y el IC 'comparable' con el resultado del contraste. Salvo que p y π sean muy diferentes, los IC serán similares.

Además se intentó comparar la variable nº de cigarrillos/día entre dos de las licenciaturas:

Nº de cigarrillos/día	N	Media (cig/día)	Desviación típica (s)
Estudiantes de Derecho	29	13	4

[4] Recuerde que para poder aplicarlos es recomendable que n·p sea grande (se suele recomendar np>5)..

Estudiantes de políticas	31	17	7

Cuestiones:

a) ¿Existen dichas diferencias? Para contestar considere que las varianzas son homogéneas y que la distribución de la variable es normal en ambos grupos.

1. Plantea la hipótesis nula y la hipótesis alternativa.
2. Calcula el valor del estadístico de contraste.
3. ¿Qué decisión tomarías para un nivel de significación $\alpha = 0{,}05$?
4. ¿Con qué nivel de significación?

Solución.

Se trata de una comparación de medias 2 grupos independientes.

Uno de los grupos es menor de 30, pero el enunciado nos informa de la normalidad de la variable en ambos grupos, por lo que podemos utilizar los test paramétricos.

```
       Mean SD  N
Group1   13  4 29
Group2   17  7 31
           F p-value Var. Equal    Sp2
Hom.Var 1.75  0.0678       TRUE 33.069
       group1 group2 diff.   SError CI.lower CI.upper       t df p-value
t-Test     13     17    -4 1.485615  -6.9738  -1.0262 -2.6925 58  0.0092
```

En una distribución t con 58 gl (en la tabla el más cercano sería del 40 gl) el punto que deja un área de 0.025 (0.05 bilateral) es -2. Nuestro punto queda a su izquierda y por tanto el p-valor es menor que 0.05.

El resultado nos permite rechazar la hipótesis nula (p-valor=0.0092) y afirmar con una confianza del 95% que los estudiantes de políticas fuman de promedio entre 1 y 6 cigarrillos más que los estudiantes de derecho.

Aunque se trata de una variable realmente discreta, el hecho de que sus valores promedios estén lejos de 0 (si lo hicieran veríamos valores truncados por la izquierda) y que las desviaciones típicas no son demasiado grandes (CV: 30-40%), sumado a que nos indican que la variable se distribuye normalmente, nos permite trabajar con ella asumiendo que las medias muestrales se distribuyen normalmente alrededor de la verdadera media poblacional.

10. Hiperémesis gravídica (vómitos del embarazo)

Se desea saber si un nuevo fármaco es eficaz para prevenir los vómitos del embarazo. Para ello se trataron 100 mujeres que manifestaron padecer dicho síntoma en el primer trimestre y se compararon con otras 140 mujeres que no siguieron dicho tratamiento pero que también padecieron el síntoma. Los resultados se recogen en la siguiente tabla:

	Vómitos	No vómitos	TOTAL
Tratadas	25	75	100
No tratadas	60	80	140
	85	155	240

Cuestiones:

a. ¿Es eficaz el fármaco?
   1. Plantea la H_0 y la H_1 para este estudio.
   2. Calcula el valor del estadístico de contraste.
   3. ¿Qué decisión tomarías? Justifica la respuesta.

Solución

Vamos a resolverlo considerando **'éxito'** el hecho de **no tener vómitos** y por tanto, si el fármaco es eficaz, deberíamos observar una proporción significativamente menor de vómitos en el grupo de intervención.

Ofrecemos la solución utilizando la z (contraste e IC) y la ji-cuadrado (solo contraste con y sin corrección de Yates) y también el p-valor resultado de realizar el test de Fisher en la misma tabla y un par de IC corregidos por diferentes métodos (mencionados en clase).
Aunque se trata de 2 grupos, vamos a resolverlo utilizando los diferentes métodos vistos en clase.

a) Wald

Utilizando la "p ponderada" podemos calcular Z_{exp} en la que p_1 es la proporción de mujeres que no padecieron vómitos en el grupo de tratadas (75/100=0,75), y p2 la proporción en el grupo de no tratadas (80/140=0,57).

La proporción ponderada:

$$p^* = \frac{75 + 80}{240} = 0.6458$$

$$q^* = (1 - 0.6458)$$

$$EE_{dp} = \sqrt{0.6458 \cdot 0.3542 \cdot \left(\frac{1}{100} + \frac{1}{140}\right)}$$

$$EE_{dp} = \sqrt{0.003921} = 0.0626$$

$$z_{exp} = \frac{0.75 - 0.5714}{0.0626} = 2.853$$

Z_{exp} es el estadístico de contraste en este experimento, que seguirá una distribución normal tipificada.

Esto siempre que se den las condiciones (np≫5 en cada grupo[6] o esperados mayores que 5 en más del 80% de casillas de la tabla de contingencia para obtener la ji-cuadrado)

Dado que los cuantiles que dejan (bilateralmente) las áreas de 0.05, 0.01 y 0.001 en una distribución normal tipificada son -1.959964 -2.575829 -3.290527, respectivamente, rechazamos la hipótesis nula (p-valor<0.01), en concreto p-valor=0.0043.

La proporción de **'no vómitos'** es significativamente inferior. Si además construimos el IC95% para la diferencia de proporciones.

$$CI(1-\alpha)\%: (p_1 - p_2) \pm z_{\alpha/2} \cdot EE_{dp} \quad \text{(Wald)}$$

$$CI95\%: 17.86 \pm 1.95 \cdot 0.00626$$

$$CI95\%: (0.059; 0.3013)$$

Aunque las proporciones cambiarían, el resultado de la diferencia y el contraste sería el mismo si lo enfocásemos considerando éxito haber padecido vómitos.

$$p^* = \frac{25+60}{240} = 0.3542$$

$$q^* = (1 - 0.6458)$$

$$EE_{dp} = \sqrt{0.3542 \cdot 0.6458 \cdot \left(\frac{1}{100} + \frac{1}{140}\right)}$$

$$EE_{dp} = \sqrt{0.003921} = 0.0626$$

$$CI95\%: -17.86 \pm 1.95 \cdot 0.00626$$

$$CI95\%: (-0.3013; -0.059;)$$

b) Ji-cuadrado (chi-square)

Otra forma de comparar proporciones consiste en la elaboración de una tabla de contingencia y la evaluación de las diferencias entre los "valores observados" y "los esperados mediante el estadístico" $\chi^2_{Pearson}$ (ji cuadrado de Pearson).
Los esperados de cada celda si suponemos que las proporciones son homogéneas (H_0), se calculan de la siguiente manera:

$$e_{ij} = \frac{\sum_{j=1}^{k} o_{ij} \cdot \sum_{i=1}^{r} o_{ij}}{N}$$

$$gl = (r-1) \cdot (k-1)$$

Posteriormente hay que calcular el valor de χ^2_{exp} de nuestro experimento mediante la fórmula:

$$\chi^2_{exp} = \sum_{i=1}^{r} \sum_{j=1}^{k} \left[\frac{\left(o_{ij} - e_{ij}\right)^2}{e_{ij}} \right]$$

$$gl = (r-1) \cdot (k-1)$$

En nuestro experimento:

```
$Observed.oij
             Vomiting Not vomiting
Treated            25           75
Non treated        60           80
```

[6] Recordad que estas "reglas" arbitrarias no son todo o nada. Solo nos indican que cuanto más lejos estemos de la condición, menos cierta será la afirmación que pretendamos soporten nuestros datos. Como vimos en clase existen mejores opciones que Wald (Agresti-Coull, Wilson…) pero su cálculo manual es más tedioso y en muchas circunstancias los IC obtenidos son muy semejantes.

```
$Expected.eij
            Vomiting Not vomiting
Treated     35.41667     64.58333
Non treated 49.58333     90.41667

$oij_eij
             Vomiting Not vomiting
Treated     -10.41667     10.41667
Non treated  10.41667    -10.41667

$oij_eij_2
            Vomiting Not vomiting
Treated     108.5069     108.5069
Non treated 108.5069     108.5069

$Sumandosji2
            Vomiting Not vomiting
Treated     3.063725     1.680108
Non treated 2.188375     1.200077
```

	prop1	prop2	p1-p2	proppond	EEp	zexp/chi2exp	df	p-value
Ztest	0.25	0.4286	-0.1786	0.3542	0.0626	-2.8526	NA	0.0043
chi2	NA	NA	NA	NA	NA	8.1323	1	0.0043
chi2_corr	NA	NA	NA	NA	NA	7.3703	1	0.0066
Fisher	NA	NA	NA	NA	NA	NA	NA	0.0060

	Point Est. Diff	CIlower 95 %	CIupper 95 %
WaldCI	-0.1786	-0.3013	-0.0559
Wald_AC	-0.1786	-0.2921	-0.0573
ScoreCI	-0.1786	-0.2925	-0.0568

$$\chi^2_{exp} = \frac{(25-35,42)^2}{35,42} + \frac{(60-49,58)^2}{49,58} + \frac{(75-64,58)^2}{64,58} + \frac{(80-90,42)^2}{90,42}$$

$$\chi^2_{exp} = \frac{108,6}{35,42} + \frac{108,6}{49,58} + \frac{108,6}{64,58} + \frac{108,6}{90,42}$$

$$\chi^2_{exp} = 3,06 + 2,19 + 1,68 + 1,2 = 8,13$$

Los grados de libertad se calculan como (r-1)·(k-1), siendo r y k las filas y columnas de la tabla de contingencia. En nuestro caso 1 grado de libertad.

Si buscamos los valores teóricos de la distribución χ^2 de Pearson para 1 grado de libertad que deja a su derecha el 5% y el 1% del área (área=0,05 y área = 0,01) de la curva obtenemos:

$$\chi^2_{0.05,1} = 3,84 \; \chi^2_{0.01,1} = 6,63 \Rightarrow \chi^2_{exp} > \chi^2_{0.01,1}$$

Por ello podemos concluir que existen diferencias significativas con una p<0.01, entre las proporciones de vómitos de los grupos, siendo mayor la proporción en las mujeres **sin** tratamiento (como se observa en la tabla resumen). O lo que es lo mismo, la probabilidad de haber encontrado unas diferencias como las observadas o mayores si la hipótesis nula fuese cierta, es menor del 1%.

En concreto pchisq(8.13,1,lower.tail=F)=**0.0043**53875

En las tablas anteriores, además de las soluciones con la z y la chi2, se recoge el valor de ji-cuadrado con la corrección de Yates, y del test de Fisher. Como se observa el p-valor es algo mayor, pero nuestra decisión no cambia.

También se incluyen los intervalos de confianza para la diferencia e proporciones utilizando la aproximación normal (Wald), una versión

corregida (Agresti-Coull) y el calculado por otro método corregido (Score). Como se puede observar los IC95% generados son bastante similares.

Con R se podría resolver de la siguiente manera.

Metemos en una matriz la tabla de contingencia.

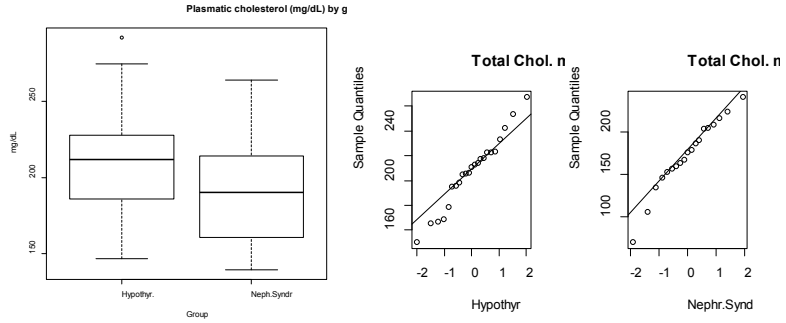
```
tabla<-matrix(c(25,60,75,80),2,2)
prop.test(tabla)
```

o también

```
chisq.test(table) #pero esto no devuelve el intervalo de confianza.
```

11. Colesterol adultos

Se quiere probar si existe diferencia en el colesterol plasmático entre sujetos con hipotiroidismo y sujetos con síndrome nefrótico. Para ello se obtuvieron dichos niveles en 23 del primer grupo y 19 del segundo. (Nota: En el test de Levene se obtiene una p=0.27). Antes de abordar el análisis, se obtuvieron los siguientes gráficos y análisis.



$Hypothyr.	$Neph.Syndr
Shapiro-Wilk normality test	Shapiro-Wilk normality test
data: X[[i]]	data: X[[i]]
W = 0.97504, p-value = 0.8071	W = 0.92232, p-value = 0.1249

Niveles de colesterol	N	Media (mg/dl)	Desviación típica (σ)
Hipotiroideos	23	210	30
Síndrome nefrótico	19	197	40

Cuestiones:

a. ¿Son significativamente diferentes los niveles en ambos grupos? Para contestar:
   1. Razona el cumplimiento de los supuestos que sean necesarios para decidir la técnica adecuada.
   2. Plantea la hipótesis nula y la hipótesis alternativa.
   3. Calcula el valor del estadístico de contraste.
   4. ¿Qué decisión tomarías para un nivel de significación α = 0,01?

Solución

De nuevo estamos ante una comparación de las medias de dos muestras. Se trata de comprobar la homogeneidad de dos medias.

H_0: $\mu_1-\mu_2=0$

H_1: $\mu_1-\mu_2\neq0$

Para ello necesitamos partir de varios supuestos:
Aunque los tamaños de ambos grupos son inferiores a 30, a la vista de los test de Shapiro-Wilk (p>0.05 en ambos grupos) y los gráficos (boxplot y Q-Q frente a normal) en cada grupo, parecer razonable aceptar que los datos no se apartan significativamente de una distribución normal, por lo que podemos utilizar los métodos que utilizamos con muestras de mayor tamaño muestral.

Las varianzas poblacionales $\sigma_1 \, y \, \sigma_2$ son desconocidas.

Debemos comparar las varianzas muestrales mediante algún contraste que nos permita observar si son suficientemente homogéneas como para asumir que proceden de la misma varianza poblacional y solo se han separado de ella por el error de muestreo (contraste de hipótesis de homogeneidad de varianzas). El enunciado del problema nos dice que en el test de Levene no se han encontrado diferencias significativas, así pues no podremos hacer una cuasivarianza ponderada, y construir el estadístico t para el contraste utilizando el siguiente error estándar.

$$EE_{dm} = \sqrt{S_p^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \text{ o}$$

$$EE_{dm} = S_p \cdot \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

y por tanto…

	Mean	**SD**	**N**
Group1	210	30	23
Group2	197	40	19

	F	p-value	Var. Equal	Sp2
Hom.Var	1.333	0.2701	TRUE	1215

	group1	**group2**	**diff.**	**SError**	**CI.lower**	**CI.upper**	**t**	**df**	**p-value**
t-Test	210	197	13	10.80618	-8.8401	34.8401	1.203	40	0.236

$$t_{exp} = \frac{210 - 197}{\sqrt{1215 \cdot \left(\frac{1}{23} + \frac{1}{19}\right)}}$$

$$t_{exp} = 1.203$$

```
2*pt(-1.203,40)
[1] 0.2360486
```

p-valor=0.2236

Buscamos en la tabla de la t de Student con 40 gl. el valor de t para un alfa del 5% bilateral (hay que buscar el que deja alfa/2), que es -2.021 (deja 0.025 a su izquierda), +2.021 (lo deja a su derecha. Nuestro estadístico de contraste deja un área mayor (p-valor>alfa o p-valor/2>alfa/2) y por tanto no podemos rechazar la hipótesis nula.

Concluimos que no hemos podido demostrar diferencias significativas entre ambos grupos y por tanto que no parece que las diferencias en el colesterol entre estos grupos sean suficientemente importantes como para afirmar que existen en las poblaciones a las que representan.

12. Unidad de diálisis

En la Unidad de diálisis de cierto hospital, los DUE al cargo de la misma, creen que en una de las dos salas que utilizan, algo va mal. Los pacientes dializados en la sala A, salen (o al menos eso les parece a ellos) con mayores concentraciones de urea en sangre que los de la sala B. Aceptando que la asignación de los pacientes es aleatoria, y por lo tanto que salen de una población homogénea, deciden llevar a cabo un estudio para comprobar esta hipótesis. Los resultados son los siguientes:

	n_j	Concentración media de urea (mg/dl)	Desviación típica
SALA A	23	32	4
SALA B	26	40	3
Test de Levene p>0.05			

Cuestiones:

a. ¿Son suficientemente diferentes las concentraciones medias de urea en ambos grupos de pacientes para afirmar que los son en la población a la que representan? Para responder a esta pregunta asuma que la distribución de la variable en cada grupo es normal.
   1. Plantea la hipótesis nula y la hipótesis alternativa que barajan los DUE.
   2. Calcula el valor del estadístico de contraste.
   3. Si son diferentes, ¿con qué seguridad se puede afirmar?

Solución:

Se trata de una comparación de grupos **no** emparejados (independientes) con tamaños muestrales menores de 30, pero que se distribuyen normalmente. El test de Levene nos indica que las varianzas son razonablemente homogéneas. Calculamos la cuasivarianza ponderada y el estadístico de contraste correspondiente.

```
       Mean SD  N
Group1   32  4 23
Group2   40  3 26
            F p-value Var. Equal     Sp2
Hom.Var 1.333  0.2427       TRUE 12.2766
       group1 group2 diff.   SError CI.lower CI.upper       t df p-value
t-Test     32     40    -8 1.002967 -10.0177  -5.9823 -7.9763 47   0.001
```

Interpretación.
Las diferencias observadas son demasiado grandes (p-valor<0.001) y nos obligan a rechazar la hipótesis nula.

Tenemos una confianza del 95% de que la media de concentración de urea sea entre 5.98 y 10 mg/dl superiores en las poblaciones representada en la sala B.

13. Antidepresivos

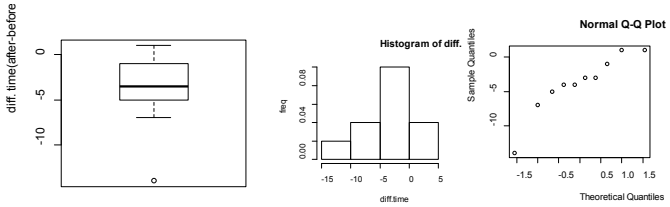
Cuestiones.

Basándonos en los datos empíricos que aparecen a continuación ¿podemos afirmar que un determinado antidepresivo disminuye el tiempo (minutos) que los pacientes tardan en realizar una tarea concreta?

Tras construir la variable diferencia, el test de Shapiro-Wilk sobre esta variable, ofrece el siguiente resultado:

Shapiro-Wilk normality test (data: time after-time before)

W = 0.87918, **p-value = 0.1277**



T. antes	34	45	31	43	40	41	33	29	41	37
T. después	29	42	32	29	36	42	26	28	38	33

Solución.

Se trata de una comparación de medias, grupos emparejados (autoemparejamiento) con tamaños muestrales pequeños.
El resultado del test de Shapiro-Wilk, no nos llevaría a rechazar la normalidad de la variable diferencia, si bien es cierto que la potencia con este tamaño muestral puede ser baja (podríamos estar no rechazando la nula, siendo falsa). Valorados en conjunto, los gráficos y el resultado del test de hipótesis, nos llevan a pensar que la variable diferencia no sigue una distribución muy anormal, por lo que lo vamos a resolver utilizando un test paramétrico.

Obtenemos la variable diferencia, su media y su desviación típica.

Antes	34	45	31	43	40	41	33	29	41	37
Después	29	42	32	29	36	42	26	28	38	33
antes-después	5	3	-1	14	4	-1	7	1	3	4

Descriptivos:

```
     vars  n mean   sd median trimmed  mad min max range  skew kurtosis   se
Ant     1 10 37.4 5.42   38.5   37.50 6.67  29  45    16 -0.16    -1.63 1.71
Desp    2 10 33.5 5.78   32.5   33.38 5.93  26  42    16  0.29    -1.58 1.83
dif     3 10 -3.9 4.36   -3.5   -3.25 2.97 -14   1    15 -0.97     0.25 1.38
```

Comparación de medias emparejadas.

Lo convertimos en un contraste sobre valor teórico de la media de las diferencias en la población.

$$H_0: \mu_{dif} = 0$$

$$H_1: \mu_{dif} \neq 0$$

	mean	t/z	0.025	se	lower	upper	texp	pvalue
ci usando t	-3.9		2.2622	1.378	-7.0173	-0.7827	-2.8302	**0.0197**

Por lo tanto podemos rechazar la hipótesis nula de igualdad y afirmar que el tiempo de respuesta es significativamente inferior en el grupo de tratamiento.
Otra forma de expresar lo mismo consistiría en afirmar que hemos observado una reducción promedio en el tiempo de realización de la tarea de 3.9 minutos y tenemos una confianza del 95% de que el promedio de dicha reducción en la población esté entre 0.783 y 7.02 minutos.

Otra cosa es la relevancia clínica de dicha reducción.

Si nuestras dudas sobre la normalidad de la distribución de la variable fuesen importantes, podríamos recurrir a la alternativa no paramétrica. En este caso el test de Wilcoxon[7].
Este es el resultado.

Wilcoxon signed rank test with continuity correction[8]

data: prob11$bfr and prob11$baft
V = 51, **p-value = 0.01859**
alternative hypothesis: true location shift is not equal to 0

Lo que nos llevaría a la misma conclusión de rechazo de la hipótesis nula.

[7] De cara al examen, recuerde que no es necesario saber resolver manualmente los test no paramétricos, pero sí debe reconocer las condiciones en las que debemos aplicarlos y debe saber interpretar su resultado.
[8] Para ejecutar esta función, hay que al menos generar los vectores. Dejamos código aquí:

```
bfr<-c(34,45,31,43,40,41,33,29,41,37)
baft<-c(29,42,32,29,36,42,26,28,38,33)
wilcox.test(bfr,baft)
```

14. Obesidad

En una consulta de endocrinología se decide llevar a cabo un estudio entre sus pacientes diabéticos tipo II para verificar la eficacia de una nueva dieta. Se desea contemplar en un intervalo de 6 meses dos variables: Índice de masa corporal y Control de glucemia. Como valoración del control de glucemia se decidió utilizar la HbA_{1c}. Ésta es un buen descriptor del perfil glucémico en los últimos 3 meses. Se obtienen los siguientes resultados:

PAC: Paciente
Vis 0: Visita inicial
Vis 6m: Visita a los 6 meses

	HbA_{1c}		IMC	
PAC	Vis 0	Vis 6m	Vis 0	Vis 6m
1	8	6	30	26
2	7	7	26	21
3	6	6	24	20
4	6	6	28	23
6	8	7	21	21
7	4	5	26	23
8	7	6	31	25
9	10	7	29	26
10	12	12	30	25
11	11	7	28	23
12	12	10	27	25
13	13	7	24	26
14	12	6	23	24
15	10	9	25	26
16	16	15	29	25
17	15	10	30	28
18	13	10	28	27
19	15	13	27	26
20	7	6	31	31
21	8	7	32	30
22	9	5	25	26
23	14	7	26	27
24	9	7	29	25
25	8	3	30	29
26	12	6	28	30
27	13	5	25	26
28	16	10	20	21
29	8	6	29	27
30	10	3	28	26
32	12	5	27	24
33	15	3	22	21
34	11	6	21	22
35	10	2	30	28

Cuestiones:

a) Obtén las medidas de tendencia central y de dispersión de ambas valores en la visita inicial (Vis 0) y en la Visita a los 6 meses (Vis 6m).

b) Estudie si hay una reducción estadísticamente significativa en los niveles de HbA_{1c} y en el IMC entre antes y después de la dieta. Para ello:

   1. Plantea las hipótesis nula y alternativa.
   2. Calcula el valor del estadístico de contraste.

3. Toma una decisión justificando tu respuesta.

Solución:

Construimos la variable diferencia (Vis.6m-Vis.0) para ambas variables. La variable HbA1c es en realidad el % de Hb que está glicosilada, pero en este problema para nosotros es variable continua cuya diferencia queremos estimar.

Hemos construido las variables diferencia tanto para HbA1c como para IMCdif[9].

PAC	HbA1c.0m	HbA1c.6m	IMC.0m	IMC.6m	HbA1cdif	IMCdif
1	8	6	30	26	-2	-4
2	7	7	26	21	0	-5
3	6	6	24	20	0	-4
4	6	6	28	23	0	-5
6	8	7	21	21	-1	0
7	4	5	26	23	1	-3
8	7	6	31	25	-1	-6
9	10	7	29	26	-3	-3
10	12	12	30	25	0	-5
11	11	7	28	23	-4	-5
12	12	10	27	25	-2	-2
13	13	7	24	26	-6	2
14	12	6	23	24	-6	1
15	10	9	25	26	-1	1
16	16	15	29	25	-1	-4
17	15	10	30	28	-5	-2
18	13	10	28	27	-3	-1
19	15	13	27	26	-2	-1
20	7	6	31	31	-1	0
21	8	7	32	30	-1	-2
22	9	5	25	26	-4	1
23	14	7	26	27	-7	1
24	9	7	29	25	-2	-4
25	8	3	30	29	-5	-1
26	12	6	28	30	-6	2
27	13	5	25	26	-8	1
28	16	10	20	21	-6	1
29	8	6	29	27	-2	-2
30	10	3	28	26	-7	-2
32	12	5	27	24	-7	-3
33	15	3	22	21	-12	-1
34	11	6	21	22	-5	1
35	10	2	30	28	-8	-2

Los descriptivos de todas ellas.

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
HbA1c.0m	1	33	10.52	3.16	10	10.52	2.97	4	16	12	0.03	-1.00	0.55
HbA1c.6m	2	33	6.97	2.87	6	6.74	1.48	2	15	13	0.82	0.54	0.50
IMC.0m	3	33	26.94	3.15	28	27.15	2.97	20	32	12	-0.54	-0.68	0.55

[9] Script para construir el dataframe en R.

```
PAC<-c(1,2,3,4,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,32,33,34,35)
HbA1c.0m<-c(8,7,6,6,8,4,7,10,12,11,12,13,12,10,16,15,13,15,7,8,9,14,9,8,12,13,16,8,10,12,15,11,10)
HbA1c.6m<-c(6,7,6,6,7,5,6,7,12,7,10,7,6,9,15,10,10,13,6,7,5,7,7,3,6,5,10,6,3,5,3,6,2)
IMC.0m<-c(30,26,24,28,21,26,31,29,30,28,27,24,23,25,29,30,28,27,31,32,25,26,29,30,28,25,20,29,28,27,22,21,30)
IMC.6m<-c(26,21,20,23,21,23,25,26,25,23,25,26,24,26,25,28,27,26,31,30,26,27,25,29,30,26,21,27,26,24,21,22,28)
probdmob<-data.frame(PAC,HbA1c.0m,HbA1c.6m,IMC.0m,IMC.6m)
probdmob$HbA1cdif<-(probdmob$HbA1c.6m-probdmob$HbA1c.0m)
probdmob$IMCdif<-(probdmob$IMC.6m-probdmob$IMC.0m)
```

IMC.6m	4	33	25.24	2.80	26	25.19	2.97	20	31	11	0.00	-0.66 0.49
HbA1cdif	**5**	**33**	**-3.55**	**3.06**	-3	-3.33	2.97	-12	1	13	-0.62	-0.31 0.53
IMCdif	**6**	**33**	**-1.70**	**2.34**	-2	-1.67	2.97	-6	2	8	-0.07	-1.28 0.41

Aunque para resolver el problema solo nos hacen falta los descriptivos de la variable diferencia.
La solución de nuevo pasa por convertirlo en un contraste de conformidad sobre valor de la media poblacional de la diferencia $\mu_{dif} = 0$.
Como el procedimiento es el mismo, resolvemos el contraste para ambas variables simultaneamente.

	mean	t/z	0.025	se	lower	upper	texp	df	pvalue
ci Dif.HbA1c	-1.697	2.0369	0.4071	**-2.5261**	**-0.8678**	**-4.1689**	32	**0.001**	
	mean	**t/z**	**0.025**	**se**	**lower**	**upper**	**texp**	**df**	**pvalue**
ci Dif.IMC	-3.5455	2.0369	0.5332	**-4.6315**	**-2.4594**	**-6.6498**	32	**0.001**	

Efectivamente el cambio en ambas variables (reducción) ha sido suficientemente importante como para rechazar la hipótesis nula.
La reducción promedio del HbA1c (es la diferencia media de un %) es **del 1.7% con un IC95% (0.87%;2.52%)**. Tenemos una confianza del 95% la reducción promedio en el %HbA1c en la población a la que representan estos pacientes, esté entre (0.87% y 2.52%). Como se observa, no incluye la hipótesis nula $H_0: \mu_{\text{dif}_{\text{HbA1c}}} = 0$.

En el caso del **IMC** este intervalo (también reducción) es -1.7(-4.63;-2.46). Tampoco incluye la hipótesis nula del contraste $H_0: \mu_{\text{dif}_{\text{IMC}}} = 0$.

En conclusión parece que la intervención ha sido eficaz en la reducción de la HbA1c y del IMC.

Usando el dataframe creado en anteriormente ('probdmob') podríamos utilizar la función t.test para obtener la misma solución tanto si se utiliza la variable diferencia ya construida (ejecutando un contraste contra valor teórico mu=0), como si se le pide el test emparejado utilizando las variable al inicio y a los 6m.

HbA1c	**IMC**
t.test(probdmob$HbA1cdif)	***t.test(probdmob$IMCdif)***
One Sample t-test	One Sample t-test
data: probdmob$HbA1cdif	data: probdmob$IMCdif
t = -6.6498, df = 32, p-value = 1.677e-07	**t = -4.1689, df = 32, p-value = 0.0002174**
alternative hypothesis: true mean is not equal to 0	alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:	95 percent confidence interval:
-4.631473 -2.459436	**-2.5261067 -0.8678327**
sample estimates:	sample estimates:
mean of x	mean of x
-3.545455	-1.69697
t.test(HbA1c.6m,HbA1c.0m,data=probdmob,paired=T)	***t.test(IMC.6m,IMC.0m,data=probdmob,paired=T)***
Paired t-test	Paired t-test
data: HbA1c.0m and HbA1c.6m	data: IMC.0m and IMC.6m
t = -6.6498, df = 32, p-value = 1.677e-07	**t = -4.1689, df = 32, p-value = 0.0002174**
alternative hypothesis: true difference in means is not equal to 0	alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:	95 percent confidence interval:
-2.459436 -4.631473	**-0.8678327 -2.5261067**
sample estimates:	sample estimates:
mean of the differences	mean of the differences
-3.545455	**-1.69697**

15. Contaminación en el agua de un río.

En un bonito pueblo de Soria dedicado a la industria papelera, la médica titular del centro de salud comarcal que cubre la zona, tiene la sensación de que la proporción de pacientes con urticaria en el último mes es especialmente elevada. Decide revisar las Historias clínicas y hacer una primera comparación de proporciones. Para ello recoge una serie de variables en su base de datos. Decide comparar la proporción de enfermos en función de diferentes factores. No sabe mucha epidemiología, pero sí algo de estadística. Cuando hace una comparación dividiendo su población en dos grupos, los que viven a menos de 100 metros del río y los que viven a más distancia, obtiene la siguiente tabla:

		Uriticaria		
		Sí	No	Total
Distancia al río	<100m	25	40	65
	>100m	10	190	200
	Total	35	230	265

Cuestiones:

a) ¿Qué conclusión se puede obtener de esta tabla respecto a la relación entre el río y la aparición de urticaria?

Solución.

```
$Observed.oij
             <100m >100m
Urticaria sí     25    40
Uriticaria no    10   190

$Expected.eij
                  <100m     >100m
Urticaria sí   8.584906  56.41509
Uriticaria no 26.415094 173.58491

$oij_eij
                  <100m     >100m
Urticaria sí   16.41509 -16.41509
Uriticaria no -16.41509  16.41509

$oij_eij_2
                 <100m    >100m
Urticaria sí  269.4553 269.4553
Uriticaria no 269.4553 269.4553

$Sumandosji2
                 <100m    >100m
Urticaria sí  31.38710 4.776298
Uriticaria no 10.20081 1.552297

           prop1 prop2  p1-p2 proppond    EEp zexp/chi2exp df p-value
Ztest     0.3846  0.05 0.3346   0.1321 0.0483       6.9279 NA   <0.001
```

chi2	NA	NA	NA	NA	NA	47.9165	1	**<0.001**
chi2_corr	NA	NA	NA	NA	NA	45.0419	1	**<0.001**
Fisher	NA	NA	NA	NA	NA	NA	NA	**<0.001**

	Point Est. Diff	CIlower 95 %	CIupper 95 %
WaldCI	0.3346	0.2399	0.4293
Wald_AC	0.3346	0.2128	0.4544
ScoreCI	0.3346	0.2204	0.4597

$$\chi^2_{exp} = \sum_{i=1}^{r}\sum_{j=1}^{k}\left[\frac{(o_{ij} - e_{ij})^2}{e_{ij}}\right]$$

$$gl = (r-1)\cdot(k-1)$$

$$\chi^2_{\exp} = \frac{(25-8{,}585)^2}{8{,}585} + \frac{(10-26{,}415)^2}{26{,}415} + \frac{(40-56{,}415)^2}{56{,}415} + \frac{(190-173{,}585)^2}{173{,}585}$$

$$\chi^2_{\exp} = 31{,}39 + 10.20 + 4.78 + 1.55 = 47.92$$

Puesto que el valor del estadístico de contraste es muy superior al del $\chi^2_{0.05,1} = 3.84$, rechazamos la hipótesis nula y podemos decir que la frecuencia de urticaria es diferente en función de vivir cerca o lejos del río. Si nos fijamos en la tabla de esperados, se observa que si la proporción en la población fuese igual independientemente de la distancia al río, esperaríamos menor número de personas con urticaria en los que viven a <100m y más en los que viven a >100 m.

La diferencia de % es del **33.5% IC95%(23.9%-42.9%**) superior entre aquellos que viven <100m. Por lo tanto podemos afirmar que la proporción es superior en los que viven cerca.

IMPORTANTE. Esto solo implica relación estadística (las dos variables no son independientes) pero no causalidad, es decir, no podemos afirmar que sea la utilización del agua del río (por ejemplo para beber o lavarse) lo que produce la urticaria, solo podemos afirmar que vivir cerca del río aumenta la probabilidad de padecer urticaria, sea por el consumo de agua o por cualquier otro factor asociado a dicha cercanía (por ejemplo el crecimiento de ciertas plantas urticantes en la zona en la que los vecinos que viven cerca del río suelen ir a pasear**)**.

16. Separación de padres y obesidad

Un estudio danés realizado en 2006 y recientemente publicado10, tenía como objetivo averiguar si la separación de los padres (definida como separación de uno de los padres biológicos, independientemente de la razón durante al menos un año antes de la edad de 17) se relacionaba con el desarrollo de obesidad en la edad adulta y si la duración de la separación se correlacionaba con el aumento en el índice de masa corporal en el periodo 2002-2006.

Para ello se escogieron gemelos monozigóticos[11] de entre 20 y 71 años de edad a partir de un registro nacional. En la encuesta realizada en 2002, se recogía información sobre altura y peso en ese año (a partir de la cual se estimó el BMI[12]).

Para el presente estudio, se consideró sobrepeso un BMI ≥ 30 kg/m^2 y normopeso un BMI entre 20 y 25 kg/m^2 en el año de la encuesta (2002). Se analizaron todas las parejas de gemelos registradas y se escogieron aquellas en las que un gemelo tenía BMI ≥ 30 kg/m^2 y el otro gemelo un BMI entre 20 y 25 kg/m^2 en el año en que se realizó la encuesta.

También se realizó una entrevista en la que se recogió información relativa a la separación de los padres biológicos junto con otras variables relacionadas con el cuidado parental. **Catorce parejas lo eran de gemelos cuyos padres NO se habían separado** y **20 parejas lo eran de gemelos cuyos padres SÍ se habían separado**. Rellene la siguiente tabla (¡Cuidado! Aunque se analiza igual, es algo diferente a la vista en teoría) y responda a las preguntas planteadas.

		gemelo2 (BMI 20-25 kg/m^2)	
		Separación no	**Separación sí**
gemelo1 BMI >30 kg/m^2)	**Separación sí**		
	Separación no		

Cuestiones.

1. ¿Qué tipo de estudio han realizado los investigadores? Enumere y explique razonadamente sus características más importantes.
2. Teniendo en cuenta el diseño escogido, ¿Se asoció a un mayor riesgo de obesidad la separación de los padres biológicos? Para responder a esta pregunta:
    a. Formalice las hipótesis estadísticas del contraste.
    b. Elija y calcule el estadístico de contraste adecuado y responda la hipótesis planteada.
3. Obtenga el estimador de la diferencia que considere oportuno y su correspondiente intervalo de confianza del 95%.**Solución.**

[10] Adaptado a partir de Petersen, J. D., K. O. Kyvik, B. L. Heitmann, and M. E. Vámosi. 2016. 'The Association between Parental Separation during Childhood and Obesity in Adulthood: A Danish Twin Study'. *Obesity Science & Practice*, n/a-n/a. doi:10.1002/osp4.79.
[11] Desarrollados a partir del mismo cigoto y por tanto con idéntica dotación cromosómica.
[12] Body Mass Index.

Estudio con emparejamiento natural. La peculiaridad con respecto al diseño explicado en clase es que generan la discordancia en el propio emparejamiento (por lo que solo vemos pares discordantes) en vez de buscarlas después. Así pues solo anotan si los que proceden de familias en los que los padres se separaron o en los que no (la pareja de gemelos no puede estar a la vez en una familia de padres separados y no separados).

$$\begin{cases} H_0: \pi_{gem.disc_{p.sep}} = \pi_{gem.disc_{p.nsep}} \\ H_1: \pi_{gem.disc_{p.sep}} \neq \pi_{gem.disc_{p.nsep}} \end{cases}$$

	obese no / sep sí	**sep no**
obese sí / tiempo sí	0	20
obese sí / tiempo no	14	0

	prob b	prob c	dif.prop	EEdp	CIlower	CIupper	z/chi2	df	p.value
Z	0.5882	0.4118	0.1765	0.1715	-0.1597	0.5126	1.0290	NA	0.3035
chi2	NA	NA	NA	NA	NA	NA	1.0588	1	0.3035
chi2Corr	NA	NA	NA	NA	NA	NA	0.7353	1	0.3912

Atendiendo a los resultados la diferencia de proporciones no es suficientemente importante para rechazar la hipótesis nula. Lo mismo se desprende de la lectura del IC. Tenemos una confianza del 95% de que la diferencia en la proporción de pares discordantes en la población de la que han salido estas muestras, sea **entre el 16% superior entre hijos de *parejas de padres no separados* y el 51.3% superior entre *hijos de padres separados*.**

Dado el tamaño muestral, en este caso es recomendable utilizar la corrección de continuidad.

$$\chi^2_{Mcnemar_{exp}} = \frac{(|b-c|-1)^2}{b+c}$$

$$\chi^2_{Mcnemar_{exp}} = \frac{(|20-14|-1)^2}{20+14} = 0.7353$$

Si no rechazábamos sin corrección, no rechazaremos con la corrección, siempre más conservadora.

17. Exposición a radiación e incidencia de cáncer.

Tras el accidente de la central de Fukushima en marzo de 2011 los trabajadores implicados en los trabajos de limpieza y desescombro fueron sometidos a un seguimiento especial de exposición a la radiación y revisiones periódicas de salud.

El nivel de exposición fue recogido mediante dosímetros homologados. En función del nivel de exposición máximo alcanzado, los trabajadores fueron clasificados en cuatro niveles de menor (I) a mayor (IV) nivel de exposición en el primer mes (cuando los niveles de emisión eran más altos).

Los trabajadores fueron seguidos desde entonces y se registró la aparición de enfermedades a lo largo del primer año. Los resultados respecto a la incidencia de cáncer de tiroides se recogen en la siguiente tabla.

	n	Nº de casos cáncer de tiroides
I	230	40
II	170	36
III	80	30
IV	42	25

Cuestiones.

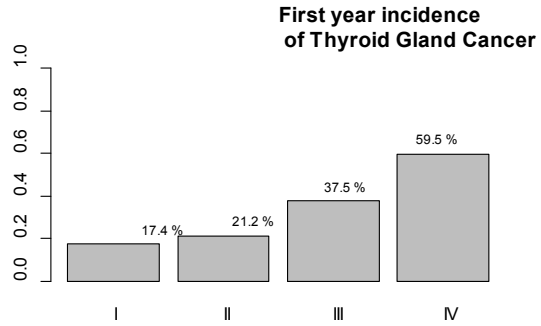
1. ¿Fue diferente la incidencia acumulada en el primer año (proporción de casos nuevos tras un año de seguimiento) en función del grado de exposición?
2. ¿Existió una tendencia lineal de aumento entre la incidencia de cáncer y el grado de exposición?

Para contestarlas interprete el resultado de los siguientes gráficos y test.

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|-------------------------|

Total Observations in Table:  522
```

	Cancer	NoCancer	Row Total
I	40	190	230
	0.174	**0.826**	**0.441**
II	36	134	170
	0.212	**0.788**	**0.326**
III	30	50	80
	0.375	**0.625**	**0.153**
IV	25	17	42
	0.595	**0.405**	**0.080**
Column Total	131	391	522
	0.251	**0.749**	




```
Statistics for All Table Factors

Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  41.6832     d.f. =  3     p =  4.683782e-09


Fisher's Exact Test for Count Data
------------------------------------------------------------
Alternative hypothesis: two.sided
p =  2.595714e-08

 Chi-squared Test for Trend in Proportions

data:  casos out of trab ,
 using scores: 1 2 3 4
X-squared = 36.239, df = 1, p-value = 1.745e-09
```

Solución.

Si la pregunta es solo si hay diferencia en las proporciones, al tener más de dos grupos utilizamos una ji-cuadrado de Pearson.

Hipótesis.

$$H_0: \mu_1 = \mu_2 = \cdots . \mu_j = \mu$$

$$H_1: Cualquier\ \mu_j \neq \mu$$

Dado el número de celdas sería bastante laborioso hacerlo manualmente, por ello se sumiinistra la salida obtenida utilizando la función *CrosTable* del paquete 'gmodels' personalizada para que muestre solo el % de la fila, el resultado del ji-cuadrado y del test de Fisher.

Ambos test nos indica que una o más proporciones son demasiado diferentes como para asumir que todos los grupos vienen de la misma población (son homogéneos.

Como se puede ver la proporción en el conjunto es del 25.1%, pero es heterogénea entre los grupos pasando del 17% en nivel de exposición I al 60% en nivel de exposición IV.

Para analizar la tendencia lineal es necesario recurrir a la ji-cuadrado de tendencia lineal (existen procedimientos alternativos como la correlación no paramétrica de Spearman).

(*) No es necesario saber construir la siguiente tabla, pero si interpretar la salida del test.

Exposición radiación	n_i	$n_i xc_i$	x_i	**inc cancer**	$x_i xc_i$	$n_i xc_i^2$
I	230	230	40	**17.4%**	40	230
II	170	340	36	**21.2%**	72	680
III	80	240	30	**37.5%**	90	720
IV	42	168	25	**59.5%**	100	672
Total	522	978	131	**25.1%**	302	2302

proporción (p)	25%
1-proporción (q)	75%
Media ponderada de clase	1.874
N	522
ji-cuadrado tend lineal	**36.24**
p	1.74499E-09

El test de tendencia lineal explora la existencia de dicha tendencia en el cambio de las proporciones. Como ya se intuye en la tabla y en el gráfico, a medida que aumenta el grado de exposición aumenta la incidencia, e interpretando el test de tendencia lineal (**X-squared = 36.239, df = 1, p-value = 1.745e-09**) dicha tendencia parece lineal.

Análisis de datos con R

Competencias a alcanzar

Hábito tabáquico.

Se está realizando un estudio sobre el consumo de tabaco en la Comunidad de Madrid. Estos son los resultados por zonas geográficas.

	Tamaño	Fumadores
Madrid Capital	25000	15%
Zona Norte	15000	12%
Zona Sur	10000	25%
Zona Este	7500	30%
Zona Oeste	6500	17%

Cuestiones:

a) Realice el contraste oportuno para valorar si existen diferencias en la proporción de fumadores entre las zonas.

```
Construimos una matriz que recoja la tabla de contingencia.
p<-c(.15,.12,.25,.30,.17)
n<-c(2500,1500,1000,750,650)
fumadores<-n*p
nofumadores<-n*(1-p)
tabfum<-matrix(c(fumadores,nofumadores),5,2)
dimnames(tabfum)<-
list(c('Capital','Norte','Sur','Este','Oeste'),c('Fuma','No fuma'))
prop.table(tabfum,1)
        Fuma No fuma
Capital 0.15    0.85
Norte   0.12    0.88
Sur     0.25    0.75
Este    0.30    0.70
Oeste   0.17    0.83
> prop.test(tabfum)

 5-sample test for equality of proportions without continuity correction

data:  tabfum
X-squared = 1597.5, df = 4, p-value < 2.2e-16
alternative hypothesis: two.sided
sample estimates:
prop 1 prop 2 prop 3 prop 4 prop 5
  0.15   0.12   0.25   0.30   0.17
```

Rechazaríamos la hipótesis nula, por lo que en una o más zonas la proporción de fumadores es diferente.
El elevado tamaño muestral de los grupos hace que incluso una pequeña diferencia nos lleve a rechazar la hipótesis nula. Otra consecuencia es

que la memoria necesaria para obtener Fisher es mayor y dependiendo de la memoria disponible en el equipo, se podrá o no realizar. Recordar que el resultado del p-valor obtenido mediante ji-cuadro con estos tamaños muestrales y proporciones no es diferente del obtenido por Fisher.

Si echamos un vistazo a la tabla de residuales[13], vemos que los grupos que más se separan

Cell Contents
N
Expected N
Chi-square contribution
N / Row Total
N / Col Total

Total Observations in Table: 6399

	Fuma	No fuma	Row Total
Capital	375	2125	2500
	445.382	2054.618	
	11.122	2.411	
	0.150	0.850	0.391
	0.329	0.404	
Norte	180	1320	1500
	267.229	1232.771	
	28.473	6.172	
	0.120	0.880	0.234
	0.158	0.251	
Sur	250	750	1000
	178.153	821.847	
	28.975	6.281	
	0.250	0.750	0.156
	0.219	0.143	
Este	225	525	750
	133.615	616.385	
	62.503	13.549	
	0.300	0.700	0.117
	0.197	0.100	
Oeste	110	539	649
	115.621	533.379	
	0.273	0.059	
	0.169	0.831	0.101
	0.096	0.102	
Column Total	1140	5259	6399
	0.178	**0.822**	

Statistics for All Table Factors

[13] Con esta función los residuos se calculan $\frac{(o_{ij}-e_{ij})}{\sqrt{e_{ij}}}$

```
Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  159.8192     d.f. =  4     p =  1.598452e-33
```

En la tabla se observa que aunque la proporción de fumadores en el conjunto es de un 17.8% las región Este (30%), la Sur (25%) y la norte (12%) son las regiones que más aportan al estadístico y por tanto las diferentes al conjunto, siéndolo en sentidos diferentes (Este y Sur mayor prevalencia, Norte menor).

En conclusión la proporción de fumadores es diferente siendo superior a la existente en el conjunto en las regiones Este y Sur y menor en el Norte. Para contestar a cuál es diferente de cuál habría que recurrir a comparaciones por pares, pero la multiplicidad conllevaría el incremento del error de tipo I.