Unidad nº7: *ANÁLISIS de RELACIONES y RELACIONES BIVARIADAS*

Ejercicios resueltos

Prueba de Bondad de Ajuste

Se observaron 80 nacimientos obtenidos del cruzamiento de 10 chanchas Landrace con el mismo padrillo, de los cuales 42 fueron rojizos, 12 negros y 26 blancos. El modelo genético supuesto en este cruzamiento prevé una distribución de colores con frecuencias 9:3:4. ¿Son los datos consistentes con el modelo teórico propuesto, con un nivel del 0,01?

Solución:

$\mathbf{H_0}$: La frecuencia de distribución de colores es 9:3:4
$\mathbf{H_1}$: La frecuencia de distribución de colores difiere de 9:3:4

$\alpha = 0{,}01$

Frecuencias	**Color**			**Total**
	Rojizo	*Negro*	*Blanco*	
Observadas	42	12	26	**80**
Esperadas	$80 \times \frac{9}{16} = 45$	$80 \times \frac{3}{16} = 15$	$80 \times \frac{4}{16} = 20$	**80**

Nota: cómo ninguna frecuencia esperada es menor a 5, no utilizamos la Corrección de Yates.

$$\chi^2_{H_0} = \sum_{i=1}^{3} \frac{(fo_i - fe_i)^2}{fe_i} = \frac{(42-45)^2}{45} + \frac{(12-15)^2}{15} + \frac{(26-20)^2}{20} = 0{,}2 + 0{,}6 + 1{,}8 = 2{,}6$$

Si buscamos 2,6 -o el valor que más se aproxime- en la tabla χ^2, con 2 gl. (Fórmulas y Tablas III, 2ª ed., p. 26), obtenemos $p_v \cong 0{,}25 > \alpha = 0{,}01$. No rechazamos H_0, es decir no hay evidencias suficientes (al 1%) para determinar que la frecuencia de distribución de colores no es 9:3:4.

Prueba de Independencia (marginales libres)

Una empresa minera hizo un estudio para verificar si el lugar de trabajo se relaciona con el grado de silicosis de los/as trabajadores/as. Para lo cual se toma una muestra aleatoria de 300 trabajadores/as y se clasifican en la tabla siguiente:

Lugar de trabajo	**Grado de silicosis**			**Total**
	I	*II*	*III*	
Oficina	42	24	30	**96**
Terreno	54	78	72	**204**
Total	**96**	**102**	**102**	**300**

Pruebe la hipótesis de que el lugar de trabajo afecta el grado de silicosis del trabajador al nivel de significación del 5%.

Solución:

$\mathbf{H_0}$: El grado de silicosis es independiente del lugar de trabajo.
$\mathbf{H_1}$: Existe alguna relación entre grado de silicosis y lugar de trabajo.

$\alpha = 0{,}05$

Ahora vamos a generar la tabla de frecuencias esperadas:

Lugar de trabajo	**Grado de silicosis**			**Total**
	I	*II*	*III*	
Oficina	$\frac{96 \times 96}{300} = 30{,}72$	$\frac{102 \times 96}{300} = 32{,}64$	$\frac{102 \times 96}{300} = 32{,}64$	**96**
Terreno	$\frac{96 \times 204}{300} = 65{,}28$	$\frac{102 \times 204}{300} = 69{,}36$	$\frac{102 \times 204}{300} = 69{,}36$	**204**
Total	**96**	**102**	**102**	**300**

Nota: cómo ninguna frecuencia esperada es menor a 5, no utilizamos la Corrección de Yates.

$$\chi^2_{H_0} = \sum_{i=1}^{2}\sum_{j=1}^{3}\frac{(fo_{ij}-fe_{ij})^2}{fe_{ij}} = \frac{(42-30{,}72)^2}{30{,}72} + \frac{(24-32{,}64)^2}{32{,}64} + \frac{(30-32{,}64)^2}{32{,}64} + \frac{(54-65{,}28)^2}{65{,}28} + \frac{(78-69{,}36)^2}{69{,}36}$$
$$+ \frac{(72-69{,}36)^2}{69{,}36} = 4{,}1419 + 2{,}2871 + 0{,}2135 + 1{,}9491 + 1{,}0763 + 0{,}1005 = 9{,}7683$$

Si buscamos dicho valor -o el que más se aproxime- en la tabla χ^2, con 2 gl. (Fórmulas y Tablas III, 2ª ed., p. 26), obtenemos $p_v \cong 0{,}01 < \alpha = 0{,}05$. Se rechaza H_0 y se concluye que existe alguna relación entre el grado de silicosis y el lugar de trabajo, al nivel del 5%.

Un investigador quiere estudiar si están asociadas la práctica deportiva y la sensación de bienestar. Para ello, toma una muestra aleatoria de 47 estudiantes. Los datos los datos obtenidos aparecen en la siguiente tabla:

Sensación de bienestar	**Práctica deportiva**		**Total**
	Si	*No*	
Si	7	10	**17**
No	6	24	**30**
Total	**13**	**34**	**47**

Docime la hipótesis del investigador, con un nivel del 10%.

Solución:

H_0: La sensación de bienestar es independiente a la práctica deportiva.
H_1: La sensación de bienestar depende de la práctica deportiva.

$\alpha = 0{,}10$

Ahora vamos a generar la tabla de frecuencias esperadas:

Sensación de bienestar	**Práctica deportiva**		**Total**
	Si	*No*	
Si	$\frac{17 \times 13}{47} = 4{,}70$	$\frac{17 \times 34}{47} = 12{,}30$	**17**
No	$\frac{30 \times 13}{47} = 8{,}30$	$\frac{30 \times 34}{47} = 21{,}70$	**30**
Total	**13**	**34**	**47**

Nota: cómo una frecuencia esperada es menor a 5 (correspondiente al 25% de dichas frecuencias), utilizamos la Corrección de Yates.

$$\chi^2_{H_0} = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(|fo_{ij}-fe_{ij}|-0{,}5)^2}{fe_{ij}} = \frac{(|7-4{,}70|-0{,}5)^2}{4{,}70} + \frac{(|10-12{,}30|-0{,}5)^2}{12{,}30} + \frac{(|6-8{,}30|-0{,}5)^2}{8{,}30} +$$
$$+ \frac{(|24-21{,}70|-0{,}5)^2}{21{,}70} = 0{,}6894 + 0{,}2434 + 0{,}3904 + 0{,}1493 = 1{,}4725$$

Si buscamos dicho valor -o el que más se aproxime- en la tabla χ^2, con 1 gl. (Fórmulas y

Tablas III, 2ª ed., p. 26), obtenemos $p_v \cong 0{,}25 > \alpha = 0{,}10$. No se rechaza H_0 y se concluye que no hay evidencias suficientes (al 0,10) para suponer que la sensación de bienestar depende de la práctica deportiva.

Prueba de Homogeneidad (marginales fijos)

El Ministerio de Salud desea verificar si la distribución proporcional del estado nutricional de niños/as no varía en tres ciudades de la región, para lo cual toma una muestra de niños/as de cada ciudad y los clasifica según el estado nutricional obteniendo los siguientes resultados:

Estado nutricional	**Ciudad**			**Total**
	1	*2*	*3*	
Obesidad	82	70	62	**214**
Sobrepeso	93	62	67	**222**
Normo peso	25	18	21	**64**
Bajo peso	16	15	18	**49**
Total	**216**	**165**	**168**	**549**

Pruebe la hipótesis planteada por el Ministerio de Salud, usando una confianza de 0,95.

Solución:

H_0: La distribución proporcional del estado nutricional en las tres ciudades es el mismo.
H_1: La distribución proporcional del estado nutricional difiere en las tres ciudades.

$\alpha = 0{,}05$

Ahora vamos a generar la tabla de frecuencias esperadas:

Estado nutricional	**Ciudad**			**Total**
	1	*2*	*3*	
Obesidad	$\frac{216 \times 214}{549} = 84{,}20$	$\frac{165 \times 214}{549} = 64{,}32$	$\frac{168 \times 214}{549} = 65{,}48$	**214**
Sobrepeso	$\frac{216 \times 222}{549} = 87{,}34$	$\frac{165 \times 222}{549} = 66{,}72$	$\frac{168 \times 222}{549} = 67{,}94$	**222**
Normo peso	$\frac{216 \times 64}{549} = 25{,}18$	$\frac{165 \times 64}{549} = 19{,}23$	$\frac{168 \times 64}{549} = 19{,}59$	**64**
Bajo peso	$\frac{216 \times 49}{549} = 19{,}28$	$\frac{165 \times 49}{549} = 14{,}73$	$\frac{168 \times 49}{549} = 14{,}99$	**49**
Total	**216**	**165**	**168**	**549**

Nota: cómo ninguna frecuencia esperada es menor a 5, no utilizamos la Corrección de Yates.

$$\chi^2_{H_0} = \sum_{i=1}^{4}\sum_{j=1}^{3}\frac{(fo_{ij}-fe_{ij})^2}{fe_{ij}} = \frac{(82-84{,}20)^2}{84{,}20} + \frac{(70-64{,}32)^2}{64{,}32} + \frac{(62-65{,}48)^2}{65{,}48} + \frac{(93-87{,}34)^2}{87{,}34} + \frac{(62-66{,}72)^2}{66{,}72} +$$

$$+ \frac{(67-67{,}94)^2}{67{,}94} + \frac{(25-25{,}18)^2}{25{,}18} + \frac{(18-19{,}23)^2}{19{,}23} + \frac{(21-19{,}59)^2}{19{,}59} + \frac{(16-19{,}28)^2}{19{,}28} + \frac{(15-14{,}73)^2}{14{,}73} + \frac{(18-14{,}99)^2}{14{,}99} =$$

$$= 0{,}0057 + 0{,}5016 + 0{,}1849 + 0{,}3668 + 0{,}3339 + 0{,}0130 + 0{,}0013 + 0{,}0787 + 0{,}1015 +$$

$$+0{,}5580 + 0{,}0049 + 0{,}6044 = 2{,}7547$$

Buscando dicho valor -o bien, el valor más cercano- en la tabla χ^2, con 6 gl. (Fórmulas y Tablas III, 2ª ed., p. 26), obtenemos $p_v \cong 0{,}90 > \alpha = 0{,}05$. Es decir que no rechazamos H_0 y concluimos que no existen diferencias significativas (al 5%) para suponer que la distribución proporcional del estado nutricional, en las tres ciudades, difiere.

Prueba de la Mediana de Mood

Se quiere probar la eficacia de una intervención psicológica para reducir el estrés en un grupo de pacientes que asisten a un centro integrativo de la ciudad de San Luis. Se propone entonces a 32 pacientes que elijan voluntariamente su participación (o no) en el grupo tratamiento y se realiza una evaluación de ambos grupos luego de la intervención, obteniéndose los siguientes resultados:

Intervención	12	17	41	18	22	30	28	40	12	5	18	18	7	24	38	19	15
No Intervención	34	22	27	28	45	9	34	29	31	19	42	35	28	29	30		

Compruebe, al 5%, si existen diferencias entre los grupos.

Solución:

Si calculamos los índices muestrales obtenemos que:

- Grupo con intervención (I): $n_I = 17$ $\bar{X}_I = 21{,}41$ $S_I = 10{,}52$ $CV_I = 49{,}13\%$
- Grupo sin intervención (NI): $n_{NI} = 15$ $\bar{X}_{NI} = 29{,}47$ $S_{NI} = 8{,}44$ $CV_{NI} = 28{,}64\%$

Cómo las medias no son representativas (ya que ambos CV superan el 20%), debemos realizar la prueba de la Mediana de Mood.

En primer lugar, se calcula la mediana combinada de ambos grupos (es decir, para n = 32) que simbolizaremos Ψ (Psi), resultando Ψ = 27,5. Luego, los datos se disponen en una tabla de contingencia teniendo en cuenta si pertenecen al grupo con intervención o al que no se le realizó la intervención y en función de la frecuencia observada de datos de n_I y n_{NI} que sean $\leq \Psi$ y los que sean $> \Psi$, quedando en este caso la disposición de frecuencias observadas:

Frecuencia observada (fo)	**Intervención**	**No Intervención**	**Total**
fo ≤ Ψ	12	4	**16**
fo > Ψ	5	11	**16**
Total	**17**	**15**	**32**

A partir de ello, se trabaja de la misma manera que una prueba Ji cuadrado de homogeneidad (marginales fijos). Siendo el planteo de hipótesis, el siguiente:

$\mathbf{H_0}$: $\theta_I = \theta_{NI}$ vs. $\mathbf{H_1}$: $\theta_I \neq \theta_{NI}$

Dónde: θ_I representa la mediana poblacional de los pacientes que reciben la intervención y θ_{NI} la mediana poblacional de los pacientes que no reciben la intervención.

Sabiendo que trabajamos al 0,05 y con 1 gl (filas-1 × columnas-1). Al calcular las frecuencias esperadas encontramos que:

Frecuencia esperada (fe)	**Intervención**	**No Intervención**	**Total**
fe ≤ Ψ	$\frac{17 \times 16}{32} = 8{,}5$	$\frac{15 \times 16}{32} = 7{,}5$	**16**
fe > Ψ	$\frac{17 \times 16}{32} = 8{,}5$	$\frac{15 \times 16}{32} = 7{,}5$	**16**
Total	**17**	**15**	**32**

Nota: cómo ninguna frecuencia esperada es menor a 5, no utilizamos la Corrección de Yates.

$$\chi^2_{H_0} = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(fo_{ij}-fe_{ij})^2}{fe_{ij}} = \frac{(12-8{,}5)^2}{8{,}5} + \frac{(4-7{,}5)^2}{7{,}5} + \frac{(5-8{,}5)^2}{8{,}5} + \frac{(11-7{,}5)^2}{7{,}5} =$$

$$= 1{,}44 + 1{,}63 + 1{,}44 + 1{,}63 = 6{,}14$$

Si buscamos 6,14 -o aquel valor que más se aproxime- en la tabla χ^2, con 1 gl. (Fórmulas y Tablas III, 2ª ed., p. 26), obtenemos $p_v \cong 0{,}01 < \alpha = 0{,}05$. Se rechaza H_0, es decir que la mediana de estrés de los pacientes que recibieron tratamiento difiere significativamente de la mediana de los pacientes que no lo recibieron, a un nivel del 0,05.

Análisis de Regresión lineal simple

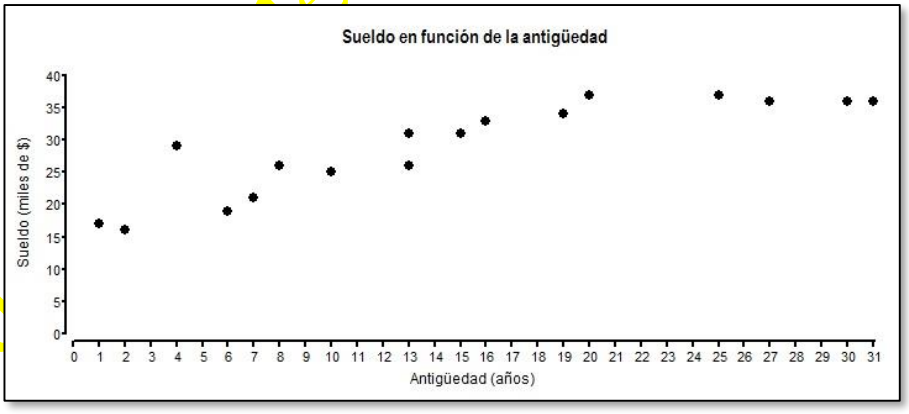
Se llevó a cabo un estudio para determinar la relación entre la antigüedad laboral -en años- (X) y el sueldo mensual -en miles de pesos- (Y) entre galenos de una determinada región. Para ello se tomó una muestra aleatoria de 17 médicos y se obtuvieron los siguientes datos:

Antigüedad	13	16	30	2	8	6	31	19	20	1	4	10	27	25	7	15	13
Sueldo	26	33	36	16	26	19	36	34	37	17	29	25	36	37	21	31	31

a) Realice un dispersograma.
b) En el supuesto de una relación lineal, utilice el método de mínimos cuadrados para calcular los coeficientes de regresión.
c) Interprete el significado de pendiente.
d) ¿La pendiente es significativa al 1%?
e) ¿Qué sueldo mensual (en promedio) predeciría usted que va a ganar un/a médico/a, con 18 años de antigüedad laboral?
f) Calcule el coeficiente de determinación, R^2. ¿Cómo interpretaría el resultado obtenido?

Solución:

a) Dispersograma (o diagrama de dispersión)



b) Para calcular los coeficientes de regresión estimados (a y b), necesitamos los siguientes datos parciales:

$$\sum_{i=1}^{17} x_i = 247;\quad \sum_{i=1}^{17} y_i = 490;\quad \sum_{i=1}^{17} x_i\, y_i = 8097;\quad \sum_{i=1}^{17} x_i^2 = 5065;\quad \sum_{i=1}^{17} y^2 = 14958$$

$$\Rightarrow b = \frac{8097 - 247 \times 490/17}{5065 - (247)^2/17} = \frac{977{,}59}{1476{,}24} = 0{,}66 \qquad a = \frac{490}{17} - 0{,}66 \times \frac{247}{17} = 19{,}23$$

$$\Rightarrow \text{Sueldo} = 19{,}23 + 0{,}66 \times (\text{Antigüedad})$$

c) De acuerdo al modelo presentado en el punto anterior y siendo la pendiente positiva e igual a 0,66, podemos decir que -en promedio- por cada año de antigüedad el/la médico/a

incrementará su salario mensual en $660.

d) $\mathbf{H_0}: \beta = 0$ vs. $\mathbf{H_1}: \beta \neq 0$

$\alpha = 0{,}01$

Siendo $S_e^2 = \frac{14958 - 19{,}23 \times 490 - 0{,}66 \times 8097}{15} = 12{,}75 \Rightarrow V(b) = \frac{17 \times 12{,}75}{17 \times 5065 - (247)^2} = \frac{216{,}75}{25096} = 0{,}0086$

$$\Rightarrow t_{H_0} = \frac{0{,}66}{\sqrt{0{,}0086}} \cong 7{,}12$$

Buscando 7,12 -o el número más próximo- en la tabla "t" de Student, con 15 gl. (Fórmulas y Tablas III, 2ª ed., p.25) obtenemos que, por ser una prueba bilateral, $p_v \cong 2 \times 0{,}0005 = 0{,}001 < \alpha = 0{,}01 \Rightarrow$ por rechazar H_0, la pendiente es significativa al 1%

e) Cómo: $\text{Sueldo} = 19{,}23 + 0{,}66 \times (\text{Antigüedad}) \Rightarrow \text{Sueldo}_{(18)} = 19{,}23 + 0{,}66 \times 18 = 31{,}11$
Esto quiere decir que el sueldo mensual promedio de un/a médico/a con 18 años de antigüedad laboral será de, aproximadamente, $31110.

f) $R^2 = \frac{(8097 - 247 \times 490/17)^2}{[5065 - (247)^2/17] \times [14958 - (490)^2/17]} = \frac{(977{,}588)^2}{1476{,}235 \times 834{,}471} = \frac{955678{,}298}{1231875{,}297} = 0{,}7758$

Esto nos dice que el 77,58% del comportamiento de los datos está explicado por la recta; el resto (22,42%) de dicho comportamiento se debe al azar.

Análisis de Correlación lineal

La siguiente tabla muestra el promedio del consumo anual de vino, en litros por habitante- (X) y las muertes por enfermedad cardíaca -cada 100000 habitantes- (Y), en ciertos países:

País	X	Y
Alemania	2,7	172
Australia	2,5	211
Austria	3,9	167
Bélgica	2,9	131
Canadá	2,4	191
Dinamarca	2,9	220
España	6,5	86
Estados Unidos	1,2	199
Finlandia	0,8	297
Francia	9,1	71

País	X	Y
Irlanda	0,7	300
Islandia	0,8	211
Italia	7,9	107
Noruega	0,8	227
Nueva Zelanda	1,9	266
Países Bajos	1,8	167
Reino Unido	1,3	285
Suecia	1,6	207
Suiza	5,8	115

a) Calcule el coeficiente de correlación entre consumo de vino y muertes por enfermedad cardíaca e interprete el resultado obtenido.

b) ¿Es, el coeficiente de correlación entre consumo de vino y muertes por enfermedad cardíaca, significativamente distinto de cero al 99%?

Solución:

a) Para calcular el coeficiente de correlación, necesitamos los siguientes datos parciales:

$$\sum_{i=1}^{19} x_i = 57{,}5 \quad \sum_{i=1}^{19} y_i = 3630 \quad \sum_{i=1}^{19} x_i^2 = 287{,}4 \quad \sum_{i=1}^{19} y_i^2 = 777726 \quad \sum_{i=1}^{19} x_i\, y_i = 8381{,}4$$

$$\Rightarrow r_{xy} = \frac{8381{,}4 - (57{,}5 \times 3630)/19}{\sqrt{[287{,}4 - (57{,}5)^2/19] \times [777726 - (3630)^2/19]}} = \frac{-2604{,}13}{\sqrt{113{,}39 \times 84203{,}95}} = \frac{-2604{,}13}{3989{,}97} = -0{,}8428$$

Ya que $r_{xy} = -0{,}8428$, estamos en presencia de una correlación negativa es decir, a medida que aumenta el consumo de vino, disminuyen las muertes por enfermedad cardíaca.

b) $\mathbf{H_0}: \rho = 0$ vs. $\mathbf{H_1}: \rho \neq 0$

$\alpha = 0{,}01$

$t_{H_0} = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}} = \frac{(-0{,}8428)\sqrt{17}}{\sqrt{1-(-0{,}8428)^2}} = \frac{-3{,}4750}{0{,}5382} = -6{,}457 \Rightarrow$ Buscando el valor 6,457 -o el que más se aproxime- en la tabla "t" de Student, con 17 gl. (Fórmulas y Tablas III, 2ª ed., p. 25), luego por ser ésta una prueba bilateral $\Rightarrow p_v \cong 2 \times 0{,}0005 = 0{,}001 < \alpha = 0{,}01$. Por ello rechazamos H_0 concluyendo que el coeficiente de correlación, entre consumo de vino y muertes por enfermedad cardíaca, es significativamente distinto de cero, con un nivel de confianza del 99%.

Nota: todos los ejercicios resueltos, presentados anteriormente, pueden realizarse utilizando el software estadístico InfoStat o la planilla de Excel "Prueba de Hipótesis (v.8.11.21)", según corresponda.

Ejercicios propuestos

1. Un estudio estableció la distribución teórica del grupo sanguíneo de una población en 49%, 10%, 6% y 35% para los grupos A, B, AB y O, respectivamente. En una ciudad determinada, se realizó un estudio con una muestra de 200 participantes encontrándose 100, 25, 15 y 60 individuos con cada grupo, respectivamente. Enuncie las hipótesis correspondientes, verifique si se cumple el modelo teórico y señale si son verdaderas o falsas las siguientes afirmaciones:

Se trata de una prueba de bondad de ajuste; la H_0 sostiene que no hay discrepancia entre la distribución teórica esperada y los resultados de cada grupo, lo que es cierto con $\alpha = 0{,}01$.	V - F
Con un nivel de confianza de 0,95 se concluye que la muestra obtenida pertenece a la misma población respecto de la cual se consideró, teóricamente, la distribución de la variable "grupo sanguíneo".	V - F
Si se acepta cometer un Error Tipo I de 5%, debe rechazarse la H_0 concluyéndose que las submuestras no son homogéneas en sus proporciones de grupos sanguíneos.	V - F
Dado que el estadístico de prueba $\chi^2_{H_0} = 3{,}47$, y al ser su $p_v \cong 0{,}25$; puede determinarse que, con un nivel de significación de 0,10, se rechazará H_0.	V - F

2. Una profesora considera que la predisposición de los estudiantes a elegir cursar en las comisiones del turno "mañana" implica mayor responsabilidad en general y se refleja en mejores resultados en las evaluaciones parciales. Luego de la primera instancia respectiva nos comenta que, en efecto, fue mayor la cantidad de aprobados y promocionales en la comisión de la mañana.

Dado que los resultados fueron los indicados en la siguiente tabla, ¿se corrobora empíricamente la suposición de la profesora?, ¿con qué nivel de significación?, ¿considera razonable tal aseveración?, ¿por qué? Enuncie las hipótesis correspondientes.

Resultado	*Turno mañana*	*Turno Tarde*
No aprobaron	77	49
Regularizaron	15	12
Promocionaron	8	4

Elija la opción que corresponda a cada afirmación:

No hay relación significativa entre las variables con un nivel de confianza del 97,5%, por lo que no se rechaza H_0 y se sostiene la independencia entre la elección de la comisión en la cual se cursará y los resultados del parcial.	V - F
Se trata de una prueba de homogeneidad significativa para $\alpha = 0{,}05$. Se rechaza H_1 por demostrarse que hay una leve relación entre las variables	V - F
Dado que el *p*-valor $\cong 0{,}75$ se sostiene que, si se acepta rechazar H_0 siendo ésta verdadera con un error de probabilidad de 0,75, hay una relación cierta entre la predisposición a elegir el turno de la mañana y el mejor resultado en la primera evaluación parcial.	V - F
La H_1 propone independencia entre las variables "resultado del examen" y "comisión elegida".	V - F

3. Luego de un entrenamiento en mnemotecnia realizado en cierta Universidad, se quiso comparar la efectividad de dos estrategias de aprendizaje: 42 estudiantes implementaron la estrategia "E_1" y 58 estudiantes pusieron en práctica la estrategia "E_2". Al finalizar la prueba, 17 de los estudiantes que aplicaron la estrategia "E_1" superaron una puntuación sobresaliente, así como también 18 de los que emplearon la estrategia "E_2".
Determine con una significación del 5% si hay diferencias significativas en la efectividad de dichas estrategias. Enuncie las hipótesis correspondientes, indique el *p*-valor y elija la opción que corresponda a cada afirmación:

Se trata de una prueba de bondad de ajuste; la H_0 sostiene que no hay discrepancia entre la distribución teórica esperada y los resultados de cada grupo, lo que es cierto al 0,05.	V - F
Con un nivel de confianza de 0,50 se concluye que hay diferencias significativas en las proporciones de superación de cada grupo, entonces no son homogéneos.	V - F
Si se llega a aceptar la probabilidad de cometer un Error Tipo I de 50%, no puede rechazarse la H_0 concluyéndose que los grupos son homogéneos en sus proporciones de éxitos.	V - F
Dado que el estadístico de prueba ($\chi^2_{H_0}$) tiene un valor de 0,95 puede determinarse que, con un nivel de significación de 0,25, se rechazará H_0.	V - F

4. Se constató que, de un total de 900 partos, nacieron 470 varones y 430 mujeres. La relación teórica entre los géneros es 1:1. ¿Concuerdan los datos obtenidos con la teoría genética, con un nivel de significación del 2,5%? Enuncie las hipótesis correspondientes y elija la opción que corresponda a cada afirmación:

La similitud entre la proporción teórica y la observada se manifiesta con un nivel de confianza de 0,95. La H_0 debe sostenerse por no encontrarse diferencias significativas entre la distribución de nacimientos observados y el modelo teórico.	V - F
Ya que el *p*-valor es menor al nivel de significación de 0,5; se concluirá que los nacimientos de la muestra coinciden con lo teóricamente esperado con un nivel de confianza de 0,5.	V - F
Las frecuencias observadas en el género de nacimiento en la muestra se alejan de lo esperado teóricamente pero no tanto como para rechazar H_0 con un nivel de confianza del 90%.	V - F
La H_1 no se sostiene con un nivel de confianza de 99%; por eso es esperable que, en 100 muestras pertenecientes a esta distribución, solo 1 vez determinemos que no se cumple con el modelo teórico siendo esta conclusión errada.	V - F

5. Dado que la asistencia a las clases teórico-prácticas de la asignatura "Estadística" es opcional, se quiso determinar si existe relación entre la asistencia a clases y los resultados del primer parcial. De 139 estudiantes que en un determinado año se presentaron a rendir en tal instancia, encontramos: entre los/as 39 estudiantes que superaron un mínimo de asistencia a clases, 25 aprobaron; entre los/as 100 estudiantes que no alcanzaron tal mínimo de asistencia, aprobaron 38. Determine, al 99%, si se demostró relación significativa

entre las variables asistencia a clases y aprobación del primer parcial. Enuncie las hipótesis, verifique el *p*-valor y elija la opción que corresponda a cada afirmación:

Dado que el investigador fija en el estudio, decidiendo a priori, cuántos/as estudiantes aprobarán y cuántos/as no, así como quiénes asisten, los marginales fijos dan la pauta de que se trata de una prueba de homogeneidad. Por ello, la distribución teórica de contraste es una χ^2 con 1 gl.	V - F
Ya que el *p*-valor < α, según un nivel de confianza del 99%, se sostiene que las variables "asistencia a clases" y "resultado de aprobación del parcial" son estadísticamente independientes.	V - F
La H_0 propone la ausencia de relación entre las variables consideradas. Como aquélla es rechazada con $\alpha = 0{,}01$ podemos afirmar que hay relación estadísticamente significativa y esto significa que, sólo en una oportunidad de cada 100, encontraríamos este resultado por Error, de acuerdo al modelo de distribución continua de probabilidad χ^2.	V - F
Se trata de una prueba de independencia cuyo estadístico de contraste presenta un *p*-valor aproximado de 0,005. Con este nivel de significación queda establecida la mutua dependencia estadística entre las variables estudiadas, esto es decir que están relacionadas.	V - F

6. Se toma una muestra aleatoria de 100 madres primerizas atendidas en un determinado hospital, cuya recuperación posparto es clasificada como *alta*, *media* o *baja* según la distribución de la Tabla nº1. Se desea determinar si en tal hospital se cumple con las especificaciones médicas previstas según la Tabla nº2. Enuncie las hipótesis involucradas y señale el *p*-valor correspondiente a la prueba efectuada.

Tabla nº1

Recuperación posparto		
Alta	*Media*	*Baja*
69	24	7

Tabla nº2

Especificaciones médicas		
Alta	*Media*	*Baja*
0,78	0,18	0,04

Elija la opción que corresponda a cada afirmación:

Se trata de una prueba de bondad de ajuste consistente en determinar si la distribución muestral observada guarda correspondencia con los valores teóricos previamente establecidos. La H_0 indica tal "ajuste" y se sostiene al 1%.	V - F
Con $\alpha = 0{,}05$ puede sostenerse que los niveles de recuperación posparto en ese hospital cumplen con las especificaciones médicas previstas; sin embargo, si dadas las implicancias prácticas se prefiere mayor precaución respecto de cometer un Error Tipo II (p/ejemplo, reduciendo el nivel de confianza al 90%), se concluirá que en ese hospital los niveles de recuperación no son los que corresponden.	V - F
Si la PH fuera contrastada según una distribución χ^2 (con 2 gl) al 0,25, la hipótesis de investigación se sostendría concluyendo que las frecuencias observadas se alejan tanto de lo esperado que podemos afirmar que es otro el modelo teórico que explica los datos encontrados.	V - F
La H_0 no puede rechazarse; ni con un nivel de confianza del 95%, ni con un nivel de significación de 0,025, ni aceptándose una probabilidad de Error Tipo I del 1%; por lo tanto, no puede sostenerse que en dicho hospital se cumplan las especificaciones médicas requeridas.	V - F

7. Una genetista realiza un cruzamiento entre arvejas lisas y amarillas, por un lado; y arvejas rugosas y verdes, por el otro, obteniendo los siguientes resultados:

Semillas	**f**
Lisas y amarillas	939
Lisas y verdes	271
Rugosas y amarillas	280
Rugosas y verdes	110
Total	**1600**

Establecer si las características halladas siguen las leyes de Mendel de acuerdo con una de las proporciones clásicas de la herencia mendeliana: 9:3:3:1. Enuncie las hipótesis correspondientes, determine el p_v y elija la opción que corresponda a cada afirmación:

Se trata de un problema de contraste de hipótesis que debe resolverse mediante una prueba de independencia y cuya H_0 será, como el nombre de la prueba indica, la no dependencia de las variables color y rugosidad.	V - F
Con un nivel de significación de 0,01 puede sostenerse más plausible la H_1 que la H_0.	V - F
Con un nivel de significación de 0,05 puede afirmarse que la distribución encontrada en la muestra no guarda correspondencia con las proporciones mendelianas; sin embargo, esto no puede sostenerse aumentando la precisión hasta un nivel de significación de 0,01.	V - F
Con un nivel de significación del 2,5% debe rechazarse H_0. Por otra parte, aceptando una probabilidad de Error Tipo I de 25% se sostiene la H_1. En ambos casos debe concluirse que la distribución observada no se ajusta a la teórica.	V - F

8. Una investigación tuvo el objetivo de verificar si existe relación entre género (masculino y femenino) y rendimiento en comprensión lingüística (CL), mediante la aplicación de una prueba específica. La muestra se constituyó de 32 estudiantes; 12 de ellos fueron varones y, entre éstos, 5 lograron completar satisfactoriamente la prueba. El resto fueron mujeres, y 12 de éstas también consiguieron tal resultado. Desea determinarse, al 5%, si hay diferencias significativas en CL asociadas al género. Enuncie las hipótesis, determine el p_v y elija la opción que corresponda a cada afirmación:

Las variables no están relacionadas; ni siquiera según un nivel de confianza de apenas 75%. Tales variables son independientes al 0,01; al 0,05; al 0,10 y hasta al 0,25.	V - F
Sabemos que no se trata de una prueba de bondad de ajuste debido a que no hay datos teóricos a priori respecto de la distribución de frecuencias esperadas.	V - F
Puede sostenerse que el género está asociado o relacionado estadísticamente con la comprensión lingüística; no obstante, tal hipótesis no es más verosímil que la conclusión que pudiese obtenerse decidiendo según el resultado de lanzar al aire una moneda.	V - F
Dado que el $p_v > 0{,}10$, la H_0 no se rechaza y, por esto, se sostiene el ajuste muestral a la teoría.	V - F

9. Se piensa que cierto rasgo humano es heredado de acuerdo con la proporción 1:2:1 para homocigota dominante, heterocigota y homocigota recesivo, respectivamente. EI examen de una muestra aleatoria simple de 200 individuos proporcionó la siguiente distribución del rasgo: dominantes, 43; heterocigotas, 118; recesivos, 39. Se desea saber si los datos proporcionan evidencia suficiente para desechar dudas sobre tal distribución del rasgo con un nivel de confianza del 95%. Enuncie las hipótesis contrastadas, diga cuál es el *p*-valor y elija la opción que corresponda a cada afirmación:

La hipótesis debe contrastarse con una prueba de independencia debido a que partimos de una hipótesis nula sosteniendo que no hay relación entre las variables genética y ambiente, es decir que queremos rechazar su independencia.	V - F
La hipótesis debe contrastarse con una prueba de bondad ajuste siguiendo una distribución χ^2_1 al 5% cuya variable pivotal, $\chi^2_{H_0} = 6{,}64$, corresponde a un p_v del 1% y por esto no puede sostenerse que la distribución observada se corresponda con el modelo teórico.	V - F
La hipótesis debe contrastarse con una prueba de bondad de ajuste siguiendo una distribución χ^2_2 al 5% cuyo estadístico de prueba ($\chi^2_{H_0} = 6{,}64$) es superior al de un nivel de significación de 0,05 pero inferior al de un nivel de significación de 0,01. Por esto, debe rechazarse H_0 con un nivel de confianza del 95% pero no si éste se aumenta al 99%.	V - F
Ya que a la variable pivotal le corresponde un p_v que permite rechazar H_0, con $P(\varepsilon I) = 0{,}05$; se concluye que las diferencias encontradas, entre la distribución de frecuencias observada y la propuesta teóricamente, son debidas a variaciones de muestreo pero, de todos modos, el ajuste es apropiado.	V - F

10. Un grupo de 140 estudiantes aceptaron voluntariamente participar en un proyecto de investigación. En una de sus etapas, se seleccionaron aleatoriamente a 90 de ellos/as para

desarrollar una serie de talleres de entrenamiento en técnicas de relajación, dejando a los/as restantes como grupo control. Una vez concluido el ciclo de talleres, se citó a los/as participantes en un anfiteatro en el que se proyectaron cortos de terror luego de activarse en cada estudiante un equipo portátil individual de biofeedback. De entre los/as que asistieron a los talleres, 32 estudiantes sostuvieron los niveles deseados de ecuanimidad, en contraste con 12 estudiantes pertenecientes al grupo control. Desea determinarse, al 90%, si las técnicas enseñadas en los talleres fueron efectivas. Enuncie las hipótesis involucradas y elija la opción que corresponda a cada afirmación:

En rigor, no queremos evaluar la relación entre dos variables sino verificar la efectividad de la técnica comparando un grupo experimental con el grupo control (así, contamos con marginales fijos). Dado que compararemos las proporciones observadas en cada grupo, emplearemos una prueba de homogeneidad para la realización del contraste de hipótesis.	V - F
Aunque la proporciones de éxito observadas en el grupo de los que aprendieron esas técnicas son mayores que las esperadas, al 90%, concluimos que tal entrenamiento no implica diferencias significativas respecto del desempeño de quienes no las manejan (incluso aunque en el grupo control las frecuencias observadas fueran menores a las esperadas).	V - F
Se plantea una H_0 sosteniendo la homogeneidad entre el grupo control y el grupo experimental respecto de las proporciones entre quienes lograron los niveles deseados y quiénes no. Esto se rechaza con una $P(\varepsilon I) = 0{,}25$.	V - F
No puede sostenerse que el aprendizaje de tales técnicas explique las diferencias encontradas, ni con un nivel de significación de 0,05 ni con uno de 0,10.	V - F

11. Se observó el color del pelo en 80 neonatos; de los cuales, 42 tuvieron pelo color castaño, 12 pelirrojos y 26 rubios. El modelo genético correspondiente a esta característica es 9:3:4. ¿Son los datos consistentes con el modelo teórico propuesto, al 99%? Enuncie las hipótesis a ser contrastadas y elija la opción que corresponda para cada afirmación:

Los marginales fijos dan la pauta de que se trata de una hipótesis que necesita contrastarse aplicando una prueba de homogeneidad, aunque también puede plantearse como una prueba de independencia con 1 gl.	V - F
Se trata de un problema que requiere del contraste de una hipótesis de nulidad que sostiene la concordancia entre los resultados empíricos y los datos teóricos respecto de la distribución de frecuencias de los caracteres estudiados.	V - F
Por el solo hecho de contar con tres categorías necesitaremos efectuar la corrección por continuidad de Yates restando 0,5 a cada valor absoluto de las diferencias antes de elevarlas al cuadrado para calcular la variable pivotal.	V - F
Con un nivel de confianza del 50%, así como también con un nivel de significación del 50%, podemos rechazar H_0. Sin embargo, esto implicará tener que aceptar $P(\varepsilon I) = 0{,}50$. Es decir que, en tales circunstancias, rechazaremos una H_0 verdadera la mitad de las veces.	V - F

12. Fueron elegidas al azar 60 estudiantes voluntarias para participar de un experimento de ubicación geoespacial. Como se disponía de 20 cascos de realidad virtual (RV), se sortearon éstos para el entrenamiento de dicha cantidad de participantes, asignándose las restantes a una práctica por medio de computadoras personales (PC). En una posterior etapa, se llevó a las estudiantes a un laberinto de arbustos y se registró quiénes lograron encontrar la salida en un tiempo inferior a un record preestablecido. De las que entrenaron con cascos de RV, 5 lograron dicha marca; así como también 7 de las que practicaron con una PC. Determine, al 10%, si el entrenamiento mediante RV marcó alguna diferencia significativa. Enuncie las hipótesis a ser contrastadas y elija la opción correcta para cada afirmación:

Puesto que el investigador decide a priori los totales que corresponderán a cada grupo, el de quienes usarán casco de RV y el de quienes usarán PC (marginales fijos), consideraremos el problema como un contraste de hipótesis a resolver mediante una prueba de homogeneidad de las proporciones de éxito encontradas en cada grupo.	V - F
Dado que más del 20% de las casillas requiere frecuencias esperadas menores a 5, deberá aplicarse la corrección de continuidad de Yates.	V - F
Aun siendo tolerantes respecto del nivel de Error Tipo I a aceptar, fijando, por ejemplo, un nivel de confianza de 75% no se justifica la implementación del entrenamiento mediante cascos de RV ya que los resultados no se alejan significativamente de los obtenidos mediante meros entrenamientos con PC.	V - F
La conclusión de este experimento es que con un nivel de confianza de 90% no puede rechazarse H_0 de homogeneidad; por lo tanto, los grupos son homogéneos respecto de las proporciones de superación del récord preestablecido.	V - F

13. El administrador de un hospital suponía que los ingresos al servicio de cirugía provenían equivalentemente de tres fuentes: *consultorio externo*; *urgencias* y *derivaciones* desde otros hospitales. Al estudiar el origen de las internaciones en los últimos tres años, encontró la siguiente distribución:

Consultorio externo	**Urgencias**	**Derivaciones**
735	684	795

¿Qué tan verosímil era la suposición del administrador del hospital? Enuncie las hipótesis a contrastar y elija la opción que corresponda a cada afirmación:

Dado que ponemos a prueba la relación entre las variables consultorio externo, urgencias y derivaciones, debemos efectuar un contraste de hipótesis mediante la prueba de independencia según una distribución χ^2_2	V - F
Sabemos que la suposición del administrador es verosímil -pero aceptando P(εI) = 0,05- debido a que tendremos que rechazar H_0 por ser el *p*-valor < α.	V - F
Si se aceptara aumentar el nivel de confianza hasta un 97,5%, entonces sí tendría razón el administrador del hospital.	V - F
Con una significación del 0,01 no logra rechazarse la H_0. Siendo así, y aceptando entonces no más que un Error Tipo I del 1%, se concluye que el administrador del hospital está equivocado.	V - F

14. En una encuesta preelectoral realizada a 500 personas se obtuvo la siguiente distribución en función de sus edades y de su intención de voto por los respectivos Partidos Políticos:

Partido Político	**Rango etario**			**Total**
	18 – 35	*35 – 50*	*50 o más*	
A	10	40	60	**110**
B	15	70	90	**175**
C	45	60	35	**140**
D	30	30	15	**75**
Total	**100**	**200**	**200**	**500**

¿Se puede afirmar, al 99%, que la intención de voto depende de la edad? Enuncie las hipótesis respectivas y elija la opción que corresponda a cada afirmación:

Dados los elevados coeficientes de variación y la asimetría de las variables para cada grupo de votantes, será preferible la aplicación de la Prueba de la Mediana de Mood con 6 gl.	V - F
La intención de voto para el "partido D" observada en el grupo etario de los más jóvenes es superior a la esperada; mientras que en el grupo de los votantes mayores fue menor que lo esperado: la aplicación de la prueba de hipótesis posibilita determinar si tal contraposición es debida a particularidades del muestreo o corresponde a un efecto sistemático.	V - F
El *p*-valor aproximado correspondiente al estadístico de prueba calculado es de 0,025.	V - F
Este estudio permite concluir que, con un nivel de confianza de 99%, la intención de voto no guarda relación con la edad del votante.	V - F

15. Los siguientes datos dan la talla, en centímetros, de 100 neonatos seleccionados al azar:

52,0	51,6	49,1	50,6	50,7	53,6	49,3	46,9	52,2	47,6	54,9	45,0	45,0
45,7	55,2	54,1	50,7	46,0	52,6	48,7	49,0	52,3	47,3	49,7	55,1	51,1
47,2	56,1	45,8	45,0	56,1	42,9	49,6	46,6	52,7	54,0	55,0	49,1	50,6
44,6	52,3	49,8	46,8	52,0	46,7	51,0	47,0	46,7	51,9	50,0	57,1	51,0
53,4	46,6	43,9	54,7	51,0	45,6	49,1	55,7	43,6	48,1	49,6	45,6	49,6
51,9	47,8	45,1	43,9	52,0	46,1	49,6	54,9	43,7	42,1	49,1	50,0	54,0
44,8	53,1	46,7	47,0	45,8	44,1	56,3	55,0	50,1	53,0	49,5	55,9	49,0
48,9	42,5	49,6	49,9	51,5	50,7	55,0	57,1	50,4				

Determine al 5% si la muestra proviene de una distribución normal, sabiendo que la media es de 50 cm y la desviación estándar de 3 cm (considere 10 intervalos). Elija la opción que corresponda a cada afirmación:

Se trata de un contraste de hipótesis en el cual se pretende contrastar si la distribución de frecuencias observada cumple una distribución teórica prevista (normal, para el caso); por esto, debemos emplear una prueba de bondad de ajuste.	V - F
Dado que el *p*-valor supera el valor de α, la H_0 se sostiene: con un nivel de confianza del 95%, puede aceptarse que la distribución muestral coincide con el modelo de distribución normal.	V - F
Si las características del estudio indicaran cuidar el Error Tipo II y se aceptara una probabilidad de 0,10 de incurrir en Error Tipo I, se rechazará H_0 y se concluirá que la distribución de frecuencias observadas no se ajusta al modelo normal.	V - F
Con un nivel de confianza del 75%, lo que es igual que decir una significación de 0,25, debe rechazarse H_0. Se contará de este modo con datos suficientes para sostener que la distribución de la muestra no se aparta significativamente de una distribución normal.	V - F

16. Las fábricas "A" y "B" produjeron el año pasado, en conjunto, un total de 450 audífonos, de los cuales 200 fueron de la fábrica "A". Se sabe además que el coeficiente de variación de la vida útil para los audífonos producidos es, en ambas industrias, mayor al 20%. Realice una prueba de la Mediana de Mood con $\alpha = 10\%$, teniendo en cuenta que en la fábrica "A", 98 unidades superaron esa mediana general Ψ, mientras que en la fábrica "B" lo hicieron 126 unidades. Plantee las hipótesis correspondientes e identifique si son verdaderos o falsos los siguientes razonamientos:

Por realizarse una prueba de hipótesis mediante la Prueba de la Mediana de Mood, es innecesario el cálculo de los grados de libertad; el resultado final será el mismo cualesquiera sean éstos.	V - F
¿Para qué se realiza una Prueba de la Mediana? Para comparar las medianas de los grupos que se contrastan y permitir determinar si alguna de las fábricas produce sistemáticamente audífonos de mayor durabilidad útil, más allá de las diferencias asociadas al muestreo particular.	V - F
¿Por qué no es válido comparar directamente las medianas, sin asociarle una prueba de hipótesis? Porque no habría fundamentos estadísticos para decidir si las diferencias encontradas son ocasionadas por azar o se deben a una relación efectiva entre las variables consideradas.	V - F
La H_0 puede rechazarse tanto con un nivel de confianza del 90% como del 95%.	V - F

17. De 430 estudiantes que cursaron Metodología I durante el año pasado, 400 son de Psicología, y de este grupo, 198 superaron el valor de Ψ. De los estudiantes restantes, pertenecientes a la carrera de Fonoaudiología, el 40% no pudo superar el valor de Ψ. Sabiendo que en ambas muestras el CV indica que los promedios no son representativos, deberá realizar una prueba de medianas para comprobar si existen diferencias entre ambos grupos de estudiantes, trabajando con $\alpha = 0{,}25$. Plantee las hipótesis y elija la respuesta correspondiente:

Mediante la H_0, se plantea que los estudiantes de ambas carreras obtienen similares valores en la variable investigada, presentando homogéneamente equivalentes recuentos de frecuencias observadas por encima y por debajo de Ψ.	V - F
Con un nivel de confianza de 0,75 debe rechazarse H_0 y concluirse que hay diferencias significativas entre los resultados de los grupos de cada carrera.	V - F
El p_v aproximado al estadístico de prueba calculado es de 0,10.	V - F
Sostener que hay diferencias entre los resultados obtenidos por estudiantes de Psicología respecto a los de Fonoaudiología, con $\alpha = 5\%$, no tiene fundamento estadístico suficiente.	V - F

18. Se desea estudiar hasta qué punto existe relación entre el tiempo de residencia de inmigrantes en nuestro país y su percepción de integración. Se tomó una muestra de 230 inmigrantes a los/as que se les evaluó en ambas variables obteniéndose las siguientes frecuencias observadas: 130 de ellos/as tuvieron más tiempo de residencia mientras que de los/as 130 con percepción de integración bajo, 90 tuvieron menos tiempo de residencia. ¿Confirman estos datos la hipótesis planteada, al 99%?

19. Se desea probar la hipótesis que los resultados en una prueba de eficiencia (medida en "insuficiente", "regular", "buena" y "muy buena") son independientes del tipo de establecimiento. Para ello se evaluó a 46 estudiantes de un establecimiento privado y 82 de un establecimiento público. De los/as 36 estudiantes que obtuvieron una calificación "insuficiente", 6 pertenecían al establecimiento privado y de los/as 46 que obtuvieron calificación "regular", 32 pertenecían al establecimiento público. Además, de los/as 12 con calificación "muy buena", 9 eran del establecimiento privado. Contrastar la hipótesis al 95%.

20. Se quiere comparar si la eficacia de tres equipos de trabajo que realizan su labor en tres plantas diferentes, es similar. Para ello se consideró el número medio de artículos por hora que se terminan en cada planta durante una semana determinada. Los resultados obtenidos han sido:

Planta A	7,3	6,9	7,2	7,8	7,3	8,1	5,7
Planta B	6,1	5,7	8,9	6,7	7,9	8,8	5,2
Planta C	6,7	7,6	8,1	7,4	7,6	8,5	5,5

¿Puede aceptarse, al 5%, dicho supuesto?

21. El número de bacterias por unidad de volumen, presentes en un cultivo después de un cierto número de horas, viene expresado en la siguiente tabla:

nº de horas	0	1	2	3	4	6
nº de bacterias	12	19	23	34	56	62

a) Realice un dispersograma del número de bacterias en función al número de horas.
b) ¿El número de bacterias aumenta o disminuye con el tiempo?
c) ¿Cuántas bacterias esperaría encontrar -en promedio- después de cinco horas?
d) ¿La pendiente es significativa al 1%?
e) Calcule el coeficiente de determinación, R^2, e interprete su significado.

22. En un grupo de 8 pacientes se miden las cantidades antropométricas: peso (en kg) y edad (en años), obteniéndose los siguientes resultados:

Edad	12	8	10	11	7	7	10	14
Peso	58	42	51	54	40	39	49	56

a) ¿Existe una relación lineal entre ambas variables?

b) Determine la recta de regresión del peso en función a la edad.
c) ¿En qué medida, por término medio, varía el peso cada año?
d) Calcule R^2 e interprete su significado.

23. Los siguientes datos corresponden a los porcentajes de mortalidad obtenidos a dosis crecientes de dos insecticidas naturales (I_1 e I_2). El experimento consistió en someter a grupos de 1000 insectos a cada una de las dosis ensayadas. Cada producto se ensayó independientemente. Los resultados de mortalidad (en %) fueron los siguientes:

Ln(dosis)	**0,001**	**0,01**	**0,05**	**0,1**	**0,15**	**0,2**	**0,25**	**0,3**	**0,4**	**0,7**	**0,9**
Mortalidad I_1	5	7	10	16	17	25	26	30	35	60	81
Mortalidad I_2	7	8	12	13	15	24	24	35	45	74	93

a) Construya un diagrama de dispersión Mortalidad vs. Ln(dosis), para cada producto.
b) De acuerdo al gráfico construido en el punto a), ¿es razonable un ajuste lineal?
c) Escriba el modelo lineal que, se supone, relaciona la mortalidad con las distintas dosis.
d) ¿Cuál es la mortalidad de insectos, para cada insecticida, a partir de una dosis de 0,50?

24. La jefa de personal de una empresa considera que hay relación entre ausentismo (en días) y edad (en años) y querría usar la edad de un/a empleado/a para predecir el número de días promedio de ausencia durante un año calendario. De una muestra aleatoria se obtuvieron los siguientes datos:

Edad	27	61	37	23	46	58	29	36	64	40
Ausentismo	15	6	10	18	9	7	14	11	8	8

a) Realice un diagrama de dispersión.
b) En el supuesto de una relación lineal, utilice el método de mínimos cuadrados para calcular los coeficientes de regresión.
c) Interprete el significado de la pendiente.
d) ¿Cuántos días (en promedio) predeciría usted que va a estar ausente un/a empleado/a de 48 años de edad?
e) Calcule el coeficiente de determinación, R^2, e interprete su significado.

25. Se realiza un estudio para establecer una ecuación mediante la cual se pueda utilizar la concentración de estrona en saliva (en pg/mL) para predecir la concentración del esteroide en plasma libre (en nmol/L). Se extrajeron los siguientes datos de 13 varones, mayores de edad:

Estrona	7,5	8,5	9	9	11	13	14	14,5	16	17	18	20	23
Esteroide	2,5	3,15	2,75	3,95	3,8	4,3	4,9	5,5	4,85	5,1	6,45	6,3	6,8

a) Estudie la posible relación lineal entre ambas variables y obtenga la ecuación que se menciona en el enunciado del problema.
b) Determine la variación de la concentración de esteroide en plasma por unidad de estrona en saliva.

26. Se ha medido el aclaramiento de creatinina (mg/dl) en pacientes tratados con Captopril tras la suspensión del tratamiento con diálisis, resultando la siguiente tabla:

Días tras la diálisis	1	5	10	15	20	25	30
Creatinina	5,7	5,2	4,8	4,5	4,2	4,0	3,8

a) Halle la expresión de la ecuación lineal que mejor exprese la variación de la creatinina,

en función de los días transcurridos tras la diálisis.

b) ¿En qué porcentaje la variación de la creatinina es explicada por el tiempo transcurrido desde la diálisis?

c) Si un individuo presenta 4,1 mg/dl de creatinina, ¿cuánto tiempo es de esperar que haya transcurrido desde la suspensión de la diálisis?

27. El contador de una serie de negocios desearía predecir el saldo (miles de dólares) de las cuentas al final del período de facturación, con base en el número de transacciones efectuadas durante dicho período. Se seleccionó una muestra de 12 cuentas, obteniendo los siguientes resultados:

Transacciones	1	2	3	3	4	5	6	7	10	10	12	15
Saldo	15	36	40	63	69	78	84	100	175	120	150	198

a) Realice un diagrama de dispersión.

b) Use el método de mínimos cuadrados para calcular los coeficientes de regresión estimados.

c) Interprete el significado de la ordenada al origen y la pendiente.

d) Pruebe la significación de la regresión al 5%.

e) Prediga el saldo promedio para una cuenta que haya tenido 14 transacciones al final del período de facturación.

f) Calcule el coeficiente de determinación e interprete su significado.

28. La administradora de una heladería desearía estudiar el efecto de la temperatura ambiente (ºF) sobre las ventas (miles de dólares) en temporada estival. Para ello, seleccionó una muestra aleatoria de 14 días, obteniendo los siguientes resultados:

ºF	63	70	73	75	80	82	85	88	90	91	92	75	98	100
Ventas	1,5	1,7	1,8	2,1	2,4	2,3	2,7	2,9	3,1	3,1	3,2	1,9	3,4	3,3

a) Realice un dispersograma.

b) En el supuesto de una relación lineal, calcule los coeficientes de regresión.

c) Interprete el significado de la ordenada al origen y la pendiente.

d) ¿La pendiente es significativa al 1%?

e) Prediga las ventas promedio, por día, cuando la temperatura es de 83 ºF.

f) Calcule el coeficiente de determinación e interprete su significado.

29. El número de personas que viven en ciertas áreas rurales fue decreciendo en los últimos años. A continuación, presentamos los datos (en miles de personas) entre 1965 y 2010:

Año	1965	1970	1975	1980	1985	1990	1995	2000	2005	2010
Población	32,1	30,5	24,4	23,0	19,1	15,6	12,4	9,7	8,9	7,2

a) Realice un dispersograma y estime la recta de mínimos cuadrados para la población rural para los diferentes años.

b) ¿Cuánto decreció la población rural durante un lapso promedio de tiempo?

c) ¿Qué porcentaje de este cambio es explicado por la regresión lineal?

d) Prediga cuántas personas vivirán en áreas rurales en 2025. ¿Es este resultado razonable? ¿Por qué?

30. En un ensayo clínico realizado tras el posible efecto hipotensor de un fármaco, se evalúa

la tensión arterial diastólica (TAD_0) en condiciones basales y tras 4 semanas de tratamiento (TAD_4), en 14 pacientes hipertensos. Se obtuvieron los siguientes valores de TAD:

TAD_0	95	100	102	104	100	95	95	98	102	96	100	96	110	99
TAD_4	85	94	84	88	85	80	80	92	90	76	90	87	102	89

a) Determine el modelo lineal que mejor explica la relación entre la TAD_0 y la que se observa luego del tratamiento.

b) Al 5%, ¿es significativo el valor, de la pendiente, obtenido en el modelo lineal?

c) ¿Cuál es el valor de TAD_4 esperado tras el tratamiento, en un paciente que presentó una TAD_0 de 105 mm de Hg?

31. Una investigadora le mide a 16 sujetos el grado de sensibilidad de un color determinado (Y) relacionado con tiempo de reacción (en segundos - X) frente a un estímulo visual. Se registraron los siguientes datos:

Tiempo	2,31	1,75	1,69	1,36	1,29	1,13	1,02	0,29	0,74	0,73	0,62	0,56	0,40	0,39	0,34	0,00
Color	83	91	91	87	93	93	96	107	100	104	97	118	113	122	107	116

a) Realice un gráfico de dispersión.

b) ¿La pendiente es significativa al 0,01?

c) Si un individuo presenta 110 puntos de sensibilidad a dicho color, ¿cuál fue su tiempo de reacción?

32. Los datos de la siguiente tabla se refieren al contenido de proteína bruta (PB) y caseína (CA) en leche en una muestra de 15 tambos de la cuenca lechera del centro del país.

PB	2,74	2,96	2,91	3,23	3,04	3,23	3,11	2,95	3,08	3,12	2,95	3,15
CA	1,87	2,07	2,09	2,28	2,04	2,30	2,17	2,04	2,16	2,23	2,07	2,24

PB	3,20	3,14	3,14
CA	2,22	2,16	2,22

a) Calcule el coeficiente de correlación de Pearson y explique el resultado obtenido en función a los datos.

b) ¿Es esta correlación estadísticamente significativa al 10%?

33. Las puntuaciones APGAR son asignadas a los recién nacidos uno y cinco minutos después del nacimiento e indican el estado general de salud del neonato. Puntuaciones de 7 a 10 son típicas e indican que el neonato sólo requiere cuidado posnatal de rutina. Puntuaciones de 4 a 6 indican que el bebé quizá necesite asistencia, mientras que puntuaciones de 3 o menos indican que requiere de asistencia inmediata para mantenerlo con vida. Suponga que se realiza un estudio para determinar la consistencia con la cual los puntos APGAR son asignados. Para tal fin, dos Obstetras observaron 12 nacimientos y asignaron independientemente los puntos APGAR a los neonatos. La siguiente tabla da las puntuaciones otorgadas por la primera Obstetra (O_1) y aquellas asignadas por la segunda Obstetra (O_2).

Obstetra	**Neonato**											
	A	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*	*L*
O_1	9	8	7	8	6	4	9	7	2	8	7	9
O_2	7	9	8	8	7	4	8	7	3	7	8	9

a) A partir de la inspección gráfica, ¿cree usted que hay una correlación positiva, negativa o nula entre las dos puntuaciones? ¿Por qué?

b) Calcule el coeficiente de correlación de Pearson para los datos e interprete el resultado obtenido.

c) ¿Es significativo el coeficiente encontrado al 0,05?

34. Catorce estudiantes de segundo año de medicina midieron la presión sanguínea del mismo paciente. Los resultados se presentan a continuación:

Sistólica	138	130	135	140	120	125	120	130	130	144	143	140	130	150
Diastólica	82	91	100	100	80	90	80	80	80	98	105	85	70	100

a) ¿Existe, estadísticamente, correlación entre los valores sistólicos y diastólicos?

b) ¿Es significativo el coeficiente encontrado al 0,05?

35. En un estudio de campo se hicieron mediciones de perímetro (en cm.) y peso (en g.) de cabezas de ajo. Los datos obtenidos fueron los siguientes:

Perímetro	12,4	12,4	12,7	9,8	12,3	10,1	11,8	11,4	9,4	11,5
Peso	32,3	29,4	30,8	15,6	29,8	16,9	28,1	23,3	14,1	25,4

a) Cómo espera que sea la asociación entre peso y perímetro: positiva, negativa o sin asociación, ¿por qué?

b) Calcule coeficiente correlación entre peso y perímetro e interprete el resultado.

c) ¿Es significativo el coeficiente encontrado al 5%?

36. El dueño de una concesionaria de autos desea medir la relación entre el ingreso familiar (en miles de u$d) y el precio de compra de automóviles nuevos (en miles de u$d). Para ello selecciona una muestra aleatoria de 10 personas que adquirieron autos nuevos, obteniendo los siguientes resultados:

Ingresos	10,2	14,4	16,3	20,0	24,3	11,6	32,8	9,4	26,7	18,3
Precio	3,6	4,1	3,9	5,2	5,1	3,9	7,8	3,4	9,1	5,0

a) Calcule el coeficiente de correlación entre el ingreso familiar y el precio de compra e interprete el resultado obtenido.

b) Al nivel del 99%, ¿hay una relación lineal significativa ente el ingreso familiar y el precio de compra?

37. Un estadístico, de una organización de estudios de consumidores, desea determinar la relación entre el precio (en u$d) de una celda fotovoltaica y la duración (en años) de las mismas. Para ello, tomó una muestra aleatoria de 14 celdas, obteniendo los siguientes resultados:

Precio	24	32	49	49	39	69	69	89	119	79	35	80	109	72
Duración	5,4	4,8	6,3	7,2	6,3	7,4	6,8	10,2	13,1	9,2	6,0	8,3	10,2	7,6

a) Calcule el coeficiente de correlación entre el precio y la duración de las celdas fotovoltaicas e interprete el resultado obtenido.

b) Con un nivel de significación de 0,05, ¿hay una relación lineal significativa entre el precio y la duración de las celdas fotovoltaicas?

38. Se quiere comprobar si existe relación entre el consumo de energía per cápita (en libras esterlinas) de un país y su producto bruto interno (PBI). Se seleccionaron al azar 12 países obteniendo los siguientes resultados:

País	Consumo per cápita	PBI
EEUU	25598	5515
Australia	12568	3370
Noruega	10227	3779
Japón	7167	2757
Venezuela	5452	1291
Brasil	1173	513

País	Consumo per cápita	PBI
Canadá	23715	4704
Dinamarca	12273	3978
Francia	9156	3810
Italia	6164	2170
Grecia	3543	1382
India	410	98

a) Calcule el coeficiente de correlación entre consumo de energía per cápita y producto bruto interno e interprete el resultado obtenido.

b) ¿Presentan los datos suficiente evidencia sobre la existencia de una relación lineal entre el consumo de energía per cápita y el PBI, al 0,01?

39. Una administradora está interesada en saber si las tasas de interés proporcionan un indicador clave para predecir el número de construcciones. Para ello consideró, durante los últimos ocho años, las tasas de interés (en %) y el número de nuevas construcciones iniciadas cada año, obteniendo los siguientes resultados:

Tasa de interés	32,5	30	32,5	37,5	42,5	47,5	50	45
nº de construcciones	2165	2984	2780	1940	1750	1535	962	1310

a) Calcule el coeficiente de correlación entre tasa de interés y número de construcciones e interprete el resultado obtenido.

b) ¿Es, el coeficiente de correlación entre tasa de interés y número de construcciones, significativamente distinto de cero al 95%?

40. Un investigador está interesado en saber si existe relación entre el índice de masa corporal (IMC) y el colesterol sérico en función de saber si puede predecir, a partir del IMC, el colesterol sérico total. Para ello tomó una muestra de 10 sujetos, obteniendo:

Colesterol total	165	155	141	228	190	155	132	170	188	150
IMC	25,9	20,1	22,2	30,7	28,0	29,4	20,2	20,7	26,3	18,2

a) Calcule el coeficiente de correlación entre colesterol total e IMC e interprete el resultado obtenido.

b) ¿Presentan los datos suficiente evidencia sobre la existencia de una relación lineal entre ambas variables, al 0,05?

41. Cuando la nicotina es absorbida por el cuerpo, se produce cotinina. Por consiguiente, la medición de cotinina podría ser un buen indicador de cuánto fuma una persona. A continuación, se incluye el reporte del número de cigarrillos fumados (por día) y las cantidades medidas de nicotina (en ng/mL), de una muestra aleatoria de 12 sujetos.

Cigarrillos	60	10	4	15	10	1	20	8	7	10	10	20
Cotinina	179	283	75,6	174	209	9,5	350	1,8	43,4	25,1	408	344

a) Determine si la cotinina es un buen indicador de cuántos cigarrillos fuma por día una persona.

b) ¿Es el coeficiente de correlación entre cigarrillos por día y cotinina, significativamente distinto de cero al 1%?

42. La siguiente tabla nos muestra el diámetro del pecho (en cm) y el peso (en kg) de osos elegidos al azar, que fueron anestesiados y medidos. Como es más difícil pesar un oso que

medir el diámetro de su pecho, la presencia de una correlación podría conducir a un método para estimar el peso a partir del diámetro del pecho.

Diámetro	66,0	114,3	137,2	124,5	88,9	104,1	104,1	124,5	96,5	78,7
Peso	36,3	156,0	188,7	157,9	75,3	99,8	118,8	163,3	92,5	65,3

a) ¿Existe una correlación lineal entre el diámetro del pecho y el peso?
b) ¿Qué la sugiere el resultado obtenido en el punto anterior?

43. En la siguiente tabla se listan los números de homicidios y los tamaños poblacionales (en cientos de miles) de grandes ciudades de Estados Unidos durante un año reciente. ¿Qué concluye usted, con una confianza del 0,99?

Homicidios	258	264	402	253	111	648	288	654	256	60	590
Población	4	6	9	6	3	29	15	38	20	6	81

44. En "The Effects of Temperature on Marathon Runner's Performance", se incluyen las altas temperaturas (en ºF) registradas junto con los tiempos (en minutos) de mujeres que ganaron la maratón de la ciudad de Nueva York en años recientes, siendo los resultados:

Temperatura	55	61	49	62	70	73	51	57
Tiempo	145,28	148,72	148,30	148,10	147,62	146,40	144,67	147,53

a) Realice un dispersograma e indique qué tipo de correlación existe entre la temperatura y el tiempo de las ganadoras. Interprete el resultado obtenido.
b) ¿Considera que los tiempos obtenidos por las ganadoras se ven afectados por la temperatura ambiental, al 5%? Interprete el resultado obtenido.

45. Un investigador hipotetiza que las personas más altas tienen pulsos más rápidos porque la sangre tiene que viajar más lejos. A continuación, se listan los pulsos (en latidos/minuto) y las estaturas (en pulgadas) de una muestra aleatoria de mujeres adultas. Las estaturas y pulsos aparecen en orden, de manera que corresponden a la misma persona.

Estatura	64,3	66,4	62,3	62,3	59,6	63,6	59,8	63,3	67,9	61,4	66,7	64,8
Pulso	76	72	88	60	72	68	80	64	68	68	80	76

a) ¿Cómo interpreta la asociación entre la estatura y el pulso?
b) ¿Es el coeficiente de correlación significativamente distinto de cero al 1%? Interprete el resultado obtenido.

46. Una compañía de seguros considera que el número de vehículos que circulan por una determinada ruta, a más de 120 km/h, puede estar asociados al número de accidentes que ocurren en ella. Dicha compañía tomó una muestra, durante 5 días, obteniendo los siguientes resultados:

nº de accidentes	5	7	2	1	9
nº de vehículos	15	18	10	8	20

a) Determine el coeficiente de correlación e interprételo.
b) ¿Dicho coeficiente es estadísticamente significativo, al 90%?

47. Por medio de un procedimiento químico llamado polarografía diferencial de pulsos, un químico midió la corriente máxima que se generó (en microamperios, μA) al agregar una solución que contenía una cantidad determinada de níquel (en partes por mil millones, pmm) a una solución amortiguadora. Los datos se presentan a continuación:

Níquel	19,1	38,2	57,3	76,2	95	114	131	150	170
Corriente máxima	0,095	0,174	0,256	0,384	0,429	0,5	0,58	0,651	0,722

a) Elabore un diagrama de dispersión con estos datos.
b) Calcule el coeficiente de correlación e interprete el resultado obtenido.
c) ¿El coeficiente obtenido en el punto anterior es significativo al 0,05?

48. El CEO de un Centro Comercial sabe que, en función de la distancia (en km) a la que está situado del centro de una ciudad, acuden los clientes (por cien). Para ello toma una muestra aleatoria de potenciales clientes y obtiene los siguientes datos:

nº de clientes	8	7	6	4	2	1
Distancia	15	19	25	23	34	40

a) Represente los datos de la tabla en un diagrama de dispersión.
b) En el supuesto de una relación lineal, calcule los coeficientes de regresión de la distancia en función al número de clientes.
c) Interprete el significado de la ordenada al origen y la pendiente.
d) ¿Considera que la pendiente es significativa al 5%?
e) ¿En qué porcentaje el modelo propuesto explica el comportamiento de ambas variables?
f) Si el centro comercial se situase a 10 km, ¿cuántos clientes puede esperar?
g) Si desea recibir a 500 clientes, ¿a qué distancia del centro de población debe situarse?

49. En la tabla adjunta se presentan el precio (en u$d) de doce libros técnicos y el número de páginas de los mismos:

Precio	3,50	8,00	2,50	3,50	1,80	5,00	3,50	7,00	5,40	7,30	3,20	3,70
Páginas	310	400	420	300	170	610	280	430	420	310	230	450

a) Estime el modelo de regresión lineal que mejor prediga el precio en relación al número de páginas.
b) ¿Cuánto de la variabilidad absoluta es explicada por el modelo obtenido en el punto anterior?
c) ¿Considera usted que la pendiente es estadísticamente significativa, al 1%?
d) ¿Cuál sería el precio promedio (en u$d) de un libro con 350 páginas?
e) ¿Cuántas páginas promedio tendrá un libro que cuesta 6 u$d?

ALGUNAS DUDAS FRECUENTES

¿Por qué, al aplicar la prueba χ^2 en el estudio de tablas de contingencia, es importante verificar si hay frecuencias observadas menores de cinco?

Porque con frecuencias observadas bajas cabe esperar frecuencias esperadas bajas. Y las frecuencias esperadas aparecen en el denominador de la fórmula con que se determina la variable pivotal. Una frecuencia esperada cero llevaría a una división por cero que no es posible efectuar y un cociente con una frecuencia muy próxima a cero (por ejemplo, 0,000007) conllevaría un sumando, en la fórmula, enorme y, así, una suma enorme que nos llevaría a rechazar la H_0 indebidamente. Dicho en términos más técnicos se incrementaría el Error Tipo I (rechazo indebido de la H_0 por ser ésta verdadera).

Dado el nivel de significación tradicional, ¿una prueba χ^2 que lleve asociado un $p_v > 0,05$ -razón por la que no podría rechazarse la H_0- nos permite estar seguros de que las dos variables son independientes?

No. Sólo nos dice que no hay pruebas suficientes para rechazar la independencia con el nivel de confianza o de significación preestablecido, o de acuerdo con el nivel de Error Tipo I que se esté dispuesto a cometer.

Una prueba χ^2 que lleve asociado un $p_v < 0,05$, ¿nos permite estar seguros de que las dos variables están relacionadas?

No. Sólo nos dice que, a la vista de las pruebas que aportan los datos, el riesgo que corremos al rechazar la hipótesis nula de independencia es bajo. Pero garantía total de que estén relacionadas las variables no tenemos. Tal como ocurre en un juicio, si el juez estima que las pruebas son suficientes para declarar culpable al acusado, así lo declara y lo penaliza (con cárcel, por ejemplo) pero certeza absoluta no hay.

¿Se puede utilizar la prueba χ^2 con variables ordinales?

Sí. Cualquier variable puede operacionalizarse de modo tal que quede reducida no más que al categórico: pero éste no da cuenta de la información (numérica, cuantitativa) asociada al orden en sí.

Al analizar una tabla de contingencia con frecuencias observadas nulas que proporciona resultados estadísticamente significativos, ¿debemos tomar alguna precaución?

Sí, porque la significación puede ser espuria (ver respuesta a la primera pregunta).

Al analizar una tabla de contingencia con frecuencias observadas nulas que proporciona resultados estadísticamente no significativos, ¿debemos tomar alguna precaución?

No, porque frecuencias observadas bajas llevan asociadas frecuencias esperadas bajas. Éstas serán consideradas en el denominador del estadístico de contraste y pueden proporcionar un valor experimental más alto que el real. Entonces: si, incluso siendo mayor de lo que debería, no ha sido suficiente para rechazar la hipótesis nula de independencia… el hecho de que se haya aumentado no hubiera tenido relevancia alguna.

Al confeccionar y analizar una tabla de contingencia que contiene frecuencias relativas (o porcentajes) y proporciona resultados estadísticamente significativos, ¿debemos tomar alguna precaución?

Trabajando con frecuencias relativas (o porcentajes), el valor del estadístico de prueba (o variable pivotal) es más bajo del que debería ser. Si, aun siendo más bajo, ha sido suficiente para rechazar la hipótesis nula de independencia… entonces no hay inconvenientes.

Al confeccionar y analizar una tabla de contingencia que contiene frecuencias relativas (o porcentajes) y proporciona resultados estadísticamente no significativos, ¿debemos tomar alguna precaución?

Sí. Trabajando con frecuencias relativas (o porcentajes), el valor del estadístico de prueba es más bajo del que debería ser. La aceptación de la hipótesis nula de independencia puede deberse a eso. En términos más precisos, podríamos decir que se incrementa el riesgo de cometer un Error Tipo II (no rechazo de la H_0 siendo ésta falsa).

Si en lugar de tener dos variables, tenemos tres (o más): ¿bastaría con estudiar tantas tablas de contingencia como resulten de la combinación en pares de variables (cruzando la variable n° 1 con la n° 2, la n° 1 con la n° 3, la n° 2 con la n° 3, etc.)?

No. Esa es una mala práctica porque puede ocurrir que la información resultante de los análisis parciales sea contradictoria (hecho conocido en la literatura especializada como la "Paradoja de Simpson": *fenómeno de confusión en el cual la intervención de una variable cambia la dirección de una asociación*).

Supongamos las siguientes tablas en la que se recoge información sobre la posible relación entre la esperanza de encontrar empleo y la duración de la desocupación, tanto para varones como para mujeres:

Varones:

Duración de la desocupación	Esperanza de encontrar empleo: Sí	Esperanza de encontrar empleo: No
Corta	90	10
Larga	9	1

$\chi^2 = 0$

Mujeres:

Duración de la desocupación	Esperanza de encontrar empleo: Sí	Esperanza de encontrar empleo: No
Corta	1	9
Larga	10	90

$\chi^2 = 0$

Si analizamos la tabla bifactorial correspondiente a los varones, obtenemos que las frecuencias esperadas bajo supuesto de independencia son exactamente iguales a las frecuencias observadas; hecho este que se corresponde con un valor del estadístico de prueba $\chi^2 = 0$ y con la conclusión de que, en varones, ambas variables pueden considerarse independientes.

El mismo análisis sobre la tabla bidimensional para las mujeres, proporciona de nuevo un valor experimental nulo del cual podemos concluir que, también en mujeres, la esperanza de encontrar empleo puede considerarse independiente de la duración de la desocupación.

Como el resultado se mantiene, tanto en hombres como en mujeres, parece que el género no

tiene mayor interés en el estudio. Sin embargo, reuniendo en la tabla las frecuencias correspondientes a ambos géneros, obtenemos la tabla siguiente:

Varones y mujeres:

Duración de la desocupación	Esperanza de encontrar empleo: Sí	Esperanza de encontrar empleo: No
Corta	91	19
Larga	19	91

$\chi^2 = 94{,}252$

El valor del estadístico de prueba χ^2 es 94,252; el cual corresponde con un *p*-valor altamente significativo ($p_v < 0{,}01$), razón por la cual debemos concluir, a la vista de este nuevo análisis, que ambas variables están claramente correlacionadas. Esto supone una clara contradicción con los dos resultados anteriores: "Paradoja de Simpson".

El problema se debe a que la presencia de la tercera variable puede llevarnos a una ponderación inadecuada de las distintas poblaciones en estudio. En este ejemplo ficticio, 100 varones llevaban desocupados un tiempo corto y sólo 10 mujeres estaban en esa situación; justo la proporción se invertía para los de larga duración. Sin embargo, esta información no había sido reflejada al calcular el valor del estadístico de contraste.

¿Qué hacer cuando tenemos tres variables en estudio?

Abordar el problema analizando las tablas trifactoriales. Esta situación es más compleja ya que no hay una única hipótesis a contrastar sino 7. Requiere del estudio de procedimientos que no veremos en la materia.

¿Un coeficiente de correlación significativo implica que entre las variables X e Y existe una relación causa efecto?

¡No!, en absoluto. Indica solamente que X e Y covarían; pero puede ser que covaríen simplemente por efecto de una tercera variable.

Por ejemplo, el número de canas y la presión arterial covarían… no porque una cause la otra sino por efecto de una tercera variable, que puede ser la edad.

¿Un coeficiente de correlación no significativo implica que entre las variables no existe una relación causa efecto?

No necesariamente. Si no están relacionadas es evidente que una no es causa de la otra; pero puede ocurrir que exista relación y no la detectemos en el estudio por falta de información (por ejemplo, por una muestra pequeña) o que la relación exista, pero no sea lineal.

¿Un coeficiente de correlación cero (o próximo a cero) implica que las variables X e Y son independientes?

No. Lo recíproco sí es cierto: cuando dos variables son independientes el coeficiente de correlación de Pearson es cero; pero puede ser cero y que las variables estén relacionadas según otro modelo.

¿Podemos fiarnos de una publicación en la que aparece una recta de regresión y no incluye ni el coeficiente de correlación ni el coeficiente de determinación?

Podemos fiarnos, pero no es recomendable dada la insuficiencia de la información presentada ya que no sabemos el poder explicativo del modelo: es decir, qué porcentaje de puntos de la nube -covarianza- vienen bien representados por dicho modelo.

Un modelo de regresión lineal con alto valor explicativo (R^2 próximo al 100%), ¿es un buen modelo de predicción?

Lo contrario sí es cierto: si no tiene alto poder explicativo no es bueno para predecir; pero puede tener alto poder explicativo y no ser bueno en la predicción debido a que el modelo se estudia en un rango de valores delimitado y fuera de ese rango la relación podría ser diferente.

Por ejemplo, para unas dosis determinadas de cierto fertilizante puede existir un modelo lineal que explique adecuadamente la relación entre la cantidad de un nutriente y el crecimiento de una planta; pero si aumentamos indefinidamente la dosis, la planta no crecerá infinitamente (como predeciríamos con el modelo lineal) sino que podría incluso detener su crecimiento, revertirlo y hasta morir.

¿Por qué en las publicaciones se usa el coeficiente de correlación en vez de la covarianza, cuando en los textos de estadística siempre se explica primero la covarianza?

La covarianza es el estadístico que captura realmente la información acerca de la relación lineal entre las variables, pero no está acotada y conserva las unidades de medida de las dos variables, lo que hace difícil su interpretación. Por el contrario, el coeficiente de correlación de Pearson recoge toda la información capturada por la covarianza y, además, es adimensional y está acotado. Todo esto facilita considerablemente la interpretación y es la razón por la que es mucho más utilizado en las publicaciones. Sin embargo, a nivel didáctico no tiene sentido explicar el coeficiente de correlación sin haber explicado la covarianza.

¿Es lo mismo el coeficiente de correlación que el coeficiente de regresión?

No. Ambos se calculan a partir de la covarianza y tienen, por lo tanto, el mismo signo, pero transmiten información muy diferente.

El coeficiente de correlación se calcula como la covarianza dividida por el producto de las desviaciones estándar de las dos variables y se utiliza para decidir si hay relación lineal entre las variables, de qué tipo es (directa o inversa) y qué magnitud tiene.

El coeficiente de regresión se calcula como la covarianza dividida por la varianza de la variable independiente y se interpreta como el incremento (o decremento) dado en la variable dependiente por cada incremento unitario en la variable independiente.

¿Una nube de puntos muy dispersa puede llevar asociado un coeficiente de correlación significativo?

Sí. Un coeficiente de correlación significativo sólo dice que existen indicios suficientes para suponer que la relación dada no es azarosa; pero esto no equivale a afirmar que será posible

encontrar un buen modelo que nos permita estimar la variable dependiente a partir del conocimiento de la variable independiente.

¿Puede ocurrir que al estudiar la relación de una variable X_1 con una variable Y resulte significativa, al estudiar la relación entre X_2 e Y, también sea significativa y al estudiar la relación entre X_1 y X_2 e Y, alguna de las dos variables X aparezca como no significativa?

Lastimosamente, sí. Si X_1 y X_2 están relacionadas entre sí, una de las dos (por ejemplo X_2) aparecería como no significativa ya que no aportaría información diferente de la ya capturada al estudiar la relación entre Y y la variable X_1.

¿Puede ocurrir que, si las variables explicativas X_1 y X_2 están fuertemente relacionadas entre sí, al estudiar el modelo de regresión que relaciona Y con X_1 y con X_2 aparezca un signo para el coeficiente de regresión contrario al real?

Lamentablemente, sí. Cuando las variables explicativas están fuertemente relacionadas se dice que hay colinealidad. En presencia de colinealidad pueden aparecer como no significativas variables que sí están realmente relacionadas con la respuesta; y puede ocurrir que el signo que aparezca en el modelo de regresión sea el contrario al esperado. Es debido a esto la importancia de controlar la colinealidad.