

Examen de la asignatura "Estadística aplicada a las ciencias sociales"

Profesor Josu Mezo.

Respuestas correctas del examen de 3 de septiembre de 2003

Pregunta nº 1 (5 puntos)

A continuación tienes una lista de variables, que proceden de una base de datos de municipios. Indica, para cada una de ellas, si es una variable cualitativa o cuantitativa, y en el segundo caso, si es discreta o continua.

- Provincia a la que pertenece - **Cualitativa**
- Número de habitantes - **Cuantitativa (Estrictamente hablando es discreta; se trata como continua: se aceptan las dos respuestas)**
- Porcentaje de la población activa en paro - **Cuantitativa; continua**
- Partido al que pertenece el alcalde - **Cualitativa**
- Número de partidos representados en el ayuntamiento - **Cuantitativa, discreta**

Pregunta nº 2 (10 puntos)

El salario mensual neto en euros de los 10 empleados de una empresa es el siguiente:

550 550 580 650 700 825 1.050 1.100 1.800 3.300

Calcula:

a) La media

Suma de todos los valores: 11.105

Partido por el número de valores: 10

Resultado: 1.110,5

b) La desviación típica

$$s_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{N}} = \sqrt{\frac{(550-1.110,5)^2 + (550-1.110,5)^2 + \dots + (3300-1.110,5)^2}{10}} = \sqrt{\frac{66449225}{10}} = \sqrt{6644922,5} = 815,23$$

c) El coeficiente de variación

$$CV_x = \frac{s_x}{|\bar{x}|} = \frac{815,23}{1.110,5} = 0,734$$

d) El coeficiente de asimetría

$$CA_x = \frac{\sum_{i=1}^N (x_i - \bar{x})^3}{Ns_x^3} = \frac{(550-1.110,5)^3 + (550-1.110,5)^3 + (580-1.110,5)^3 + \dots + (3300-1.110,5)^3}{10 \cdot 815,16^3}$$
$$CA_x = \frac{10.132.268490}{5.416.701210} = 1,8705$$

e) La mediana

Si se ordenan todos los valores de mayor a menor, el valor intermedio es la mediana. Como tenemos un número par de valores, la mediana es la media de los dos valores intermedios, es decir: $(700+825)/2 = 762,5$

f) La mediana de las desviaciones absolutas

Desviaciones absolutas (valor-mediana) son: 212,5 212,5 182,5 112,5 62,5 62,5 287,5 337,5 1037,5 2537,5

Ordenadas de menor a mayor: 62,5 62,5 112,5 182,5 212,5 212,5 287,5 337,5 1037,5 2537,5

La mediana de las desviaciones absolutas es la media de los dos valores centrales de esa serie. Pero como los dos valores centrales son iguales, la mediana es igual a cualquiera de ellos, es decir 212,5

Pregunta nº 3 (10 puntos)

Comenta los resultados del ejercicio anterior:

- Qué podríamos decir de esa empresa, si tuviéramos sólo los datos calculados en el ejercicio, no los datos primarios
- Qué podemos decir en relación a la diferencia de valores entre la media y la mediana
- ¿Cómo debe interpretarse el coeficiente de asimetría?
- Otros comentarios que consideres oportunos

Con los datos del ejercicio anterior puede decirse que es una empresa con una distribución de salarios bastante desigual. Esto aparece indicado por la desviación típica, que es muy alta: el coeficiente de variación es casi 3/4 de la media, lo que indica mucha dispersión.

También el coeficiente de asimetría de 1,87 nos da a entender que hay una fuerte dispersión, con una “cola hacia la derecha” muy grande, es decir que hay algunos sueldos que se apartan mucho de todos los demás, hacia arriba.

Además la diferencia considerable entre la media y la mediana indican también una distribución asimétrica. Probablemente la mediana sea más representativa de la mayor parte de los valores de la serie que la media.

Pregunta nº 4 (5 puntos)

En la tabla siguiente tienes la distribución de frecuencias absolutas de los turismos vendidos en España en el año 2001, por marcas.

- 1) Elabora la columna con la distribución de frecuencias relativas
- 2) Averigua cuál es la moda

Marca	Coches vendidos (miles)	
Citroën	155,2	0,1079
Peugeot	156,9	0,1091
Renault	180,3	0,1254
Ford	122,7	0,0853
Opel	161,2	0,1121
Seat	158,3	0,1101
Volkswagen	111,3	0,0774
Otros	392	0,2726
Suma	1437,9	1,0000

La moda es “otros”, puesto que es la categoría con una frecuencia relativa mayor (0,27) (Si excluyéramos esa categoría, pues en realidad es una agregación de otras más pequeñas, la moda sería “Renault”, pues es la que tiene la frecuencia relativa entre las verdaderas marcas.)

Pregunta nº 5 (5 puntos)

Los histogramas del gráfico siguiente representan la distribución de la renta per cápita en dos países. ¿Qué podemos decir sobre la distribución de la riqueza en esos países a la vista de los dos histogramas?

Podemos decir varias cosas:

- El país A es más rico, en promedio, puesto que la moda de su renta está en torno a los

26.000 euros, y la del país B está en torno a los 19.000

- Los dos países parecen tener una distribución de la renta relativamente equitativa: tienen una distribución bastante simétrica, con la inmensa mayoría de la población centrada en torno a la moda (que seguramente no será muy distinta de la mediana y de la media)
- No obstante, en los dos países hay unas poquitas personas que tienen rentas muy altas
- Ahora bien, la distribución es más equitativa en A que en B, puesto que la dispersión de los valores alrededor de la moda es menor en A que en B, es decir, casi toda la población está agrupada en valores relativamente cercanos al valor de la moda; mientras que en el país B la dispersión es mayor, es decir, hay más diferencias entre la población

Pregunta nº 6 (5 puntos)

Las operarios de las excavadoras de la empresa “CYDSA” tiene una tasa de accidentes laborales de 6,5 accidentes por 10.000 horas trabajadas, por encima de la media del sector (que es 5,4, con una desviación típica de 1,1). Por su parte, los conductores de camiones de la empresa tienen una tasa de 8,5 accidentes por millón de kilómetros recorridos, también por encima de la media del sector (que es 7,6, con una desviación típica de 1,2). ¿Comparativamente, cuál de las dos secciones de la empresa tiene peor récord de seguridad?

Para comparar las dos tasas es necesario utilizar el valor estandarizado de las tasas de accidentes de los dos grupos de trabajadores.

El valor estandarizado (Tasa-Media/Desv. típica) de los que trabajan con excavadoras es 1,00 y el valor estandarizado es de los que conducen camiones es 0,75.

Por tanto, los que operan excavadoras tienen un peor record de seguridad.

	Excavadoras	Camiones
Tasa	6,5	8,5
Media	5,4	7,6
Desv típica	1,1	1,2
Valor estandarizado	1,00	0,75

Pregunta nº 7 (12 puntos, 4 por cada letra)

La siguiente tabla presenta la distribución conjunta (imaginaria) de frecuencias relativas de dos variables referidas a países del mundo: la variable GOL, que se refiere al número de golpes de estado que han sufrido en los últimos 30 años, y la variable CIV que se refiere al número de guerras civiles que han padecido en ese mismo tiempo.

CIV (número de guerras civiles en 30 años)	GOL (número de golpes de estado en los últimos 30 años)				
	0	1	2	3	4
0	0,38	0,06	0,04	0,00	0,02
1	0,02	0,05	0,09	0,07	0,01
2	0,00	0,02	0,06	0,05	0,03
3	0,01	0,01	0,03	0,03	0,02

Averigua:

a) La distribución marginal de las variables GOL y CIV

CIV (número de guerras civiles en	GOL (número de golpes de estado en los últimos 30 años)					Marginal CIV
	0	1	2	3	4	

30 años)						
	0	0,38	0,06	0,04	0	0,02
	1	0,02	0,05	0,09	0,06	0,01
	2	0	0,02	0,06	0,05	0,03
	3	0,01	0,01	0,03	0,03	0,02
Marginal		0,41	0,14	0,22	0,14	0,08
GOL						

b) El número medio de golpes de estado y de guerras civiles que han tenido los países en esos 30 años

Número medio de golpes de estado: $(0 \cdot 0,41) + (1 \cdot 0,14) + (2 \cdot 0,22) + (3 \cdot 0,14) + (4 \cdot 0,08) = 1,32$

Número medio de guerras civiles: $(0 \cdot 0,5) + (1 \cdot 0,23) + (2 \cdot 0,16) + (3 \cdot 0,1) = 0,85$

c) La distribución de frecuencias relativas de los golpes de estado de los países que no han tenido ninguna guerra civil

	0	1	2	3	4
	$(0,38/0,5) =$	$(0,06/0,5) =$	$(0,04/0,5) =$	0	$(0,02/0,5) =$
	0,76	0,12	0,08		0,04

Pregunta nº 8 (6 puntos)

El gráfico siguiente representa el gráfico de dispersión de dos variables. Cada punto representa uno de los 100 casos de los que se han tomado medidas. El coeficiente de correlación es 0,16. Interpreta la gráfica y el coeficiente de correlación.

La gráfica muestra claramente que existe una relación entre las dos variables, pero es una relación no lineal, es decir, que cuando la variable representada en el eje de abscisas (horizontal) aumenta de valor, la variable representada en el eje de ordenadas, en un primer tramo, disminuye de valor, y en un segundo tramo aumenta de valor.

Las relaciones no lineales no son bien capturadas por medidas como el coeficiente de correlación, que es una relación lineal. Por esa razón, el coeficiente de correlación de 0,16 puede dar lugar a interpretaciones erróneas: el coeficiente parece decir que no hay apenas relación entre las dos variables, cuando lo que sucede es que la relación es no lineal. Si dividiéramos a la variable representada en el eje de abscisas en dos tramos, de 0 a 30 y de 30 al final, y calculáramos de nuevo los coeficientes de correlación con la variable representada en el eje de ordenadas, seguramente veríamos que tenemos dos coeficientes de correlación, uno negativo y otro positivo, muy altos.

Pregunta nº 9 (9 puntos, 3 por cada letra)

La ecuación de la recta de regresión que pone en relación los valores de las variables MOR (Tasa de mortalidad infantil de un país en tanto por mil) y REN (Renta per cápita de ese país, expresada en miles de dólares) es la siguiente: $MOR = 165,4 - (8,23 \cdot REN)$. El coeficiente de correlación es $r = -0,65$.

a) Explica el significado de la ecuación

La ecuación quiere decir que hay una relación lineal negativa entre la renta per cápita y la mortalidad infantil, o sea, que cuanto más alta es la renta per cápita de un país, menor es la mortalidad infantil. En concreto, la recta significa que la mejor predicción de la tasa de mortalidad infantil de un país, a partir del dato de su renta per cápita sería el resultado de restar a 165,4 (en tanto por mil), 8,23 puntos por cada 1.000 dólares de renta per cápita.

b) ¿Cuál es la predicción, según la regresión, de la mortalidad infantil en un país con una renta per cápita de 10.000 dólares?

$$MOR = 165,4 - (8,23 \cdot REN)$$

$$MOR = 165,4 - (8,23 \cdot 10) = 165,4 - 82,3 = 83,1$$

c) ¿Cuál es la reducción del error de predicción que se produce al utilizar la ecuación para predecir un valor de MOR, en lugar de utilizar el valor medio de MOR?

Si tomamos como medida del error de predicción la desviación típica, en el caso de usar el valor medio; y el error típico, o desviación típica residual, en el caso de usar la recta de regresión, la relación entre ambos valores es:

$$\frac{s_r}{s_y} = \sqrt{1-r^2} = \sqrt{1-(-0,65)^2} = \sqrt{1-0,4225} = \sqrt{0,5775} = 0,7595$$

Es decir, que el error típico es 0,759 veces la desviación típica, o dicho de otra manera, que el error de predicción, utilizando la recta de regresión para hacer la predicción es un 24,1% menor que el error de predicción utilizando la media como predicción.

Pregunta nº 10 (15 puntos, 5 por cada letra)

La tabla que sigue representa la renta per cápita en dólares corrientes de 4 países del sur de Europa en 3 años distintos.

Año	España	Grecia	Italia	Portugal
1980	4.340	3.890	5.240	2.160
1990	10.920	6.000	16.850	4.890
2000	14.960	11.960	20.010	11.060

Elabora, una nueva tabla que contenga:

- a) Un número índice para cada país, que exprese la variación de la renta per cápita, tomando como base el año 1980

Año	España	Grecia	Italia	Portugal
1980	100	100	100	100
1990	252	154	322	226
2000	345	307	382	512

- b) Un número índice de la renta per cápita conjunta de los cuatro países, con agregación ponderada (índice de Laspeyres), basándote en que en 1980 la población de los 4 países era: 37,4 millones en España; 9,4 en Grecia; 56,9 en Italia y 9,9 en Portugal.

Cálculo del índice de ponderación de cada país:

Población	37,4	9,4	56,9	9,9	113,6
Renta total	37,4*4.340=	36.566	298.156	21.384	518.422
(millones de dólares)	162.316				
Ponderación	162.316/518.422=	0,31	0,58	0,04	

Índice ponderado:

1980:100

1990: (252*0,31)+(154*0,07)+(322*0,58)+(226*0,04)=283,94

$$2000: (345*0,31)+(307*0,07)+(382*0,58)+(512*0,04)=370,35$$

c) Interpreta los resultados obtenidos en la nueva tabla.

La tabla muestra claramente que el crecimiento relativo de los países ha sido diferente, y además a ritmos muy distintos.

El país cuya renta ha crecido más en el conjunto del periodo ha sido Portugal, que ha multiplicado por más de cinco su renta, y el que menos Grecia, que la ha multiplicado por algo más de 3. España e Italia están en posiciones intermedias, con Italia cerca de 4 veces su renta inicial y España en torno a las 3,5 veces.

Sin embargo, la tabla permite ver también hacia 1990 las cosas eran bastante distintas: el país que más había crecido había sido Italia, con gran diferencia sobre los demás, al multiplicar la renta por más de 3, mientras que los demás países no pasaban del 2,5 (en el caso de España). En ese primer periodo Portugal estaba en un discreto tercer puesto, al multiplicar la renta por 2,3 aproximadamente.

Por tanto, puede decirse que la década de los ochenta fue muy buena para Italia, mientras que en los noventa este país se estancó bastante, y en cambio Portugal dio un gran acelerón.

España ha crecido casi tanto como Italia, pero a un ritmo más regular, casi por igual en los ochenta y los noventa, y Grecia ha crecido el que menos, y también bastante regularmente, parecido en los ochenta y noventa.

Pregunta nº 11 (9 puntos, 3 por cada letra)

La distribución de frecuencias de la variable “número de asignaturas suspendidas al final del curso” entre los alumnos de los institutos de una ciudad es la siguiente.

Número de suspensos	Frecuencia relativa
0	0,44
1	0,21
2	0,14
3	0,09
4	0,06
5	0,04
6	0,02

a) ¿Cuál es la probabilidad de que un alumno sacado al azar tenga dos o más asignaturas suspensas?

La probabilidad es o bien la suma de la probabilidad de que tenga 2, 3, 4, 5 ó 6, o bien, más sencillo 1 menos la probabilidad de que tenga 0 o 1 asignaturas suspensas, es decir: $1-(0,44+0,21)=1-0,65=0,35$

b) Si creamos la variable aleatoria “alumnos con 3 suspensos en muestras de 100 alumnos” ¿cuál será el valor medio de la variable?

El valor medio será el número de elementos de cada muestra por la probabilidad de que tengan 3 suspensos, es decir: $100*0,09=9$

c) Si creamos la variable aleatoria “número medio de suspensos en muestras de 100 alumnos” ¿cuál será el valor medio de esa variable?

El número medio de suspensos en las muestras de 100 alumnos será el mismo que en la población, es decir $(0*0,44)+(1*0,21)+(2*0,14)+(3*0,09)+(4*0,06)+(5*0,04)+(6*0,02)=1,32$

Pregunta nº 12 (9 puntos, 3 por cada letra)

Imagina que una encuesta realizada a una muestra aleatoria de 1.200 personas, entre la población adulta de una ciudad, ha obtenido que la proporción de los que se oponen a la construcción de un gran centro comercial en la ciudad es 0,52.

- a) Calcula el error típico de estimación

El error típico de estimación de una proporción se calcula con la fórmula

$$ETE = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0,52 * 0,48}{1200}} = \sqrt{\frac{0,2496}{1200}} = 0,0144$$

- b) Calcula el intervalo de confianza dentro del cual debe estar la proporción de los que se oponen a la construcción del centro comercial dentro de la población en su conjunto, con un nivel de confianza del 95,5%.

Con el 95,5% de confianza, la proporción en la población se encuentra en el intervalo comprendido entre por la proporción en la muestra menos 2 veces el ETE y la proporción en la muestra más 2 veces el ETE. En este caso, entonces, el valor en la población estará en un intervalo entre $0,52 - (2 * 0,0144)$ y $0,52 + (2 * 0,0144)$, es decir entre 0,4911 y 0,5488.

- c) Con los datos obtenidos de las respuestas anteriores, explica si tiene razón el jefe de la oposición en el ayuntamiento al decir que “la mayoría de los vecinos se opone a la construcción del centro comercial”.

Estrictamente hablando no tiene razón, porque dado el margen de error de la encuesta, se estima que la proporción de la población que está en contra podría ser sólo de 0,4911, es decir, menos de la mitad de los vecinos. Aunque es cierto también que muchos valores dentro del intervalo de confianza están por encima del valor 0,50, no es posible decir con seguridad (ni siquiera al 95%) que la mayoría de los vecinos estén en contra.