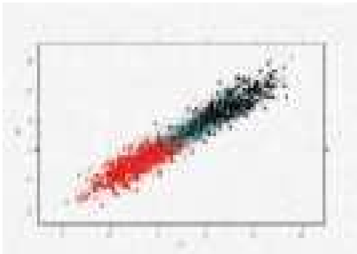
ESTADÍSTICA (SISTEMAS)

Profesores: Hilario Navarro. Jorge Martín



DEPARTAMENTO DE ESTADÍSTICA, INVESTIGACIÓN OPERATIVA Y CÁLCULO NUMÉRICO

Soluciones a los problemas de examen.
Primera prueba presencial. Curso 2003-2004

Problema 1. Primera semana. Partiendo de 10 observaciones del par (X, Y), y con el fin de estudiar la asociación entre dichas variables, se calcularon la recta de regresión de Y sobre X y el coeficiente de correlación muestral, resultando:

$$Y = 25.5 + 3.6X \quad ; \quad r = 0.90,$$

respectivamente. También se sabe que la varianza de X es 16. Se pide:

(a) Calcular la varianza de la variable Y.

(b) ¿Cuánto vale la covarianza entre ambas variables?

Solución

(a) La pendiente (*pend*) de la recta de regresión de Y sobre X se relaciona con la varianza de Y a través de la siguiente igualdad

$$pend = \frac{cov_{x,y}}{v_x} = r\sqrt{\frac{v_y}{v_x}}$$

Con los datos del enunciado resulta

$$v_y = \left(\frac{pend}{r}\right)^2 \times v_x = \left(\frac{3.6}{0.9}\right)^2 \times 16 = 256$$

(b) Para el cálculo de la covarianza se puede utilizar

$$cov_{x,y} = r\sqrt{v_x v_y}$$

ó bien,

$$cov_{x,y} = pend \times v_x$$

De cualquier modo, se obtiene que

$$cov_{x,y} = 57.6$$

□

Problema 2. Primera semana. Con el fin de ejecutar un proceso se selecciona uno de tres periféricos A, B y C. Las probabilidades de escoger cada uno de ellos son: 0.5 para A, 0.3 para B y 0.2 para C. Como resultado de la elección, se pueden producir perturbaciones que detienen la ejecución del proceso. Esto ocurre el 10 % de las veces si el periférico seleccionado fue A, el 20 % si fue B y el 15 % si fue C.

(a) Hallar la probabilidad de que el proceso no se ejecute.

(b) Si el proceso se ha ejecutado, ¿cuál es la probabilidad de que lo haya hecho desde A ó B?

Solución

(a) Vamos a denotar por A, B, C y E los siguientes sucesos:

$$\begin{aligned} A &= \text{“}\textit{El periférico seleccionado es el } A\text{”} \\ B &= \text{“}\textit{El periférico seleccionado es el } B\text{”} \\ C &= \text{“}\textit{El periférico seleccionado es el } C\text{”} \\ E &= \text{“}\textit{El proceso se ejecuta}\text{”} \end{aligned}$$

Las probabilidades de los sucesos A, B y C son las probabilidades de selección de cada uno de los periféricos, que vienen dadas por:

$$P(A) = 0.5 \quad , \quad P(B) = 0.3 \quad , \quad P(C) = 0.2.$$

El proceso no se ejecutará cuando se produzca una perturbación que lo detenga. Puesto que las probabilidades de que ocurran estas perturbaciones en cada uno de los periféricos vienen dadas por 0.1, 0.2 y 0.15, se tendrá que:

$$P(E^c|A) = 0.1 \quad , \quad P(E^c|B) = 0.2 \quad \text{y} \quad P(E^c|C) = 0.15.$$

Consecuentemente, aplicando el teorema de la probabilidad total se obtiene la probabilidad pedida de que el proceso no se ejecute, dada por

$$P(E^c) = P(E^c|A)P(A) + P(E^c|B)P(B) + P(E^c|C)P(C)$$

$$= 0.1 \times 0.5 + 0.2 \times 0.3 + 0.15 \times 0.2 = 0.14.$$

(b) El enunciado nos dice que el proceso se ha ejecutado. Dada esta información, se pide hallar la probabilidad de que la ejecución se hubiera realizado desde el periférico A ó B. La probabilidad pedida es una probabilidad condicionada por el suceso que recoge tal información —el suceso E—. Por tanto, tendremos que calcular $P(A \cup B|E)$. Se sabe que:

$$P(A \cup B|E) = \frac{P((A \cup B) \cap E)}{P(E)}$$

$$= \frac{P(A \cap E) + P(B \cap E)}{P(E)} = \frac{P(E|A)P(A) + P(E|B)P(B)}{P(E)}.$$

La probabilidad del suceso E se obtiene del apartado anterior:

$$P(E) = 1 - P(E^c) = 1 - 0.14 = 0.86.$$

Las probabilidades condicionadas del numerador son:

$$P(E|A) = 1 - P(E^c|A) = 0.9 \quad \text{y} \quad P(E|B) = 1 - P(E^c|B) = 0.8.$$

Por tanto,

$$P(A \cup B|E) = \frac{0.9 \times 0.5 + 0.8 \times 0.3}{0.86} = \frac{0.45 + 0.24}{0.86} = \frac{69}{86} \approx 0.8023256.$$

Otra manera de llegar al mismo resultado sería mediante el cálculo de la probabilidad del complementario $P(C|E)$, la cual viene dada por

$$P(C|E) = \frac{P(C \cap E)}{P(E)} = \frac{P(E|C)P(C)}{P(E)} = \frac{0.85 \times 0.2}{0.86} = \frac{17}{86} \approx 0.1976744,$$

de donde $P(A \cup B|E) = 1 - P(C|E) = 1 - \dfrac{17}{86} = \dfrac{69}{86} \approx 0.8023256.$

□

Problema 3. Primera Semana. Una muestra aleatoria de 200 dígitos proporcionó los datos que se recogen en la siguiente tabla de frecuencias

Dígito	Frecuencia
0	10
1	20
2	19
3	21
4	21
5	15
6	21
7	22
8	25
9	26

Contraste la hipótesis de que la muestra procede de una distribución en la que los 10 dígitos son igualmente probables.

Solución

Estamos ante una situación que el Texto Base identifica como "contraste de la bondad del ajuste (primer caso)". La forma de proceder es calcular el valor del estadístico

$$\chi^2 = \sum_{i=1}^{10} \frac{(O_i - e_i)^2}{e_i}$$

y comparar con el dato que nos proporciona la distribución χ^2_9 (9 grados de libertad) para el nivel de significación seleccionado.

Respecto al cálculo del estadístico, el enunciado nos proporciona la frecuencia observada (O_i) en una muestra aleatoria de tamaño 200, mientras que la frecuencia esperada (e_i) bajo la hipótesis de que los 10 dígitos fuesen igualmente probables es, en cada caso, 20. Entonces,

$$(O_i - e_i)^2 : 100, 0, 1, 1, 1, 25, 1, 4, 25, 36$$

y

$$\sum_{i=1}^{10} \frac{(O_i - e_i)^2}{e_i} = \frac{194}{20} = 9.7$$

Entre los datos auxiliares encontramos $\chi^2_{9;0.05} = 16.92$; entonces, dado que

$$9.7 < 16.92$$

las desviaciones observadas no son significativas, al nivel $\alpha = 0.05$, y por tanto, aceptamos la hipótesis de que *en la población los* 10 *dígitos son equiprobables.*

$\square$

Problema 4. Primera semana. De una población normal se extrae una muestra aleatoria de tamaño 10, obteniéndose las siguientes observaciones:

$$1, 2, 5, 1, 3, 7, 4, 2, 3, 2$$

(a) Obtener una estimación insesgada de la media poblacional.

(b) Contrastar la hipótesis de que la media poblacional es 3.5, frente a la alternativa de que es menor. Tómese un nivel de significación de $\alpha = 0.01$.

Solución

(a) Se sabe que la media muestral $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ es un estimador insesgado de la media poblacional μ, ya que $E\{\bar{X}\} = \frac{1}{n}\sum_{i=1}^{n} E\{X_i\} = \mu$.

Para la muestra extraída la estimación viene dada por

$$\bar{x} = \frac{1+2+5+1+3+7+4+2+3+2}{10} = 3.$$

(b) La muestra obtenida ha proporcionado un valor para la media muestral de $\bar{x} = 3$. Se trata de comprobar si ese valor proporciona evidencia estadística en favor de la hipótesis de que la media poblacional es $\mu = 3.5$ o por el contrario apoya la alternativa de que es inferior.

El problema se plantea en términos de un contraste de hipótesis en el que

$$H_0 : \mu = 3.5 \qquad\qquad H_1 : \mu < 3.5.$$

Dado que la variable aleatoria poblacional sigue una distribución normal de media y varianza desconocidas, el contraste se lleva a cabo utilizando la cantidad pivotal

$$\frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1},$$

donde

- μ_0: la media poblacional bajo H_0. En este caso $\mu_0 = 3.5$.
- $\bar{X}$: media muestral.
- S^2: cuasivarianza muestral.
- n: tamaño muestral.

La región crítica del test —conjunto de todas las muestras para las que se rechaza H_0— viene dada por

$$\mathcal{R} = \left\{(X_1, X_2, \ldots, X_n) : \frac{\bar{X} - \mu_0}{S/\sqrt{n}} < c\right\}$$

donde el punto crítico c se determina con la condición del nivel de significación, la cual viene dada por:

$$P_{H_0}((X_1, X_2, \ldots, X_n) \in \mathcal{R}) = P\left(\frac{\bar{X} - \mu_0}{S/\sqrt{n}} < c\right) = P(t_{n-1} < c) = 0.01.$$

En este caso, para un tamaño muestral de $n = 10$, el punto crítico vale $c = -2.821$, ya que para una distribución t con 9 grados de libertad se tiene que $P(t_9 < -2.821) = 0.01$ (véase la figura 1, en la que el área de la región sombreada coincide con el nivel de significación $\alpha = 0.01$). Por tanto, la región crítica $\mathcal{R}$ está definida por $\mathcal{R} = \left\{\frac{\bar{X} - \mu_0}{S/\sqrt{n}} < -2.821\right\}$.

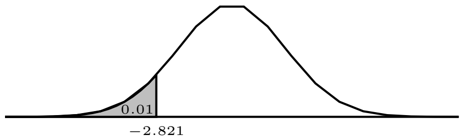


Figura 1: Punto crítico para una t_9 ($\alpha = 0.01$)

Para la muestra extraída, se sabe que $\bar{x} = 3$ y que

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{(1-3)^2 + (2-3)^2 + \cdots + (2-3)^2}{9} = \frac{32}{9},$$

de donde se sigue que

$$\frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{3 - 3.5}{\sqrt{32/90}} = -1.5 \times \sqrt{\frac{10}{32}} \notin \mathcal{R}.$$

Consecuentemente, la muestra extraída no proporciona evidencia estadística como para rechazar H_0 y mantendremos la afirmación de que la media de la población es 3.5.

□

Problema 1. Segunda Semana. Con el fin de describir un conjunto de 10 observaciones del par de variables (X, Y), se calcularon la recta de regresión de Y sobre X y la correspondiente varianza residual, resultando:

$$Y = 23.5 - 10.2X \quad ; \quad Varianza\ residual = 9,$$

respectivamente. Se pide:

(a) ¿Cuál es el signo de la covarianza? (Justifique la respuesta sin calcular dicho dato)

(b) Si la varianza residual representa el 19 % de la varianza de Y, ¿cuánto vale el coeficiente de correlación muestral?

Solución

(a) La relación entre la covarianza y la pendiente (*pend*) de la recta de regresión de Y sobre X se pone de manifiesto en la siguiente igualdad

$$cov_{x,y} = pend \times v_x$$

Como la varianza es siempre positiva, covarianza y pendiente tendrán siempre el mismo signo. En este caso la pendiente es -10.2; por tanto, la covarianza es negativa.

(b) El enunciado nos adelanta que

$$\frac{Varianza\ residual}{v_y} = 0.19$$

Entonces, dado que

$$\frac{Varianza\ residual}{v_y} = 1 - r^2$$

resulta

$$r^2 = 0.81$$

y, como consecuencia

$$r = -0.9$$

<u>Nota</u>: Observe que el signo del coeficiente de correlación es el mismo que el de la covarianza.

□

Problema 2. Segunda semana. Se lanza un dado, y a continuación, tantas monedas como puntuación se obtuvo en el lanzamiento del dado.

(a) Hallar la probabilidad de que se obtenga alguna cara.

(b) Si se ha obtenido alguna cara, ¿cuál es la probabilidad de que el resultado del dado fuera un uno?

Solución

(a) Denotaremos por X la variable aleatoria que da la puntuación del dado y por A el suceso “ *Obtener alguna cara* ”. Vamos a razonar por cálculo de la probabilidad del complementario de A.

Las probabilidades a priori de las distintas puntuaciones del dado son

$$P(X = i) = \frac{1}{6} \quad \text{con} \quad i = 1, 2, 3, 4, 5, 6.$$

Por otro lado, si en el lanzamiento del dado se obtuvo la puntuación i, el suceso complementario de A ocurrirá cuando se obtengan i cruces al lanzar la moneda; con lo cual

$$P(A^c|X = i) = \left(\frac{1}{2}\right)^i \quad : \quad i = 1, 2, 3, 4, 5, 6.$$

Del teorema de la probabilidad total se sigue la probabilidad del suceso complementario de A:

$$P(A^c) = \sum_{i=1}^{6} P(A^c|X = i)P(X = i) = \frac{1}{6}\sum_{i=1}^{6}\left(\frac{1}{2}\right)^i$$

$$= \frac{1}{6}\frac{(1/2)^6(1/2) - (1/2)}{(1/2) - 1} = \frac{1}{6}\left(\frac{2^6 - 1}{2^6}\right) = \frac{21}{128},$$

de donde se obtendrá que $P(A) = 1 - P(A^c) = 1 - \dfrac{21}{128} = \dfrac{107}{128}$.

(b) La información dada sobre el resultado del experimento aleatorio: “ *se ha obtenido alguna cara* ” modifica las probabilidades a priori de las distintas puntuaciones del dado. Nos están pidiendo recalcular la probabilidad de que $X = 1$, conocida dicha información, es decir, hallar la probabilidad a posteriori $P(X = 1|A)$.

Este es un ejemplo típico de aplicación de la regla de Bayes para el que

$$P(X = 1|A) = \frac{P(\{X = 1\} \cap A)}{P(A)} = \frac{P(A|X = 1)P(X = 1)}{P(A)}$$

$$= \frac{[1 - P(A^c|X = 1)]P(X = 1)}{P(A)} = \frac{\left(1 - \frac{1}{2}\right)\frac{1}{6}}{\frac{107}{128}} = \frac{32}{321}.$$

□

Problema 3. Segunda semana. El análisis de la varianza de una variable respuesta Y, medida bajo una serie de condiciones, proporcionó los datos que se recogen en la siguiente tabla

Fuente de variación	SC	GL	CM	F
Entre grupos	348	3	—	—
Dentro de los grupos	—	8	—	

(SC: Suma de cuadrados; GL: Grados de libertad; CM: Cuadrado medio; F: Estadístico)

Sabiendo que la varianza total es 428, complete las posiciones que faltan en la tabla (marcadas con —) y obtenga conclusiones sobre el comportamiento de Y en los grupos considerados.

Solución

Como la varianza total es 428 y la $SC(\text{“}Entre\text{”}) = 348$, resulta

$$SC(\text{“}Dentro\text{”}) = 428 - 348 = 80$$

Además,

$$\begin{aligned} CM(\text{“}Entre\text{”}) &= 348/3 = 116 \\ CM(\text{“}Dentro\text{”}) &= 80/8 = 10 \\ F(\text{“}Estadístico\text{”}) &= 116/10 = 11.6 \end{aligned}$$

Una vez completada la tabla, observamos que el valor obtenido para el estadístico F satisface la desigualdad

$$F_{3,8;0.05} = 9.28 < 11.6$$

Este hecho nos conduce al rechazo, con un nivel de significación del 5 %, de la hipótesis nula, que establece que *la respuesta media es la misma en los* 4 *grupos considerados.*

□

Problema 4. Segunda Semana. De una población con distribución normal se extrae una muestra aleatoria de tamaño 11, de la cual se obtiene una varianza muestral de 5.

(a) Obtener una estimación insesgada de la varianza poblacional.

(b) Contrastar la hipótesis de que la varianza poblacional es 4.5, frente a la alternativa de que es mayor. Tómese un nivel de significación de $\alpha = 0.05$.

Solución

(a) Se sabe que un estimador insesgado de la varianza poblacional es la cuasivarianza muestral definida mediante

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n}{n-1}\, v_x^2,$$

donde v_x^2 es la varianza muestral —cuyo valor es proporcionado por el enunciado—.

Por tanto, la estimación insesgada vendrá dada por $S^2 = \dfrac{11}{10} \times 5 = 5.5.$

(b) El enunciado pide realizar un contraste de hipótesis para la varianza de la población σ^2, con una hipótesis nula que postula que la varianza es 4.5 frente una alternativa que afirma que es mayor:

$$H_0 : \sigma^2 = 4.5 \qquad\qquad H_1 : \sigma^2 > 4.5.$$

El objetivo es comprobar si la evidencia proporcionada por la muestra, descrita en términos de la cuasivarianza muestral, corrobora la hipótesis nula o, por el contrario, la desacredita. Para ello, utilizaremos la cantidad pivotal que corresponde a la situación planteada.

Dado que la distribución poblacional es normal, llevaremos a cabo el contraste empleando el estadístico:

$$\frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1},$$

donde

- σ_0^2: varianza poblacional bajo H_0. En este caso $\sigma_0^2 = 4.5$.
- S^2: cuasivarianza muestral. Para la muestra extraída se obtuvo en el apartado anterior que $S^2 = 5.5$.
- n: tamaño muestral. En este caso $n = 11$.

La región crítica del contraste —conjunto de muestras para las que se rechaza H_0— viene dada por

$$\mathcal{R} = \left\{ (X_1, X_2, \ldots, X_n) : \frac{(n-1)S^2}{\sigma_0^2} > c \right\},$$

donde el punto crítico c se determina con la condición del nivel:

$$P_{H_0}((X_1, X_2, \ldots, X_n) \in \mathcal{R}) = P\left(\frac{(n-1)S^2}{\sigma_0^2} > c \right) = P(\chi^2_{n-1} > c) = 0.05.$$

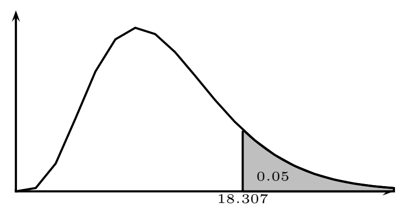


Figura 2: Punto crítico para una χ^2_{10} ($\alpha = 0.05$)

En este caso, para un tamaño muestral de $n = 11$, el punto crítico vale $c = 18.307$, ya que para una distribución χ^2 con 10 grados de libertad se cumple que $P(\chi^2_{10} > 18.307) = 0.05$ —véase la región sombreada en la figura 2—.

Para la muestra extraída, el valor de la cuasivarianza se sabe que es $S^2 = 5.5$; con lo cual $\dfrac{(n-1)S^2}{\sigma_0^2} = \dfrac{10 \times 5.5}{4.5} < 18.307$, es decir, la muestra no pertenece a la región crítica. Por tanto, la decisión que adoptaremos será aceptar, con un nivel de significación de $\alpha = 0.05$, la hipótesis nula de que el valor de la varianza poblacional es 4.5.

□