2º final 2016-2, problema 3: $f_X(2)$ aparece como 0.066; el valor verificado es $\frac{1}{15} = \mathbf{0.0667}$.

C.4 · Unidad 4 · Modelos probabilísticos comunes: binomial, geométrica, Pascal, hipergeométrica, Poisson, uniforme, exponencial y normal

1er final 2011-1, problema 3 b): la variable se define como «número de pozos **no** productivos» con Binomial($n = 10,\ p = 0.2$), pero $p = 0.2$ es la probabilidad de un pozo **productivo**. Lo coherente es contar los pozos productivos y pedir $P(U = 0) = 0.8^{10}$. El valor numérico, **0.1074**, es correcto.

2º final 2015-1, problema 4 b): la resolución calcula $P(5 < T < 25) = 0.522$, que no es condicional. Lo pedido es $P(T < 25 \mid T > 5) = P(T < 20) = \mathbf{0.7135}$.

2º final 2016-1, problema 4: en la aproximación normal se estandariza restando 39 en lugar de $\mu = 35$: $P(-0.77 \le Z \le 0.15) = 0.3390$. Con $\mu = 35$, $P(0.46 \le Z \le 1.39) = \mathbf{0.2405}$, coherente con el valor exacto **0.2490** (la resolución da 0.2489 porque suma términos redondeados).

2º final 2011-1, problema 4: la corrección por continuidad no corresponde a las desigualdades del enunciado.

- a) «Menor que 32» es $X \le 31$ y se aproxima con $P(Y \le 31.5) = \mathbf{0.0778}$; la resolución calcula $P(Y \le 32.5) = 0.1056$.
- b) «Mayor que 49» es $X \ge 50$ y se aproxima con $P(Y \ge 49.5) = \mathbf{0.0571}$; la resolución calcula $P(Y \ge 48.5) = 0.0778$.
- c) «Por lo menos 35 pero menor que 47» es $35 \le X \le 46$ y se aproxima con $P(34.5 \le Y \le 46.5) = \mathbf{0.6811}$; la resolución usa 47.5 como extremo y obtiene 0.713.

1er final 2015-1, problema 6: la resolución aproxima la binomial con la normal sin corrección por continuidad: $P(Z \ge 2.07) = 0.0192$. Con la corrección, $P(Y \ge 27.5) = \mathbf{0.0262}$; el valor exacto es 0.0295.

C.5 · Unidad 5 · Variables aleatorias conjuntas: marginales, condicionales, independencia, covariancia y correlación

1er final 2018-1, problema 4 a):

- la marginal aparece como «$f(x = 0) = 3$»; es $f_X(0) = \frac{3}{16}$;
- el texto anuncia la celda «$x = 1,\ y = 1$», pero los valores usados son los de la celda $(0, 1)$. La conclusión (no son independientes) es correcta.

2º final 2013-1, problema 6 c): la probabilidad se escribe $P(X \le 2.5 \mid Y = 2)$, pero la integral y el resultado corresponden a 15 minutos de espera: lo correcto es $P(X \le 2.25 \mid Y = 2) = 0.25$.

1er final 2013-1, problema 5 b): la marginal $f_Y(y) = \frac{1}{6640}(80y + 3240)$ aparece con el dominio «$0 \le x \le 2$»; es $0 \le y \le 2$.

3. $n = 36 \ge 30$: por el TLC, $\bar{X} \sim N\left(1.5,\ \frac{1.25}{36}\right)$, con $\sigma_{\bar{X}} = \sqrt{0.034722} = 0.1863$.
4. $z_a = \frac{1.45-1.5}{0.1863} = -0.2683 \to -0.27$; $z_b = \frac{1.75-1.5}{0.1863} = 1.3416 \to 1.34$.
5. $P(1.45 < \bar{X} < 1.75) = F_Z(1.34) - F_Z(-0.27) = 0.9099 - 0.3936 = \mathbf{0.5163}$.

Ejercicio 6.5.7 · Variancia muestral: ji cuadrada

Adaptado de: 1er final 2017-1, problema 6.

Las edades de los habitantes de una ciudad tienen distribución normal con variancia $\sigma^2 = 6$. Se toma una muestra aleatoria de 25 observaciones. Encuentre la probabilidad de que la muestra tenga una variancia s^2 mayor que 9.1.

Resolución.

1. La población es normal: $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(24)}$, con $\nu = 25 - 1 = 24$.
2. $\chi^2 = \frac{(24)(9.1)}{6} = \frac{218.4}{6} = 36.4$.
3. $P(S^2 > 9.1) = P(\chi^2_{(24)} > 36.4)$.
4. En la tabla, fila $\nu = 24$: $\chi^2_{0.05,24} = 36.415 \approx 36.4$, así que el área a la derecha es **0.05** (valor exacto: 0.0502).
5. **Resultado:** la probabilidad de que la variancia muestral pase de 9.1 es aproximadamente **0.05**.

6.6 Ejercicios propuestos

Resuélvelos sin ver la solución. Las resoluciones completas están en el **Apéndice E, sección E.6**.

Ejercicio propuesto 6.6.1

Adaptado de: 1er final 2011-1, problema 4.

Una fábrica de focos para equipos de refrigeración afirma que tienen una vida útil de 60 meses con desviación estándar de 6 meses. Un supervisor prueba una muestra de 64 focos. ¿Cuál es la probabilidad de encontrar una muestra con vida útil promedio menor que 58 meses?

Resultado: $P(Z < -2.67) = \mathbf{0.0038}$.

→ Resolución en el Apéndice E, p. 105.

Ejercicio propuesto 6.6.2

Adaptado de: 1er final 2013-1, problema 6.

Un soldador debe colocar un cordón de soldadura a 40 piezas. El tiempo por pieza tiene media 5 minutos y desviación estándar 3 minutos. Empieza a las 6:00 p.m. y, para terminar antes de las 9:00 p.m., el tiempo medio por cordón debe ser como máximo 4.5 minutos. ¿Cuál es la probabilidad de que termine a tiempo?

Resultado: $P(\bar{X} < 4.5) = P(Z < -1.05) = \mathbf{0.1469}$.

→ Resolución en el Apéndice E, p. 106.

Examen	Problemas (unidad: tema)
1er final 2016-2	U1: datos sin agrupar · U2: equipos (Bayes) · U3: variable discreta; distribución exponencial · U5: tabla conjunta · U6: media muestral · U7: predicción
2º final 2016-2	U1: *cv* · U2: tabla de contingencia · U3: función de distribución discreta · U4: Pascal · U5: densidad uniforme en un rectángulo · U6: media muestral · U7: ¿relación lineal?
1er final 2017-1	U1: media, mediana, moda y sesgo · U2: directivos (Bayes) · U3: densidad y F_X por tramos · U5: fusibles (tabla conjunta) · U4: binomial y Pascal · U6: ji cuadrada · U7: r e interpretación
2º final 2017-1	U1: ojiva y media · U2: alarma (Bayes) · U3: impresoras vendidas en dos días · U5: densidad conjunta · U4: binomial · U6: media muestral · U7: r^2
1er final 2018-1	U1: completar la tabla · U2: opción múltiple al azar (Bayes) · U3: densidad triangular · U5: tabla conjunta · U4: uniforme · U6: media muestral · U7: ¿relación lineal?
2º final 2018-1	U1: histograma, moda y media · U2: máquinas (Bayes) · U5: densidad conjunta · U4: Poisson; normal · U6: media muestral · U7: regresión

(Esqueleto: los temas de cada problema se revisan contra los exámenes al reescribir cada unidad.)

6.7 Errores frecuentes

1. **Usar σ en lugar de $\sigma/\sqrt{n}$.** El promedio usa el error estándar. Con σ sola se resuelve el problema de **un** individuo (Unidad 4).
2. **Confundir σ^2 con σ.** $N(100, 15^2)$ da $\sigma = 15$; $N(\mu = 106, \sigma = 12)$ da $\sigma = 12$; «variancia 225» da $\sigma = 15$.
3. **No justificar la normalidad de $\bar{X}$.** Escribir «población normal» o «$n \geq 30$, por el TLC» vale puntos.
4. **Aplicar el TLC con $n < 30$ y distribución desconocida.** En ese caso la probabilidad no puede calcularse con la información dada.
5. **Equivocar el lado de la tabla.** La normal da el área a la izquierda: «mayor que» es $1 - F_Z(z)$. La ji cuadrada y la t dan el área a la **derecha**.
6. **Sumar en lugar de restar** en «entre a y b»: es $F_Z(z_b) - F_Z(z_a)$, el mayor menos el menor.
7. **Tomar $F_Z(-z)$ como $F_Z(z)$.** $F_Z(-1.37) = 0.0853$, no 0.9147. Si $z < 0$, el área izquierda es menor que 0.5.
8. **Redondear z a un decimal.** Se redondea a dos: 2.1082 es 2.11, no 2.1.
9. **Olvidar el factor de corrección** cuando el muestreo es sin reemplazo de una población finita conocida.
10. **Ji cuadrada con n grados de libertad.** Son $\nu = n - 1$. Y si se dan desviaciones s, hay que elevarlas al cuadrado antes de transformar.
11. **Resultado imposible.** Una probabilidad negativa, mayor que 1, o $P(\bar{X} > c) > 0.5$ con $c > \mu$ indican un «1 −» de más o de menos.

6.8 Formulario de emergencia de la unidad

- $\mu_{\bar{X}} = \mu \cdot \sigma^2_{\bar{X}} = \frac{\sigma^2}{n} \cdot \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \cdot$ sin reemplazo: $\sigma^2_{\bar{X}} = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$
- Población normal: $\bar{X}$ normal con cualquier n · desconocida y $n \geq 30$: TLC · desconocida y $n < 30$: no se puede
- $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \cdot$ suma: $\sum X_i \sim N(n\mu, n\sigma^2)$
- $P(\bar{X} < c) = F_Z(z) \cdot P(\bar{X} > c) = 1 - F_Z(z) \cdot P(a < \bar{X} < b) = F_Z(z_b) - F_Z(z_a) \cdot F_Z(-z) = 1 - F_Z(z)$
- M muestras: $M \cdot p \cdot$ muestras independientes: $P(A)P(B)$
- $\chi^2 = \frac{(n-1)S^2}{\sigma^2}$, $\nu = n - 1$, tabla con área a la derecha
- $T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$, $\nu = n - 1$, tabla con área a la derecha

6.9 Procedimiento recomendado en el examen

1. La definición de X y de $\bar{X}$ con palabras, y los datos μ, σ (con raíz si se dio σ^2) y n.
2. La justificación: «población normal» o «$n \geq 30$, por el teorema del límite central».
3. El modelo $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ con números y el error estándar con 4 decimales.
4. La estandarización con z a dos decimales y un dibujo de la campana con el área sombreada.
5. La lectura de tabla, el resultado con 4 decimales y una frase de conclusión.

A.7 Manejo de datos bivariados (Unidad 7)

Recta estimada: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

- $SS_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$ · $SS_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$ · $SS_{xy} = \sum xy - \frac{\sum x \sum y}{n}$.
- $\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$ · $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ · la recta pasa por $(\bar{x}, \bar{y})$.
- $r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}}$ · $r^2 = \frac{SS_{xy}^2}{SS_{xx} SS_{yy}} = 1 - \frac{SEC}{SS_{yy}}$ (proporción de la variación de y explicada por la recta) · $SEC = \sum y^2 - \hat{\beta}_0 \sum y - \hat{\beta}_1 \sum xy$.
- Cov $= SS_{xy}/n$ · $\hat{\sigma}^2 = \frac{SEC}{n-2}$ · prueba de $H_0 : \beta_1 = 0$: $T_0 = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2 / SS_{xx}}}$; se rechaza si $|t_0| > t_{\alpha/2, n-2}$.
- Interpretación: $r > 0$ relación directa; $r < 0$ inversa; $|r|$ cercano a 1, fuerte; $r \approx 0$, sin relación lineal.

A.8 Recordatorios para el examen

Define cada evento o variable con palabras («Sea X el número de…»), escribe el modelo ($X \sim$ …), justifica el uso de cada teorema (límite central, independencia), arrastra 4 decimales, encierra la respuesta y escribe una frase de interpretación.

Unidad 7 · Manejo de datos bivariados: recta de mínimos cuadrados, correlación y determinación

↑ Volver al índice

Objetivo (programa de la asignatura). El alumno analizará la relación que existe entre dos variables a partir de la información obtenida por el ajuste de regresión y sus coeficientes de correlación.

Instrucciones de emergencia

- **Frecuencia:** aparece en los 18 finales de 2011-1 a 2018-1, casi siempre como último problema. Vale de 10 a 20 puntos. En 2016-1 vino como opción múltiple.
- **Tiempo:** de 2 a 3 horas para esta unidad.
- **Ruta corta:**
  1. Lee la sección 7.1 (15 min).
  2. Sigue el ejemplo 7.4 tapando la resolución.
  3. Resuelve los ejercicios 7.5 con cronómetro (15 min cada uno).

7.1 Conceptos y expresiones

Datos bivariados. Se observan n parejas (x_i, y_i) de dos variables medidas en el mismo individuo. Interesan dos preguntas: ¿se relacionan en línea recta? y, si es así, ¿qué recta las describe para **estimar** y a partir de x?

- x es la variable **independiente** o explicativa: la que se conoce, con la que se predice («en función de…», «influye en…»).
- y es la variable **dependiente** o de respuesta: la que se quiere estimar. «Peso en función de la estatura» quiere decir x = estatura, y = peso.

Diagrama de dispersión. Un punto (x_i, y_i) por pareja, con x en el eje horizontal. Los puntos **no se unen**. Su forma anticipa el signo y la fuerza de la relación.

Sumas de cuadrados. Con la tabla de columnas x, y, x^2, y^2 y xy:

$$SS_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}, \qquad SS_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}, \qquad SS_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n}.$$

- SS_{xx} y SS_{yy} son siempre positivas; SS_{xy} puede ser negativa.
- $SS_{xx} = \sum (x_i - \bar{x})^2$ y $SS_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$: son las mismas sumas escritas en forma de desviaciones.

Recta de mínimos cuadrados. El modelo es $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. La recta de mínimos cuadrados es la que hace mínima la suma de los cuadrados de los errores verticales $\sum (y_i - \hat{y}_i)^2$. Sus coeficientes son

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \text{ (pendiente)}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \text{ (ordenada al origen)}.$$

- Forma equivalente que usan algunas resoluciones: $\hat{\beta}_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$. Si una resolución escribe $y = a + bx$, a es la ordenada y b la pendiente.

platicúrtica.
- Sesgo positivo: $x_{mo} < \tilde{x} < \bar{x}$; negativo: $\bar{x} < \tilde{x} < x_{mo}$. Relación empírica $\bar{x} - x_{mo} \approx 3(\bar{x} - \tilde{x})$. El coeficiente de Pearson $3(\bar{x} - \tilde{x})/s$ solo confirma el signo; el coeficiente de sesgo es a_3.
- Ojiva: (frontera superior, F_i o F_i^*), desde (primera frontera, 0). Histograma: barras pegadas sobre las fronteras.
- Datos sin agrupar: $\bar{x} = \frac{\sum x_i}{n}$; mediana en la posición $\frac{n+1}{2}$ (n impar) o promedio de los dos centrales (n par); $s^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1}$.

A.2 Probabilidad (Unidad 2)

- $P(\bar{A}) = 1 - P(A) \cdot P(A \cup B) = P(A) + P(B) - P(A \cap B) \cdot P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C) \cdot$ De Morgan: $\overline{A \cup B} = \bar{A} \cap \bar{B}$.
- $P(A \mid B) = \frac{P(A \cap B)}{P(B)} \cdot P(A \cap B) = P(B)P(A \mid B) = P(A)P(B \mid A) \cdot$ independencia: $P(A \cap B) = P(A)P(B)$.
- Probabilidad total: $P(A) = \sum_i P(B_i)P(A \mid B_i) \cdot$ **Bayes:** $P(B_r \mid A) = \frac{P(B_r)P(A|B_r)}{\sum_i P(B_i)P(A|B_i)} \cdot P(\bar{A} \mid B) = 1 - P(A \mid B)$.
- Regla de la multiplicación: $P(A \cap B \cap C) = P(A)P(B \mid A)P(C \mid A \cap B)$.
- Sistemas con componentes independientes: **serie** $\prod p_i$; **paralelo** $1 - \prod(1 - p_i)$.
- Conteo: $P_r^n = \frac{n!}{(n-r)!} \cdot \binom{n}{r} = \frac{n!}{r!(n-r)!} \cdot$ con repetición $n^r \cdot$ objetos idénticos $\frac{n!}{n_1!\cdots n_k!} \cdot$ arreglos circulares $(n-1)!$.

A.3 Variable aleatoria (Unidad 3)

f_X: función de probabilidad o de densidad; F_X: función de distribución.
- $\sum f_X(x) = 1$ o $\int f_X = 1 \cdot F_X(x) = P(X \le x) = \sum_{t \le x} f_X(t)$ o $\int_{-\infty}^{x} f_X(t)\,dt$ (por tramos: 0, fórmula, 1).
- Continua: $f_X = F_X'$; $P(a \le X \le b) = F_X(b) - F_X(a)$; $P(X = a) = 0$. Discreta: $f_X(x_i) =$ salto de F_X en x_i; $P(a \le X \le b) = F_X(b) - F_X(a) + f_X(a)$.
- $\mu_X = E(X) = \sum x f_X$ o $\int x f_X \cdot E[g(X)] = \sum g(x) f_X$ o $\int g(x) f_X \cdot \sigma_X^2 = E(X^2) - \mu_X^2 \cdot E(aX + b) = a\mu + b \cdot \mathrm{Var}(aX + b) = a^2\sigma^2$.
- $CV_X = \sigma_X/\mu_X \cdot \alpha_3 = \mu_3/\sigma^3$, $\alpha_4 = \mu_4/\sigma^4$, con $\mu_r = E[(X - \mu)^r] \cdot$ mediana: $F_X(\tilde{x}) = 0.5 \cdot$ moda: máximo de f_X.
- Chebyshev: $P(\mu - k\sigma < X < \mu + k\sigma) \ge 1 - \frac{1}{k^2}$.

A.4 Modelos probabilísticos (Unidad 4)

Modelo	$f_X(x)$	μ	σ^2
Binomial(n, p)	$\binom{n}{x} p^x q^{n-x}$, $x = 0, \ldots, n$	np	npq
Hipergeométrica(r, n, N)	$\binom{r}{x}\binom{N-r}{n-x}/\binom{N}{n}$	$n\frac{r}{N}$	$n\frac{r}{N}\left(1 - \frac{r}{N}\right)\frac{N-n}{N-1}$
Poisson(λ)	$\frac{\lambda^x e^{-\lambda}}{x!}$; en un intervalo t, $\lambda \to \lambda t$	λ	λ
Geométrica(p)	$q^{x-1}p$, $x = 1, 2, \ldots$	$1/p$	q/p^2
Pascal(r, p)	$\binom{x-1}{r-1} p^r q^{x-r}$, $x = r, r+1, \ldots$	r/p	rq/p^2

Si el enunciado dice…	Se pide…	Procedimiento
«coeficiente de determinación», «¿qué porcentaje explica el modelo?», «¿es bueno el modelo?»	r^2	7.3.3
«si x vale…, ¿cuánto vale y?», «estime», «pronóstico»	$\hat{y}(x_0)$ y aviso de extrapolación	7.3.4
«diagrama de dispersión», «trace la recta»	gráfica	7.3.5
solo se dan las sumas $\sum x$, $\sum y$, $\sum x^2$, $\sum y^2$, $\sum xy$	ir directo a las SS	7.3.1
se dan r o r^2 y piden interpretar	interpretación sin cuentas	7.3.3
opción múltiple sobre r, r^2, covariancia o la recta	reglas rápidas	7.3.6

7.3 Procedimientos

7.3.1 Tabla y sumas de cuadrados

1. **Define** las variables con unidades: «X: [variable independiente]», «Y: [variable que se estima]». Cuenta n, el número de **parejas**.
2. Haz la tabla con las columnas x, y, x^2, y^2 y xy y un renglón de sumas. Si el enunciado da las sumas, cópialas identificando cuál es cuál.
3. Calcula SS_{xx}, SS_{yy} y SS_{xy}, cada una en su renglón y sustituida.
4. Comprueba que $SS_{xx} > 0$ y $SS_{yy} > 0$. Si sale negativa, hay un error de cuenta (casi siempre $(\sum x)^2$ confundido con $\sum x^2$).

7.3.2 Recta de mínimos cuadrados

1. $\bar{x} = \frac{\sum x}{n}$ y $\bar{y} = \frac{\sum y}{n}$.
2. $\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$, **sin redondear** (6 cifras significativas).
3. $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$.
4. Escribe la recta con números: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.
5. Comprueba que $\hat{y}(\bar{x}) = \bar{y}$.

7.3.3 Correlación y determinación

1. $r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$.
2. $r^2 = \frac{SS_{xy}^2}{SS_{xx}SS_{yy}}$ (mejor así que elevando un r ya redondeado).
3. Interpreta con una frase que tenga los tres elementos:
   - **signo:** directa o inversa, con el significado en el problema;
   - **fuerza:** fuerte, moderada o débil (tabla de la sección 7.1);
   - **porcentaje:** «el modelo explica el __ % de la variación de y».

7.3.4 Predicción

1. Verifica que x_0 esté dentro del intervalo de los x observados.

- Cov $= \dfrac{SS_{xy}}{n} = \dfrac{\sum xy}{n} - \bar{x}\bar{y}$
- $|r| \geq 0.8$ fuerte · 0.5 a 0.8 moderada · < 0.5 débil · «explica el $100r^2$ %»
- Predicción solo dentro del intervalo de los x observados

7.9 Procedimiento recomendado en el examen

1. La definición de X y Y con unidades, y n.
2. La tabla de cinco columnas con sumas, o las sumas dadas identificadas.
3. Las tres SS, cada una en su renglón, con la comprobación $SS_{xx}, SS_{yy} > 0$.
4. $\hat{\beta}_1$, $\hat{\beta}_0$ y la recta; la comprobación $\hat{y}(\bar{x}) = \bar{y}$.
5. r y r^2 según lo que se pida, y una frase con signo, fuerza y porcentaje explicado.
6. Si hay predicción: $\hat{y}(x_0)$ y, si corresponde, el aviso de extrapolación.

Si no se encuentra el camino, la tabla con sus sumas y las tres SS sustituidas ya valen buena parte del problema.

- a) Trace el diagrama de dispersión.
- b) Calcule el coeficiente de correlación lineal.
- c) Obtenga la recta de regresión y trácela junto con el diagrama.
- d) Calcule el coeficiente de determinación e interprételo.

Datos. X: número del mes; Y: número de pacientes programados; $n = 5$.
Planteamiento. Tabla de sumas, las tres SS, luego r, la recta y r^2.

x	y	x^2	y^2	xy
3	100	9	10 000	300
4	150	16	22 500	600
5	200	25	40 000	1000
6	210	36	44 100	1260
7	220	49	48 400	1540
25	**880**	**135**	**165 000**	**4700**

Resolución.

1. a) Cinco puntos que suben rápido y luego se aplanan: relación directa, cercana a una recta.
2. $SS_{xx} = 135 - \dfrac{25^2}{5} = 135 - 125 = 10$.
3. $SS_{yy} = 165\,000 - \dfrac{880^2}{5} = 165\,000 - 154\,880 = 10\,120$.
4. $SS_{xy} = 4700 - \dfrac{(25)(880)}{5} = 4700 - 4400 = 300$.
5. b) $r = \dfrac{300}{\sqrt{(10)(10\,120)}} = \dfrac{300}{318.1195} = \mathbf{0.9430}$: relación lineal **directa y fuerte**.
6. c) $\hat{\beta}_1 = \dfrac{300}{10} = 30$; $\bar{x} = 5$, $\bar{y} = 176$; $\hat{\beta}_0 = 176 - 30(5) = 26$. La recta es $\hat{y} = 26 + 30x$.
7. Para trazarla: $\hat{y}(3) = 116$ y $\hat{y}(7) = 236$; pasa por $(\bar{x}, \bar{y}) = (5, 176)$.
8. d) $r^2 = \dfrac{300^2}{(10)(10\,120)} = \dfrac{90\,000}{101\,200} = \mathbf{0.8893}$.
9. **Resultado:** el modelo explica el **88.93 %** de la variación del número de pacientes. Cada mes se programan en promedio 30 pacientes más. Pronóstico: $\hat{y}(8) = 266$, $\hat{y}(9) = 296$ y $\hat{y}(10) = 326$ pacientes, con la advertencia de que son extrapolaciones (fuera de $3 \leq x \leq 7$).

7.5 Ejercicios de examen (resueltos)

Ejercicio 7.5.1 · Pesticidas y cascarón: predicción

Adaptado de: 1er final 2016-2, problema 7.

En un estudio sobre los efectos de los pesticidas en el espesor de los cascarones de ciertas aves se obtuvieron los datos siguientes:

Residuos en la yema x [ppm]	117	65	393	98	112
Espesor del cascarón y [mm]	0.49	0.52	0.37	0.53	0.49

Ejercicio propuesto 7.6.2

Adaptado de: 2º final 2011-1, problema 6.

Para estudiar la relación entre la velocidad (en pies por segundo) y el ritmo de zancadas (pasos por segundo) de 11 corredores de maratón se tienen las sumas: $\sum$ velocidad = 205.4, $\sum$ velocidad² = 3880.08, $\sum$ ritmo = 35.16, $\sum$ ritmo² = 112.681, $\sum$ (velocidad)(ritmo) = 660.13. Obtenga la recta de regresión para pronosticar el ritmo de zancadas a partir de la velocidad.

Resultado: $\hat{y} = 1.6939 + 0.0805x$.

→ Resolución en el Apéndice E, p. 107.

Ejercicio propuesto 7.6.3

Adaptado de: 1er final 2013-1, problema 7.

Una compañía de detergentes estudia la relación entre las ventas y (cientos de unidades por mes) y el número x de comerciales transmitidos por día en una muestra de siete ciudades. Se sabe que $r^2 = 0.736$. Interprete el resultado.

Resultado: el modelo explica el 73.6 % de la variación de las ventas; $r = +\sqrt{0.736} = 0.858$: relación lineal directa y fuerte.

→ Resolución en el Apéndice E, p. 107.

Ejercicio propuesto 7.6.4

Adaptado de: 2º final 2014-1, problema 7.

En siete regiones se midió el tránsito vehicular x (automóviles por hora) y el contenido de ozono y en la vegetación a la orilla de las carreteras: x = 294, 402, 103, 416, 573, 216, 334; y = 26.1, 37.2, 4.6, 24.8, 38.7, 74.4, 34.3.

- a) Obtenga la recta de regresión.
- b) Calcule el coeficiente de correlación e interprételo.

Resultado: $\hat{y} = 27.1832 + 0.0213x$; $r = 0.1527$: asociación lineal muy débil.

→ Resolución en el Apéndice E, p. 107.

Ejercicio propuesto 7.6.5

Adaptado de: 2º final 2016-2, problema 7.

En un estudio hidráulico se desea saber si el porcentaje de estrechamiento x de un río influye en el porcentaje de elevación y de su nivel de agua: x = 2.5, 5, 7.5, 10, 12.5, 15, 17.5, 20; y = 2, 4, 7, 9, 10, 14, 15, 18. ¿Es posible considerar que existe un comportamiento lineal entre las variables? Justifique.

Resultado: $r = 0.9948$: sí, relación lineal directa muy fuerte ($\hat{y} = -0.25 + 0.9x$).

→ Resolución en el Apéndice E, p. 108.

Ejercicio propuesto 7.6.6

Adaptado de: 1er final 2018-1, problema 7.

Se desea saber si el número de horas de sueño x influye en el peso y (kg) de las personas.

Ejercicio 7.5.3 · Vendedores: recta y correlación con sumas

Adaptado de: 2º final 2015-1, problema 7.

El presidente de una agencia de autos supone que el tiempo que un vendedor pasa con un cliente (minutos) tiene una relación positiva con el monto de la compra. Con los datos de nueve vendedores se obtuvieron las sumas: minutos 874, montos 3565, (minutos)² 91 330, (montos)² 2 147 175, (minutos)(montos) 358 852.

- a) Estime la recta que permite estimar el monto de las compras en función de los minutos.
- b) ¿Existe una relación lineal entre el monto y el tiempo invertido? Justifique.

Resolución.

1. X: minutos con el cliente; Y: monto de la compra; $n = 9$.
2. $SS_{xx} = 91\,330 - \frac{874^2}{9} = 6454.8889$.
3. $SS_{yy} = 2\,147\,175 - \frac{3565^2}{9} = 735\,038.8889$.
4. $SS_{xy} = 358\,852 - \frac{(874)(3565)}{9} = 12\,650.8889$.
5. a) $\hat{\beta}_1 = \frac{12\,650.8889}{6454.8889} = 1.959893$; $\bar{x} = 97.1111$, $\bar{y} = 396.1111$; $\hat{\beta}_0 = 396.1111 - 1.959893(97.1111) = 205.7838$. La recta es $\hat{y} = 205.78 + 1.96x$.
6. b) $r = \frac{12\,650.8889}{\sqrt{(6454.8889)(735\,038.8889)}} = \frac{12\,650.8889}{68\,881.016} = \mathbf{0.1837}$.
7. **Resultado:** la relación lineal es **muy débil** ($r^2 = 0.0337$: el tiempo explica apenas el 3.4 % de la variación del monto). Los datos no apoyan la suposición del presidente.

Ejercicio 7.5.4 · Probetas: coeficiente de determinación

Adaptado de: 2º final 2017-1, problema 7.

Se estudia la relación entre el esfuerzo aplicado a una probeta de plástico y el tiempo que tarda en fracturarse:

Esfuerzo x [N]	17	100	41	120	15	91	18	27
Duración y [s]	8	7	7	6.5	7.5	6.5	7.5	8

Obtenga el coeficiente de determinación del modelo que describe la duración en función del esfuerzo e indique qué se puede concluir.

Resolución.

1. X: esfuerzo [N]; Y: duración [s]; $n = 8$.
2. Sumas: $\sum x = 429$, $\sum y = 58$, $\sum x^2 = 35\,929$, $\sum y^2 = 423$, $\sum xy = 2958$.
3. $SS_{xy} = 2958 - \frac{(429)(58)}{8} = -152.25$; $SS_{xx} = 35\,929 - \frac{429^2}{8} = 12\,923.875$; $SS_{yy} = 423 - \frac{58^2}{8} = 2.5$.
4. $r^2 = \frac{(-152.25)^2}{(12\,923.875)(2.5)} = \frac{23\,180.0625}{32\,309.6875} = \mathbf{0.7174}$.
5. **Resultado:** el modelo explica el **71.74 %** de la variación de la duración; es un ajuste aceptable. Como $SS_{xy} < 0$, $r = -\sqrt{0.7174} = -0.8470$: a mayor esfuerzo, menor duración.

Ejercicio propuesto 6.6.3

Adaptado de: 1er final 2016-1, problema 6.

El precio de las acciones de cierta empresa tiene valor esperado de \$200 000 y desviación estándar de \$100 000. Si se compran 30 acciones, determine la probabilidad de que el precio promedio de las acciones compradas sea menor que \$175 000.

Resultado: $P(Z < -1.37) = \mathbf{0.0853}$.

→ Resolución en el Apéndice E, p. 106.

Ejercicio propuesto 6.6.4

Adaptado de: 2º final 2016-1, problema 5.

Una población de profesionales tiene un salario promedio de \$25 000 con desviación estándar de \$2135. Si se toma una muestra de tamaño 100, ¿cuál es la probabilidad de que el promedio de sus salarios sea menor que \$24 500?

Resultado: $P(Z < -2.34) = \mathbf{0.0096}$.

→ Resolución en el Apéndice E, p. 106.

Ejercicio propuesto 6.6.5

Adaptado de: 2º final 2016-2, problema 6.

La cantidad de llenado de las botellas de refresco de una fábrica es normal, con media de 2 litros y desviación estándar de 0.05 litros. Si se selecciona una muestra aleatoria de 25 botellas, ¿cuál es la probabilidad de que el llenado promedio esté entre 1.99 y 2 litros?

Resultado: $P(-1 \leq Z \leq 0) = \mathbf{0.3413}$.

→ Resolución en el Apéndice E, p. 106.

Ejercicio propuesto 6.6.6

Adaptado de: 1er final 2018-1, problema 6.

De una población normal con $\mu_X = 106$ y $\sigma_X = 12$ se extrae una muestra aleatoria de tamaño 25. Determine la probabilidad de que el promedio de la muestra sea mayor que 110.

Resultado: $P(Z > 1.67) = \mathbf{0.0475}$.

→ Resolución en el Apéndice E, p. 106.

Ejercicio propuesto 6.6.7

Adaptado de: 2º final 2013-1, problema 4.

Se toma una muestra aleatoria de 10 observaciones de una población normal con variancia $\sigma^2 = 42.5$. Calcule la probabilidad aproximada de obtener una desviación estándar muestral entre 3.14 y 8.94.

Resultado: $P(2.088 < \chi^2_{(9)} < 16.925) \approx 0.99 - 0.05 = \mathbf{0.94}$.

→ Resolución en el Apéndice E, p. 106.

Apéndice C · Fe de erratas

↑ Volver al índice

Diferencias entre las resoluciones publicadas de los exámenes y los valores verificados en este cuaderno.

C.1 · Unidad 1 · Análisis estadístico de datos muestrales: tablas de frecuencias, tendencia central, dispersión y sesgo

2º final 2014-1, problema 1: la resolución concluye primero «la muestra aleatoria no es buena» y después «con 14.79 % es buena la muestra». Con $cv = 0.1479 \leq 0.15$, la muestra **es homogénea**.

1er final 2015-1, problema 1: la media aparece como 3.15 kg; el valor verificado es **3.157 kg** (≈ 3.16).

1er final 2018-1, problema 1:

- la última frontera aparece como «0.4445–0.4595»; es **0.4545–0.4595**;
- la columna titulada «frecuencia relativa» contiene en realidad la **frecuencia acumulada relativa** F_i^* (0.006, 0.040, 0.172…).

2º final 2012-1, problema 1: la variancia aparece redondeada como 0.003 y la desviación estándar como $\sqrt{0.003} = 0.051$. Sin redondear, $s^2 = 0.002574$ y $s = \mathbf{0.0507}$ ($\sqrt{0.003}$ daría 0.0548). El valor de s publicado es correcto; el de s^2 está redondeado de más.

2º final 2016-1, problema 1: la mediana se obtiene interpolando en $\frac{n+1}{2} = 18$, lo que da 13.5. Para datos agrupados se interpola en $\frac{n}{2} = 17.5$: $\tilde{x} = \mathbf{13.333}$. La conclusión (sesgo a la izquierda) no cambia.

C.2 · Unidad 2 · Fundamentos de la teoría de la probabilidad: probabilidad condicional, probabilidad total y Bayes

1er final 2014-1, problema 2:

- a) el resultado aparece truncado como 0.535; el valor verificado es **0.5357**;
- b) la resolución plantea bien $\frac{0.687}{0.972}$, pero da como resultado 0.0707. El valor verificado es **0.7068**: el punto decimal está recorrido.

2º final 2015-1, problema 2 b): el resultado aparece como 0.9996; con $1 - (0.07)^3 = 0.999657$, el valor verificado es **0.9997**.

C.3 · Unidad 3 · Variables aleatorias: función de probabilidad, de densidad y de distribución, media y variancia

1er final 2016-2, problema 3: la media aparece como 1.1999 y la variancia como 0.559939, porque $\frac{1}{30}$ se redondea a 0.0333 antes de sumar. Los valores verificados son $E(X) = \mathbf{1.2}$, $E(X^2) = 2$, $\sigma_X^2 = \mathbf{0.56}$ y $\sigma_X = \mathbf{0.7483}$.

Si no se encuentra el camino, la definición de $\bar{X}$, la justificación del TLC y la expresión de z sustituida ya valen buena parte del problema.

Apéndice B · Qué vino en cada final (2011-1 a 2018-1)

↑ Volver al índice

Examen	Problemas (unidad: tema)
1er final 2011-1	U2: máquinas que producen tornillos (Bayes) · U3: valor esperado a partir de F_X · U4: geométrica y Pascal · U6: media muestral · U5: densidad conjunta · U7: recta y r (6 problemas)
2º final 2011-1	U2: sistema en serie y paralelo · U3: utilidad esperada y variancia · U5: tabla conjunta, condicional y covariancia · U4: binomial aproximada por la normal · U6: media muestral de una uniforme discreta · U7: recta y r (6 problemas)
1er final 2012-1	U1: datos sin agrupar, dos muestras · U2: fumadores (Bayes) · U3: distribución del mayor y de la suma · U4: binomial · U5: densidad conjunta · U6: media muestral con población finita · U7: recta, pronóstico y r^2
2º final 2012-1	U1: tabla a partir de los datos · U2: aseguradora (Bayes) · U3: constante k, F_X y mediana · U4: normal · U4: exponencial · U5: tabla conjunta e independencia · U7: recta, r y r^2
1er final 2013-1	U1: ojiva, media, mediana, moda y cv · U2: opción múltiple (Bayes) · U3: densidad y F_X · U3: valor esperado de un costo · U5: densidad conjunta · U6: media muestral · U7: interpretación de r^2
2º final 2013-1	U1: cv y coeficiente de simetría · U2: pantallas (Bayes) · U3: función de una variable discreta · U6: ji cuadrada · U4: Poisson · U5: uniforme conjunta · U7: recta y r^2
1er final 2014-1	U1: media, mediana, moda y sesgo · U2: impresoras (Bayes) · U3: constantes con media y variancia dadas · U4: Poisson y exponencial · U5: densidad conjunta · U6: media muestral · U7: recta, r y pronóstico
2º final 2014-1	U1: cv y mediana · U2: eventos en diagrama de Venn · U3: densidad simétrica · U4: exponencial con media dada · U5: densidad conjunta · U6: muestras con y sin reemplazo · U7: recta y r
1er final 2015-1	U1: histograma de pesos · U2: petróleo (Bayes) · U3: variable aleatoria · U4: normal; binomial aproximada por la normal · U5: variables conjuntas · U7: r a partir de sumas
2º final 2015-1	U1: media, mediana, ojiva y s · U2: sensores (probabilidad de eventos) · U3: constante k y probabilidad condicional · U4: exponencial sin memoria · U5: densidad en región triangular · U6: dos muestras · U7: recta y r a partir de sumas
1er final 2016-1	U1: cv a partir de un histograma · U2: formaciones geológicas (Bayes) · U3: variable discreta, σ_X · U4: Poisson · U5: tabla conjunta · U6: media muestral · U7: opción múltiple
2º final 2016-1	U1: ¿sesgada? · U2: probabilidad condicional · U3: F_X en una gráfica · U4: modelo nombrado · U6: media muestral · U5: densidad conjunta · U7: opción múltiple

- La recta **siempre pasa por** $(\bar{x}, \bar{y})$.
- **Interpretación de la pendiente:** por cada unidad que aumenta x, y cambia en promedio $\hat{\beta}_1$ unidades.

Coeficiente de correlación lineal.

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}\, SS_{yy}}}, \qquad -1 \le r \le 1.$$

- **Signo:** $r > 0$, relación **directa** (si x sube, y sube); $r < 0$, relación **inversa**.
- **Fuerza:** cuanto más cerca de ± 1, más se alinean los puntos. Guía usada en las resoluciones publicadas:

$\lvert r\rvert$	Conclusión
de 0.8 a 1	relación lineal fuerte: «sí existe relación lineal»
de 0.5 a 0.8	relación lineal moderada
menor que 0.5	relación lineal débil: «no puede considerarse lineal»

- r no tiene unidades y no cambia si se intercambian x y y; la recta sí cambia.
- $r \approx 0$ significa «sin relación **lineal**», no «sin relación». Y un $\lvert r\rvert$ grande no prueba que x cause y.

Coeficiente de determinación.

$$r^2 = \frac{SS_{xy}^2}{SS_{xx}\, SS_{yy}}, \qquad 0 \le r^2 \le 1.$$

- Es la proporción de la variación de y que explica la recta: «el modelo explica el $100r^2$ % de la variación de y».
- Si se da r^2, entonces $r = \pm\sqrt{r^2}$ con el **signo de la pendiente**.

Covariancia muestral. $\text{Cov}(x, y) = \frac{SS_{xy}}{n} = \frac{\sum xy}{n} - \bar{x}\,\bar{y}$. Tiene el mismo signo que SS_{xy}, $\hat{\beta}_1$ y r.

Predicción. $\hat{y}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$. Si x_0 está fuera de los datos, es una **extrapolación**: la recta no está respaldada ahí y la estimación no es recomendable.

7.2 Cómo reconocer el problema

Si el enunciado dice…	Se pide…	Procedimiento
«coeficiente de correlación», «¿existe relación lineal?», «¿comportamiento lineal?»	r y su interpretación	7.3.3
«recta de regresión», «mínimos cuadrados», «modelo para estimar y en función de x»	$\hat{\beta}_0$ y $\hat{\beta}_1$	7.3.2

Modelo	$f_X(x)$	μ	σ^2
Uniforme(a, b)	$\frac{1}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponencial(λ)	$\lambda e^{-\lambda t}$; $F = 1 - e^{-\lambda t}$	$1/\lambda$	$1/\lambda^2$
Gamma(α, β)	$\frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)}$	$\alpha\beta$	$\alpha\beta^2$
Normal $N(\mu, \sigma^2)$	$Z = \frac{X-\mu}{\sigma}$	μ	σ^2

- Exponencial: $P(T > t) = e^{-\lambda t}$; **sin memoria**: $P(T > t + s \mid T > t) = P(T > s)$; T es el tiempo entre eventos de un proceso de Poisson(λ). «Media de 16 años» significa $\lambda = 1/16$.
- Aproximación normal a la binomial (si $np > 5$ y $p \le \frac{1}{2}$, o $nq > 5$ y $p > \frac{1}{2}$): $P(x_1 \le X \le x_2) \approx P\left(\frac{x_1 - 0.5 - np}{\sqrt{npq}} \le Z \le \frac{x_2 + 0.5 - np}{\sqrt{npq}}\right)$. «Al menos 28» se toma como 27.5; «más de 28», como 28.5.
- Aproximación de Poisson a la binomial: $\lambda = np$ (si $n \ge 20$ y $p \le 0.05$). Binomial en lugar de hipergeométrica: si $n < 0.1N$, con $p = r/N$. Factor de corrección por población finita: $\frac{N-n}{N-1}$.
- Combinación lineal de normales independientes: $aX + bY \sim N(a\mu_X + b\mu_Y,\ a^2\sigma_X^2 + b^2\sigma_Y^2)$.

A.5 Variables aleatorias conjuntas (Unidad 5)

- Marginales: $f_X(x) = \sum_y f_{XY}$ o $\int f_{XY}\, dy \cdot f_Y(y) = \sum_x f_{XY}$ o $\int f_{XY}\, dx$ (límites según la región).
- Condicionales: $f_{X|y_0}(x) = \frac{f_{XY}(x, y_0)}{f_Y(y_0)} \cdot f_{Y|x_0}(y) = \frac{f_{XY}(x_0, y)}{f_X(x_0)}$.
- $P(a < X < b,\ c < Y < d) = \int_a^b \int_c^d f_{XY}\, dy\, dx \cdot E(Y \mid X = x) = \int y\, f_{Y|x}(y)\, dy$ (curva de regresión).
- Independencia: $f_{XY} = f_X f_Y$ para todo (x, y) (requiere una región rectangular). Independientes implica Cov = 0; Cov = 0 **no implica independencia.**
- $E(XY) = \sum\sum xy f_{XY}$ o $\iint xy f_{XY} \cdot \text{Cov}(X, Y) = \sigma_{XY} = E(XY) - \mu_X\mu_Y \cdot \rho = \frac{\sigma_{XY}}{\sigma_X\sigma_Y} \in [-1, 1] \cdot \rho^2$: coeficiente de determinación.
- $E(aX + bY) = a\mu_X + b\mu_Y \cdot \text{Var}(aX + bY) = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\,\sigma_{XY}$.

A.6 Distribuciones muestrales (Unidad 6)

- $\bar{X} \sim N\left(\mu_{\bar{X}} = \mu_X,\ \sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n}\right)$ si la población es normal (cualquier n) o si $n \ge 30$ (teorema del límite central). Con población desconocida y $n < 30$ no se puede aplicar.
- $Z = \frac{\bar{X} - \mu_X}{\sigma_X/\sqrt{n}} \cdot$ error estándar $\sigma_{\bar{X}} = \sigma/\sqrt{n} \cdot$ sin reemplazo y población finita: $\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$.
- Suma: $\sum X_i \sim N(n\mu,\ n\sigma^2) \cdot$ diferencia de medias: $Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$.
- $\frac{(n-1)S_{n-1}^2}{\sigma^2} \sim \chi^2_{(n-1)} \cdot T = \frac{\bar{X} - \mu}{S_{n-1}/\sqrt{n}} \sim t_{(n-1)}$ (población normal, σ desconocida, $n < 30$); con $n \ge 30$ se usa $S \approx \sigma$ y Z.
- Proporción muestral: $\hat{p} \sim N(p,\ pq/n)$.

2. $\hat{y}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$, con los coeficientes sin redondear.
3. Si x_0 está fuera del intervalo, o si r^2 es muy bajo, escribe que la estimación **no es recomendable**.

7.3.5 Diagrama de dispersión y recta

1. Ejes con nombre y unidades: x horizontal, y vertical.
2. Un punto por pareja, sin unirlos.
3. Para la recta, evalúa $\hat{y}$ en el menor y el mayor x y une esos dos puntos. Debe pasar por $(\bar{x}, \bar{y})$.

7.3.6 Reglas rápidas (opción múltiple e interpretación)

1. $r^2 = (r)^2$: con $r = 0.9165$, $r^2 = 0.84$. Un enunciado que llame «determinación» a 0.9165 está confundiendo r con r^2.
2. $-1 \le r \le 1$ y $0 \le r^2 \le 1$: una opción fuera de esos intervalos es falsa.
3. Signo de la covariancia = signo de SS_{xy} = signo de $\hat{\beta}_1$ = signo de r.
4. La recta pasa por $(\bar{x}, \bar{y})$: sustituye $\bar{x}$ y comprueba que sale $\bar{y}$.
5. Con r^2 dado, $r = \pm\sqrt{r^2}$ con el signo de la pendiente.

7.4 Ejemplo resuelto

Adaptado de: 1er final 2014-1, problema 7.

El director del área de cirugía de un hospital tiene el número de pacientes programados en los meses de marzo a julio y desea un pronóstico, por regresión lineal de mínimos cuadrados, para los tres meses siguientes ($x = 8$, 9 y 10).

Mes	marzo	abril	mayo	junio	julio
x	3	4	5	6	7
y (pacientes)	100	150	200	210	220

Apéndice A · Formulario y tablas

↑ Volver al índice

El examen permite llevar un **formulario** y las **tablas** de las distribuciones. Este apéndice reúne todas las expresiones del cuaderno, en el orden de las unidades; cabe en dos hojas escritas a mano.

Tablas que se necesitan (se llevan aparte):

- **Normal estándar:** $F_Z(z) = P(Z \le z)$, área a la izquierda; $F_Z(-z) = 1 - F_Z(z)$.
- **t de Student:** $t_{\alpha,\nu}$ con ν grados de libertad.
- **Ji cuadrada:** área a la derecha, $P(X^2 \ge \chi^2_{\alpha,\nu}) = \alpha$.
- La distribución F no se usa en este curso.

Recordatorios generales.

- La frecuencia acumulada F_i se suma desde la primera clase; la ojiva se traza en la frontera superior; todo cuantil queda entre el dato mínimo y el máximo.
- Los estadísticos se calculan siempre sobre las marcas x_i, nunca sobre las frecuencias: $\bar{x} = \sum x_i f_i / n$.
- En $P(A \mid B)$ el denominador es el grupo «dado»: «de los ___, ¿qué fracción…?».
- Convenciones de este cuaderno: $a_3 = m_3/s^3$; cv como cociente (≤ 0.15 homogéneos, sin multiplicar por 100); Cov $= SS_{xy}/n$ en regresión; $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

A.1 Estadística descriptiva (Unidad 1)

Datos agrupados: m clases, x_i marca de clase, f_i frecuencia, F_i frecuencia acumulada, $f_i^* = f_i/n$, c longitud de clase.

- Número de clases $\approx \sqrt{n}$ (de 5 a 20); c = rango/número de clases; fronteras = límites ± media unidad; $x_i = \frac{\text{fr. inf+fr. sup}}{2}$.
- n = último F = número de datos. c = frontera superior − frontera inferior (o marca − marca anterior): 10–19 da $c = 10$, no 9.
- $\bar{x} = \frac{1}{n}\sum x_i f_i \cdot \tilde{x} = L_{inf} + \frac{n/2 - F_{ant}}{f_{med}} c$ (interpolación en la ojiva en $n/2$; clase mediana: la primera con $F_i \ge n/2$) $\cdot\ x_{mo} = L_{inf} + \frac{\Delta_1}{\Delta_1+\Delta_2} c$, con $\Delta_1 = f_{mo} - f_{ant}$ y $\Delta_2 = f_{mo} - f_{sig}$ (o $x_{mo} \approx$ marca de la clase modal). f_{med}, Δ_1 y Δ_2 usan frecuencias absolutas.
- Cuantiles: igual que $\tilde{x}$ con $kn/4$ (Q_k), $kn/10$ (D_k) o $kn/100$ (P_k). Rango intercuartil $R_Q = Q_3 - Q_1$; desviación cuartil $Q = \frac{Q_3 - Q_1}{2}$; rango percentil $R_P = P_{90} - P_{10}$.
- $s^2_{n-1} = \frac{1}{n-1}\sum (x_i - \bar{x})^2 f_i = \frac{\sum x_i^2 f_i - n\bar{x}^2}{n-1}$ (muestra; columna siempre positiva) $\cdot\ s^2_n = \frac{1}{n}\sum (x_i - \bar{x})^2 f_i$ (población) $\cdot\ s^2_{n-1} = \frac{n}{n-1} s^2_n$ · estimación rápida $s \approx R/4$.
- Desviación media: $DM = \frac{1}{n}\sum |x_i - \bar{x}| f_i$.
- $cv = s/\bar{x}$: ≤ 0.15 datos homogéneos; > 0.15 heterogéneos.
- Momentos: $m'_r = \frac{1}{n}\sum x_i^r f_i$ (respecto al origen), $m_r = \frac{1}{n}\sum (x_i - \bar{x})^r f_i$ (respecto a la media, entre n, no $n-1$); $m_3 = m'_3 - 3m'_1 m'_2 + 2m'^3_1$.
- **Sesgo** $a_3 = m_3/s^3$, con $s = s_{n-1}$: $a_3 > 0$ cola a la derecha; $a_3 < 0$ cola a la izquierda; $a_3 = 0$ simétrica. **Curtosis** $a_4 = m_4/s^4$, contra 3: > 3 leptocúrtica, $= 3$ mesocúrtica, < 3

Si los residuos de pesticidas son 105 ppm, ¿cuánto vale el espesor del cascarón?

Resolución.

1. X: residuos de pesticida [ppm]; Y: espesor [mm]; $n = 5$.
2. Sumas: $\sum x = 785$, $\sum y = 2.4$, $\sum x^2 = 194\,511$, $\sum y^2 = 1.1684$, $\sum xy = 343.36$.
3. $SS_{xx} = 194\,511 - \frac{785^2}{5} = 71\,266$; $SS_{xy} = 343.36 - \frac{(785)(2.4)}{5} = -33.44$.
4. $\hat{\beta}_1 = \frac{-33.44}{71\,266} = -0.00046923$; $\bar{x} = 157$, $\bar{y} = 0.48$.
5. $\hat{\beta}_0 = 0.48 - (-0.00046923)(157) = 0.553669$, así que $\hat{y} = 0.553669 - 0.00046923x$.
6. 105 ppm está dentro del intervalo de los datos (65 a 393): $\hat{y}(105) = 0.553669 - 0.00046923(105) =$ **0.5044 mm**.
7. Complemento: $SS_{yy} = 0.0164$ y $r = \frac{-33.44}{\sqrt{(71\,266)(0.0164)}} = -0.978$. Relación **inversa** muy fuerte: más pesticida, cascarón más delgado.

Ejercicio 7.5.2 · Basquetbolistas: recta, determinación y extrapolación

Adaptado de: 2º final 2013-1, problema 7; repetido en el 1er final 2015-1, problema 7.

Las estaturas (en cm) y pesos (en kg) de 10 jugadores de baloncesto son:

Estatura x	186	189	190	192	193	193	198	201	203	205
Peso y	85	85	86	90	87	91	93	103	100	101

Sumas: $\sum x = 1950$, $\sum y = 921$, $\sum x^2 = 380\,618$, $\sum y^2 = 85\,255$, $\sum xy = 179\,971$.

- a) Estime la recta de regresión de los pesos en función de la estatura.
- b) Calcule el coeficiente de determinación e interprételo.
- c) Obtenga el peso estimado de un jugador que mide 208 cm.

(En el 1er final 2015-1 se dan solo las sumas y la recta, y se pregunta si existe relación lineal: es el coeficiente r del paso 6.)

Resolución.

1. X: estatura [cm]; Y: peso [kg]; $n = 10$.
2. $SS_{xx} = 380\,618 - \frac{1950^2}{10} = 368$; $SS_{yy} = 85\,255 - \frac{921^2}{10} = 430.9$; $SS_{xy} = 179\,971 - \frac{(1950)(921)}{10} = 376$.
3. a) $\hat{\beta}_1 = \frac{376}{368} = 1.021739$; $\bar{x} = 195$, $\bar{y} = 92.1$.
4. $\hat{\beta}_0 = 92.1 - 1.021739(195) = -107.1391$. La recta es $\hat{y} = -107.14 + 1.0217x$.
5. b) $r^2 = \frac{376^2}{(368)(430.9)} = \frac{141\,376}{158\,571.2} =$ **0.8916**: el modelo explica el **89.16 %** de la variación del peso; es un modelo adecuado.
6. $r = +\sqrt{0.8916} =$ **0.9442** (positivo, como la pendiente): relación lineal directa y fuerte; a mayor estatura, mayor peso.
7. c) 208 cm está **fuera** del intervalo de los datos (186 a 205 cm): es una extrapolación y la estimación no es recomendable. Si se calcula, $\hat{y}(208) = 105.38$ kg, con esa advertencia.

En una muestra de 10: x = 5, 8, 7, 10, 6, 9, 10, 6, 6, 8; y = 69, 69, 58, 54, 54, 65, 95, 62, 57.5, 30. ¿Puede considerarse que existe un comportamiento lineal? Justifique.

Resultado: $r = 0.2217$: la relación lineal no es buena; no puede considerarse lineal.

→ Resolución en el Apéndice E, p. 108.

Ejercicio propuesto 7.6.7

Adaptado de: 2º final 2016-1, problema 7; mismos datos que el ejercicio 7.5.7.

Con los datos del experimento químico del ejercicio 7.5.7 ($\sum xy = 10\,794$, $\bar{x} = 33.89$, $\bar{y} = 31.67$, recta $y = 0.6012x + 11.291$, $r = 0.9165$), ¿cuál de las afirmaciones es **incorrecta**?

a) Existe una relación lineal entre la temperatura y la producción · b) La recta pasa por el punto (33.89, 31.67) · c) La covariancia de X y Y es positiva · d) El coeficiente de determinación es 0.9165.

Resultado: d): 0.9165 es r; el coeficiente de determinación es $r^2 = 0.84$.

→ Resolución en el Apéndice E, p. 108.

7.7 Errores frecuentes

1. **Intercambiar x y y.** r no cambia, pero la recta sí. Lee «en función de», «a partir de», «influye en». Cuidado cuando y viene en el renglón de arriba de la tabla.
2. **Redondear antes de tiempo.** Con $\hat{\beta}_1$ redondeado a tres cifras, $\hat{\beta}_0$ cambia en la segunda o tercera decimal. Guarda 6 cifras hasta el final.
3. **Calcular $\hat{\beta}_0$ antes que $\hat{\beta}_1$.** La ordenada necesita la pendiente.
4. **Confundir r con r^2.** «Correlación» es r; «determinación» o «porcentaje explicado» es r^2.
5. **Perder el signo.** r, $\hat{\beta}_1$ y la covariancia salen del mismo SS_{xy} y tienen el mismo signo. Al obtener r de r^2, el signo es el de la pendiente.
6. **SS_{xx} o SS_{yy} negativa.** Es error de cuenta: casi siempre $(\sum x)^2$ en lugar de $\sum x^2$ o al revés.
7. **Contar mal n.** n es el número de parejas que realmente se tienen, aunque el enunciado diga otra cifra.
8. **Mezclar n y $n-1$.** Las tres SS se dividen entre n dentro de la resta; la covariancia de los datos es SS_{xy}/n.
9. **Predecir fuera del intervalo de los datos** sin advertir que es extrapolación.
10. **Concluir causalidad.** Un r alto indica asociación lineal, no que x provoque y.
11. **Unir los puntos del diagrama de dispersión.** Van sueltos; la única línea es la recta ajustada.

7.8 Formulario de emergencia de la unidad

- $SS_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} \cdot SS_{yy} = \sum y^2 - \frac{(\sum y)^2}{n} \cdot SS_{xy} = \sum xy - \frac{\sum x \sum y}{n}$
- $\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \cdot \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \cdot \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \cdot$ pasa por $(\bar{x}, \bar{y})$
- $r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}} \cdot r^2 = \frac{SS_{xy}^2}{SS_{xx} SS_{yy}} \cdot r = \pm\sqrt{r^2}$ con el signo de $\hat{\beta}_1$

Ejercicio 7.5.5 · Datos leídos de una gráfica

Adaptado de: 1er final 2017-1, problema 7.

Se estudia la relación entre el esfuerzo aplicado a una probeta de plástico y su duración antes de fracturarse. Los datos vienen en una gráfica con el esfuerzo x [N] en el eje horizontal y la duración y [s] en el vertical. Los puntos son (6.5, 148), (6.5, 91), (7, 100), (7, 41), (7.5, 20), (7.5, 15), (8, 27) y (8, 17). Indique el valor del coeficiente de correlación e interprételo.

Resolución.

1. X: esfuerzo [N]; Y: duración [s]. La gráfica tiene 8 puntos: $n = 8$.
2. Sumas: $\sum x = 58$, $\sum y = 459$, $\sum x^2 = 423$, $\sum y^2 = 43\,509$, $\sum xy = 3155$.
3. $SS_{xy} = 3155 - \frac{(58)(459)}{8} = -172.75$; $SS_{xx} = 423 - \frac{58^2}{8} = 2.5$; $SS_{yy} = 43\,509 - \frac{459^2}{8} = 17\,173.875$.
4. $r = \frac{-172.75}{\sqrt{(2.5)(17\,173.875)}} = \frac{-172.75}{207.2069} = \mathbf{-0.8337}$.
5. **Resultado:** relación lineal **inversa y fuerte**: a mayor esfuerzo aplicado, menor duración de la probeta.

Ejercicio 7.5.6 · Vendedores y televisores: ¿se puede extrapolar?

Adaptado de: 2º final 2018-1, problema 6.

El gerente de una tienda de televisores observa las ventas en 10 días y el número de vendedores que había cada día:

Vendedores x	2	2	2	3	3	3	4	4	4	3
Televisores vendidos y	3	5	10	5	10	12	5	10	10	8

- a) Obtenga por mínimos cuadrados el modelo que permita estimar las ventas en función del número de vendedores.
- b) ¿Es recomendable el modelo para estimar las ventas cuando haya un solo vendedor? Justifique.

Resolución.

1. X: número de vendedores; Y: televisores vendidos; $n = 10$.
2. Sumas: $\sum x = 30$, $\sum y = 78$, $\sum x^2 = 96$, $\sum y^2 = 692$, $\sum xy = 241$.
3. $SS_{xy} = 241 - \frac{(30)(78)}{10} = 7$; $SS_{xx} = 96 - \frac{30^2}{10} = 6$; $SS_{yy} = 692 - \frac{78^2}{10} = 83.6$.
4. a) $\hat{\beta}_1 = \frac{7}{6} = 1.166667$; $\bar{x} = 3$, $\bar{y} = 7.8$; $\hat{\beta}_0 = 7.8 - 1.166667(3) = 4.3$. El modelo es $\hat{y} = 4.3 + 1.1667x$.
5. b) $x = 1$ está fuera del intervalo observado (2 a 4 vendedores): sería una extrapolación.
6. Además, $r = \frac{7}{\sqrt{(6)(83.6)}} = 0.3125$ y $r^2 = \mathbf{0.0977}$: el modelo explica menos del 10 % de la variación de las ventas.
7. **Resultado:** el modelo **no es recomendable** para un solo vendedor, por extrapolación y por su bajo coeficiente de determinación.

Ejercicio 7.5.7 · Opción múltiple: porcentaje explicado

Adaptado de: 1er final 2016-1, problema 7.

En un experimento químico se quiere saber si existe relación entre la temperatura ambiente x [°C] y la producción de la reacción y [g]:

x	10	15	25	30	35	40	45	50	55
y	17.3	25.4	24.5	26.1	26.2	39.4	39.6	37.6	48.9

La gráfica del enunciado muestra la recta $y = 0.6012x + 11.291$ y $r = 0.9165$; además $\bar{x} = 33.89$ y $\bar{y} = 31.67$. ¿Qué porcentaje del comportamiento de los datos explica el modelo?

a) El modelo no explica los datos · b) 83.99 % · c) 91.65 % · d) 100 % · e) No se puede determinar con los datos proporcionados.

Resolución.

1. «Porcentaje explicado» es el coeficiente de **determinación** r^2, no r.
2. $r^2 = (0.9165)^2 = 0.8400$, es decir, 84 %.
3. **Resultado:** inciso **b) 83.99 %**. El inciso c), 91.65 %, es el valor de r expresado como porcentaje.
4. Comprobación con los datos: $SS_{xx} = 1888.889$, $SS_{yy} = 812.84$, $SS_{xy} = 1135.667$; $\hat{\beta}_1 = 0.6012$, $\hat{\beta}_0 = 11.2915$, $r = 0.9165$ y $r^2 = 0.8400$.

7.6 Ejercicios propuestos

Resuélvelos sin ver la solución. Las resoluciones completas están en el **Apéndice E, sección E.7**.

Ejercicio propuesto 7.6.1

Adaptado de: 1er final 2011-1, problema 6.

Se midió el contenido de alquitrán (x, en mg) y de monóxido de carbono (y, en mg) de seis marcas de cigarros: (16.0, 16.6), (8.0, 10.2), (15.1, 14.4), (15.8, 17.5), (17.0, 18.5) y (8.6, 10.6).

- a) Elabore el diagrama de dispersión.
- b) Obtenga la recta de regresión.
- c) ¿Corroboran los resultados que un aumento de alquitrán conlleva un aumento de monóxido de carbono?
- d) ¿Es bueno el modelo? ¿Qué porcentaje del monóxido de carbono queda explicado por el modelo?

Resultado: $\hat{y} = 3.1736 + 0.8541x$; sí (pendiente y $r = 0.9658$ positivos); $r^2 = 0.9328$: el modelo explica el 93.28 %, es bueno.

→ Resolución en el Apéndice E, p. 107.