Probabilidad y estadística

MAT 041, Primer semestre

Francisco Cuevas Pacheco

14 de noviembre de 2022

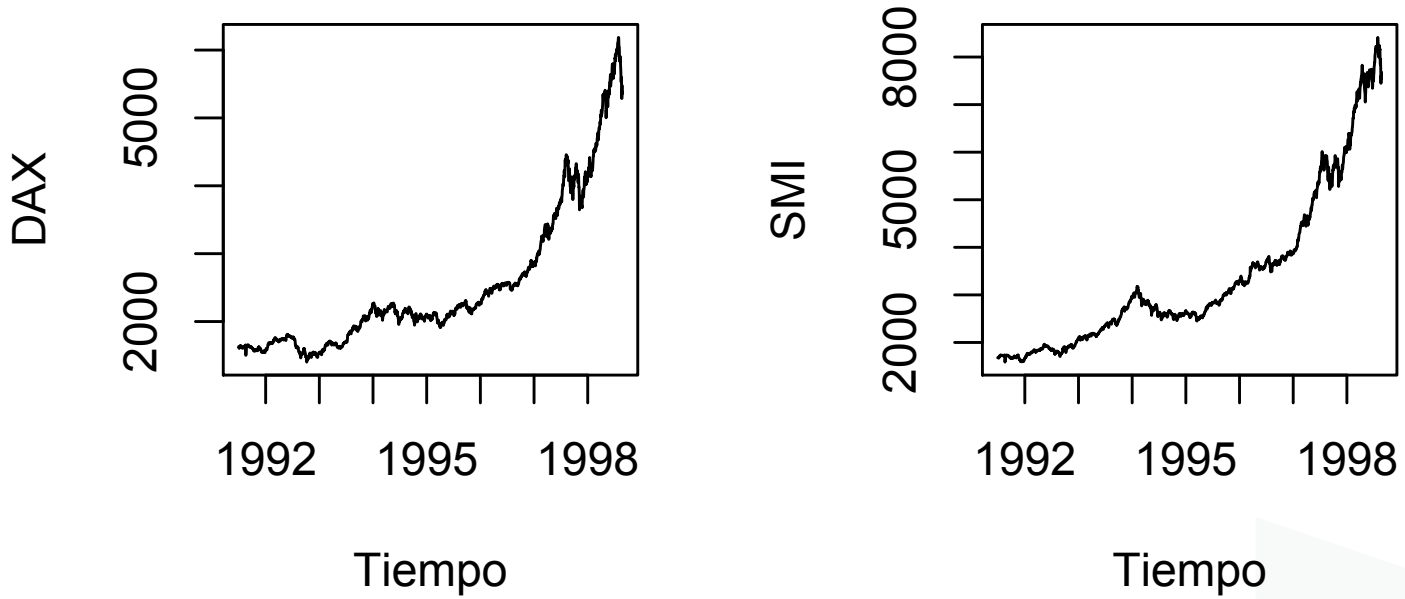
Contenidos

- ✓ Introducción
- ✓ Covarianza y Correlación
- ✓ Tablas de Contingencia
- ✓ Regresión Lineal Simple
- ✓ Otros Indicadores de Asociación

Estadística Descriptiva Multivariada

- En estadística descriptiva multivariada se estudian, de manera simultánea, múltiples variables.
- Nos centramos en el caso de dos variables (caso bivariado).
- No sólo es interesante analizar el comportamiento de cada variable de manera individual (como en el caso univariado), sino que también capturar posibles interacciones entre ellas.

Precios de cierre diarios de dos índices bursátiles europeos (DAX y SMI) entre 1991 y 1998.



La covarianza mide la asociación lineal entre las variables X e Y.

Definición (Covarianza)

Asuma que disponemos de una muestra bivariada de la forma

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

Se define la covarianza entre las variables X e Y como

$$\begin{aligned} cov(X, Y) &= \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y}) \\ &= \frac{1}{n} \sum_{i=1}^{n} X_i Y_i - \overline{X}\,\overline{Y} \end{aligned}$$

Caso Especial

Se cumple la relación

$$\text{cov}(X, X) = S_X^2.$$

Observación

- Si las variables están directamente asociadas, $\text{cov}(X, Y) > 0$.
- Si las variables están inversamente asociadas, $\text{cov}(X, Y) < 0$.
- Si las variables no tienen una asociación lineal, $\text{cov}(X, Y) = 0$.

La correlación es una versión normalizada de la covarianza.

Definición (Correlación)

Sean X e Y dos variables estadísticas y supongamos que disponemos de las observaciones

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

El coeficiente de correlación (de Pearson) se define como

$$r = \frac{cov(X, Y)}{S_X S_Y} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}.$$

Observación

La desigualdad de Cauchy-Schwarz implica que el coeficiente de correlación satisface la desigualdad $-1 \leq r \leq 1$.

- $\boxed{r = 1}$ significa asociación lineal directa perfecta entre las variables.
- $\boxed{r = -1}$ significa asociación lineal inversa perfecta entre las variables.
- $\boxed{r = 0}$ significa ausencia de asociación lineal entre X e Y (podría existir asociación de otro tipo, por ejemplo, asociación cuadrática).

Observación

A modo de ilustración, la correlación entre los índices DAX y SMI, vistos anteriormente, es 0.991.

Considere una muestra conjunta de dos variables X e Y. El objetivo es tabular esta información, lo que da lugar a las tablas de contingencia.

Dividimos la muestra en r clases A_i, $i = 1, \ldots, r$, con respecto a la variable X y en s clases B_j, $j = 1, \ldots, s$, con respecto a la variable Y.

Definición (Frecuencia Absoluta Conjunta)

La frecuencia absoluta conjunta de la modalidad $A_i B_j$, denotada por n_{ij}, se define como la cantidad de individuos o elementos de la muestra que pertenecen a las clases A_i y B_j.

Definición (Frecuencia Relativa Conjunta)

La frecuencia relativa conjunta de la modalidad A_iB_j, denotada por f_{ij}, se define por

$$f_{ij} = \frac{n_{ij}}{n},$$

donde $i = 1, \ldots, r$, y $j = 1, \ldots, s$.

Observación

Desde las definiciones anteriores podemos ver que

- $\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij} = n$
- $\sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij} = 1$

Una tabla de contingencia tiene el siguiente aspecto:

X / Y	B_1	B_2	$\ldots$	B_s	Total
A_1	n_{11}	n_{12}	$\ldots$	n_{1s}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	$\ldots$	n_{2s}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_r	n_{r1}	n_{r2}	$\ldots$	n_{rs}	$n_{r\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$\ldots$	$n_{\cdot s}$	n

Definición (Frecuencias Marginales)

Las frecuencias absolutas marginales (denotadas por $n_{i\cdot}$ y $n_{\cdot j}$) y las frecuencias relativas marginales (denotadas por $f_{i\cdot}$ y $f_{\cdot j}$) se definen de la siguiente manera:

$$n_{i\cdot} = \sum_{j=1}^{s} n_{ij}, \qquad f_{i\cdot} = \frac{n_{i\cdot}}{n}, \qquad i = 1, \ldots, r.$$

$$n_{\cdot j} = \sum_{i=1}^{r} n_{ij}, \qquad f_{\cdot j} = \frac{n_{\cdot j}}{n}, \qquad j = 1, \ldots, s.$$

Definición (Frecuencias Condicionales)

Las frecuencias relativas condicionales asociadas a una tabla de contingencia se definen como sigue:

$$f_{i|j} = \frac{f_{ij}}{f_{.j}} = \frac{n_{ij}/n}{n_{.j}/n} = \frac{n_{ij}}{n_{.j}}.$$

Indicadores para Datos Tabulados

- El **promedio marginal** de X:

$$\overline{X} = \sum_{i=1}^{r} f_{i\cdot}\, \mathcal{M}_{A_i}$$

donde $\mathcal{M}_{A_i}$ es la marca de clase de A_i.

- La **varianza marginal** de X:

$$S_X^2 = \sum_{i=1}^{r} f_{i\cdot}(\mathcal{M}_{A_i} - \overline{X})^2$$

Indicadores para Datos Tabulados

- El **promedio marginal** de Y:

$$\overline{Y} = \sum_{j=1}^{s} f_{\cdot j}\, \mathcal{M}_{B_j}$$

donde $\mathcal{M}_{B_j}$ es la marca de clase de B_j.

- La **varianza marginal** de Y:

$$S_Y^2 = \sum_{j=1}^{s} f_{\cdot j}(\mathcal{M}_{B_j} - \overline{Y})^2$$

Indicadores para Datos Tabulados

- El **promedio** de X **condicionado** a que Y pertenece a B_j:

$$\overline{X}_j = \sum_{i=1}^{r} f_{i|j}\, \mathcal{M}_{A_i}$$

- La **varianza** de X **condicionada** a que Y pertenece a B_j:

$$V_j(X) = \sum_{i=1}^{r} f_{i|j}(\mathcal{M}_{A_i} - \overline{X}_j)^2$$

Indicadores para Datos Tabulados

- El **promedio** de Y **condicionado** a que X pertenece a A_i:

$$\overline{Y}_i = \sum_{j=1}^{s} f_{j|i}\, \mathcal{M}_{B_j}$$

- La **varianza** de Y **condicionada** a que X pertenece a A_i:

$$V_i(Y) = \sum_{j=1}^{s} f_{j|i}(\mathcal{M}_{B_j} - \overline{Y}_i)^2$$

- La **descomposición de la varianza** de X:

$$V(X) = \sum_{j=1}^{s} f_{\cdot j} V_j(X) + \sum_{j=1}^{s} f_{\cdot j} (\overline{X}_j - \overline{X})^2.$$
$$= \textbf{Varianza Intra} + \textbf{Varianza Inter}$$

- La **descomposición de la varianza** de Y:

$$V(Y) = \sum_{i=1}^{r} f_{i \cdot} V_i(Y) + \sum_{i=1}^{r} f_{i \cdot} (\overline{Y}_i - \overline{Y})^2.$$
$$= \textbf{Varianza Intra} + \textbf{Varianza Inter}$$

- La **covarianza** está dada por

$$\text{cov}(X, Y) = \sum_{i=1}^{r} \sum_{j=1}^{s} f_{ij} \mathcal{M}_{A_i} \mathcal{M}_{B_j} - \overline{X}\,\overline{Y}.$$

Ejercicio

En un esfuerzo por obtener el máximo rendimiento en una reacción química, un experto analiza los valores de las siguientes variables:

T: Temperatura (en ºC).
P: Porcentaje de material convertido al producto deseado.

Se resumen los datos para una muestra de tamaño $n = 20$:

T/P	40 - 50	50 - 60	60 - 70	70 - 80	Total
160 - 170	3	1	0	0	4
170 - 180	0	3	5	0	8
180 - 190	0	2	3	3	8
Total	3	6	8	3	20

(a) *Calcule los promedios y varianzas marginales de cada variable. ¿Cuál variable es más homogénea?*

(b) *Calcule el porcentaje promedio de material convertido, dado que la temperatura es superior a 170ºC.*

(c) *¿Existe evidencia de asociación lineal entre las variables?*

Solución:

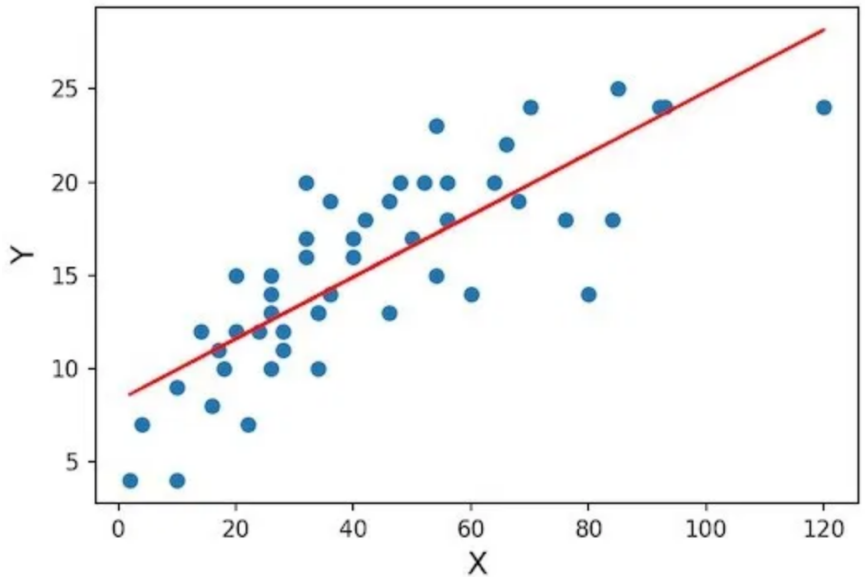
(a) Los promedios y varianzas marginales son: $\overline{T} = 177$, $\overline{P} = 60.5$, $S_T^2 = 56$ y $S_P^2 = 84.75$. Para determinar cuál variable es más homogénea, calculamos los coeficientes de variación (CV), $\text{CV}_T = 0.042$ y $\text{CV}_P = 0.15$. Por lo tanto, las temperaturas son más homogéneas.

(b) El porcentaje promedio de material convertido, condicionado a que T es superior a 170°C, está dado por

$$\overline{P}_{cond} = \frac{0 \times 45 + 5 \times 55 + 8 \times 65 + 3 \times 75}{16} = 63.75.$$

(c) La covarianza está dada por $\text{cov}(T, P) = 49$. Luego, la correlación es $r = 0.71$. Es razonable pensar que existe asociación lineal directa entre las variables.

Sea Y una variable respuesta que queremos explicar en términos de una covariable X (también denominada variable independiente). Disponemos de una muestra de pares ordenados $(X_1, Y_1), \ldots, (X_n, Y_n)$.



Planteamos el siguiente modelo de regresión lineal simple:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

donde β_0 y β_1 son parámetros del modelo a ser determinados en función de los datos y ϵ_i es un error aleatorio asociado (por ejemplo) al error de medición.

El problema se reduce a encontrar la recta que *mejor* se ajusta a los puntos observados en el plano.

Formalmente, debemos hallar β_0 y β_1 que minimizan la función

$$g(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 = \sum_{i=1}^{n} \epsilon_i^2.$$

En otras palabras, estamos minimizando la suma de los errores al cuadrado. Al resolver el sistema de ecuaciones

$$\begin{cases} \dfrac{\partial g(\beta_0, \beta_1)}{\partial \beta_0} = 0 \\ \\ \dfrac{\partial g(\beta_0, \beta_1)}{\partial \beta_1} = 0 \end{cases}$$

se obtienen los estimativos $\widehat{\beta}_0$ y $\widehat{\beta}_1$.

Teorema

Dado un conjunto de observaciones en el plano, $(X_1, Y_1), \ldots, (X_n, Y_n)$*, la recta de regresión que minimiza el error cuadrático está caracterizada por el intercepto*

$$\boxed{\widehat{\beta_0} = \overline{Y} - \widehat{\beta}_1 \overline{X}}$$

y por la pendiente

$$\boxed{\widehat{\beta}_1 = \frac{cov(X, Y)}{S_X^2}}$$

Observación

Si X_0 es una nueva observación, entonces la recta de regresión estimada nos permite proporcionar una predicción para la variable Y,

$$\widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 X_0.$$

Definición (Residuos)

Las cantidades

$$e_i = Y_i - \widehat{Y}_i$$

se llaman residuos asociados al modelo de regresión lineal simple, y siempre es cierto que

$$\sum_{i=1}^{n} e_i = 0.$$

Ejercicio

Considere el pH (X) y el porcentaje de arsénico removido (Y) para una muestra de agua.

X	Y
7.01	60
7.11	67
7.12	66
7.24	52
7.94	50
7.94	45
8.04	52
8.05	48
8.07	40
8.90	23
8.94	20
8.95	40
8.97	31
8.98	26
9.85	9
9.86	22
9.86	13
9.87	7

(a) *Hallar la correlación entre el pH y el porcentaje de arsénico removido. ¿Existe evidencia de asociación lineal?*

(b) *Hallar la recta de regresión lineal simple asociada a este conjunto de datos.*

(c) *Predecir el valor del porcentaje de arsénico removido para un pH de* 7.5.

Solución:

(a) La correlación es -0.95, lo que indica una fuerte asociación lineal inversa.

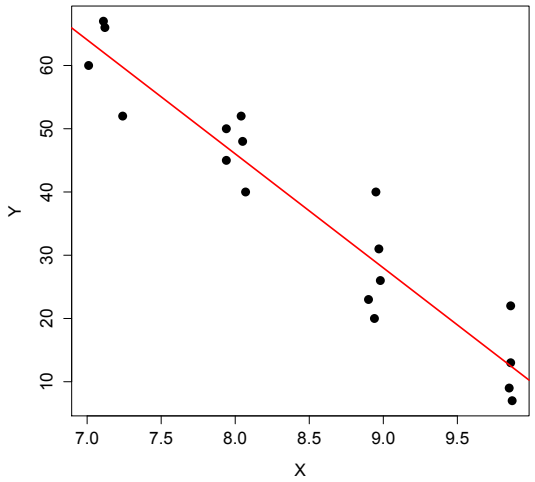
(b) Las estimaciones del intercepto y de la pendiente están dadas por

$$\widehat{\beta}_0 = 190.27 \quad \text{y} \quad \widehat{\beta}_1 = -18.03$$

Solución (continuación):

(c) Para $X_0 = 7.5$, tenemos la predicción

$$\begin{aligned} \widehat{Y}_0 &= \widehat{\beta}_0 - \widehat{\beta}_1 X_0 \\ &= 190.27 - 18.03 \times 7.5 \\ &= 55.04. \end{aligned}$$



Ejemplo

Rising Hills Manufacturing Inc. desea estudiar la relación entre el número de trabajadores, X, y el número de mesas, Y, producidas en su planta de Redwood Falls. Ha tomado una muestra aleatoria de 10 horas de producción. Se han obtenido las siguientes combinaciones (x, y) de puntos:

$(12, 20)$	$(30, 60)$	$(15, 27)$	$(24, 50)$	$(14, 21)$
$(18, 30)$	$(28, 61)$	$(26, 54)$	$(19, 32)$	$(27, 57)$

Calcule la covarianza y el coeficiente de correlación. Analice brevemente la relación entre el número de trabajadores y el número de mesas producidas por hora. Los datos se encuentran en el fichero de datos Rising Hills.

Si la dirección decide emplear 25 trabajadores, estime el número esperado de meses que es probable que se produzcan.

AED: Resumen

Veremos un repaso de AED viendo cada uno de los pasos recomendados para llevar a cabo un análisis completo.

AED: Etapas para el caso multivariado

1. Identificar los tipos de variables y sus mediciones
2. Análisis gráfico apropiado de las variables de forma individual y en conjunto:
   - Tabla de contigencia (agrupar datos).
   - Tabla bivariada.
   - Buscar posible outliers.
   - Detectar simetría y curtosis.
   - Identificación de relaciones entre las variables.
3. Proponer y calcular indicadores adecuados para cada variable.
   - Simetría y curtosis.
   - Tendencia central.
   - Variabilidad.

5. Indicadores de asociación.
6. Análisis marginal (según sea el caso)
7. Estratificación (según sea el caso)
    - Indicadores comparativos
    - Promedio ponderado
    - Varianza Dentro y Entre.