

Clase 2:

Estadística Bivariada

Raúl Bastidas

Estadística Descriptiva Bivariada

Supongamos que tenemos dos variables estadísticas X e Y . Además, supongamos que tomamos una muestra aleatoria de tamaño n y que dividimos la variable X en r clases A_i , $i = 1, \dots, r$ y la variable Y en s clases B_j , $j = 1, \dots, s$.

Observación

La toma de la muestra se hace de manera conjunta.

Definición: Se define la frecuencia absoluta conjunta n_{ij} de la modalidad $A_i B_j$ por la cantidad de individuos que pertenecen a la clase A_i y B_j , al mismo tiempo.

Definición Se define la frecuencia relativa conjunta f_{ij} de la modalidad $A_i B_j$ por:

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

Observación

De las definiciones se puede ver que:

$$\textcircled{C} \sum_{i=1}^r \sum_{j=1}^s n_{ij} = n$$

$$\textcircled{C} \sum_{i=1}^r \sum_{j=1}^s f_{ij} = 1$$

Tabla de Contingencia

Para trabajar con datos agrupados se puede construir una tabla de contingencia.

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

Definición Se definen las frecuencias absolutas marginales por:

$$n_{i.} = \sum_{j=1}^s n_{ij} \quad i = 1, \dots, r$$

$$n_{.j} = \sum_{i=1}^r n_{ij} \quad j = 1, \dots, s$$

Definición Se definen las frecuencias relativas marginales por:

$$f_{i.} = \frac{n_{i.}}{n} \quad i = 1, \dots, r$$

$$f_{.j} = \frac{n_{.j}}{n} \quad j = 1, \dots, s$$

Observación

Las frecuencias marginales nos sirven para recuperar la información de las variables X e Y .

Definición Se definen las frecuencias relativas condicionales por:

$$f_{i/j} = \frac{f_{ij}}{f_{.j}} = \frac{n_{ij}}{n_{.j}} \quad j = 1, \dots, s$$

Observación

Hasta ahora se han definido tres tipos de frecuencias, estas son: conjunta, marginal y condicional.

Definición: Suponga que se han muestreado de forma conjunta n individuos. Suponga que m_{X_i} es la marca de clase del intervalo i -ésimo para la variable X y m_{Y_j} es la marca de clase del intervalo j -ésimo para la variable Y . Entonces se definen:

1. Promedio marginal de X :

$$\overline{X} = \sum_{i=1}^r f_{i.} m_{X_i}$$

2. Promedio marginal de Y :

$$\overline{Y} = \sum_{j=1}^s f_{.j} m_{Y_j}$$

3. Varianza marginal de X :

$$s_X^2 = \sum_{i=1}^r f_{i.}(m_{X_i} - \bar{X})^2$$

4. Varianza marginal de Y :

$$s_Y^2 = \sum_{j=1}^r f_{.j}(m_{Y_j} - \bar{Y})^2$$

5. Promedio de X condicionada a B_j :

$$\overline{X_j} = \sum_{i=1}^r f_{i/j} m_{X_i}$$

6. Promedio de Y condicionada a A_i :

$$\overline{Y}_i = \sum_{j=1}^s f_{j/i} m_{Y_j}$$

7. Varianza de X condicionada a B_j :

$$v_j(X) = \sum_{i=1}^r f_{i/j} (m_{X_i} - \overline{X}_j)^2$$

8. Varianza de Y condicionada a A_i :

$$v_i(Y) = \sum_{j=1}^s f_{j/i} (m_{Y_j} - \overline{Y}_i)^2$$

9. Descomposición de la varianza de X :

$$V(X) = \sum_{j=1}^s f_{.j} V_j(X) + \sum_{j=1}^s f_{.j} (\bar{X}_j - \bar{X})^2$$

10. Descomposición de la varianza de Y :

$$V(Y) = \sum_{i=1}^r f_{i.} V_i(Y) + \sum_{i=1}^r f_{i.} (\bar{Y}_i - \bar{Y})^2$$

Observación

Las medidas de forma también pueden ser calculadas a partir de una tabla de contingencia.

Covarianza

Definición Sean X e Y dos variables estadísticas. Supongamos que disponemos de una muestra $\{(X_i, Y_i)\}_{i=1}^n$. Se define la covarianza entre X e Y por:

$$\text{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

Observación

- © Si $\text{Cov}(X, Y) \geq 0$ entonces las variables están directamente asociadas.
- © Si $\text{Cov}(X, Y) \leq 0$ entonces las variables están inversamente asociadas.
- © Si $\text{Cov}(X, Y) = 0$ entonces no tienen una asociación lineal.

Observación

- © Se puede probar que: $\text{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^n X_i Y_i - \bar{X} \bar{Y}$
- © Si $\alpha, \beta \in \mathbb{R}^+$, entonces $\text{Cov}(\alpha X, \beta Y) = \alpha \beta \text{Cov}(X, Y)$
- © Si los datos están agrupados en una tabla de contingencia, entonces la covarianza se puede calcular por:

$$\text{Cov}(X, Y) = \sum_{i=1}^r \sum_{j=1}^s f_{ij} (m_{X_i} - \bar{X})(m_{Y_j} - \bar{Y})$$

- © También se puede probar que para datos agrupados se cumple que:

$$\text{Cov}(X, Y) = \sum_{i=1}^r \sum_{j=1}^s f_{ij} m_{X_i} m_{Y_j} - \bar{X} \bar{Y}$$

Correlación

Definición Sean X e Y dos variables estadísticas. Supongamos que disponemos de una muestra $\{(X_i, Y_i)\}_{i=1}^n$. Se define la correlación entre X e Y por:

$$r = \frac{\text{Cov}(X, Y)}{\sqrt{S_X^2 S_Y^2}}$$

Observación

- © $|r| \leq 1$
- © Si $r = 1$ entonces hay asociación lineal directa perfecta.
- © Si $r = -1$ entonces hay asociación lineal inversa perfecta.
- © Si $r = 0$ entonces no tienen una asociación lineal.

Regresión Lineal Simple

Consideremos la variable Y una variable a ser explicada (variable respuesta) y X una variable predictora (Explicativa o independiente). Supongamos que disponemos de los datos:

$$\{(X_i, Y_i)\}_{i=1}^n,$$

entonces se plantea el siguiente modelo de regresión lineal:

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

donde β_0 y β_1 son parámetros a ser modelados en función de los datos y ϵ_i es un error aleatorio.

Regresión Lineal Simple

El cuadrado de las distancias entre los puntos observados y los puntos que provee el modelo viene dado por:

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

Entonces se quiere la mejor recta tal que minimice $g(\beta_0, \beta_1)$. Esto es equivalente a resolver:

$$\frac{\partial g(\beta_0, \beta_1)}{\partial \beta_0} = 0$$

$$\frac{\partial g(\beta_0, \beta_1)}{\partial \beta_1} = 0$$

Teorema Dado un conjunto de datos en el plano $\{(X_i, Y_i)\}_{i=1}^n$ y el modelo de regresión lineal anterior. Entonces la mejor recta se puede representar por:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Donde:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\beta}_1 = \frac{\text{Cov}(X, Y)}{S_X^2}$$

Definición: Se define el coeficiente de determinación del modelo de regresión lineal por:

$$R^2 = \frac{\text{Cov}(X, Y)^2}{S_X^2 S_Y^2} = r^2$$

Observación

- © $0 \leq R^2 \leq 1$
- © El coeficiente R^2 determina la calidad del modelo para replicar los resultados.