

Estadística Descriptiva Bivariada

Dr. Cristian Carvajal Muquillaza

Universidad de Playa Ancha



Universidad de
Playa Ancha
Facultad de Ciencias
Naturales y Exactas

**DEPARTAMENTO
DISCIPLINARIO
DE MATEMÁTICA
Y ESTADÍSTICA**



Contenido

- 1 Preliminares
- 2 Medidas de Dependencia
- 3 Medidas de Asociación
- 4 Modelo de Regresión
- 5 Inferencia en el Modelo de Regresión



Universidad de
Playa Ancha
Facultad de Ciencias
Naturales y Exactas

DEPARTAMENTO
DISCIPLINARIO
DE MATEMÁTICA
Y ESTADÍSTICA

Estadística Descriptiva Bivariada

Estadística Descriptiva Bivariada

La **Estadística Descriptiva Bivariada** aborda el estudio de los sucesos en los que intervienen dos variables simultáneamente, por ejemplo:

- Edad y Género,
- Escolaridad e Ingreso,
- Peso y Estatura,
- Estado civil y Nacionalidad,



Estadística Descriptiva Bivariada

Tabla de Contingencia

Sean las variables A y B , tal que, $A = a_1, \dots, a_I$ y $B = b_1, \dots, b_J$, entonces:

		Variable B					
		b_1	...	b_j	...	b_J	
Variable A	a_1	n_{11}	...	n_{1j}	...	n_{1J}	$n_{1.}$
	...						
	a_i	n_{i1}	...	n_{ij}	...	n_{iJ}	$n_{i.}$
	...						
	a_I	n_{I1}	...	n_{Ij}	...	n_{IJ}	$n_{I.}$
		$n_{.1}$		$n_{.j}$		$n_{.J}$	$n_{..}$

Estadística Descriptiva Bivariada

Tabla de Contingencia: **Distribuciones Marginales**

Sean las variables A y B , tal que, $A = a_1, \dots, a_I$ y $B = b_1, \dots, b_J$, entonces:

1 Variable A:

$$n_{i.} = \sum_{j=1}^J n_{ij}$$

2 Variable B:

$$n_{.j} = \sum_{i=1}^I n_{ij}$$



Estadística Descriptiva Bivariada

Tabla de Contingencia: **Distribuciones Condicionales**

Sean las variables A y B , tal que, $A = a_1, \dots, a_I$ y $B = b_1, \dots, b_J$, entonces:

- 1 Variable A condicionada a $B = B_J$:

$$n_{i|j} = n_{ij} | n_{.j}$$

- 2 Variable B condicionada a $A = A_I$:

$$n_{j|i} = n_{ij} | n_{i.}$$



Medidas de Dependencia: Variables Cualitativas

Contraste de Independencia Chi-Cuadrado:

Suponga una tabla de contingencia $I \times J$, y suponga que se quiere contrastar la Hipótesis Nula de Independencia, es decir:

$$H_0 : p_{ij} = p_{i.} \cdot p_{.j} \quad V/S \quad H_1 : p_{ij} \neq p_{i.} \cdot p_{.j}$$

Para contrastar la hipótesis utilizamos la cantidad pivotal de Pearson, que se define:

$$X^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}$$

Donde

$$\hat{m}_{ij} = \frac{n_{i.} \times n_{.j}}{n}$$

Medidas de Dependencia: Variables Cualitativas

Contraste de Independencia Chi-Cuadrado:

Bajo la Hipótesis Nula el estadístico X^2 se distribuye asintóticamente Chi-Cuadrado con $(I - 1) \times (J - 1)$ grados de libertad, es decir:

$$X^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}} \sim \chi^2_{(I-1) \times (J-1)}$$

Entonces se rechaza la Hipótesis Nula de Independencia al nivel α , cuando se verifica

Se Rechaza H_0

$$X_{obs}^2 \geq \chi^2_{\alpha; (I-1) \times (J-1)}$$

Medidas de Asociación: Variables Cuantitativas

Covarianza:

Sean $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ una muestra de n datos bivariados, entonces la covarianza es un estadístico que permite medir la fuerza de asociación entre las variables **X** e **Y**, y se define:

$$Cov(X, Y) = S_{XY} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

- Si $S_{XY} > 0 \Rightarrow$ Asociación directa
- Si $S_{XY} < 0 \Rightarrow$ Asociación inversa
- Si $S_{XY} = 0 \Rightarrow$ No existe Asociación

Medidas de Asociación: Variables Cuantitativas

Covarianza: Propiedades

Sean $\mathbf{X}$, $\mathbf{Y}$ variables aleatorias y $\mathbf{a}$, $\mathbf{b}$ constantes, entonces:

- 1 $Cov(X, a) = 0$
- 2 $Cov(X, X) = Var(X)$
- 3 $Cov(X, Y) = Cov(Y, X)$
- 4 $Cov(aX, bY) = abCov(X, Y)$
- 5 $Cov(X + a, Y + b) = Cov(X, Y)$



Medidas de Asociación: Variables Cuantitativas

Coeficiente de Correlación de Pearson:

Sean $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ una muestra de n datos bivariados, entonces el Coeficiente de Correlación de Pearson r_{xy} es un estadístico que mide el grado de asociación lineal entre dos variables cuantitativas, y se define:

$$r_{xy} = \frac{S_{xy}}{S_x \cdot S_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}}$$

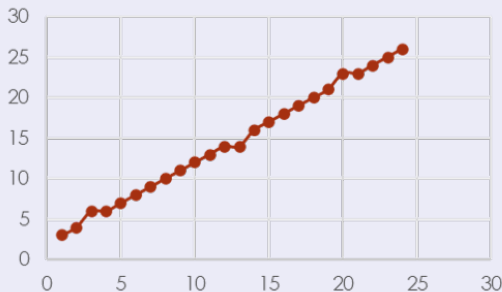
Obs: $-1 \leq r_{xy} \leq 1$



Medidas de Asociación: Variables Cuantitativas

Coeficiente de Correlación de Pearson:

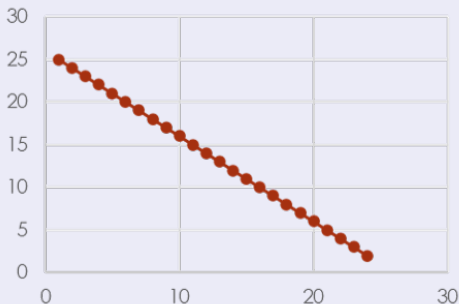
Si $r_{xy} = 1$ entonces existe una correlación lineal directa perfecta entre **X** e **Y**.



Medidas de Asociación: Variables Cuantitativas

Coeficiente de Correlación de Pearson:

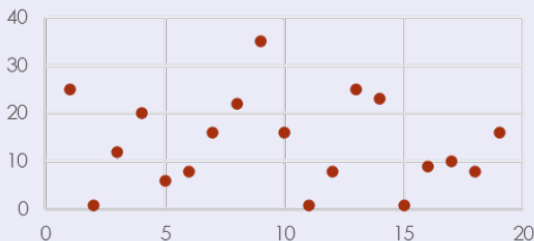
Si $r_{xy} = -1$ entonces existe una correlación lineal inversa perfecta entre **X** e **Y**.



Medidas de Asociación: Variables Cuantitativas

Coeficiente de Correlación de Pearson:

Si $r_{xy} = 0$ entonces No existe una correlación lineal entre **X** e **Y**.



Facultad de Ciencias
Naturales y Exactas

AMBITO
NARIO
DE MATEMÁTICA
Y ESTADÍSTICA

Medidas de Asociación: Variables Cuantitativas

Coefficiente de Correlación de Pearson:

Es poco probable que r_{xy} tome exactamente uno de estos tres valores $-1, 0, 1$, lo que se da en la práctica es que tome un valor que pertenezca al intervalo $[-1, 1]$, entonces:



Medidas de Asociación: Variables Cuantitativas

Ejemplo:

Se quiere estudiar la relación existente entre el grado de percepción de los estudiantes del clima escolar y el rendimiento en la asignatura de matemática, los resultados de la encuesta se ve en la tabla adjunta:

Percepción	10	8	4	12	7	8	15	3
Nota	5.2	5.0	3.4	6.5	5.0	6.0	6.8	4.5

- 1 Construir una gráfica de dispersión de los datos,
- 2 Calcular e interpretar r_{xy} .



Modelo de Regresión

Modelo de Regresión Lineal Simple

El modelo de regresión lineal simple tiene la siguiente expresión:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

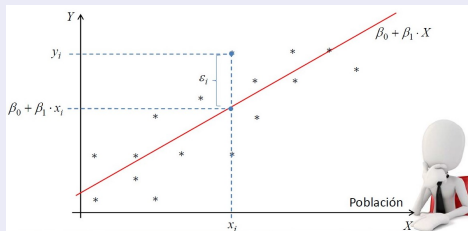
En donde β_0 es la ordenada en el origen, β_1 es la pendiente de la recta y ε es una variable que incluye un conjunto grande de factores, cada uno de los cuales influye en la respuesta sólo en pequeña magnitud, a la que llamaremos error.



Modelo de Regresión

Modelo de Regresión Lineal Simple

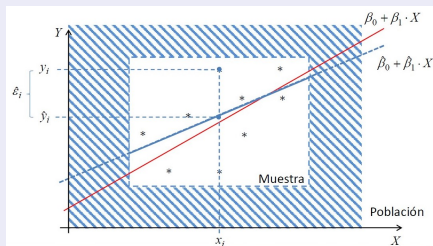
Gráficamente, un modelo de regresión lineal se representa como una recta, tal como se observa en la figura. Asumiremos que se trata de una representación á nivel poblacional.



Modelo de Regresión

Modelo de Regresión Lineal Simple

En la práctica no se puede observar el modelo poblacional, sino que sólo una parte de éste a través de una muestra, lo cual llamaremos modelo estimado.



Modelo de Regresión

Estimación del Modelo de Regresión Lineal Simple

Para estimar el modelo, debemos estimar los coeficientes β_0 y β_1 , para esto utilizaremos el método de estimación denominado “Mínimos Cuadrados Ordinarios”(MCO), el cual intenta minimizar la suma de los cuadrados de los errores (ε_i^2).

Asumiendo que los ε_i son v.a.i.i.d con $\mathbb{E}[\varepsilon_i] = 0$ y $Var[\varepsilon_i] = \sigma^2$, para todo i . Entonces, considerando $i = 1, \dots, n$:

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i$$



Modelo de Regresión

Estimación del Modelo de Regresión Lineal Simple

Se plantea el siguiente proceso de optimización:

$$\begin{aligned}\hat{\beta}_0, \hat{\beta}_1 &= \arg \min \left(\sum_{i=1}^n \varepsilon_i^2 \right) \\ &= \arg \min \left(\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \right) \\ &= \arg \min \left(\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \right) \\ &= \arg \min \Psi\end{aligned}$$

$$\sum_{i=1}^n 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-1) = 0$$

$$\sum_{i=1}^n y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = 0$$

$$\sum_{i=1}^n y_i - \sum_{i=1}^n \hat{\beta}_0 - \sum_{i=1}^n \hat{\beta}_1 x_i = 0$$

$$n\bar{y} - n\hat{\beta}_0 - n\hat{\beta}_1\bar{x} = 0$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$



Universidad de
Playa Ancha
Facultad de Ciencias
Naturales y Exactas

DEPARTAMENTO
DISCIPLINARIO
DE MATEMÁTICA
Y ESTADÍSTICA

Modelo de Regresión

$$\sum_{i=1}^n 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-x_i) = 0$$

$$\sum_{i=1}^n x_i Y_i - x_i \hat{\beta}_0 - \hat{\beta}_1 x_i^2 = 0$$

$$\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \hat{\beta}_0 - \sum_{i=1}^n \hat{\beta}_1 x_i^2 = 0$$

$$\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 \sum_{i=1}^n x_i^2 = 0$$



Modelo de Regresión

$$\hat{\beta}_1 = \frac{n\bar{x}\bar{y} - \sum_{i=1}^n x_i y_i}{n\bar{x}^2 - \sum_{i=1}^n x_i^2}$$



Universidad de Playa Ancha

Inferencia en el Modelo de Regresión

Intervalo de confianza para β_0

A partir del pivote, se plantea

$$P \left(-t_{\alpha/2} \leq \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x - \bar{x})^2} \right]}} \leq t_{\alpha/2} \right) = 1 - \alpha$$



Inferencia en el Modelo de Regresión

Intervalo de confianza para β_0

Despejando β_0 , tenemos

$$I.C_{(1-\alpha)100}(\beta_0) = \hat{\beta}_0 \pm t_{(\alpha/2;n-2)} \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x - \bar{x})^2} \right]}$$

es decir, el $I.C_{(1-\alpha)100}(\beta_0)$ está dado por:

$$\left[\hat{\beta}_0 - t_{\left(\frac{\alpha}{2};n-2\right)} \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x - \bar{x})^2} \right]}, \hat{\beta}_0 + t_{\left(\frac{\alpha}{2};n-2\right)} \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x - \bar{x})^2} \right]} \right]$$

Inferencia en el Modelo de Regresión

Intervalo de confianza para β_1

A partir del pivote, se plantea

$$P\left(-t_{\alpha/2} \leq \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum (x - \bar{x})^2}}} \leq t_{\alpha/2}\right) = 1 - \alpha$$



Inferencia en el Modelo de Regresión

Intervalo de confianza para β_0

Despejando β_1 , tenemos

$$I.C_{(1-\alpha)100}(\beta_1) = \hat{\beta}_1 \pm t_{(\alpha/2;n-2)} \sqrt{\frac{\hat{\sigma}^2}{\sum (x - \bar{x})^2}}$$

es decir, el $I.C_{(1-\alpha)100}(\beta_1)$ está dado por:

$$\left[\hat{\beta}_1 - t_{\left(\frac{\alpha}{2};n-2\right)} \sqrt{\frac{\hat{\sigma}^2}{\sum (x - \bar{x})^2}}; \hat{\beta}_1 + t_{\left(\frac{\alpha}{2};n-2\right)} \sqrt{\frac{\hat{\sigma}^2}{\sum (x - \bar{x})^2}} \right]$$

Inferencia en el Modelo de Regresión

Intervalo de confianza para σ^2

Despejando σ^2 , tenemos

$$I.C_{(1-\alpha)100}(\sigma^2) = \left[\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\chi^2_{(\alpha/2; n-2)}}, \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\chi^2_{(1-\alpha/2; n-2)}} \right]$$



Análisis de varianza del MRLS

Consideremos el test

$$H_0 : \beta_1 = 0 \quad v/s \quad H_1 : \beta_1 \neq 0$$

Estas hipótesis prueban la contribución de la *variable regresora* X para explicar la respuesta Y , una vez que H_0 sea verdadera, esto es $\beta_1 = 0$, esa contribución no es significativa. En ese caso o el MRLS no es adecuado o la variable regresora escogida no contribuye a la explicación de la variable Y .



Análisis de varianza del MRLS

Así para un test de nivel α rechazamos H_0 , si

$$\frac{\hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2}{\hat{\sigma}^2} >_1 F_{(n-2)}(\alpha),$$

del cuantil $(1 - \alpha)$ de la distribución F con 1 y $(n - 2)$ grados de libertad.



ANOVA

Definiciones

- $\sum_{i=1}^n (y_i - \bar{y})^2$: **Suma de cuadrados totales (SQT)**, Variación total de Y en torno a su media.
- $\sum_{i=1}^n (y_i - \hat{y})^2$: **Suma de cuadrados del error (SQE)**, Variación total de Y en torno a la recta.
- $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$: **Suma de cuadrados de la regresión (SQR)**, Variación de las esperanzas específicas de Y, dado x, en torno a su media.

TABLA ANOVA

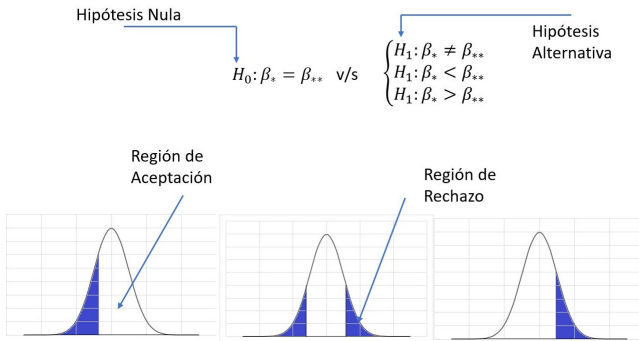
- 1 Primero calculamos SQT .
- 2 Calculamos SQR .
- 3 Obtenemos la SQE por sustracción: $SQT - SQR$.
- 4 En la cuarta columna hacemos la división de la suma de los cuadrados por los grados de libertad.
- 5 En la quinta columna obtenemos el valor de la estadística de prueba, .

Así rechazamos $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$, si $F_0 >_1 F_{n-2;1-\alpha}$, de la distribución F .



Inferencia en el Modelo de Regresión

Test de Hipótesis



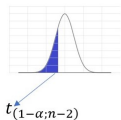
Inferencia en el Modelo de Regresión

Test de Hipótesis para β_0

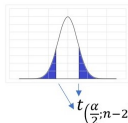
$$H_0: \beta_0 = \beta_{00}$$

v/s

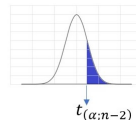
$$H_1: \beta_0 < \beta_{00}$$



$$H_1: \beta_0 \neq \beta_{00}$$



$$H_1: \beta_0 > \beta_{00}$$



$$T = \frac{\widehat{\beta}_0 - \beta_{00}}{\sqrt{\widehat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x - \bar{x})^2} \right]}}$$

ias
as

DEPARTAMENTO
DISCIPLINARIO
DE MATEMÁTICA
Y ESTADÍSTICA

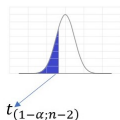
Inferencia en el Modelo de Regresión

Test de Hipótesis para β_1

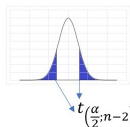
$$H_0: \beta_1 = \beta_{10}$$

v/s

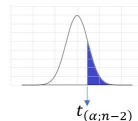
$$H_1: \beta_1 < \beta_{10}$$



$$H_1: \beta_1 \neq \beta_{10}$$



$$H_1: \beta_1 > \beta_{10}$$



$$T = \frac{\widehat{\beta}_1 - \beta_{10}}{\sqrt{\frac{\widehat{\sigma}^2}{\sum (x - \bar{x})^2}}}$$

as
is

DEPARTAMENTO
DISCIPLINARIO
DE MATEMÁTICA
Y ESTADÍSTICA

Modelo de Regresión

Ejemplo: *Modelo de Regresión Lineal simple*

Se quiere realizar un estudio sobre la relación entre el tiempo diario dedicado al estudio y el rendimiento en la asignatura de matemática, los resultados de la encuesta son los siguientes

Tiempo	2,5	3,1	4,7	1,5	3,6	5,3
Rendimiento	3,5	4,1	5,3	2,8	3,8	5,4

- Construir una gráfica de dispersión de los datos,
- Calcular e interpretar r_{xy} ,
- Determine el modelo de regresión lineal,

