Estadística II

Introducción - Contenido del curso

Kevin Pérez - Ingeniero de Sistemas - Estadístico, (E) Maestria en Ciencia de Datos.
Universidad Cooperativa de Colombia.

Capítulo I - Correlación

1. Análisis de Correlación
2. Diagramas de Dispersión
3. Correlación de Pearson
4. Correlación de Spearman
5. Relación de variables correlacionadas
6. Aplicaciones generales y en sicología.

Capítulo II - Regresión

1. Análisis regresional
2. Tipos de Variables (dependientes e Independientes)
3. Tipos de Regresión
4. Regresión Lineal Simple
5. Aplicaciones del modelo de Regresión Lineal Simple (MCO) (Verosimilitud)
6. Homocedasticidad
7. Aplicaciones generales y en sicología.

Bibliografía – Referenciación

- Ferris J. Ritchey. Estadísticas para las Ciencias Sociales (El potencial de la imaginación estadística).. Editorial Mc Graw Hill. 2007
- Hernández Arroyo, Emil. Manual de Estadística: Handbook of Statistics. Bogotá. Universidad Cooperativa de Colombia. 2006
- Horra Navarro, Julián. Estadística aplicada. Madrid. Editorial Diaz de Santos. 2003
- Levin Jack, Fundamentos de estadística en la investigación social 2da edición, Harla. 1992
- Martínez Bencardino, Ciro. Estadística y Muestreo. Ecoe Ediciones. Bogotá. 2003

Capítulo I - Correlación

En este capítulo se ustudiará una de las más importantes y útiles herramientas del análisis estadístico: La `Correlación`, esta técnica ilustra la forma como pueden analizarse las relaciones entre dos variables para encontrar patrones que conlleven al planteamiento de hipótesis que luego se deban verificar.

Los términos ***asociación***, ***correlación***, ***contingencia***, ***concordancia*** y otros similares, se suelen utilizar como equivalentes muy a menudo. No obstante, haciendo un uso más correcto de la terminología estadística, aún con significado semejante, se puede considerar:

- correlación de variables propiamente dichas, o sea, medidas en escala de intervalo.
- concordancia de ordenaciones, entendiéndose como tales las denominadas variables ordinales, y
- asociación o contingencia de variables nominales o atributos.

Capítulo I - Covarianza

La `covarianza` es una medida de variabilidad conjunta entre un par de variables X y Y medidas sobre un conjunto de n individuos:

$$S_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n-1}$$

Este coeficiente juega un importante papel en el estudio de la relación lineal entre las variables, S_{xy} es una medida simétrica y se puede leer como la suma de los productos de las desviaciones de X por las desviaciones de Y con respecto a sus medias respectivas.

- Se puede afirmar que la covarianza detecta la relación lineal entre las variables y el sentido de ésta, pero no distingue entre la no presencia de relación

- Aún para el estudio de relaciones lineales la covarianza adolece de ciertos problemas, como el de venir acompañada de las unidades de las variables y el de depender del número de observaciones.

Capítulo I - Coeficiente de Pearson

Para obviar las carencias de la covarianza se introduce el coeficiente de correlación lineal o ***coeficiente de correlación de Pearson.*** Suponga dos variables aleatorias X y Y, el coeficiente decorrelación lineal es definido por:

En general, disponemos de muestras de pares de datos $(x_1, y_1), \ldots, (x_n, y_n)$ y se define el coeficiente de correlación *muestral (estimador)* como:

$$r_{xy} = \frac{S_{xy}}{S_x S_y}$$

- Con $-1 \leq r_{xy} \leq 1$. Que es una medida adimensional, ordinal y tiene el signo de S_{xy} por lo que cuando la relación lineal entre X e Y es exacta y directa, es decir, todos los puntos se encuentran sobre una recta con pendiente positiva, vale 1, cuando es exacta e inversa, es decir, todos los puntos se encuentran sobre una recta con pendiente negativa, vale −1 y cuando no hay relación lineal 0; con un análisis lógico para las posiciones intermedias. Cuando r vale cero, se dice que las variables están *no correlacionadas.*

Capítulo I - Coeficiente de correlación de Spearman

Este coeficiente se utiliza para medir la relación entre dos sucesiones de valores ordinales. Es el coeficiente de *correlación de Pearson* para las llamadas variables *cuasi–cuantitativas*, discretas, o bien, para aquellas cuantitativas que han sido transformadas en ordinales (n primeros números naturales para cada variable) tiene la forma

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

Donde:

- r_s es el coeficiente de correlación por rangos de Spearman
- d_i es la diferencia entre el valor ordinal de la variable X y el de la variable Y en el elemento i-é$simo$
- n es el tamaño de la muestra

Se verifica para este coeficiente como en el anterior que $-1 \leq r_s \leq 1$

Capítulo I - Coeficiente de Spearman

Si hay un gran número de elementos con el mismo valor en alguna de las dos variables, es decir, si hay muchos empates, es conveniente recurrir a las correcciones de este coeficiente. Quedando el coeficiente como

$$r_s = \frac{x^2 + y^2 - \sum_{i=1}^{n} d_i^2}{2\sqrt{x^2 y^2}}$$

Con:

$$x^2 = \frac{n^3 - 3}{12} - \sum_{i=1}^{n} T_{x_i}, \qquad T_{x_i} = \frac{t_{x_i}^3 - t_{x_i}}{12}$$

$$y^2 = \frac{n^3 - 3}{12} - \sum_{i=1}^{n} T_{y_i}, \qquad T_{y_i} = \frac{t_{y_i}^3 - t_{y_i}}{12}$$

Capítulo I - Coeficiente de Spearman

Donde:

- t_{x_i} es el número de empates en el rango i de la variable X
- t_{y_i} es el número de empates en el rango i de la variable Y

Sus características e interpretación son similares a las del coeficiente de correlación de Pearson.

Capítulo I