MA34B – Estadística
Asociación Entre Variables

Prof. Rodrigo Abt B.

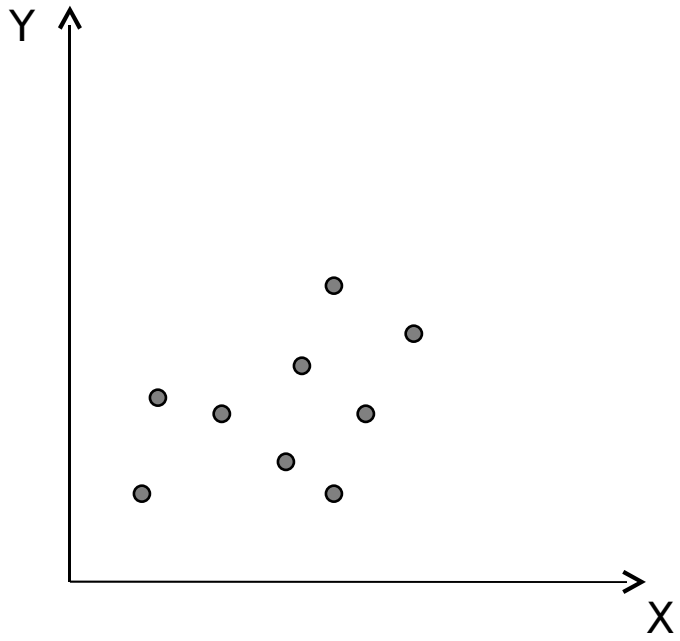
rabt@dim.uchile.cl

Introducción

- Hasta el momento nos hemos centrado en el estudio estadístico de una sola variable; sin embargo, en la realidad es usual encontrarse con problemas que involucran más de una.
- El investigador entonces puede tratar de determinar la relación que existe entre ellas, y la clase de la misma.
- Una asociación entre dos variables es la expresión de la influencia que ejercen los cambios de una variable en los cambios de la otra.
- Es importante distinguir entre asociación y causalidad. En el primer caso, si no se conoce la naturaleza del problema, es muy difícil hacer aseveraciones que tengan un sentido real aplicable.
- A diferencia de la asociación, la causalidad requiere de un juicio de valor respecto del fenómeno observado y los cambios experimentados por las variables.

Caso: Dos Variables Cuantitativas

- Haciendo uso de los recursos del cálculo, una manera sencilla de observar un patrón de asociación entre dos variables cuantitativas es graficándolas:



El Coeficiente De Correlación Lineal

- Una forma de medir si existe “dependencia” entre observaciones es utilizar la covarianza muestral, la cual se define como:

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

- El problema de la covarianza como medida de asociación es que es sensible a la escala y unidades de medida de las variables involucradas.
- Para descontar ese efecto, se procede a dividir por las respectivas desviaciones estándar muestrales de ambas variables, obteniéndose así el coeficiente de correlación lineal:

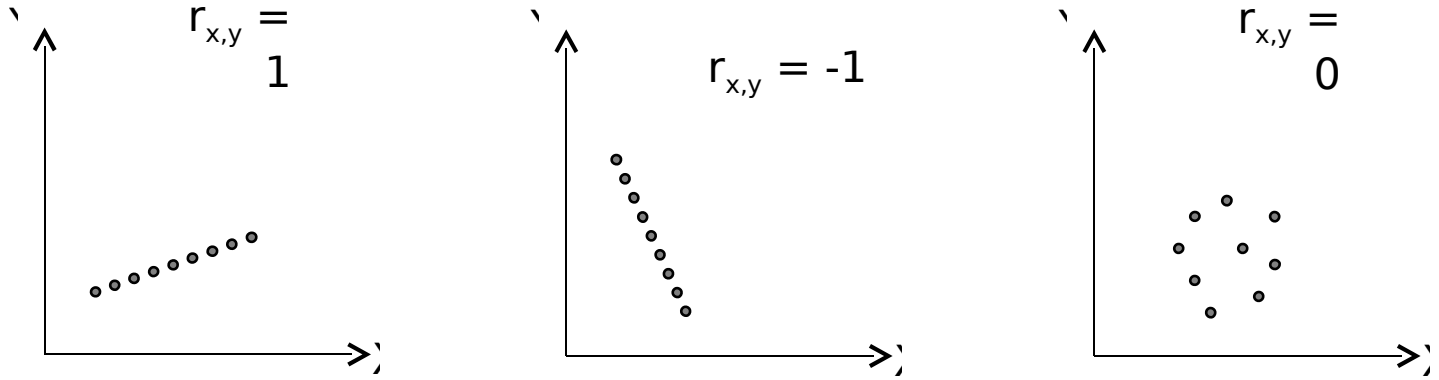
$$r_{x,y} = \frac{Cov(x,y)}{S_x S_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Propiedades

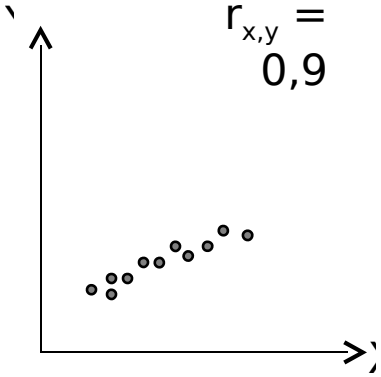
- El coeficiente de correlación es adimensional
- $|r_{x,y}|<=1$
- Si $r_{x,y}=1$, entonces existe una relación estrictamente lineal con pendiente positiva.
- Si $r_{x,y}=-1$, entonces existe una relación estrictamente lineal con pendiente negativa.
- Si $r_{x,y} \rightarrow 1$, entonces existe tendencia lineal positiva.
- Si $r_{x,y} \rightarrow -1$, entonces existe tendencia lineal negativa.
- Si Si $r_{x,y}=0$, entonces no existe tendencia **LINEAL**

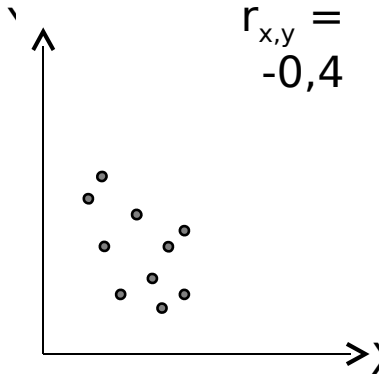
Importante (1)

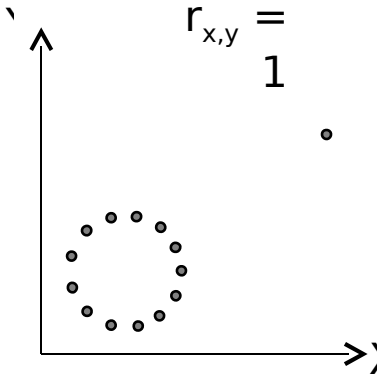
- Es muy importante tener cuidado con la interpretación del coeficiente de correlación lineal, ya que un dato atípico o un patrón diferente puede producir resultados equívocos.
- Ejemplos:

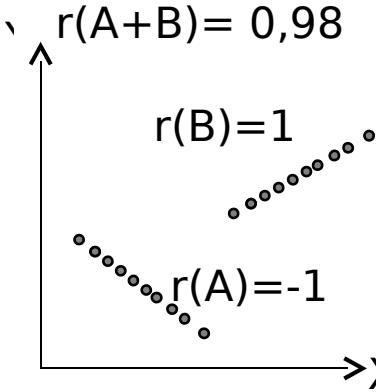


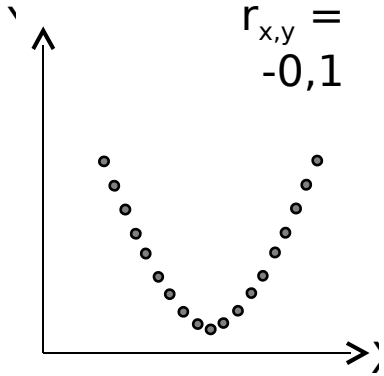
Importante (2)

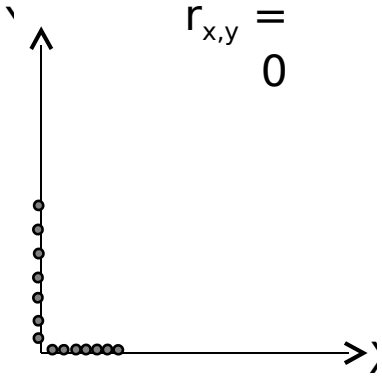












Caso Multivariado

- Cuando se tienen más de dos variables, las correlaciones se presentan en un matriz R denominada “Matriz de Correlaciones”, en que cada término r_{ij} representa la correlación entre la variable i y la j.

$$R = \begin{pmatrix} 1 & r_{12} & \dots & r_{1p} \\ \vdots & \vdots & \dots & \vdots \\ r_{p1} & r_{p2} & \dots & 1 \end{pmatrix}$$

- Esta matriz tiene unos en la diagonal.
- Es cuadrada, simétrica y semidefinida positiva

Comentarios

- Cuando se tienen dudas sobre una relación lineal, se puede recurrir a dos estrategias:
  - Utilizar la correlación lineal entre f(X) e Y, en que f representa la relación funcional que se pretende estudiar.
  - Utilizar otro indicador más general que indique existencia de relación funcional.

Caso: Una Cualitativa y Una Cuantitativa

- Cuando una de las variables es nominal u ordinal, el coeficiente de correlación ya no es aplicable, por lo que es necesario un indicador alternativo para el estudio de la relación entre las mismas.
- Sea X una variable cualitativa con “q” categorías o modalidades, e Y una variable cuantitativa.
- Podemos entonces clasificar las observaciones de Y de acuerdo a la modalidad que toman para X.
- De acuerdo a lo anterior, podemos llamar y_{1j}, y_{2j},..., y_{njj} a las observaciones toma la variable Y sobre la modalidad “j” de X.

Variabilidad De Las Observaciones (1)

- La variabilidad total de la variable Y se puede obtener como:

$$S_y^2 = \frac{1}{n} \sum_{j=1}^{q} \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2$$

- Además, como las observaciones de Y se clasifican en grupos, es posible encontrar la variabilidad dentro de cada grupo:

$$w_j^2 = \frac{1}{n_j} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$$

Variabilidad De Las Observaciones (2)

- Además podemos calcular la media ponderada de las variabilidades intra-grupo como:

$$w^2 = \sum_{j=1}^{q} \frac{n_j}{n} w_j^2$$

- La variabilidad total ENTRE los grupos se puede calcular como:

$$b^2 = \sum_{j=1}^{q} \frac{n_j}{n} (\bar{y}_j - \bar{y})^2$$

Identidad Fundamental

- Es posible demostrar que la variabilidad total de las observaciones es igual a la suma de la variabilidad media ponderada intra-grupos y la variabilidad entre-grupos:

$$S_y^2 = b^2 + w^2$$

- Si w^2 es 0, es significa que todos los w_j^2=0, y por ende todas las observaciones son iguales a la media del grupo. Por lo tanto, se puede predecir el valor de cualquier observación Y conociendo la modalidad. En este caso existirá una relación funcional entre X e Y.
- Por otro lado, si b^2 es 0, los promedios de cada grupo son iguales al promedio general, y por ende, no se puede conocer el valor de Y con la modalidad de X, es decir, no hay relación funcional de X e Y.
- Con esto se deduce un índice para medir el grado de relación funcional:

$$\eta^2_{Y|X} = \frac{b^2}{S_y^2}$$

Propiedades

- Este coeficiente toma valores entre 0 y 1.
- Si $\eta_{Y|X}^2 = 1$ hay relación funcional estricta.
- Si $\eta_{Y|X}^2 = 0$ no hay relación entre X e Y.

Análisis de Varianza (caso efectos fijos)(1)

- Retomemos el problema anterior, en que tenemos un conjunto de n observaciones de una v.a., Y repartida en q categorías. Y_{ij} representa entonces la i-ésima (i=1,...,nj) observación en la categoría j (j=1,...,q).
- Sea

$$T = \sum_{j=1}^{q} \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2$$

una medida la variabilidad total de las observaciones.

- La variabilidad al interior de los grupos se puede escribir como:

$$W = \sum_{j=1}^{q} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$$

Análisis de Varianza (caso efectos fijos)(2)

- La variabilidad aportada por los grupos:

$$B = \sum_{j=1}^{q} n_j (\bar{y}_j - \bar{y})^2$$

- Se puede comprobar entonces que T = B + W, es decir, que la variabilidad aportada intra-grupos más la variabilidad aportada entre grupos es igual a la variabilidad total de las observaciones.
- Si suponemos normalidad de las observaciones, i.e., $Y_{ij} \sim N(\mu_j, \sigma^2)$, entonces se puede mostrar que:

$$\frac{W}{\sigma^2} \sim \chi^2_{n-q} \qquad \frac{B}{\sigma^2} \sim \chi^2_{q-1}$$

- Entonces se tiene que:

$$F = \frac{B/(q-1)}{W/(n-q)} \sim F_{q-1,n-q}$$

El Estadístico F-Fisher

- En este caso el estadístico F sirve para contrastar la hipótesis de que no hay diferencias entre los grupos, es decir, se contrasta $H_0 : \mu_j = \mu \; \forall$ j, es decir, que todas las medias teóricas son iguales.
- Si $F > F_{q-1,n-q}(\alpha)$ entonces se rechaza H_0 al nivel de significación dada.

Ejemplo

- Supongamos que queremos ver si existen diferencias de absorción de humedad en 5 mezclas de concreto. Para ello se toman seis muestras para cada tipo de mezcla y se tabulan los resultados como se muestra a continuación.

	m1	m2	m3	m4	m5
	551	595	639	417	563
	457	580	615	449	631
	450	508	511	517	522
	731	583	573	438	613
	499	633	648	415	656
	632	517	677	555	679
Medias	553,33	569,33	610,5	465,17	610,67

- Deseamos entonces testear:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_5$$

H_1 : Al menos dos de las medias no son iguales

La Tabla ANOVA

- Para calcular el estadístico F que utizaremos para el contraste es útil construir una tabla ANOVA:

Columna opcional de P-Valor

Fuente	Suma de Cuadrados (SC)	Grados de Libertad (gl)	Cuadrados Medios (CM)	F
Entre grupos	B	q-1	CM(B)= B/q-1	F=CM(B)/CM(W)
Intra grupos	W	n-q	CM(W)=W/n-q	
Total	**T**	**n-1**		

- En este caso:

Fuente	Suma de Cuadrados (SC)	Grados de Libertad (gl)	Cuadrados Medios (CM)	F
Entre grupos	85,356	5-1	21,339	4,301
Intra grupos	124,021	30-5	4,961	
Total	**209,377**	**30-1**		

Solución y observaciones

- A un nivel α=5%, se tiene que $F > F_{4,25}(0.05) = 2.76$, por lo que se rechaza H_0. Se concluye que hay diferencias entre las mezclas de concreto.
- Discusión:
  - La prueba ANOVA supone que las varianzas son iguales para todos los grupos.
  - La prueba es robusta frente a pequeñas variaciones en los tamaños de los grupos.
  - Solo permite determinar si existen o no diferencias en entre los grupos y no a la magnitud de las mismas.