Part I

Descripción estadística de dos variables

Introducción

- Si para un mismo individuo observamos simultáneamente k características obtendremos como resultado una variable estadística **k**-dimensional.
- Nos ocuparemos del estudio de las variables estadísticas bidimensionales ($k = 2$).
- Representaremos por (X, Y) la variable bidimensional estudiada, donde X e Y son las variables unidimensionales correspondientes a las primera y segunda características, respectivamente.
- El estudio de cada variable bidimensional particular (X, Y) variará según las variables unidimensionales X e Y sean cuantitativas o cualitativas y, de ser cuantitativas, según sean continuas o discretas.

Distribución de frecuencias conjunta

- Sea $(x_1, y_1), \ldots, (x_n, y_n)$ una muestra de n observaciones de una variable estadística bidimensional (X, Y), tal que que X presenta k modalidades $c_1, c_2, \ldots, c_k$ e Y presenta l modalidades $c'_1, c'_2, \ldots, c'_l$, y sea n_{ij} el número de individuos de la muestra que presentan la modalidad c_i de X y la modalidad c'_j de Y $(i = 1, 2, \ldots, k;\ j = 1, 2, \ldots, l)$.
- El número n_{ij} se conoce como la frecuencia absoluta del par (c_i, c'_j).
- La frecuencia relativa de (c_i, c'_j) es el número $f_{ij} = n_{ij}/n$, que representa la proporción de individuos de la muestra que presentan el par (c_i, c'_j) $(i = 1, 2, \ldots, k;\ j = 1, 2, \ldots, l)$.

Estadística I

Mario Francisco

Variable estadística bidimensional

Distribuciones de frecuencias

Representaciones gráficas

Medidas características

Regresión lineal. Correlación

Propiedades

- $0 \leq n_{ij} \leq n \ (i = 1, 2, \ldots, k;\ j = 1, 2, \ldots, l)$.
- $0 \leq f_{ij} \leq 1 \ (i = 1, 2, \ldots, k;\ j = 1, 2, \ldots, l)$.
- $\displaystyle\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij} = n_{11} + n_{12} + \cdots + n_{kl} = n.$
- $\displaystyle\sum_{i=1}^{k}\sum_{j=1}^{l} f_{ij} = f_{11} + f_{12} + \cdots + f_{kl} = 1.$

Distribuciones de frecuencias

Distribución de frecuencias **conjunta**

$X \setminus Y$	c'_1	c'_2	$\dots$	c'_j	$\dots$	c'_l
c_1	n_{11} (f_{11})	n_{12} (f_{12})	$\dots$	n_{1j} (f_{1j})	$\dots$	n_{1l} (f_{1l})
c_2	n_{21} (f_{21})	n_{22} (f_{22})	$\dots$	n_{2j} (f_{2j})	$\dots$	n_{2l} (f_{2l})
$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$
c_i	n_{i1} (f_{i1})	n_{i2} (f_{i2})	$\dots$	n_{ij} (f_{ij})	$\dots$	n_{il} (f_{il})
$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$
c_k	n_{k1} (f_{k1})	n_{k2} (f_{k2})	$\dots$	n_{kj} (f_{kj})	$\dots$	n_{kl} (f_{kl})

Estadística I

Mario Francisco

Variable estadística bidimensional

Distribuciones de frecuencias

Representaciones gráficas

Medidas características

Regresión lineal. Correlación

Distribuciones marginales

- Llamaremos distribuciones marginales a las distribuciones de frecuencias unidimensionales de las variables X e Y.
- Para la variable X, las frecuencias absoluta y relativa de la modalidad c_i $(i = 1, 2, \ldots, k)$ son, respectivamente,

$$n_{i.} = n_{i1} + n_{i2} + \cdots + n_{il} \quad \text{y} \quad f_{i.} = \frac{n_{i.}}{n} = f_{i1} + f_{i2} + \cdots + f_{il}$$

- Para la variable Y, las frecuencias absoluta y relativa de la modalidad c'_j $(j = 1, 2, \ldots, l)$ son, respectivamente,

$$n_{.j} = n_{1j} + n_{2j} + \cdots + n_{kj} \quad \text{y} \quad f_{.j} = \frac{n_{.j}}{n} = f_{1j} + f_{2j} + \cdots + f_{kj}$$

Tabla de frecuencias marginales absolutas (relativas) para X

X	c_1	c_2	$\cdots$	c_i	$\cdots$	c_k
	$n_{1\cdot}(f_{1\cdot})$	$n_{2\cdot}(f_{2\cdot})$	$\cdots$	$n_{i\cdot}(f_{i\cdot})$	$\cdots$	$n_{k\cdot}(f_{k\cdot})$

Tabla de frecuencias marginales absolutas (relativas) para Y

Y	c'_1	c'_2	$\cdots$	c'_j	$\cdots$	c'_l
	$n_{\cdot 1}(f_{\cdot 1})$	$n_{\cdot 2}(f_{\cdot 2})$	$\cdots$	$n_{\cdot j}(f_{\cdot j})$	$\cdots$	$n_{\cdot l}(f_{\cdot l})$

Distribuciones condicionadas

- Para cada modalidad c'_j de Y, la frecuencia absoluta de la modalidad c_i de X condicionada a $Y = c'_j$ es

$$n_{i/j} = n(X = c_i / Y = c'_j) = n_{ij} \quad (i = 1, 2, \ldots, k)$$

- La frecuencia relativa de c_i de X condicionada a $Y = c'_j$ es

$$f_{i/j} = f(X = c_i / Y = c'_j) = \frac{n_{ij}}{n_{.j}} = \frac{f_{ij}}{f_{.j}} \quad (i = 1, 2, \ldots, k)$$

- Análogamente, se obtienen las frecuencias, absolutas y relativas, de cada modalidad de Y condicionada por cada modalidad de X.

Distribuciones de frecuencias

Tabla de distribuciones de X condicionada por $Y = c_j'$

$X/Y = c_j'$	c_1	$\cdots$	c_i	$\cdots$	c_k
	$n_{1/j}(f_{1/j})$	$\cdots$	$n_{i/j}(f_{i/j})$	$\cdots$	$n_{k/j}(f_{k/j})$

Tabla de distribuciones de Y condicionada por $X = c_i$

$X/Y = c_j'$	c_1	$\cdots$	c_i	$\cdots$	c_k
	$n_{1/j}(f_{1/j})$	$\cdots$	$n_{i/j}(f_{i/j})$	$\cdots$	$n_{k/j}(f_{k/j})$

Representaciones gráficas

- Diagrama de dispersión. Es la construcción gráfica que resulta de representar en un sistema de ejes de coordenadas los pares de observaciones (x_i, y_i) $(i = 1, 2, \ldots, n)$ de la variable bidimensional (X, Y).
- Histograma. En un sistema de ejes tridimensional, se construye una cuadrícula en el plano (X,Y) y sobre estas prismas rectangulares cuyos volúmenes deberán ser proporcionales a las frecuencias absolutas (relativas) de las modalidades (c_i, c'_j) definidas por cada rectángulo.
- Representación gráfica de distribuciones condicionadas. Las distribuciones (unidimensionales) condicionadas pueden representarse en gráficos conjuntos, donde se incluyen los diagramas de barras, histogramas o diagramas de caja relativos a una variable condicionada a los valores particulares de otra.

Medidas características

Momentos

- Sea $\{(x_i, y_i)\}_{i=1}^{n}$ una muestra de una variable cuantitativa (X, Y). El momento con respecto al origen de orden (r, s) $(r \geq 0, s \geq 0)$ de la variable (X, Y) es el número $a_{rs} = \frac{1}{n} \sum_{i=1}^{n} x_i^r y_i^s$
- El momento con respecto a las medias de orden (r, s) $(r \geq 0, s \geq 0)$ es $m_{rs} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^r (y_i - \overline{y})^s$
- Las medias marginales de X e Y son, respectivamente, $a_{10} = \overline{x}$ y $a_{01} = \overline{y}$; siendo $m_{20} = s_x^2$ y $m_{02} = s_y^2$ las varianzas marginales de X e Y, respectivamente.

Covarianza

En el caso particular $r = 1$, $s = 1$ se obtiene el momento m_{11}, conocido como la covarianza de (X, Y)

$$Cov(X, Y) = s_{xy} = m_{11} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

que puede interpretarse como una medida de la relación lineal entre las variables X e Y.

Covarianza. Propiedades

- La covarianza de (X, Y) es igual a la de (Y, X): $s_{xy} = s_{yx}$.
- La covarianza de (X, X) es igual a la varianza marginal de X: $s_{xx} = s_x^2$.
- Si a, b, c, d son constantes cualesquiera, la covarianza de (U, V), con $U = a + bX$ y $V = c + dY$, es $s_{uv} = bd s_{xy}$.
- Para el cálculo en la práctica de la covarianza de (X, Y) puede utilizarse la relación

$$s_{xy} = a_{11} - a_{10}a_{01} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y}$$

Estadística I

Mario Francisco

Variable estadística bidimensional

Distribuciones de frecuencias

Representaciones gráficas

Medidas características

Regresión lineal. Correlación

Vector de medias. Matriz de varianzas-covarianzas

- Llamaremos vector de medias de una variable (X, Y) al vector $(\overline{x}, \overline{y})$.
- Llamaremos matriz de varianzas-covarianzas de la variable (X, Y) a la matriz

$$S = \begin{pmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{pmatrix}$$

- S es una matriz simétrica y definida no negativa, cuyo determinante (no negativo) $|S| = s_x^2 s_y^2 - s_{xy}^2$ proporciona una medida de la variación conjunta de las variables X e Y, que se conoce como la varianza generalizada de (X, Y).

Regresión lineal. Correlación

- Es interesante investigar la posible existencia de alguna relación de dependencia entre las variables y la construcción de algún modelo matemático que permita describir dicha relación.
- Entre las variables X e Y existe una relación de dependencia exacta si conociendo el valor de una se conoce exactamente el valor de la otra.
- X e Y son variables independientes si una variable no contiene información sobre la otra.
- La variable X contiene cierta información (incompleta) acerca de la variable Y, pudiéndose predecir aproximadamente el valor de Y a partir del conocimiento del valor que ha tomado X mediante la construcción de lo que llamaremos modelos de regresión; en estas situaciones diremos que X e Y son variables dependientes.

Introducción a los modelos de regresión

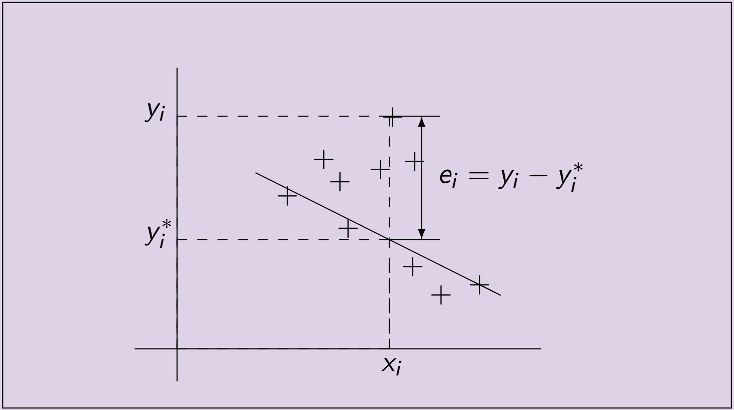
- Se construyen modelos matemáticos, llamados modelos de regresión, para explicar la relación de dependencia entre una variable respuesta, Y, y una o más variables explicativas, X.
- Modelos de regresión simple a aquellos que consideran una única variable explicativa y modelos de regresión múltiple a los que consideran dos o más variables explicativas.
- Se construyen modelos $Y = m(X) + E$ donde E es el error de observación y m la función de regresión.
- los modelos paramétricos de regresión, en los que se supone que m tiene una forma predeterminada y los modelos no paramétricos de regresión, en los que se formulan condiciones muy generales acerca de m.

El método de mínimos cuadrados

- Sean $\{(x_i, y_i)\}_{i=1}^{n}$ una muestra de una variable (X, Y) y supongamos que se desea ajustar un modelo paramétrico de la forma $y = m(x; \theta_1, \ldots, \theta_k)$ al diagrama de dispersión de (X, Y), donde $\theta_1, \ldots, \theta_k$ son constantes desconocidas a determinar.
- El método de mínimos cuadrados selecciona los valores de $\theta_1, \ldots, \theta_k$ del modelo que mejor se ajusta al diagrama de dispersión de (X, Y), es decir, los valores $\hat{\theta}_1, \ldots, \hat{\theta}_k$ que minimizan la función

$$M(\theta_1, \ldots, \theta_k) = \frac{1}{n} \sum_{i=1}^{n} (y_i - m(x_i; \theta_1, \ldots, \theta_k))^2$$

Error ($e_i = y_i - y_i^*$) al predecir el valor observado (y_i) por el valor calculado mediante el modelo teórico ($y_i^* = m(x_i) = a + bx_i$).



El modelo de regresión lineal simple

- Supongamos que la muestra $\{(x_i, y_i)\}_{i=1}^n$ de una variable (X, Y) tienden a alinearse en torno a una recta de ecuación $y = m(x) = a + bx$.
- Por mínimos cuadrados se determinan los valores de a y b que minimizan

$$M(a, b) = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - a - bx_i)^2$$

- Obteniéndose

$$\hat{a} = \overline{y} - \hat{b}\overline{x} \quad \text{y} \quad \hat{b} = \frac{s_{xy}}{s_x^2}$$

El modelo de regresión lineal simple

- La ecuación de la recta de regresión de Y sobre X, sustituyendo los valores de $\hat{a}$ y $\hat{b}$ por las expresiones anteriores, es

$$y - \overline{y} = \frac{s_{xy}}{s_x^2}(x - \overline{x})$$

- La recta de regresión de X sobre Y es

$$x - \overline{x} = \frac{s_{xy}}{s_y^2}(y - \overline{y})$$

- Llamaremos coeficientes de regresión a $\beta_{yx} = s_{xy}/s_x^2$ y $\beta_{xy} = s_{xy}/s_y^2$.

Predicción con el modelo de regresión lineal simple

- El valor de la predicción será el resultado de evaluar la función de regresión ajustada en el valor particular de la variable o variables explicativas.
- La predicción de Y con el modelo de regresión lineal simple en un valor particular x_0 es $\hat{y}_0 = \hat{a} + \hat{b}x_0$.
- Esta forma de realizar las predicciones podrá producir resultados razonablemente satisfactorios siempre que el valor x_0 pertenezca al intervalo comprendido entre el mínimo y el máximo de los valores observados en la muestra de la variable explicativa X.

Adecuación del modelo de regresión lineal simple: la razón de correlación

- Se desea conocer si el modelo ajustado es adecuado para describir la relación de dependencia entre X e Y.
- Suele utilizarse el coeficiente de determinación,

$$R^2 = 1 - \frac{s_R^2}{s_y^2} = \frac{s_y^2 - s_R^2}{s_y^2},$$

  donde s_y^2 es varianza total y $s_R^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ es la varianza no explicada.
- Para la recta de regresión, $R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = \beta_{xy}\beta_{yx}$.
- Valores de R^2 cercanos a uno indican un buen ajuste y cercanos a cero, malo.

El coeficiente de correlación lineal

- El propósito del estudio de la correlación es la construcción de medidas del grado de dependencia o asociación entre variables estadísticas, siendo la más popular de éstas el coeficiente de correlación lineal.
- Dadas observaciones $\{(x_i, y_i)\}_{i=1}^n$ de una variable bidimensional (X, Y) se define el coeficiente de correlación lineal de X e Y como el número

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

que es una medida del grado de dependencia lineal entre las variables X e Y.

El coeficiente de correlación lineal. Propiedades

- El coeficiente de correlación lineal es una medida adimensional.
- $-1 \leq r_{xy} \leq 1$.
- Si $r_{xy} = \pm 1$ significa que existe dependencia lineal exacta entre X e Y. Si $r_{xy} = 0$ significa que no existe dependencia lineal entre X e Y (incorreladas).
- $r_{xy} = r_{yx}$. Si $r_{xy} > 0$ la recta de regresión es creciente y si $r_{xy} < 0$ decreciente.
- $r_{xx} = 1$.
- Si $a,\ b,\ c,\ d$ son constantes, dadas nuevas variables $U = a + bX$ y $V = c + dY$ se verifica que $r_{uv} = r_{xy}$, si $bd > 0$, y $r_{uv} = -r_{xy}$, si $bd < 0$.

Estadística I

Mario Francisco

Variable estadística bidimensional

Distribuciones de frecuencias

Representaciones gráficas

Medidas características

Regresión lineal. Correlación

Matriz de correlaciones

- A partir de la definición anterior del coeficiente de correlación lineal, se define la matriz de correlaciones de (X, Y) como la matriz simétrica

$$C = \begin{pmatrix} 1 & r_{xy} \\ r_{xy} & 1 \end{pmatrix}$$

- En general, se define la matriz de correlaciones de una variable aleatoria k-dimensional $(X_1, X_2, \ldots, X_k)$ como la matriz simétrica $C = (r_{ij})$, donde r_{ij} $(1 \leq i \leq k,\ 1 \leq j \leq k)$ es el coeficiente de correlación lineal entre X_i y X_j.

Otros modelos de regresión

- El modelo multiplicativo o potencial. Se trata de ajustar la función de regresión $y = m(x) = ax^b$.
- El modelo recíproco o inverso. La función de regresión a ajustar es $y = m(x) = 1/(a + bx)$.
- El modelo exponencial. Se trata de ajustar la función de regresión $y = m(x) = ab^x$.
- El modelo de regresión polinómica. Generaliza el caso de la regresión L.S., siendo la función de regresión a ajustar un polinomio de grado p, $y = m(x) = \theta_0 + \theta_1 x + \cdots + \theta_p x^p$.
- El modelo de regresión lineal múltiple. Es la generalización del modelo de regresión L.S. al caso en que se disponga de k variables explicativas, tratándose de ajustar la función de regresión $y = m(x_1, \ldots, x_k) = \theta_0 + \theta_1 x_1 + \cdots + \theta_k x_k$.

Regresión no paramétrica

- En el contexto de la regresión no paramétrica no se asume ninguna forma predefinida para la función m (únicamente se suponen ciertas hipótesis muy generales acerca de ésta) e interesa estimar su valor en un punto cualquiera x.
- La idea consiste en elegir un entorno de amplitud a con centro el punto x, tomándose para $\hat{m}(x)$ una media ponderada de las observaciones y_i tales que el valor de la primera coordenada está en dicho entorno. Es decir, $\hat{m}(x) = \sum_{i=1}^{n} w(a, x, x_i) y_i$, donde $w(a, x, x_i)$ es el peso asociado al valor observado y_i, que, en general, se elegirá teniendo en cuenta la distancia entre x y x_i (a mayor distancia, menor peso).