TEMA 9 – DISTRIBUCIONES BIDIMENSIONALES

9.1 – NUBES DE PUNTOS. CORRELACIÓN

Si tenemos un colectivo de n individuos y en ellos estudiamos dos variables x e y. Si conocemos los valores de las variables para cada uno de los individuos:

El conjunto de pares de valores (x_1,y_1), (x_2,y_2), ..., (x_n,y_n) se llama **distribución bidimensional**. Si interpretamos cada par de valores como las coordenadas de un punto, el conjunto de todos ellos se llama **nube de puntos** o **diagrama de dispersión.**

La **correlación** viene a representar la relación que existe entre esas dos variables para los n individuos. Puede ser más o menos fuerte según lo apretados que estén los puntos de la nube en torno a una recta que marca la tendencia y se llama **recta de regresión**

Si la pendiente de la recta de regresión es positiva o negativa, la **correlación** se llama **positiva** o **negativa**, respectivamente.

9.2 – MEDIDAS DE CORRELACIÓN

La correlación entre dos variables más o menos fuerte, positiva o negativa, se aprecia mediante el grado de apretura de los puntos de la nube. Vamos a confeccionar una fórmula que sirva para obtener su valor de forma numérica e inequívoca.

CENTRO DE GRAVEDAD DE UNA DISTRIBUCIÓN BIDIMENSIONAL

Media de la variable x : $\bar{x} = \dfrac{\sum x_i.f_i}{n}$ $\qquad n = \sum f_i$

Media de la variable y : $\bar{y} = \dfrac{\sum y_i.f_i}{n}$

El punto $(\bar{x}, \bar{y})$ se llama **centro de gravedad** de la distribución

DESVIACIONES TÍPICAS

Desviación típica de la variable x : $\sigma_x = \sqrt{\dfrac{\sum x_i^{\,2}.f_i}{n} - \bar{x}.^2}$

Desviación típica de la variable y : $\sigma_y = \sqrt{\dfrac{\sum y_i^{\,2}.f_i}{n} - \bar{y}.^2}$

COVARIANZA

Se llama **covarianza** al parámetro: $\sigma_{xy} = \dfrac{\sum x_i.y_i.f_i}{n} - \bar{x}.\bar{y}$

CORRELACIÓN

El valor de la **correlación** entre las dos variables de una distribución bidimensional viene dado por la expresión: $r = \frac{\sigma_{xy}}{\sigma_x . \sigma_y}$

- No tiene dimensiones
- El valor de r está comprendido entre –1 y 1
  - Si la correlación es perfecta (puntos de la nube alineados), entonces |r| = 1
  - Si la correlación es fuerte |r| es próximo a 1
  - Si la correlación es débil |r| es próximo a 0

9.3 – RECTA DE REGRESIÓN

MÉTODO DE LOS MÍNIMOS CUADRADOS

Partimos de la nube de puntos (x_1,y_1), (x_2,y_2), ..., (x_n,y_n). Hemos de encontrar la recta que mejor de ajuste a la nube. Consideramos todas las posibles rectas y = a + bx y nos quedamos con aquella para la cual los cuadrados de las distancias, d_i, sumen lo menos posible, es decir, para la cual $\sum d_i^{\,2}$ es mínimo.

De este modo se llega a lo siguiente:

- La recta buscada pasa por el centro de gravedad de la distribución $(\bar{x}, \bar{y})$
- La pendiente es $m_{yx} = \frac{\sigma_{xy}}{\sigma_x^2}$

La recta que hace mínima la suma $\sum d_i^{\,2}$ tiene por ecuación : $y - \bar{y} = \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x})$ se llama **recta de regresión de Y sobre X**

A la pendiente, $\frac{\sigma_{xy}}{\sigma_x^2}$, se le llama **coeficiente de regresión.**

El signo del coeficiente de correlación y el del coeficiente de regresión coinciden, pero aquí termina la coincidencia: puede ser que la recta de regresión tenga pendiente alta y, sin embargo, el coeficiente de correlación sea bajo o al contrario.

LA RECTA DE REGRESIÓN PARA HACER ESTIMACIONES

La recta de regresión se amolda a la nube de puntos y describe, grosso modo, su tendencia. Por eso, a partir de la recta de regresión obtenemos, de forma aproximada, el valor esperado de y para cierto valor de x; o viceversa. A estos valores se les llama estimaciones.

$\hat{y}(x_0)$ es el valor estimado de y, correspondiente a $x = x_0$ sobre la recta de regresión.

$\hat{x}(y_0)$ es el valor estimado de x, correspondiente a $y = y_0$ sobre la recta de regresión.

- Las estimaciones siempre se realizan aproximadamente y en términos de probabilidad, es probable que si $x = x_0$, entonces y valga, aproximadamente: $\hat{y}(x_0)$

- La aproximación es tanto mejor cuanto mayor sea |r|, pues para valores de r próximos a 1 o a –1, los puntos están muy próximos a la recta.
- Las estimaciones solo deben hacerse dentro del intervalo de valores utilizados o muy cerca de ellos.

9.4 – HAY DOS RECTAS DE REGRESIÓN

Como ya hemos visto, la recta de regresión de Y sobre X es : $y - \bar{y} = \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x})$

Si el criterio que siguiéramos para ajustar la recta a la nube de puntos fuera hacer mínima la suma de los cuadrados de las diferencias de abscisas del punto obtendríamos la recta de regresión de X sobre Y : $x - \bar{x} = \frac{\sigma_{xy}}{\sigma_y^2}(y - \bar{y})$

El número $\frac{\sigma_{xy}}{\sigma_y^2}$ se llama **coeficiente de regresión de X sobre Y**. No es la pendiente de la recta, sino su inversa.

POSICIONES DE LAS DOS RECTAS DE REGRESIÓN

- Cuando la correlación es casi nula, las dos rectas forman un ángulo muy grande (próximo a 90º)
- Si la correlación es fuerte, el ángulo que forman las dos rectas es pequeño.
- Si |r| es próximo a 1, las rectas son casi coincidentes.
- La recta que nosotros asignamos a ojo a una nube de puntos es aproximadamente, la bisectriz de las dos rectas de regresión.

9.5 – TABLAS DE DOBLE ENTRADA

Las distribuciones de una variable, cuando el número de observaciones es pequeño, se dan, simplemente, enumerando los datos de forma ordenada. Pero cuando el número de datos es grande, se recurre a la tabla de frecuencias.

Del mismo modo, en las distribuciones bidimensionales, cuando hay pocos pares de valores se procede enumerándolos. Si algún par está repetido se pone dos veces. Pero cuando el número de datos es grande se recurre a las tablas de doble entrada.

En cada casilla se pone la frecuencia correspondiente al par de valores que definen esa casilla.

La representación gráfica de estas distribuciones se hace:
- Hinchando los puntos proporcionalmente a su frecuencia
- Levantando barras de alturas proporcionales a las frecuencias de las correspondientes casillas.