# Tema 2: Distribuciones unidimensionales

- Distribución de frecuencias:

| $X$ | $n$ |
|---|---|
| $x_1$ | $n_1$ |
| $x_2$ | $n_2$ |
| $\vdots$ | $\vdots$ |

frequenci

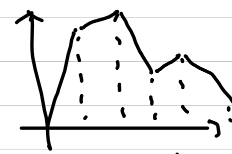
- Recorrido de una variable = range valores $= \max x_i - \min x_i$.
- Frecuencia relativa $= f_i = n_i/n \quad (n = \sum n_i)$

Frecuencia absoluta acumulada $= N_i = n_1 + \dots + n_i$

Frecuencia relativa acumulada $= F_i = \frac{N_i}{n} = f_1 + \dots + f_i$

## 2.3 - Representación gráfica

(1) Diagrama de barras
(2) Histograma (área $\propto$ frecuencia)
(3) Polígono de frecuencia



↳ Ojiva: cuando utilizamos frec. acumuladas

(4) Diagrama de tallo y hoja

| | |
|---|---|
| 0 | 2 |
| 1 | 4 5 |
| 2 | 5 8 |
| 3 | 9 5 8 |
| 4 | 7 3 4 |

Tema 3: Características de la distribución de frecuencias

Medidas:
- Posición (media, mediana, moda)
- Dispersión (grado de concentración)
- Forma (ej. curtosis, informan sobre aspecto gráfico).

## 3.1 - Medidas de posición

- Media aritmética $\bar{x} = \sum x_i n_i / n$
- Mediana ← Me
- Moda ← Mo (si es por intervalos, la media del intervalo más común = frecuencia/longitud del inter)
- Media geométrica $G = \sqrt[n]{\prod x_i^{n_i}}$
- Media armónica $H = \dfrac{n}{\sum n_i/x_i}$
- Cuantiles: valores de la variable que dividen la distribución de frecuencias en cierto número de partes iguales).

## 3.2 - Medidas de dispersión

- Recorrido $R = \max x_i - \min x_i$
- Recorrido intercuartílico $= Q_3 - Q_1$ ← 3er y 1er cuartil
- Desviación mediana $= D_{ME} = \dfrac{\sum |x_i - Me| n_i}{n}$
- Varianza $= s^2 = \dfrac{\sum (x_i - \bar{x})^2 n_i}{n}$ $\left( = \mathbb{E}[(X-\mu)^2] \right)$.
- Desviación típica $= s = \sqrt{s^2} = \sigma$
  - ↳ Teorema de Chebyshev $P(|X-\mu| \geq k\sigma) \leq 1/k^2$.
  - ↳ i.e. probabilidad de que nos desviemos $k\sigma$ de la media es menor que $1/k^2$.
- Coeficiente de apertura: $CA = \max x_i / \min x_i$
- Recorrido semi-intercuartílico: $R_{SI} = \dfrac{Q_3 - Q_1}{Q_3 + Q_1}$.
- Variación mediana: $V_{ME} = D_{ME}/Me$.
- Coeficiente de variación de Pearson: $C_V = S/|\bar{x}|$.

## 3.3 Medidas de forma

- Coeficiente de asimetría de Pearson

$$A_p = \frac{3(\bar{x} - Me)}{S} \left( \frac{\bar{x} - Mo}{S} \right)$$

  - $A_p = 0$ : simétrica
  - $A_p > 0$ : asimétrica +ve / derecha
  - $A_p < 0$ : asimétrica −ve / izquierda

- Coeficiente de asimetría de Fisher

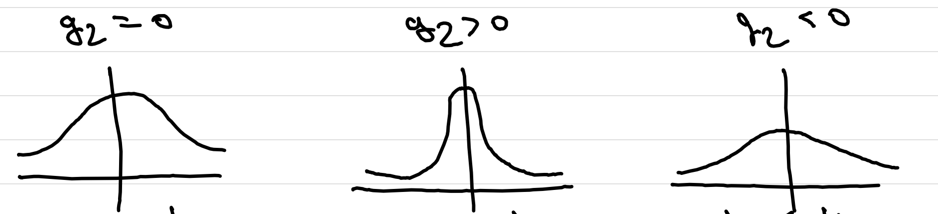
$$g_1 = \frac{m_3}{S^3} = \frac{\frac{1}{n}\sum (x_i - \bar{x})^3 n_i}{\left[\sqrt{\frac{1}{n}\sum (x_i - \bar{x})^2 n_i}\right]^3} = \frac{\mathbb{E}[(X-\bar{X})^3]}{\mathbb{E}[(X-\bar{X})^2]^{3/2}}$$

  - $g_1 = 0$ : simétrica
  - $g_1 > 0$ : asimétrica +ve
  - $g_1 < 0$ : asimétrica −ve

- La curtosis se refiere a la mayor o menor concentración de valores alrededor de la zona central de la distribución.

$$g_2 = \frac{m_4}{S^4} - 3 = \frac{\frac{1}{n}\sum (x_i - \bar{x})^4 n_i}{\left(\frac{1}{n}\sum (x_i - \bar{x})^2 n_i\right)^2} - 3$$

$g_2 = 0$ $\quad$ $g_2 > 0$ $\quad$ $g_2 < 0$



mesocúrtica $\quad$ leptocúrtica $\quad$ platicúrtica

## 3.4 Cambios de origen y escala

$x_i^* = a + x_i$ $\qquad$ $x_i^* = b x_i$

- Tipificación: operación $z_i = \dfrac{x_i - \bar{x}}{s}$.
- Escala derivada: supongamos que un profe quiere que las notas de un examen tengan una media de 5,5 y desviación de 2,2. Deberá encontrar $z_i$ y luego hacer $z_i^* = 5,5 + 2,2 z_i$.

## Apéndice:

- Momentos respecto a la media $\rightsquigarrow$ $m_1 = 0$, $m_2 = s^2$

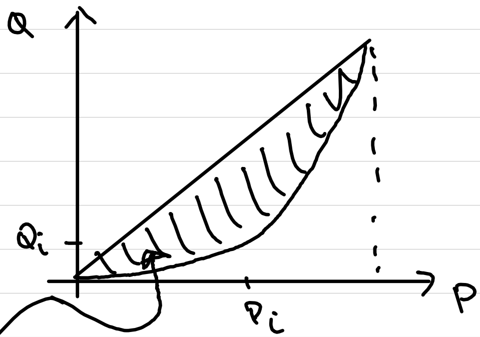
$$m_r = \frac{1}{n}\sum_i (x_i - \bar{x})^r n_i$$

# Tema 4 - Medidas de concentración:

· Grado de concentración: ej: cómo se reparte la renta de un país.

* concentración máx: $x_1 = \dots = x_{n-1} = 0$, $x_n = X$
* concentración mín: $x_i = X/n$.

## 4.1 Curva de Lorentz:




$$P_i = \frac{n_1 + \dots + n_i}{n} \times 100$$

$$Q_i = \frac{x_1 n_1 + \dots + x_i n_i}{n} \times 100$$

(ie. $Q_i$ es cuánta renta tiene en total una fracción $P_i$ de la población).

## 4.2 - Índice de Gini

$$= \frac{\text{Área sombreada}}{\text{Área del triángulo}} = \frac{\sum (P_i - Q_i)}{\sum P_i} = I_G$$

* $I_G = 0$ : igualitario
* $I_G = 1$ : muy desigual.

# Tema 5 - Distribuciones bidimensionales

## 5.5 Medidas de dispersión y posición

* $\bar{x} = \frac{1}{n}\sum_{i,j} x_i n_{ij}$ , $\bar{y} = \frac{1}{n}\sum_{i,j} y_j n_{ij}$

$S_x^2 = \frac{1}{n}\sum_{i,j}(x_i-\bar{x})^2 n_{ij}$ , $S_y^2 = \frac{1}{n}\sum_{i,j}(y_j-\bar{y})^2 n_{ij}$

$S_{xy} = \frac{1}{n}\sum_{i,j}(x_i-\bar{x})(y_j-\bar{y}) n_{ij}$

covarianza.

* C-S nos da $-S_x S_y \le S_{xy} \le S_x S_y$ .

## 5.6 Independencia

* Independencia si $\frac{n_{ij}}{n} = \frac{n_{i*}}{n} \times \frac{n_{j*}}{n}$

$\Leftrightarrow f_{ij} = f_{i*} \times f_{*j}$ .

La independencia $\Rightarrow S_{xy} = 0$ .

# Tema 6 – Regresión y correlación simple

## 6.1 Covariación

* Coeficiente de correlación lineal : $r = \frac{S_{xy}}{S_x S_y}$

$-1 \leq r \leq 1$ por C-S.

* Regresión : intentamos determinar una función $f$ tal que $y = f(x)$.

## 6.2 Regresión lineal simple (RLS)

* $y_i$ — $y$ observado; $y_i^* = a + bx_i$ — $y$ ajustado $\rightsquigarrow e_t = y_t - y_t^*$ residuo

$$\therefore y_i = y_i^* + e_i = a + bx_i + e_i.$$

* Para ajustar intentamos minimizar $\sum_{i=1}^{n} e_i^2$.

$\Leftrightarrow \min_{a,b} \sum_{i=1}^{n} (y_i - a - bx_i)^2$

$\frac{\partial}{\partial a} : -2\sum_{i=1}^{n} (y_i - a - bx_i) = 0$

$\frac{\partial}{\partial b} : -2\sum_{i=1}^{n} (y_i - a - bx_i)x_i = 0$

$\Leftrightarrow$ Se deducen las ecuaciones normales

$$\sum_i y_i = na + b\sum_i x_i \quad (1)$$
$$\sum_i y_i x_i = a\sum_i x_i + b\sum x_i^2$$

Por lo tanto deducimos $\boxed{b = \frac{S_{xy}}{S_x^2}, \quad a = \bar{y} - b\bar{x} \quad (1)}$

* Medidas de bondad de ajuste

• Varianza residual $S_e^2 := \frac{\sum_i e_i^2}{n}$ ← si es alta, mal asunto

$$\underset{\text{varianza total}}{S_y^2} = \underset{\text{var. explicada}}{S_{y^*}^2} + \underset{\text{var. residual}}{S_e^2}$$

$S_{y^*}^2 = \frac{1}{n}\sum_i (y_i^* - \bar{y})^2$, $\quad S_e^2 = \frac{1}{n}\sum (y_i - y_i^*)^2$.

• Coeficiente de determinación : $R^2 = \frac{S_{y^*}^2}{S_y^2}$ ← alto = bien.

$= S_{xy}^2 / S_x^2 S_y^2$.

## 6.3 Ajuste de funciones no lineales

* $y = ax^b$ (utilizar $\ln y = \ln a + b\ln x$ y luego resolver como si fuera RLS).

* Exponenciales $y = ab^{x_i}$, $e^{y_i} = ax_i^b$.

* Hiperbólicos, $y = a + \frac{b}{x_i}$ (utilizar $z = \frac{1}{x}$).

# Tema 7 - Regresión múltiple

## 7.1 - Regresión lineal múltiple

- $Y = b_1 + b_2 X_2 + \cdots + b_k X_k$

  $\hookrightarrow \; y_i = b_1 + b_2 x_{2i} + \cdots + b_k x_{ki}$

$$\therefore Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = Xb + e = \begin{pmatrix} 1 & x_{21} & \cdots & x_{k1} \\ \vdots & \vdots & & \\ 1 & x_{2n} & \cdots & x_{kn} \end{pmatrix} \begin{pmatrix} b_1 \\ \vdots \\ b_k \end{pmatrix} + \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$$

Las ecuaciones normales nos dan

$$\boxed{b = (X^T X)^{-1} X^T Y}$$

- Varianza residual $= s_e^2 = \dfrac{\|e\|^2}{n} = \dfrac{\|Y - bX\|^2}{n} = \cdots = \dfrac{\|Y\|^2 - b^T X^T Y}{n}$

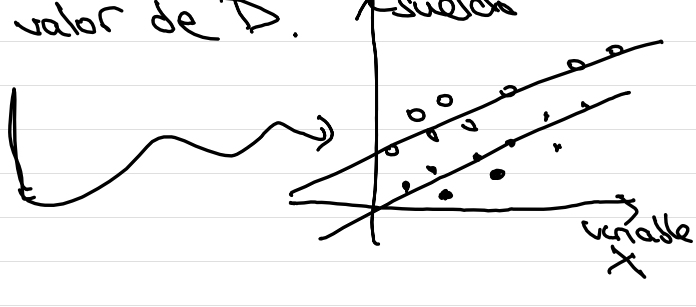
- Coeficiente de determinación múltiple : $R^2 = 1 - \dfrac{s_e^2}{s_Y^2}$ donde

$$s_Y^2 = \frac{\|Y\|^2 - n\bar{Y}^2}{n}.$$

## 7.2 Regresión lineal con variables ficticias

- A veces es útil añadir un 'fixed effect': añadimos una variable discreta (por ejemplo $D = \begin{cases} 0 & \text{varón} \\ 1 & \text{mujer} \end{cases}$ en una regresión para averiguar el sueldo en una compañía). Son los famosos dummy variables. Simplemente ajustamos el intercept: creamos ~~d~~ rectas una para cada valor de D.

~~Todos los da~~



# Tema 8 - Estadística de atributos

- A veces tenemos que estudiar atributos nominales (las modalidades no admiten ordenación, eg: colores).
- Si tenemos dos atributos podemos hacer una tabla de contingencia:

| $A_i \backslash B_j$ | $B_1$ | $B_2$ | $\cdots$ | $B_h$ |
|---|---|---|---|---|
| $A_1$ | $n_{11}$ | $n_{12}$ | $\cdots$ | $n_{1h}$ |
| $A_2$ | $n_{21}$ | $n_{22}$ | $\cdots$ | $n_{2h}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $A_k$ | $n_{k1}$ | $\cdots$ | | $n_{kh}$ |

Si queremos demostrar la independencia de los atributos, necesitamos

$$\frac{n_{ij}}{n} = \frac{n_{i*}}{n} \times \frac{n_{*j}}{n}$$

- Cuando los atributos no son independientes podemos analizar el grado de relación.

* Coeficiente de contingencia:

Si $n_{ij}^* = \frac{n_{i*} \times n_{j*}}{n^2} \cdot n$ son las frecuencias teóricas

$= \frac{n_{i*} \times n_{j*}}{n}$. Entonces,

$$\chi^2 = \sum_{i,j} \frac{(n_{ij}^* - n_{ij})^2}{n_{ij}^*}$$

$0$ = independ.
$\uparrow$ = más dep.

* El coeficiente de contingencia de Pearson es $C = \sqrt{\frac{\chi^2}{n+\chi^2}}$, $0 \leq C \leq 1$

* Si las modalidades del atributo A y B están ordenadas, es posible definir

$$r = \frac{S_{xy}}{S_x S_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}}$$

(donde $x_i$ es el rango / orden de $A_i$)

$$= \cdots$$

using $\sum x_i = 1 + \cdots + n = \frac{n(n+1)}{2}$ ...

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

Es el rango de correlación por rangos de Spearman.

# Tema 9 - Series temporales

## 9.3 Componentes:

Las fluctuaciones de una serie temporal $Y_t$ es el resultado de cuatro componentes básicos:

(1) Tendencia, T : largo
(2) Ciclo, C : medio
(3) Estación, E : corto
(4) Residuo, R : no poseen un carácter periódico

- Aditiva $Y_t = T_t + C_t + E_t + R_t$
- Multiplicativa $Y_t = T_t C_t E_t R_t$
  - $\ln Y_t = \ln T_t + \ln C_t + \dots$

Solemos simplificar a $Y_t = T_t + E_t$.

## 9.4 Detección de componentes:

- 'A ojo'

# Tema 10 - Análisis de la tendencia:

## 10.1 - Métodos de la media móvil:

- Calculamos la media en intervalos de p valores:

$\overline{y}(1) = \frac{y_1 + \cdots + y_p}{p}$ , $\overline{y}(2) = \frac{y_2 + \cdots + y_{p+1}}{p}$ , ...

Si $p$ = periodo, $\overline{y}$ elimina la parte estacional.

¿A qué punto temporal asignamos $\overline{y}$?

→ Asimétrica: en el punto final del intervalo, ie. $\overline{y}_p = \frac{1}{p}(y_1 + \cdots + y_p)$.

→ Centrada: en el medio del intervalo, ie. $\overline{y}_{\frac{p+1}{2}} = \frac{1}{p}(y_1 + \cdots + y_p)$

↑ en este caso quizás $\frac{p+1}{2} \notin \mathbb{N}$

así que tenemos que hacer $\overline{\overline{y}}_{\frac{p}{2}+1} = \frac{\overline{y}_{\frac{p+1}{2}} + \overline{y}_{\frac{p+3}{2}}}{2}$.

## 10.2 - Método de ajuste mínimo-cuadrático

* Ajustamos una ecuación matemática para encontrar la tendencia. Quizás tengamos que utilizar las medias anuales.

↳ ⚠ Las medidas de bondad de ajuste no pueden alcanzar valores adecuados pues los datos con estacionalidad se alejarán mucho de la ecuación de tendencia.

# Tema 11 - Análisis de estacionalidad

## 11.1 Método de la razón a la tendencia

- Bajo hipótesis multiplicativa.

$$IBVE_t := \frac{Y_t}{T_t} = E_t \times R_t$$

↑ índices brutos de variación estacional

$$IVE_t = \sum_{j=1}^{n} \frac{IBVE_j}{n} \quad , \quad \begin{array}{l} i=1,2,3,4 \text{ (trimestres)} \\ n = \text{número de años}. \end{array}$$

También podemos normalizar, $\widetilde{IVE}_t = \frac{IVE_t}{\overline{IVE}}$.

El $IVE_t$ aproxima la estacionalidad.

## 11.2 Método de la diferencia a la tendencia

- Bajo hipótesis aditiva (lo mismo que 11.1)

$$DBVE_t = Y_t - T_t = E_t + R_t$$

↑ diferencias brutas de var. estacional

$$DVE_i = \sum_{j=1}^{n} \frac{DBVE_j}{n} \quad , \quad \widetilde{DVE}_i = DVE_i - \overline{DVE}.$$

## 11.3 Desestacionalización

$$D_t = \frac{Y_t}{\widetilde{IVE}_t} \quad \text{o} \quad D_t = Y_t - \widetilde{DVE}_t$$

# Tema 12 - Números Índices

- Los números índices facilitan una medida que permite la comparación de magnitudes en observaciones históricas.

- Si tenemos una única variable, el índice se construye tomando un valor base ($Y_0$). Luego, $I_t = Y_t / Y_0 \times 100$.

- Las tasas de variación so $\frac{Y_t - Y_{t-1}}{Y_{t-1}} \times 100$.

- A veces tenemos varias variables, por lo que si queremos recoger todos estos índices haremos $I_t = \frac{\sum_{i=1}^{n} Y_{it}/Y_{i0} \times 100}{n} = \frac{\sum_{i=1}^{n} I_{it}}{n}$

  índice complejo →

  A veces ponderamos los índices $\frac{\sum I_{it} w_i}{\sum w_i}$

- Ejemplo: Índices de Laspeyres

  Intentamos ponderar las cantidades y precios de diferentes productos

  $$P_L = \frac{\sum_i P_{it} q_{i0}}{\sum_i P_{i0} q_{i0}}, \quad Q_L = \frac{\sum_i q_{it} P_{i0}}{\sum q_{i0} P_{i0}}$$

  Índices de Paasche

  $$P_P = \frac{\sum_i P_{it} q_{it}}{\sum P_{i0} q_{it}}, \quad Q_P = \frac{\sum_i q_{it} P_{it}}{\sum q_{i0} P_{it}}$$

## 12.4 Operaciones con números índices

- Cambio de base: si queremos cambiar el año base para hacer comparaciones con años anteriores debemos multiplicar todo $I_{t/0}$ por

| $t$ | $I_{t/0}$ | $I_{t/h}$ |
|---|---|---|
| 0 | $I_{0/0}$ | — |
| 1 | $I_{1/0}$ | — |
| ⋮ | ⋮ | |
| $h$ | $I_{h/0}$ | $I_{h/h}$ |
| $h+1$ | — | $I_{h+1/h}$ |
| ⋮ | — | ⋮ |

  $k = I_{h/h} / I_{h/0} = 100 / I_{h/0}$.

  i.e. $I_{t/h} = I_{t/0} * k$, $t = 0, 1, \dots, h-1$.

- Deflacción de series económicas: el análisis de la evolución temporal de magnitudes económicas expresadas en unidades monetarias se ve afectada por la subida de los precios.

  precios corrientes / nominales: $VN_t = \sum_{i=1}^{n} P_{it} q_{it}$

  Precios constantes / reales: $VR_t = \sum_{i=1}^{n} P_{i0} q_{it}$.

  El proceso $VN_t \rightsquigarrow VR_t$ se conoce como deflactación, y basta con dividir por el índice deflactor $P_P = \sum_i P_{it} q_{it} / \sum_i P_{i0} q_{it}$

- Repercusión: a veces es interesante saber el efecto de un cambio específico sobre el índice complejo (ej: efecto de un $\Delta$ en la electricidad sobre el IPC).

  Repercusión absoluta: $R_i = \frac{\Delta I_i w_i}{\sum w_i}$

  o relativa: $Rr_i = \frac{R_i}{I_{t-1}} \times 100$.

  Participación: $P_i = \frac{R_i}{\sum P_i} \times 100 = \frac{R_i}{I_t - I_{t-1}} \times 100$.