BABD

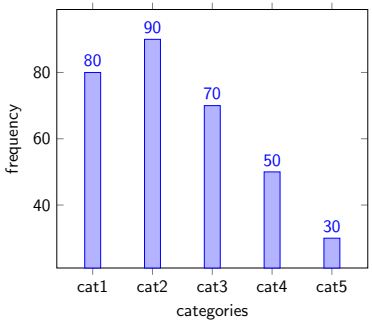
Masters in Business Analytics and Big Data

Exploratory Data Analysis

Mauricio Soto - mauricioabel.soto@polimi.it

POLITECNICO DI MILANO
GRADUATE SCHOOL OF BUSINESS

Graphical analysis categorical attribute

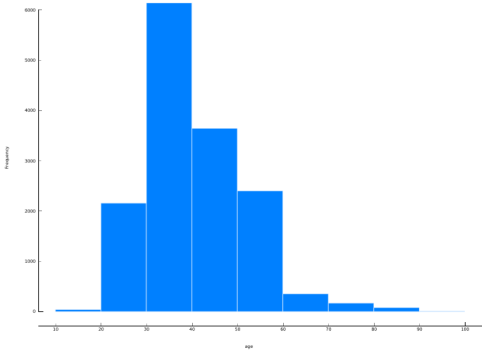


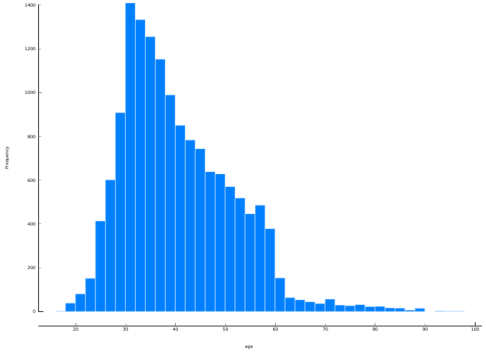
$$f_h = \frac{e_h}{m} = \frac{card\{i \in \mathcal{M} : x_i = cat_h\}}{m}$$

for large samples

$$f_h \approx P(x = cat_h)$$

Graphical analysis numerical attribute





Central tendency

- Mean:

$$\bar{x} = \frac{1}{m} \sum_{i=1}^{m} x_i$$

- Median:

$$x^{\text{med}} = x_{(m+1)/2}, \qquad x^{\text{med}} = (x_{m/2} + x_{m/2+1})/2$$

- Mode
- Midrange:

$$x^{\text{midr}} = (x_{\max} + x_{\min})/2$$

- Geometric mean=

$$\bar{\mu}_{\text{geom}} = \sqrt[m]{\prod_i x_i}$$

Measure of dispersion - numerical

- Range:

$$x_{\max} - x_{\min}$$

- Mean absolute deviation :

$$MAD = \frac{1}{m} \sum_{i=1}^{m} |x_i - \bar{\mu}|$$

- Sample variance:

$$\bar{\sigma}^2 = \frac{1}{m-1} \sum_{i=1}^{m} (x_i - \bar{\mu})^2$$

- Sample standard deviation :

$$\bar{\sigma} = \sqrt{\bar{\sigma}^2}$$

- Coefficient of Variation:

$$CV = 100 \frac{\bar{\sigma}}{\mu}$$

Measure of dispersion - categorical

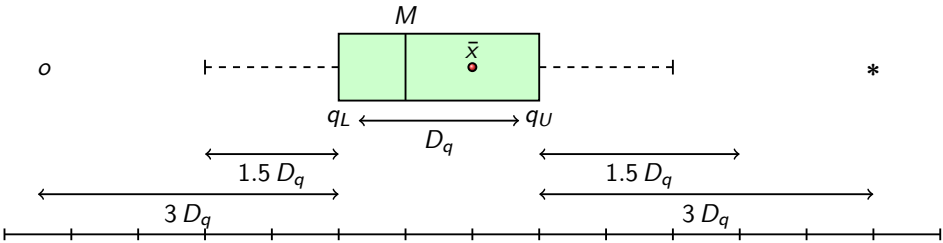
- **Gini index:**

$$Gini = 1 - \sum_{h=1}^{H} f_h^2 \qquad \in [0, (H-1)/H]$$

- **Entropy index:**

$$Entropy = -\sum_{h=1}^{H} f_h \log_2 f_h \qquad \in [0, log_2 H]$$

Box-plot



- Interquartile range $D_q = q_U - q_L = q_{0.75} - q_{0.25}$
- internal lower edge$= q_L - 1.5\, D_q$
- external lower edge$= q_L - 3\, D_q$

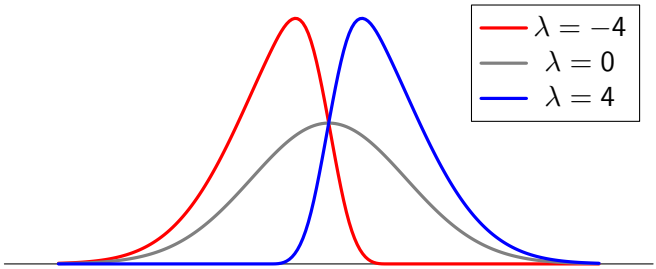
Measure relative location

- **Mead-Mean**: Mean of values between q_L and q_U
- **Trimmed-Mean**: Mean of values between q_p and $q_{(1-p)}$
- **Winsorized-Mean**: Map values smaller (bigger) than $q_p(q_{(1-p)})$ to $q_p(q_{(1-p)})$ and then compute the mean

Asymmetry

$$\bar{\mu}_3 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \bar{\mu})^3, \qquad \text{Skewness} = I_{as} = \frac{\bar{\mu}_3}{\bar{\sigma}^3}$$

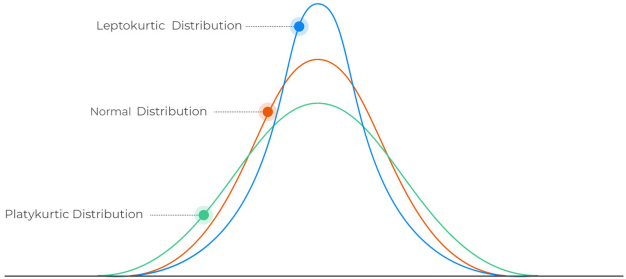
- $I_{as} > 0$ right asymmetry
- $I_{as} < 0$ left asymmetry
- $I_{as} = 0$ symmetric



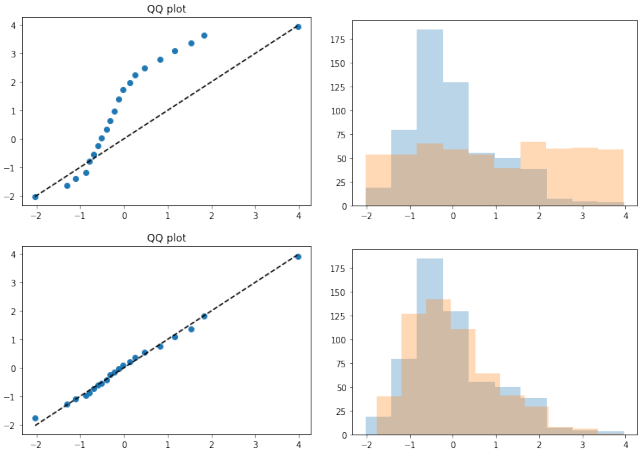
Empirical density

$$\bar{\mu}_4 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \bar{\mu})^4, \qquad \text{Kurtosis} = I_{kurt} = \frac{\bar{\mu}_4}{\bar{\sigma}^4} - 3$$

- $I_{kurt} > 0$ Hypernormal
- $I_{kurt} < 0$ Hyponormal
- $I_{kurt} = 0$ Normal



Comparing distributions - QQ-plots



Measure of correlation

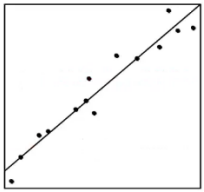
- **Sample covariance**:

$$cov(a_j, a_k) = \frac{1}{m-2} \sum_{i=1}^{m} (x_{ij} - \bar{\mu}_j)(x_{ik} - \bar{\mu}_k)$$

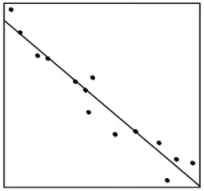
- **Sample LINEAR correlation**:

$$r_{jk} = \frac{cov(a_j, a_k)}{\bar{\sigma}_j \, \bar{\sigma}_j} \qquad \in [-1, 1]$$

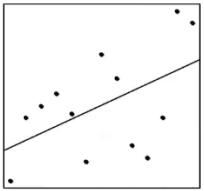
Linear Correlation



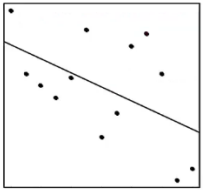
(a) r=0.96



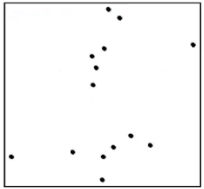
(b) r=-0.96



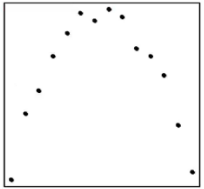
(c) r=0.6



(d) r=-0.6



(e) r=0.0



(f) r=0.0

Correlation on categorical attributes

Contingency tables

	family		
area	0	1	totale
1	2	4	6 (f_1)
2	4	2	6
3	2	5	7
4	3	3	6
totale	11 (g_1)	14 (g_2)	25

Two attributes are independent if