EPISODE 2 RECAP

TYPES OF DATA

BINARY

CATEGORICAL → NOMINAL (region, class)

CATEGORICAL → ORDINAL (military ranks)

NUMERICAL → COUNT DATA (children, phones)

NUMERICAL → CONTINUOUS DATA (return, weight)

VISUALISATION

- few values → frequency table, bar plot
- many values → histograms, box plot

LOCATION MEASURES (numerical data)

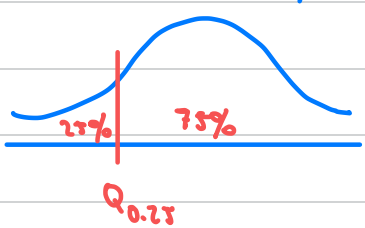
MEAN : $\frac{1}{n}\sum_{i=1}^{n} x_i$ — robust wrt extreme values

MEDIAN : value in the middle of ordered data

QUANTILE OF ORDER e.g. 0.25 (also called 25th percentile, 1st quartile)

↳ ≈ 25% obs fall on its left

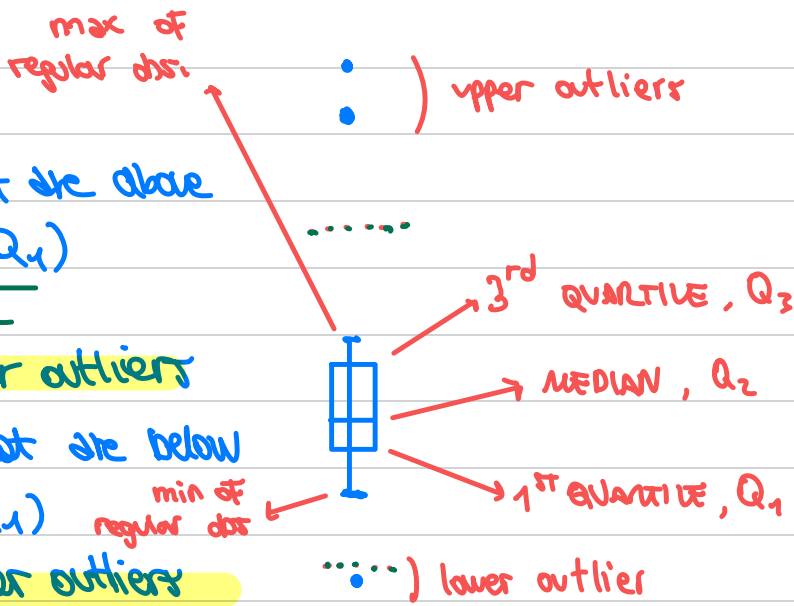
≈ 75% obs fall on its right



e.g. 90th percentile of italian salaries (gross, per year) is : 40k

VARIATION MEASURES (coming up!)

BOXPLOT



- Observations that are above $Q_3 + 1.5(Q_3 - Q_1)$ (IQR) are called **upper outliers**
- Observations that are below $Q_1 - 1.5(Q_3 - Q_1)$ are called **lower outliers**
- Observations that are not outliers are called **regular observations**

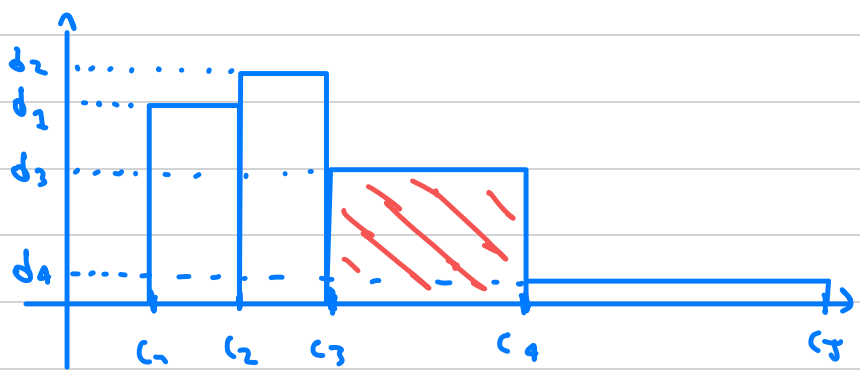
DISTRIBUTION SKEWNESS / SYMMETRY : 4 indications

	RIGHT / POSITIVE SKEW	LEFT / NEGATIVE SKEW	SYMMETRY
①	$(Q_3 - Me) > (Me - Q_1)$	$(Q_3 - Me) < (Me - Q_1)$	$(Q_3 - Me) \approx (Me - Q_1)$
②	upper whisker > lower whisker	upper whisker < lower whisker	upper whisker ≈ lower whisker
③	more upper outliers wrt lower outliers	more lower outliers wrt upper outliers	similar number of lower and upper outliers
④	Mean > Median	Mean < Median	Mean ≈ Median

VISUALISATION OF SINGLE NUMERICAL VARIABLE

HISTOGRAM

classes	abs. freq.	$p_k = \frac{f_k}{n}$ rel. freq.	$w_k = c_{k+1} - c_k$ class widths	$d_k = \frac{p_k}{w_k}$ freq. densities	$m_k = \frac{c_k + c_{k+1}}{2}$ mid. points
$[c_1, c_2)$	f_1	p_1	w_1	d_1	m_1
$[c_2, c_3)$	f_2	p_2	w_2	d_2	m_2
$[c_3, c_4)$	f_3	p_3	w_3	d_3	m_3
$[c_4, c_5)$	f_4	p_4	w_4	d_4	m_4



- The heights of a bar is the frequency density of the corresponding class. This way the area of a bar corresponds to the class relative frequency
- When the classes have same width, we can also use absolute or relative freq. as bar heights.

- From a histogram, or a distribution table for data grouped in classes, we can also get an approximation of the sample summary statistics such as mean, quantiles, variance and standard deviation.
- For example from the previous table

  MEAN $\approx \frac{1}{n} \sum_k f_k m_k$

  VARIANCE $\approx$ (coming up!)

  MEDIAN $\approx$ value that splits the histogram in two region of same area

  .
  .
  .

- All the APPROXIMATIONS above are obtained by assuming that all data are UNIFORMELY DISTRIBUTED within each interval class