Outline

- Data preprocessing
- Decomposing a dataset: instances and features
- Basic data descriptors
- Feature spaces and proximity (similarity, distance) measures
- Feature transformation for text data

Univariate descriptors 1/5

Let $x_1,\dots,x_n$ be a random sample of an attribute X.

Measures of **central tendency** of X include (most common numerical descriptive measure):

- (Arithmetic) **mean/ center/ average**:

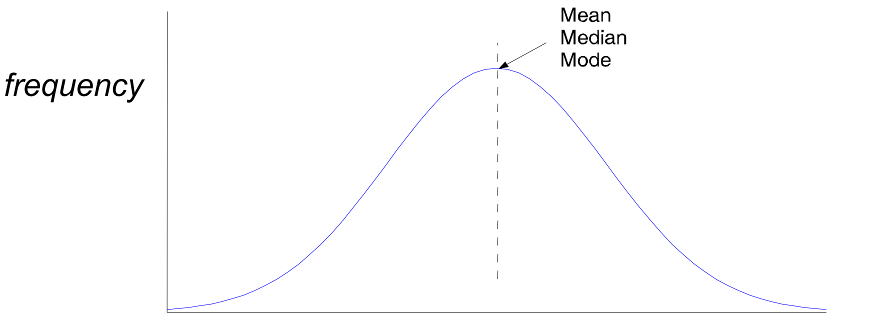
$$\bar{x} = \frac{1}{n}\sum_{1}^{n} x_i$$

- **Weighted average**:

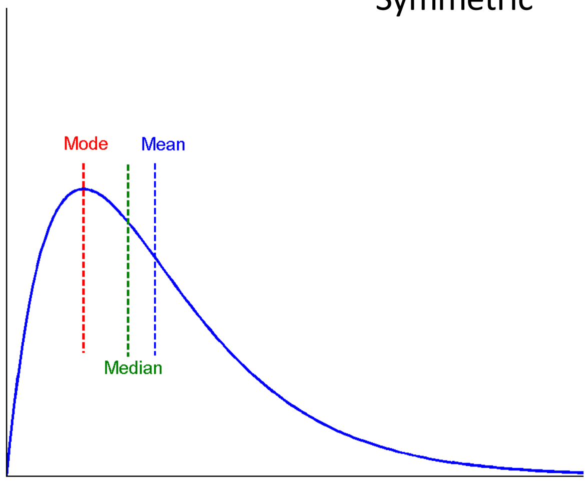
$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

- **Median**: the central element in ascending ordering
  - Middle value if odd number of values, or average of the middle two values otherwise
- **Mode**: Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal

Univariate descriptors 2/5



Symmetric



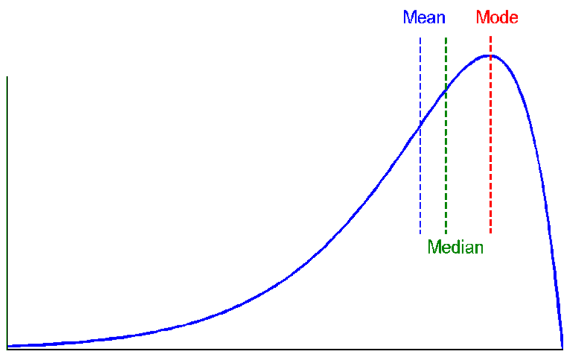
Positively skewed

Mean is susceptible to the influence of outliers.

Median is less influenced by outliers.

Mode can also be used for the description of non-numeric (e.g. categorical) variables.

Generally **better**: describing variables by more than one univariate descriptor,
e.g. mean + median indicates type of skew.



Negatively skewed

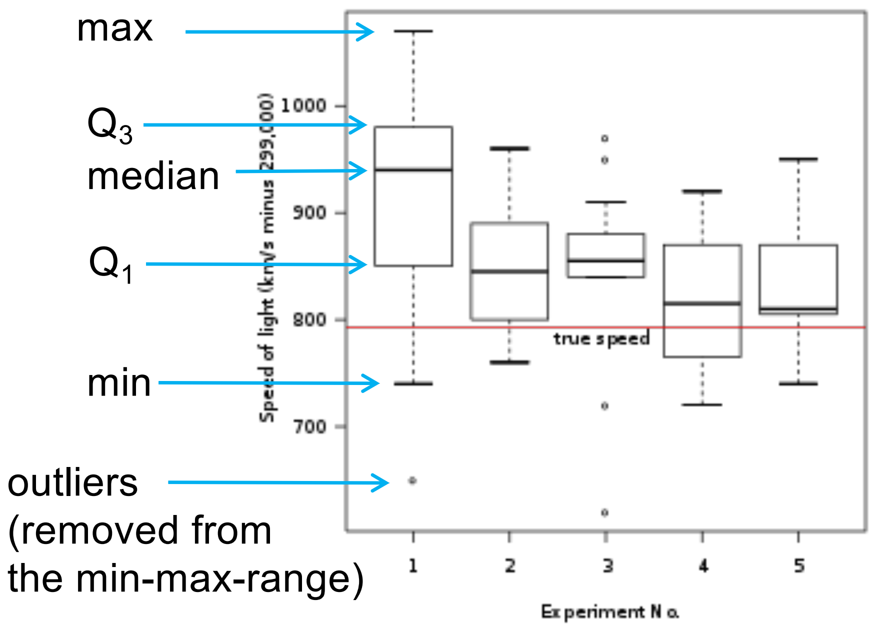
Univariate descriptors 3/5

Let $x_1,\ldots,x_n$ be a random sample of an attribute *X*. The **degree** to which X values **tend to spread** is called **dispersion** or **variance** of X :

- Range: max value – min value
  - Q_1 (25th percentile), Q_3 (75th percentile)
  - Median is the 50th percentile
- 5 number summary: min, Q_1, median, Q_3, max
  - Boxplots to visualize them
- Variance σ^2:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}[\sum_{i=1}^{n}x_i^2 - \frac{1}{n}(\sum_{i=1}^{n}x_i)^2]$$

- Standard deviation σ: $\sigma = \sqrt{\frac{1}{n}\cdot\sum_{i=1}^{n}(x_i - \bar{x})^2}$

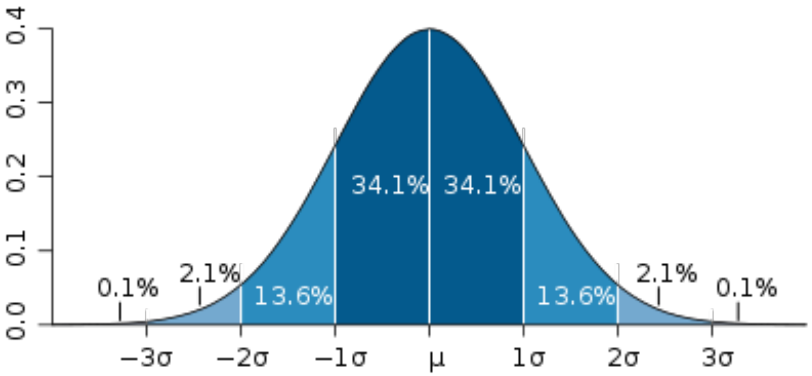


Box plot of data from the Michelson–Morley experiment
[source: https://en.wikipedia.org/wiki/Box_plot]

Univariate descriptors 4/5

Example: The normal distribution curve

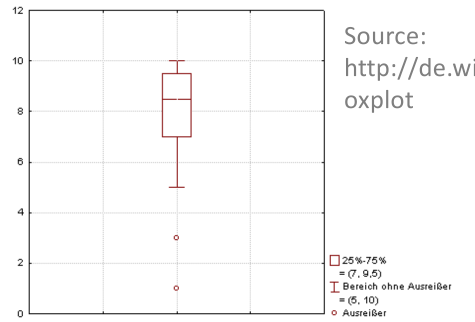
- ~68% of values drawn from a normal distribution are from μ-σ to μ+σ
- ~95% of the values lie from μ-2σ to μ+2σ
- ~99.7% of the values are from μ-3σ to μ+3σ



Source: http://en.wikipedia.org/wiki/Normal_distribution

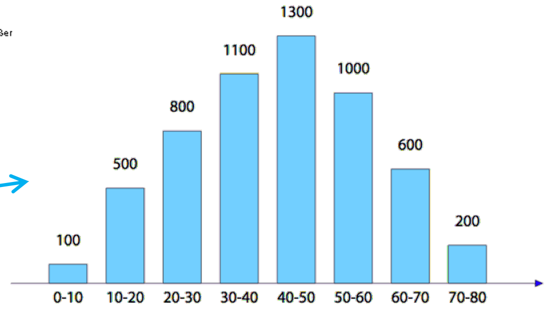
Univariate descriptors 5/5

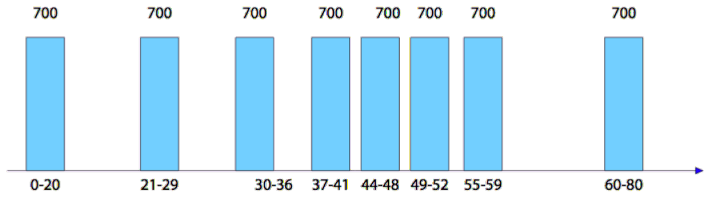
Let $x_1,\ldots,x_n$ be a random sample of an attribute X. For visual inspection of X, several types of charts are useful, e.g.:



Source: http://de.wikipedia.org/wiki/Boxplot

- Boxplots
  - 5 number summary
- Histograms:
  - Summarizes the distribution of X
  - X axis: attribute values, Y axis: frequencies
  - Absolute frequency: for each value *a*, h(*a*) = # occurrences of *a* in the sample
  - Relative frequency: f(*a*) = h(*a*)/*n*
- Different types of histograms, e.g.:
  - Equal width:
    - It divides the range into *N* intervals of equal size
  - Equal frequency/ depth:
    - It divides the range into *N* intervals, each containing approximately same number of samples
    - Inuition: High frequencies more important than low frequencies
      → higher resolution for high frequent values





Source: https://db.inf.uni-tuebingen.de/staticfiles/teaching/ws1011/db2/db2-selectivity.pdf

Bivariate descriptors 1/5

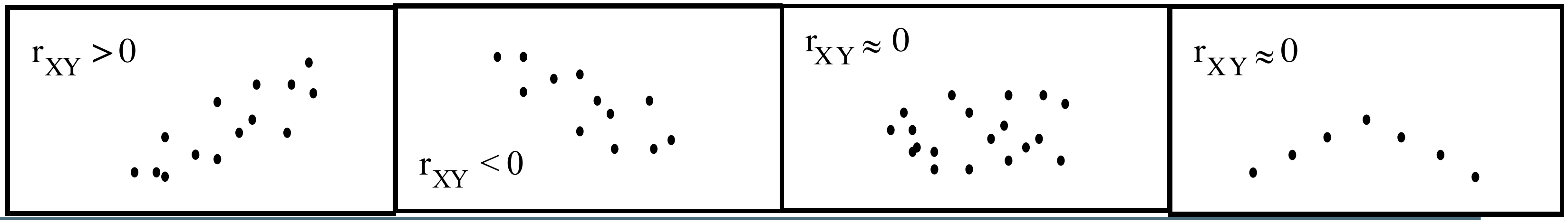
- Given two attributes X, Y one can measure how strongly they are correlated
  - For numerical data → correlation coefficient
  - For categorical data → χ^2 (chi-square)

Bivariate descriptors 2/5: for numerical features

- Correlation coefficient (also called Pearson's product moment coefficient) :

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\cdot(y_i - \bar{y})}{n\sigma_X\sigma_Y} = \frac{\sum_{i=1}^{n}(x_i y_i) - n\bar{x}\bar{y}}{n\sigma_X\sigma_Y}$$

  - n: # tuples; x_i, y_i: the values in the i^{th} tuple for X, Y
- value range: $-1 \leq r_{XY} \leq 1$
- the higher $|r_{XY}|$ the stronger the (anti-)correlation
  - $r_{XY} > 0$ positive correlation
  - $r_{XY} < 0$ negative correlation
  - $r_{XY} \sim 0$ no correlation/ independent



Bivariate descriptors 3/5: for numerical features

- Visual inspection of correlation

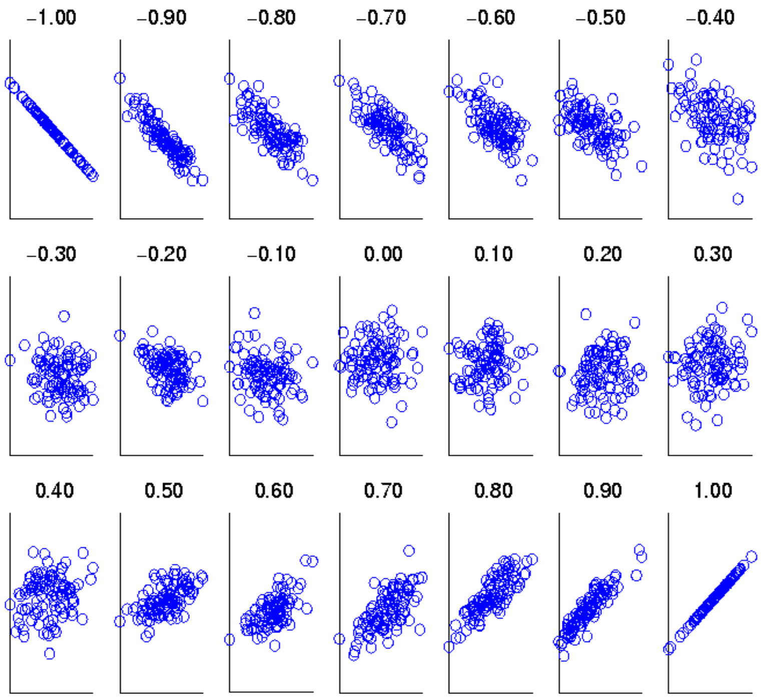


Figure 5.11. Scatter plots illustrating correlations from -1 to 1.

Bivariate descriptors 4/5: for categorical features

- Contingency table
  - For categorical/ nominal features $X=\{x_1, \ldots, x_c\}$, $Y=\{y_1, \ldots, y_r\}$
  - Represents the absolute frequency h_{ij} of each combination of values (x_i, y_J) and the marginal frequencies h_i, h_j of X, Y.

Attribute Y

Attribute X	Medium-term unemployment	Long-term unemployment	Total
No education	19	18	37
Teaching	43	20	63
Total	62	38	100

- Chi-square χ^2 test

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

o_{ij}:observed frequency
e_{ij}: expected frequency

$$e_{ij} = \frac{h_i h_j}{n}$$

Bivariate descriptors 4/5: for categorical features

- Chi-square example

o_{ij} observed frequencies e_{ij} expected frequencies

	Play chess	Not play chess	Sum (row)
Like science fiction	250 (90)	200 (360)	450
Not like science fiction	50 (210)	1000 (840)	1050
Sum(col.)	300	1200	1500

- X^2 (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- It shows that like_science_fiction and play_chess are correlated in the group

Outline

- Data preprocessing
- Decomposing a dataset: instances and features
- Basic data descriptors
- Feature spaces and proximity (similarity, distance) measures
- Feature transformation for text data