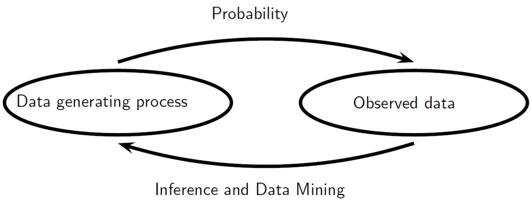
Statistical Methods for Data Science

Lesson 12 - Numerical summaries

Salvatore Ruggieri

Department of Computer Science
University of Pisa
salvatore.ruggieri@unipi.it

Condensed observations



- Probability models governs some random phenomena
- Confronted with a new phenomenon, we want to learn about the randomness that is associated with it
- Record observations $x_1, \ldots, x_n$ (a dataset)
- Can be too many: need to condense for easy comprehension and processing
- Numerical summaries:
  - Univariate: sample/empirical mean, median, standard deviation, quantiles, MAD
  - Multi-variate: Pearson's, Spearman's, Kendall's correlation coefficients

Sample summaries

- **Main idea:** translate summaries of distributions to samples
- **Purpose:** Sample summaries should be estimators of the summaries on the generating distribution
- Measures of centrality
  - *Sample mean*:

$$\bar{x}_n = \frac{x_1 + \ldots + x_n}{n} \qquad\qquad E[X], \mu$$

  - *Median* for sorted $x_1, \ldots, x_n$:

$$Med(x_1, \ldots, x_n) = \begin{cases} x_{\frac{n}{2}+1} & \text{if } n \text{ is odd} \\ (x_{\frac{n}{2}} + x_{\frac{n}{2}+1})/2 & \text{if } n \text{ is even} \end{cases} \qquad\qquad F^{-1}(0.5)$$

E.g., $Med(2, 3, 4) = 3$ and $Med(2, 3, 4, 5) = 3.5$

Measures of variability

- *Sample variance*:

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 \qquad\qquad Var(X), \sigma^2$$

  Why $n-1$ and not n?

- *Sample standard deviation*:

$$s_n = \sqrt{s_n^2} \qquad\qquad \sqrt{Var(X)}, \sigma$$

- Median of absolute deviations (MAD):

$$MAD(x_1, \ldots, x_n) = Med(|x_1 - Med(x_1, \ldots, x_n)|, \ldots, |x_n - Med(x_1, \ldots, x_n)|)$$

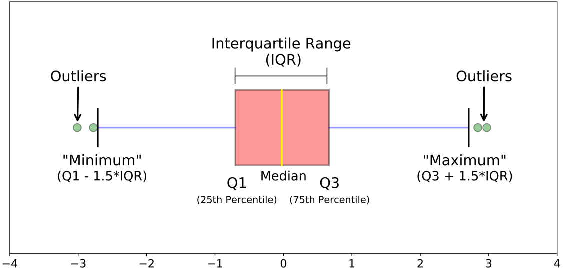
  - What is $MAD(X)$ for $X \sim F$?
  - For F symmetric and $E[X] = 0$, $MAD(X) = F^{-1}(0.75)$. Hence, $\sigma = c_F \cdot MAD$

Order statistics

- The order statistics consist of the same elements in the dataset, but in ascending order
- Let $x_{(1)}, \ldots, x_{(n)}$ be `sort(`$x_1, \ldots, x_n$`)`
- Empirical quantiles:
$$q(\frac{i-1}{n-1}) = x_{(i)}$$
E.g., for $2, 3, 4, 5, 6$, $q(0) = 2$, $q(0.25) = 3$, $q(0.5) = 4$, $q(0.75) = 5$, $q(1) = 6$
- What is $q(p)$ when p is not in the form above?
$$q(p) = x_{(k)} + \alpha(x_{(k+1)} - x_{(k)})$$
where $k = \lfloor p \cdot (n-1) + 1 \rfloor$ and $\alpha = p \cdot (n-1) + 1 - k$ (remainder)
- This is type=7 in R `quantile` function. There are 9 variants!
- The definition in the textbook is type=6

See R script

The box-and-whisker plot



- Axis here is with reference to a standard Normal distribution
- **See John Tukey** (designed FFT, coined 'bit' & 'software', and visionary of **data science**)

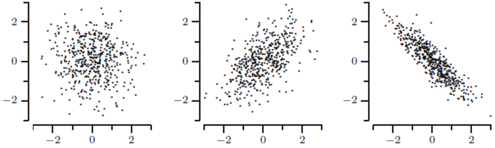
Correlation coefficients: Pearson's r

- **Correlation** is a bivariate analysis that measures the strength of association between two variables and the direction of the relationship.

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X) \cdot Var(Y)}} = \frac{E[(X-\mu_X)\cdot(Y-\mu_Y)]}{\sigma_X \cdot \sigma_Y}$$

- Pearson's (linear/product-moment) correlation coefficient: *[support in $[-1,1]$]*

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})}{(n-1)\cdot s_x \cdot s_y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \cdot \sum_{i=1}^{n}(y_i-\bar{y})^2}}$$



Uncorrelated **Positively correlated** **Negatively correlated**

- Computational cost is $O(n)$

Correlation coefficients: Spearman's ρ

- Pearsons's r asseses linear relationships over continuous values
- Let $rank(x)$ be the ranks of x_i's
  - For $x = 7, 3, 5$, $rank(x) = 3, 1, 2$
- Spearman's correlation coefficient is the Pearson's coefficient over the ranks:

$$\rho = r(rank(x), rank(y)) \qquad \frac{Cov(rank(X), rank(Y))}{\sqrt{Var(rank(X)) \cdot Var(rank(Y))}}$$

  - Only if all ranks in $rank(x)$ and $rank(y)$ are distinct:

$$\rho = 1 - \frac{6\sum_{i=1}^{n}(rank(x)_i - rank(y)_i)^2}{n \cdot (n^2 - 1)}$$

- Spearman's correlation assesses monotonic relationships (whether linear or not)
- Computational cost is $O(n \cdot \log n)$

Correlation coefficients: Kendall's τ

- Spearman's is a measure of rank correlation, i.e., degree of similarity between the sample ranks of two variables. We don't use it if e.g., Y is binary valued
- Kendall's is another such measure: *[support in $[-1, 1]$]*

$$\tau_{xy} = \frac{2\sum_{i<j} sgn(x_i - x_j) \cdot sgn(y_i - y_j)}{n \cdot (n-1)} \qquad E[sgn(X_1 - X_2) \cdot sgn(Y_1 - Y_2)]$$

  Fraction of concordant pairs minus discordant pairs, i.e., probability of observing a difference between concordant and discordant pairs.
- Correction τ_b accounting for ties, i.e., $x_i = x_j$ or $y_i = y_j$ *[implemented by `cor` in R]*
- Computational cost is $O(n^2)$

See R script

Correlation coefficients: Somers' D

- An asymmetric Kendall's:

$$D = \frac{\tau_{xy}}{\tau_{yy}} = \frac{\sum_{i<j} sgn(x_i - x_j) \cdot sgn(y_i - y_j)}{\sum_{i<j} sgn(y_i - y_j)^2}$$

  i.e., fraction of concordand pairs minus discordant pairs conditional to unequal values of y
- Example with probabilistic classifiers
  - $x =$ probabilities of positive classification, i.e., `predict_proba(...)[,1]`
  - y true class
  - D is the Gini index of classifier performances
  - related to AUC of ROC curve:

$$D = 2 \cdot AUC - 1 \qquad AUC = \frac{D}{2} + 0.5 = \frac{\tau_{xy}}{2 \cdot \tau_{yy}} + 0.5$$

See R script