Master Program in *Data Science and Business Informatics*

Statistics for Data Science

Lesson 16 - Numerical summaries

Salvatore Ruggieri

Department of Computer Science
University of Pisa, Italy
salvatore.ruggieri@unipi.it
• Probability models govern some random phenomena
• Confronted with a new phenomenon, we want to learn about the randomness associated with it
  ▶ Parametric (efficient) vs non-parameteric (general) methods
• Record observations $x_1, \ldots, x_n$ (a dataset)
• $n$ can be large: 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 (plug-in method):** translate summaries of empirical distribution (sample) $F_n$ to estimate summaries of the generating distribution $F$

- Measures of centrality
  - *Sample mean:*
    \[ \bar{x}_n = \frac{x_1 + \ldots + x_n}{n} \]
  - *Median* for sorted $x_1, \ldots, x_n$:
    \[ Med(x_1, \ldots, x_n) = \begin{cases} 
      x_{\frac{n}{2}+1} & \text{if } n \text{ is odd} \\
      (x_{\frac{n}{2}} + x_{\frac{n}{2}+1})/2 & \text{if } n \text{ is even}
    \end{cases} \]

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

- **Sample variance:**
  \[
  s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2
  \]
  \[
  \text{Var}(X), \quad \sigma^2
  \]
  Divide by \(n-1\) for a sample, and by \(n\) for a population!  
  \[\text{[Bessel's correction]}\]

- **Sample standard deviation:**
  \[
  s_n = \sqrt{s_n^2}
  \]
  \[
  \sqrt{\text{Var}(X)}, \quad \sigma
  \]

- **Median of absolute deviations (MAD):**
  \[
  \text{MAD}(x_1, \ldots, x_n) = \text{Med}(|x_1 - \text{Med}(x_1, \ldots, x_n)|, \ldots, |x_n - \text{Med}(x_1, \ldots, x_n)|)
  \]
  - For \(X \sim F\), the population MAD is \(Md = G^{-1}(0.5)\) where \(|X - F^{-1}(0.5)| \sim G\)
  - For \(F\) symmetric, \(Md = F^{-1}(0.75) - F^{-1}(0.5)\).
  - \(Md\) is a more robust-to-outlier measure of scale than standard deviation
  
  **See R script**
Order statistics

- The order statistics consist of the same elements in the dataset, but in ascending order.
- Let $x_1, \ldots, x_n$ be $\text{sort}(x_1, \ldots, x_n)$.
- Empirical quantiles:
  \[ q\left(\frac{i - 1}{n - 1}\right) = x_i \]
  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)
Association and correlation

• Bivariate analysis of joint distribution of $X$ and $Y$ or of a sample $(x_1, y_1), \ldots, (x_n, y_n)$

• **Association**: one variable provides information on the other
  ▶ $X \perp Y$ independent: zero information
  ▶ $Y = f(X)$ deterministic association: maximum information

• **Correlation**: the two variables show an increasing/decreasing trend
  ▶ $X \perp Y$ implies $Cov(X, Y) = 0$
  ▶ the converse is not always true

• **Coefficient or measure of association/correlation**: determine the strength of association/correlation between two variables and the direction of the relationship
Measures of association

- Dimension: level of measurement
  - Ordinal: discrete but ordered, e.g., 0, 1, 2 for “low”, “medium”, “severe” risks
  - Nominal: discrete without any order, e.g., 0, 1, 2 for “bus”, “car”, “train” transportation

- Other dimensions: symmetric/asymmetric, one way/two way

- See [Khamis, 2008] for a guide to the selection
- See [Berry et al., 2018] for extensive introduction

- See mhahsler.github.io for a list of measures in association rule mining $X \Rightarrow Y$
Linear correlation of continuous r.v.: Pearson’s $r$

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

- Pearson’s (linear/product-moment) correlation coefficient: $[\text{support in } [-1, 1]]$

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

- Computational cost is $O(n)$
Rank correlation of continuous/ordinal r.v.: Spearman’s $\rho$

- Pearson’s $r$ assesses linear relationships over continuous values
- Let $\text{rank}(x)$ be the ranks of $x_i$’s
  - For $x = 7, 3, 5$, $\text{rank}(x) = 3, 1, 2$
- Spearman’s correlation coefficient is the Pearson’s coefficient over the ranks:
  
  $$
  \rho = r(\text{rank}(x), \text{rank}(y)) = \frac{\text{Cov}(\text{rank}(X), \text{rank}(Y))}{\sqrt{\text{Var}(\text{rank}(X)) \cdot \text{Var}(\text{rank}(Y))}}
  $$

  - In case of no ties in $x$ and $y$:
    
    $$
    \rho = 1 - \frac{6 \sum_{i=1}^{n} (\text{rank}(x)_i - \text{rank}(y)_i)^2}{n \cdot (n^2 - 1)}
    $$

- Spearman’s correlation assesses monotonic relationships (whether linear or not)
- Computational cost is $O(n \cdot \log n)$
Rank correlation of continuous/ordinal r.v.: Kendall’s $\tau$

- Spearman’s applies when $Y$ (or also $X$) is ordinal
  - Association between age and fraction of exams passed ("0-5", "6-10", "11-15", "+15")
- Kendall’s $\tau_a$ is another (more robust) rank measure:  
  \[ \tau_{xy} = \frac{2 \sum_{i<j} \text{sgn}(x_i - x_j) \cdot \text{sgn}(y_i - y_j)}{n \cdot (n - 1)} \]
  Fraction of concordant pairs minus discordant pairs, i.e., probability of observing a difference between concordant and discordant pairs.
- Correction $\tau_b$ accounting for ties, i.e., $x_i = x_j$ or $y_i = y_j$  
  \[ E[\text{sgn}(X_1 - X_2) \cdot \text{sgn}(Y_1 - Y_2)] \]
  [implemented by cor in R]
- Computational cost is $O(n^2)$

See R script
Rank correlation of continuous and binary r.v.: Somers’ D

- X continuous and Y binary.
- An asymmetric Kendall’s:

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

i.e., fraction of concordant 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:

\[ D = 2 \cdot AUC - 1 \quad AUC = \frac{D}{2} + 0.5 = \frac{\tau_{xy}}{2 \cdot \tau_{yy}} + 0.5 \]

See R script
\[ Gini = D = \frac{A}{A + B} \]

\[ AUC = A \]
Association between nominal variables: Thiel’s U

Mutual information and NMI

\[ I(X, Y) = \sum_{a,b} p_{X|Y}(a, b) \log \frac{p_{X|Y}(a, b)}{p_X(a)p_Y(b)} \]

\[ NMI = \frac{I(X, Y)}{\min\{H(X), H(Y)\}} \in [0, 1] \]

- Uncertainty coefficient (also called entropy coefficient or Thiel’s U):
  \[ U_{sym} = \frac{I(X, Y)}{(H(X) + H(Y))/2} \]
  \[ U_{asym} = \frac{I(X, Y)}{H(X)} \]

where \( p_{X|Y} \) is the empirical joint p.m.f., and \( p_X, p_Y \) the empirical marginal p.m.f.’s

- \( U_{asym} \) what fraction of \( X \) can be predicted by \( Y \)
Association between nominal variables: $\chi^2$-based

• Several other measures based on Pearson $\chi^2$ (introduced later)
  ▶ Contingency coefficient $C$
  ▶ Cramer’s $V$
  ▶ $\phi$ coefficient (or MCC, Matthews correlation coefficient)
  ▶ Tschuprov’s $T$
  ▶ ...

See R script
Optional references

Harry Khamis (2008)

Measures of Association: How to Choose?


*Springer.*