

Master Program in *Data Science and Business Informatics*

Statistics for Data Science

Lesson 16 - Numerical summaries

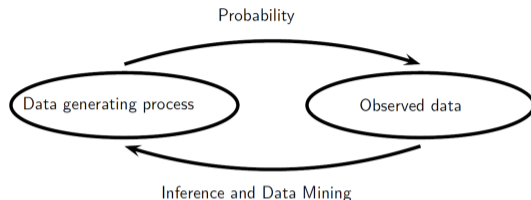
Salvatore Ruggieri

Department of Computer Science

University of Pisa, Italy

salvatore.ruggieri@unipi.it

Condensed observations: numerical summaries



- Probability models governs some random phenomena
- Confronted with a new phenomenon, we want to learn about the randomness associated with it
 - ▶ Parametric (efficient) vs non-parametric (general) methods
- Record observations x_1, \dots, 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:*

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

- ▶ *Median* for sorted x_1, \dots, x_n :

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

E.g., $\text{Med}(2, 3, 4) = 3$ and $\text{Med}(2, 3, 4, 5) = 3.5$

Measures of variability

- *Sample variance:*

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

Divide by $n - 1$ for a sample, and by n for a population!

[Bessel's correction]

- *Sample standard deviation:*

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

- Median of absolute deviations (*MAD*):

$$\text{MAD}(x_1, \dots, x_n) = \text{Med}(|x_1 - \text{Med}(x_1, \dots, x_n)|, \dots, |x_n - \text{Med}(x_1, \dots, x_n)|)$$

- ▶ 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

Order statistics and empirical quantiles

- Let $x_{\langle 1 \rangle}, \dots, x_{\langle n \rangle}$ be $\text{sort}(x_1, \dots, x_n)$. We call $x_{\langle i \rangle}$ the i -th order statistics.
 - ▶ The order statistics consist of the same elements in the dataset, but in ascending order
- Distribution quantiles $q_p = \inf_x \{P(X \leq x) \geq p\} = \inf_x \{F(x) \geq p\}$ *[See Lesson 08]*
- Empirical quantiles: $q(p) = \inf_x \{F_n(x) \geq p\} = \inf_x \{|\{i \mid x_i \leq x\}|/n \geq p\}$
 - ▶ Type 6 (book [T]): for $p = i/(n+1)$ *[There are 9 variants, see help(quantile)]*

$$q(p) = x_{\langle p \cdot (n+1) \rangle} = x_{\langle i \rangle}$$

□ E.g., for 2, 3, 4, 5, 6, $q(.167) = 2$, $q(.333) = 3$, $q(0.5) = 4$, $q(0.667) = 5$, $q(.833) = 6$

- ▶ Type 7 (default in R): for $p = (i-1)/(n-1)$

$$q(p) = x_{\langle p \cdot (n-1) + 1 \rangle} = x_{\langle i \rangle}$$

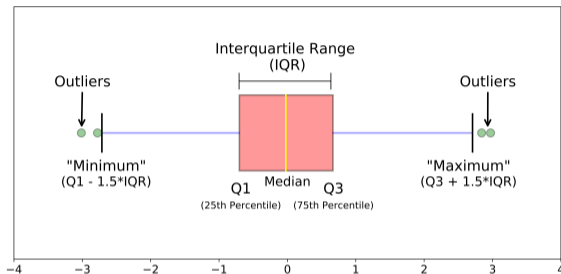
□ E.g., for 2, 3, 4, 5, 6, $q(0) = 2$, $q(0.25) = 3$, $q(0.5) = 4$, $q(0.75) = 5$, $q(1) = 6$

- What is $q(p)$ when $p \cdot (n+1)$ is not an integer?

$$q(p) = x_{\langle k \rangle} + \alpha(x_{\langle k+1 \rangle} - x_{\langle k \rangle})$$

where $k = \lfloor p \cdot (n+1) \rfloor$ and $\alpha = p \cdot (n+1) - k$ (remainder)

The box-and-whisker plot

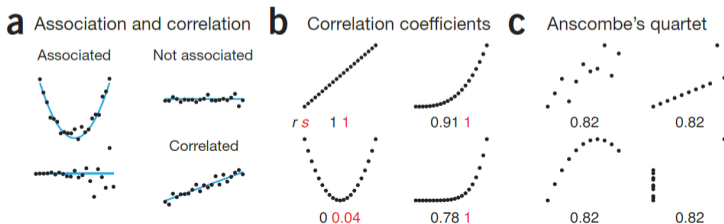


- Axis here is with reference to a standard Normal distribution
- **See John Tukey** (designed FFT, coined 'bit' & 'software', and visionary of **data science**)

See R script

Association and correlation

- Bivariate analysis of joint distribution of X and Y or of a sample $(x_1, y_1), \dots, (x_n, y_n)$
- *Association*: one variable provides information on the other
 - ▶ $X \perp\!\!\!\perp Y$ independent, i.e., $P(X|Y) = P(X)$: zero information
 - ▶ $Y = f(X)$ deterministic association with f invertible: maximum information
- *Correlation*: the two variables show an increasing/decreasing trend
 - ▶ $X \perp\!\!\!\perp Y$ implies $\text{Cov}(X, Y) = 0$
 - ▶ 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

Variable Y	Variable X		
	Nominal	Ordinal	Continuous
Nominal	ϕ or λ	Rank biserial	Point biserial
Ordinal	Rank biserial	τ_b or Spearman	τ_b or Spearman
Continuous	Point biserial	τ_b or Spearman	Pearson or Spearman

ϕ = phi coefficient, λ = Goodman and Kruskal's lambda,
 τ_b = Kendall's τ_b .

- 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
- See [[Khamis, 2008](#)] for a guide to the selection
- See [[Berry et al., 2018](#)] for extensive introduction
- See [mhahsler.github.io](https://github.com/mhahsler) for a list of measures in association rule mining $X \Rightarrow Y$

Linear correlation of continuous r.v.: Pearson's r

- Bivariate analysis of joint distribution of X and Y or of a sample $(x_1, y_1), \dots, (x_n, y_n)$
- Sample covariance:

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

$$\text{Cov}(X, Y) = E[(X - \mu_X) \cdot (Y - \mu_Y)]$$

- Apply plug-in method to correlation between X and Y :

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

- Pearson's (linear/product-moment) correlation coefficient:

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

- Support in $[-1, 1]$ due to Schwarz's inequality: $|s_{xy}| \leq s_x \cdot s_y$
- Computational cost is $O(n)$

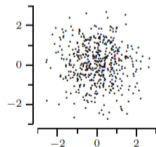
Linear correlation of continuous r.v.: Pearson's r

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

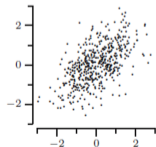
- Pearson's (linear/product-moment) correlation coefficient:

[support in $[-1, 1]$]

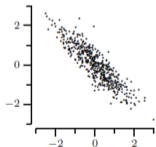
$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}}$$



Uncorrelated



Positively correlated



Negatively correlated

r	Interpretation of Linear Relationship
0.8	Strong positive
0.5	Moderate positive
0.2	Weak positive
0.0	No relationship
-0.2	Weak negative
-0.5	Moderate negative
-0.8	Strong negative

Rank correlation of continuous/ordinal r.v.: Spearman's ρ

- Pearson's r assesses *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:

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

- ▶ In case of no ties in x and y :

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

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

Rank correlation of continuous/ordinal r.v.: Kendall's τ

- 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 τ_a is another (more robust) rank measure: *[support in $[-1, 1]$]*

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

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

Rank correlation of continuous and binary r.v.: Somers' D

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

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

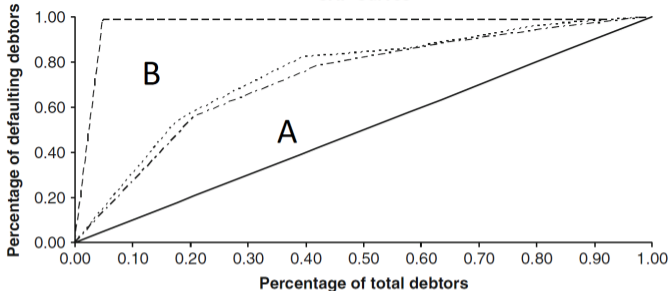
i.e., fraction of concordant pairs minus discordant pairs conditional to unequal values of y

- 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:

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

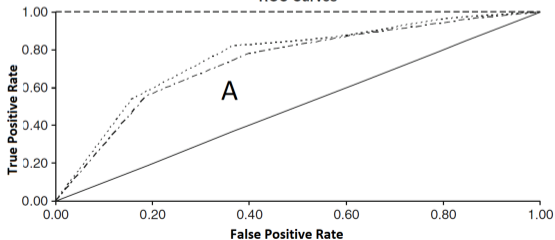
See R script

CAP Curves



$$Gini = D = A / (A + B)$$

ROC Curves



$$AUC = A + 1/2$$

Association between nominal variables: Thiel's U

- Recall from Lesson 11

Mutual information and NMI

$$I(X, Y) = \sum_{a,b} p_{XY}(a, b) \log \frac{p_{XY}(a, b)}{p_X(a)p_Y(b)} \quad NMI = \frac{I(X, Y)}{\min \{H(X), H(Y)\}} \in [0, 1]$$

- Uncertainty coefficient (also called entropy coefficient or Thiel's U) :

$$U_{sym} = \frac{I(X, Y)}{(H(X) + H(Y))/2} \quad U_{asym} = \frac{I(X, Y)}{H(X)}$$

where p_{XY} is the *empirical joint p.m.f.*, and p_X, p_Y are the *empirical marginal p.m.f.'s*

- U_{asym} what fraction of X can be predicted by Y

Association between nominal variables: χ^2 -based

- Several other measures based on Pearson χ^2 (introduced later)
 - ▶ Contingency coefficient C
 - ▶ Cramer's V
 - ▶ ϕ coefficient (or MCC, Matthews correlation coefficient)
 - ▶ Tschuprov's T
 - ▶ ...

See R script

Optional references



Harry Khamis (2008)

Measures of Association: How to Choose?

J. of Diagnostic Medical Sonography, Vol. 24, Issue 3, pages 155–162.



Kenneth J. Berry, Janis E. JohnstonPaul, and W. Mielke, Jr. (2018)

The Measurement of Association: A Permutation Statistical Approach.

Springer.