National Ph.D. Program in *Artificial Intelligence for Society*

Statistics for Machine Learning

Lesson 05 - Law of Large Numbers and Central Limit Theorem. Graphical and Numerical Summaries.

Andrea Pugnana, Salvatore Ruggieri

Department of Computer Science
University of Pisa, Italy
andrea.pugnana@di.unipi.it salvatore.ruggieri@unipi.it

Markov's inequality

Notation. Indicator function: $\mathbb{1}_{\varphi}(x) = \begin{cases} 1 & \text{if } \varphi(x) \\ 0 & \text{otherwise} \end{cases}$

- Link expectation to probability of events
- $E[\mathbb{1}_{X \geq \alpha}] = \sum_a \mathbb{1}_{X \geq \alpha}(a) p_X(a) = \sum_{a \geq \alpha} p_X(a) = P_X(X \geq \alpha)$

- Question: how much probability mass is near the expectation?

  Markov's inequality. Assume $X \geq 0$, and $\alpha > 0$:

  $$P(X \geq \alpha) \leq \frac{E[X]}{\alpha}$$

  Proof. Take expectations of $\alpha \mathbb{1}_{X \geq \alpha} \leq X$. □

- For a non-negative r.v., the probability of a large value is inversely proportional to the value

  Corollary. Assume $X \geq 0$, $E[X] > 0$ and $k > 0$. We have: $P(X \geq kE[X]) \leq \frac{1}{k}$

Chebyshev's inequality

- Question: how much probability mass is near the expectation?

CHEBYSHEV'S INEQUALITY. For an arbitrary random variable Y and any $a > 0$:

$$\mathrm{P}(|Y - \mathrm{E}[Y]| \geq a) \leq \frac{1}{a^2}\mathrm{Var}(Y)\,.$$

Proof. Let $X = (Y - E[Y])^2$ and $\alpha = a^2$. By Markov's inequality:

$$P(|Y - E[Y]| \geq a) = P((Y - E[Y])^2 \geq a^2) \leq \frac{E[(Y - E[Y])^2]}{a^2} = \frac{1}{a^2}Var(Y)$$

□

Chebyshev's inequality

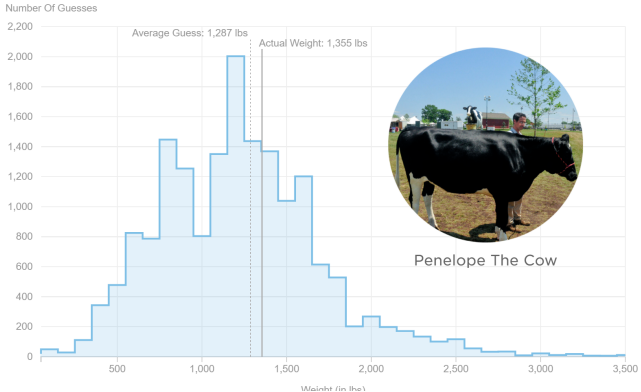
- **"$\mu \pm$ a few σ" rule:** Most of the probability mass of a random variable is within a few standard deviations from its expectation!
- Let $\mu = E[Y]$ and $\sigma^2 = Var(Y) > 0$. For $k > 0$ (and hence $a = k\sigma > 0$):

$$P(|Y - \mu| < k\sigma) = 1 - P(|Y - \mu| \geq k\sigma) \geq 1 - \frac{1}{k^2\sigma^2} Var(Y) = 1 - \frac{1}{k^2}$$

- For $k = 2, 3, 4$, the RHS is $3/4, 8/9, 15/16$
- Chebyshev's inequality is sharp when nothing is known about X, but in general it is a large bound!

Averages vary less

- $E[X]$ is a key summary of a distribution! How to "guess" it?
- Guessing the weight of a cow



- **The Wisdom of Crowds**

Expectation and variance of an average

- Let $X_1, X_2, \ldots, X_n$ be independent r. v. for which $E[X_i] = \mu$ and $Var(X_i) = \sigma^2$

$$\bar{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

Expectation and variance of an average. If $\bar{X}_n$ is the average of n independent random variables with the same expectation μ and variance σ^2, then

$$\mathrm{E}\left[\bar{X}_n\right] = \mu \quad \text{and} \quad \mathrm{Var}\left(\bar{X}_n\right) = \frac{\sigma^2}{n}.$$

- Notice that $X_1, \ldots, X_n$ are not required to be identically distributed!

The (weak) law of large numbers

- Apply Chebyshev's inequality to $\bar{X}_n$

$$P(|\bar{X}_n - \mu| > \epsilon) \leq \frac{1}{\epsilon^2} Var(\bar{X}_n) = \frac{\sigma^2}{n\epsilon^2}$$

- For $n \to \infty$, $\sigma^2/(n\epsilon^2) \to 0$

> THE LAW OF LARGE NUMBERS. If $\bar{X}_n$ is the average of n independent random variables with expectation μ and variance σ^2, then for any $\varepsilon > 0$:
>
> $$\lim_{n\to\infty} \mathrm{P}\big(|\bar{X}_n - \mu| > \varepsilon\big) = 0.$$

- probability that $\bar{X}_n$ is far from μ tends to 0 as $n \to \infty$! *[**Convergence in probability**]*
- It holds also if σ^2 is infinite (proof not included)
- Notice (again!) that $X_1, \ldots, X_n$ are not required to be identically distributed!

Recovering probability of an event

Objective: We want to know $p = P(a < X \leq b)$

- Run n independent measurements
- Model the results as $X_1, \ldots, X_n$ random variables
- Define the indicator variables, for $i = 1, \ldots, n$:

$$Y_i = \mathbb{1}_{a<X_i\leq b} = \begin{cases} 1 & \text{if } a < X_i \leq b \\ 0 & \text{otherwise} \end{cases}$$

- Y_i's are independent *[by propagation of independence]*
- $E[Y_i] = P(a < X \leq b) = p$ and $Var(Y_i) = p(1-p)$
- Defined $\bar{Y}_n = \frac{Y_1+\ldots+Y_n}{n}$, by the law of large numbers:

$$\lim_{n\to\infty} P(|\bar{Y}_n - p| > \epsilon) = 0$$

- Frequency counting of values $(a, b]$ (e.g., in histograms) is a prob. estimation method!

Estimating conditional probability

Objective: estimate $p = P(C = c|A = a) = P(A = a, C = c)/P(A = a) = p_{ac}/p_a$

- Run n independent measurement
- Model the results as $(A_1, C_1), \ldots, (A_n, C_n)$
- Using the approach of previous slide (but with the **strong LLN**):
  - for $Y_i = \mathbb{1}_{A_i=a, C_i=c}$: $P(\lim_{n\to\infty} \bar{Y}_n = p_{ac}) = 1$ where $p_{ac} = P(A = a, C = c)$
  - for $Z_i = \mathbb{1}_{A_i=a}$: $P(\lim_{n\to\infty} \bar{Z}_n = p_a) = 1$ where $p_a = P(A = a)$
- if $\bar{Z}_n \neq 0$, from previous two statements: (limit of a ratio is the ratio of the limits)

$$P(\lim_{n\to\infty} \frac{\bar{Y}_n}{\bar{Z}_n} = \frac{p_{ac}}{p_c}) = 1$$

- Sample usage: almost everywhere in Machine Learning
- Issues: when n is small (e.g., rare values)

Hoeffding bound

Theorem (Hoeffding bound)

If $\bar{X}_n$ is the average of n independent r.v. with expectation μ and $P(a \leq X_i \leq b) = 1$, then for any $\epsilon > 0$

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq 2e^{-2n\epsilon^2/(b-a)^2}$$

- For bounded support, a tight upper bound!
- When $a = 0, b = 1$ (e.g., Bernoulli trials):

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq 2e^{-2n\epsilon^2}$$

- **Other concentration inequalities**.

The central limit theorem

- Let $X_1, X_2, \ldots, X_n$ be independent r. v. for which $E[X_i] = \mu$ and $Var(X_i) = \sigma^2$

$$\bar{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n} \quad E[\bar{X}_n] = \mu \quad Var(\bar{X}_n) = \frac{\sigma^2}{n}$$

- Can we derive the distribution of $\bar{X}_n$?
- Assume $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ with μ and σ^2 known. We have:

$$\bar{X}_n \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}) \qquad Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

- Interestingly, the same conclusion extends to any other distribution!

The central limit theorem

THE CENTRAL LIMIT THEOREM. Let $X_1, X_2, \ldots$ be any sequence of independent identically distributed random variables with finite positive variance. Let μ be the expected value and σ^2 the variance of each of the X_i. For $n \geq 1$, let Z_n be defined by

$$Z_n = \sqrt{n}\,\frac{\bar{X}_n - \mu}{\sigma};$$

then for any number a

$$\lim_{n\to\infty} F_{Z_n}(a) = \Phi(a),$$

where Φ is the distribution function of the $N(0,1)$ distribution. In words: the distribution function of Z_n converges to the distribution function Φ of the standard normal distribution.

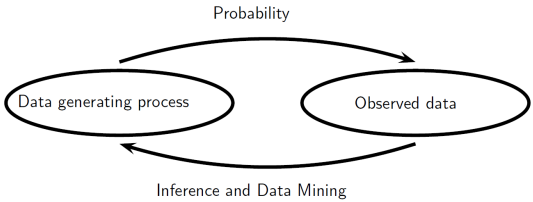
- It extends to not identically distributed r.v.'s [**Lindeberg's condition**]
- Why is it so frequent to observe a normal distribution?
  - Sometime it is the average/sum effects of other variables, e.g., as in "noise"
  - This justifies the common use of it to stand in for the effects of unobserved variables

How large should n be?

- How fast is the convergence of Z_n to $\mathcal{N}(0,1)$?
- The approximation might be poor when:
  - n is small
  - X_i is asymmetric, bimodal, or discrete
  - the value to test (0.6 in our example) is far from μ

the myth of $n \geq 30$

Condensed observations: graphical 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-parameteric (general) methods
- Record observations $x_1, \ldots, x_n$ (a dataset)
- n can be large: need to condense for easy visual comprehension
- Graphical summaries:
  - Univariate: empirical distribution functions, histograms, kernel density estimates
  - Multi-variate: kernel density estimates, scatter plots

The empirical CDF

- A r.v. X is completely characterized by its CDF F
- Record observations $x_1, \ldots, x_n$ (a dataset)
- Empirical cumulative distribution function (ECDF):

$$F_n(x) = \frac{|\{i \in [1, n] \mid x_i \leq x\}|}{n}$$

- Empirical complementary cumulative distribution function (ECCDF): $\bar{F}_n(x) = 1 - F_n(x)$
- Estimating F through F_n **[Glivenko-Cantelli Thm]**

$$P(\lim_{n \to \infty} \sup_x |F(x) - F_n(x)| = 0) = 1$$

allow for estimating other quantities by plugging F_n in the place of F, e.g., $E[X]$ as

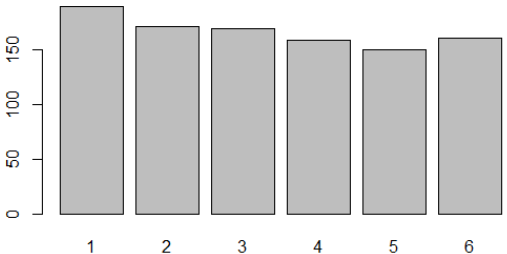
$$E[X] = \sum_a a \cdot P(X = a) \approx \sum_a a \cdot \frac{|\{i \mid x_i = a\}|}{n} = \frac{1}{n} \sum_i x_i$$

- What about p.m.f. and d.f.?

p.m.f.: Barplots

- For discrete data, barplots provide frequency counts for values
  - approximate the p.m.f. due to the law of large numbers

$$P(X = a) \approx \frac{|\{i \mid x_i = a\}|}{n}$$



- For continuous data, frequency counting of distinct values do not work. Why?

d.f.: Histograms

- Histograms provide frequency counts for ranges of values.
- Split the support to m intervals, called *bins*:

$$B_1, \ldots, B_m$$

  where the length $|B_i|$ is called the *bin width*
- Count observations in each bin and normalize them:

$$A_i = \frac{|\{j \in [1, n] \mid x_j \in B_i\}|}{n} \approx P(X \in B_i)$$

- Plot bars whose **area** is proportional to A_i

$$A_i = |B_i| \cdot H_i \qquad H_i = \frac{|\{j \in [1, n] \mid x_j \in B_i\}|}{n|B_i|}$$

Choice of the bin width

- Bins of equal width:

$$B_i = (r + (i-1)b, r + ib] \quad \text{for } i \in [1, m]$$

where $r \leq$ minimum point and b is the bin width

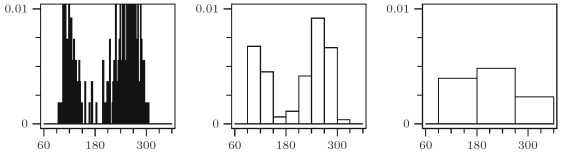


Fig. 15.2. Histograms of the Old Faithful data with different bin widths.

- Mean Integrated Square Error (MISE), for $\hat{f}$ density estimation of f:

$$MISE = E[\int (\hat{f}(t) - f(t))^2 dt] = \int \int (\hat{f}(t) - f(t))^2 f(x_1) \dots f(x_n) dt dx_1 \dots dx_n$$

- Scott's normal reference rule (minimize MISE for Normal density):

$b = 3.49 \cdot s \cdot n^{-1/3}$, where $s = \hat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$ is the sample standard deviation

Choice of the bin width

- $b = 2 \cdot IQR \cdot n^{-1/3}$, where $IQR = Q_3 - Q_1$ *[Freedman-Diaconis' choice]*
  - It replaces $3.49 \cdot s$ in the Scott's rule by $2 \cdot IQR$ (more robust to outlier)
  - Q_3 is 75% percentile of $x_1, \ldots, x_n$
  - Q_1 is 25% percentile of $x_1, \ldots, x_n$
- Variable bin width
  - Logarithmic binning in power laws
- Alternative strategy: number of bins given equal bin width b: **[other methods]**
  - $m = \left\lceil \frac{\max x_i - \min x_i}{b} \right\rceil$
  - $m = \lceil \sqrt{n} \rceil$
  - Sturges's formula: $m = \lceil \log_2 n \rceil + 1$

d.f.: Kernels

- Problem with histograms: as m increases, histograms become unusable
- Idea: estimate density function by putting **a pile (of sand)** around each observation
- Kernels state the shape of the pile
  - Epanechnikov $\frac{3}{4}(1-t^2)$ for $-1 \leq t \leq 1$
  - Triweight $\frac{35}{32}(1-t^2)^3$ for $-1 \leq t \leq 1$
  - Normal $\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$ for $-\infty < t < \infty$

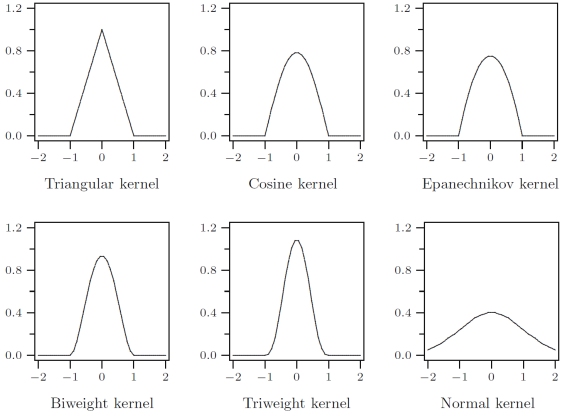


Fig. 15.4. Examples of well-known kernels K.

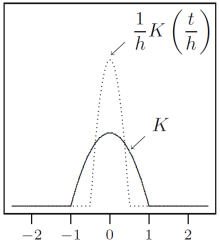
Kernel density estimation (KDE)

A Kernel is a function $K : \mathbb{R} \to \mathbb{R}$ such that

- K is a probability density, i.e., $K(t) \geq 0$ and $\int_{-\infty}^{\infty} K(t)dt = 1$
- K is symmetric, i.e., $K(-t) = K(t)$
- [sometime, it is required that] $K(t) = 0$ for $|t| > 1$, i.e., support is $[-1, 1]$

A bandwidth h is a scaling factor over the support of K from $[-1, 1]$ to $[-h, h]$

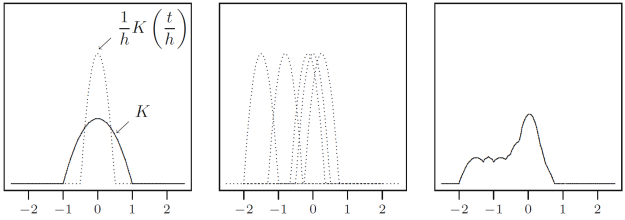
- h controls for how the probability density extends around 0
- if $X \sim K(t)$, then $hX \sim \frac{1}{h}K(\frac{t}{h})$ *[Change-of-units transformation, see Lesson 09]*



Change-of-units transformation. Let X be a continuous random variable with distribution function F_X and probability density function f_X. If we change units to $Y = rX + s$ for real numbers $r > 0$ and s, then

$$F_Y(y) = F_X\left(\frac{y-s}{r}\right) \quad \text{and} \quad f_Y(y) = \frac{1}{r}f_X\left(\frac{y-s}{r}\right).$$

Kernel density estimation (KDE)



Kernel and scaled kernel | Shifted kernel | Kernel density estimate

Let $x_1, \ldots, x_n$ be the observations

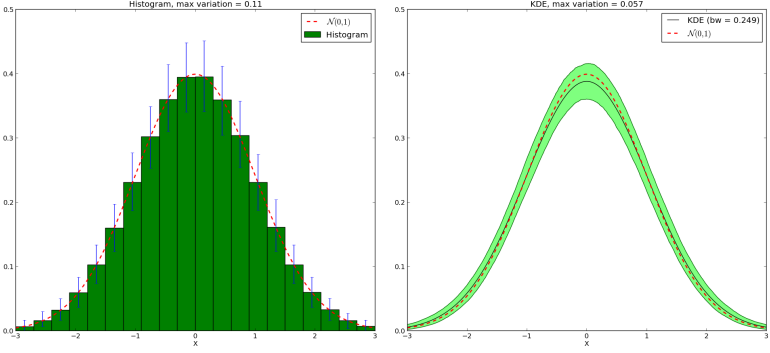
- if $X \sim K$, then $hX + x_i \sim \frac{1}{h}K(\frac{t-x_i}{h})$ *[Change-of-units transformation]*
- *K scaled and shifted* at x_i, with support $[x_i - h, x_i + h]$

The *kernel density estimate* is defined as the mixture of scaled and shifted kernel densities:

$$f_{n,h}(t) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{t - x_i}{h})$$

- It is a probability density function! **[Prove it!]**

Histograms vs KDE



- KDE has less variability than histograms!

Choice of the bandwidth

- **Note.** The choice of the kernel is not critical: different kernels give similar results
- **A problem.** The choice of the bandwith h is critical (and it may depend on the kernel)
- Mean Integrated Squared Error (MISE) is

$$E[\int_{-\infty}^{\infty}(f_{n,h}(t)-f(t))^2dt] = \int\int_{-\infty}^{\infty}(f_{n,h}(t)-f(t))^2 f(x_1)\dots f(x_n)dtdx_1\dots dx_n$$

  where $f(t)$ is the true density function and observations are independent
- For $f(t)$ being the Normal density, the MISE is minimized for

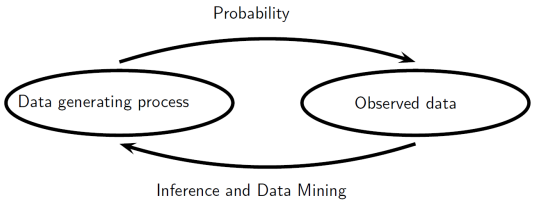
$$h = (\frac{4}{3})^{\frac{1}{5}}\cdot s\cdot n^{-\frac{1}{5}}$$

[normal reference method]

Kernel density estimation (KDE)

- **A problem.** The choice of the bandwith h is critical (and it may depend on the kernel)
- Automatic selection of h
  - Plug-in selectors (iterative bandwith selection)
  - Cross-validation selectors (part of data for estimation and part for evaluation)
- **Another problem.** When the support is finite, symmetric kernels give meaningless results
- Boundary kernels
  - Kernel (truncation) and renormalization
  - Linear (combination) kernel
  - Beta boundary kernels
  - Reflective kernels (density=0 at boundaries)
- See [Khamis, 2008] for a complete book on KDE

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-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 (useful for automated processing):
  - 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 F_n of a sample of realizations to estimate summaries of the generating distribution F

- *Sample mean*:

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

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

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

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

Measures of variability

- *Sample variance*:

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

  Divide by $n-1$ for a sample, and by n for a population! *[Bessel's correction]*

- *Sample standard deviation*:

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

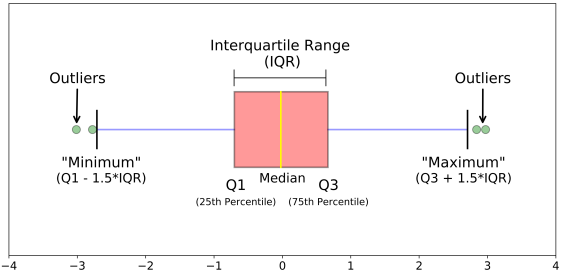
Order statistics and empirical quantiles

- Let $x_{\langle 1 \rangle}, \ldots, x_{\langle n \rangle}$ be sorted $x_1, \ldots, x_n$. 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 $q_p = \inf_x \{P(X \leq x) \geq p\} = \inf_x \{F(x) \geq p\}$
- Empirical quantiles: $q(p) = \inf_x \{F_n(x) \geq p\} = \inf_x \{|\{i \mid x_i \leq x\}|/n \geq p\}$
- 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)

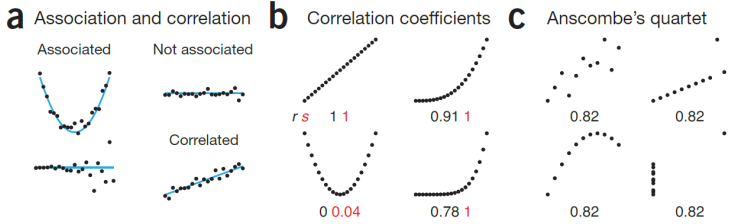
The box-and-whisker plot



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

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
  - $X \perp\!\!\!\perp Y$ 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 $Cov(X, Y) = 0$
  - 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 X		
Variable Y	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** 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 $(x_1, y_1), \ldots, (x_n, y_n)$
- Sample covariance:

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y}) \qquad Cov(X, Y) = E[(X - \mu_X) \cdot (Y - \mu_Y)]$$

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

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

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

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

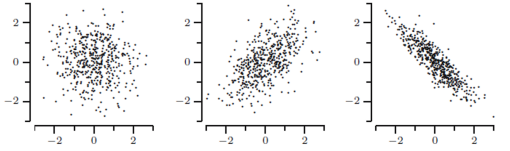
- Support in $[-1, 1]$ due to Cauchy–Schwarz's inequality: $|s_{xy}| \leq s_x \cdot s_y$ }
- Computational cost is $O(n)$

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

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

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

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



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

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

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

  - 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)
- Spearman's applies when Y (or also X) is ordinal
  - E.g., association between age and education level ("high-school", "bachelor", "master", ...)
- Computational cost is $O(n \cdot \log n)$

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

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

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

  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 $y_i = y_j$
  - Correction to divide by the number of pairs for which $sgn(x_i - x_j) \cdot sgn(y_i - y_j) \neq 0$
- Computational cost is $O(n^2)$

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

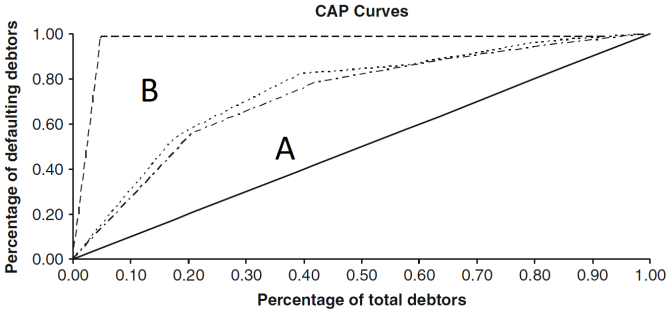
- X continuous and Y binary.
- Somers'D is an asymmetric Kendall's:

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

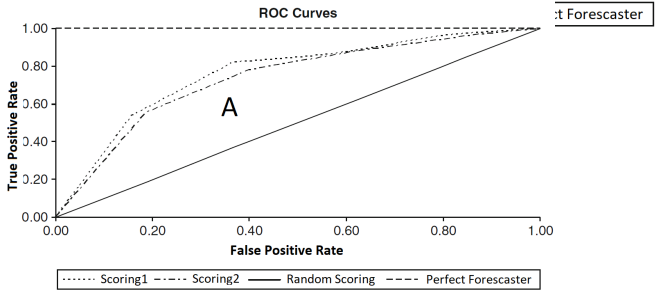
  i.e., fraction of concordand pairs minus discordant pairs conditional to unequal values of y

- Example with probabilistic classifiers *[more in future lessons]*
  - x = confidence prediction of being positive, i.e., `predict_proba(...)[,1]` in Python
  - y true class
  - D is the Gini index of classifier performances
  - related to AUC of ROC curve:

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



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



$$AUC = A + 1/2$$

Association between nominal variables: Thiel's U

Mutual information and NMI

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

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

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

  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 in future lessons)
  - Contingency coefficient C
  - Cramer's V
  - ϕ coefficient (or MCC, Matthews correlation coefficient)
  - Tschuprov's T
  - ...

Optional reference


David W. Scott (2015)
Multivariate density estimation: Theory, practice, and visualization.
John Wiley & Sons, Inc.

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.