Statistics for Data Science 24-25

Notes

University of Pisa

M.Sc. in Data Science and Business Informatics

Contents

1 Probability

Probability (on a finite sample space)

A probability function P on a finite sample space assigns to each event $A \in \Omega$ a number $P(A) \in [0, 1]$ such that

- $P(\Omega) = 1$;
- $P(A \cup B) = P(A) + P(B)$ if A and B are disjoint.

$P(A)$ is called probability that event A occurs.

Probability (on an infinite sample space)

A probability function P on an infinite sample space assigns to each event $A \in \Omega$ a number $P(A)$ such that

- $P(\Omega) = 1$;
- $P(A_1 \cup A_2 \cup A_3 \cup \dots) = P(A_1) + P(A_2) + P(A_3) + \dots$ if $A_1, A_2, A_3, \dots$ are disjoint.

Properties:

- $P(A^c) = 1 - P(A)$
- $P(\emptyset) = 0$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $A \subseteq B \implies P(A) \leq P(B)$

Conditional probability

The conditional probability of A given C is given by

$$P(A|C) = \frac{P(A \cap C)}{P(C)}$$

provided $P(C) > 0$ (it is otherwise undefined).

A consequence of this definition is the **multiplication rule**: $P(A \cap C) = P(A|C) \cdot P(C) = P(C|A) \cdot P(A)$.

Law of total probability

Let $C_1, C_2, \dots, C_n$ be a partition of Ω (i.e., they are disjoint and their union is Ω). Then, given any event $A \in \Omega$, its probability can be computed as

$$P(A) = P(A|C_1) \cdot P(C_1) + P(A|C_2) \cdot P(C_2) + \dots + P(A|C_n) \cdot P(C_n)$$

Bayes' rule

Let $C_1, C_2, \ldots, C_n$ be a partition of Ω and A be an event in Ω. Then, the probability of C_i given A is given by

$$P(C_i|A) = \frac{P(A|C_i) \cdot P(C_i)}{P(A|C_1) \cdot P(C_1) + P(A|C_2) \cdot P(C_2) + \cdots + P(A|C_n) \cdot P(C_n)}$$

Two events A and B are **independent** if $P(B) = 0$, or:

- $P(A \cap B) = P(A) \cdot P(B)$, or, equivalently,
- $P(A|B) = P(A)$.

If A and B are independent, also any combination of their complements is independent.
In general, events $A_1, A_2, \ldots, A_n$ are independent if for any subset $I \subseteq \{1, 2, \ldots, n\}$:

$$P\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} P(A_i)$$

This means that any possible subset of events in the collection is independent (since pairwise independence among individual events is not enough).

Two events A and B are **conditionally independent** given and event C ($P(C) > 0$) if $P(B|C) = 0$, or $P(A|B \cap C) = P(A|C)$. Since conditional probability is a probability, the definition is identical to the one above but conditioned on C.

2 Random variables

A **discrete random variable** takes a finite number of values, or a countably infinite number of values. Each discrete r.v. is described by a probability mass function and a cumumlative distribution function.

Probability mass function (PMF)

The PMF p of a discrete random variable X is a function $p : \mathbb{R} \to [0, 1]$, defined by

$$p(a) = P(X = a) \text{ for } -\infty < a < \infty$$

A **continuous random variable** takes any value in a continuous range (finite or infinite). Each continuous r.v. is described by a probability density function and a cumulative distribution function.

Probability density function (PDF)

A random variable X is countinuous if for some function $f : \mathbb{R} \to \mathbb{R}$ and any numbers a, b, with $a < b$,

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

where $f(x) \geq 0$ for all x and $\int_{-\infty}^{\infty} f(x)dx = 1$. f is called probability density function (PDF) of X.

Cumulative distribution function (CDF)

The CDF of a discrete random variable X is a function $F : \mathbb{R} \to [0, 1]$, defined by

$$F(a) = P(X \leq a) = \sum_{x \leq a} p(x) \quad \text{for } -\infty < a < \infty$$

The CDF of a continuous random variable X is a function $F : \mathbb{R} \to [0, 1]$, defined by

$$F(a) = P(X \leq a) = \int_{-\infty}^{a} f(x)dx \quad \text{for } -\infty < a < \infty$$

The **complementary cumulative distribution function** (CCDF) of a random variable is defined as $1 - F(a) = P(X > a)$.

Given two discrete random variables, we can define their **joint probability mass function** $p : \mathbb{R}^2 \in [0, 1]$, defined as

$$p(a, b) = P(X = a, Y = b) \text{ for } -\infty < a, b < \infty$$

For continuous random variables, we can similarly define the **joint probability density function** $f : \mathbb{R}^2 \to \mathbb{R}$, defined as

$$P(a_1 \leq X \leq b_1, a_2 \leq Y \leq b_2) = \int_{a_1}^{b_1} \int_{a_2}^{b_2} f(x, y)\ dy\ dx$$

The **joint cummulative distribution function** is defined as $F(a, b) = P(X \leq a, Y \leq b)$. For discrete random variables, this is calculated as

$$F(a, b) = P(X \leq a, Y \leq b) = \sum_{x \leq a} \sum_{y \leq b} p(x, y)$$

For continuous random variables, this is calculated as

$$F(a, b) = P(X \leq a, Y \leq b) = \int_{-\infty}^{a} \int_{-\infty}^{b} f(x, y)\ dy\ dx$$

The **marginal PMF** of a discrete r.v. X is

$$p_X(a) = P(X = a) = \sum_{y} p(a, y)$$

while the **marginal PDF** of a continuous r.v. X is

$$f_X(a) = \int_{-\infty}^{\infty} f(a, y)\ dy$$

In both cases, the **marginal distribution function** of X is

$$F_X(a) = P(X \leq a) = \lim_{b \to \infty} F_{XY}(a, b)$$

Conditional distribution of random variables

Let X and Y be two random variables, and P_{XY} their joint distribution. The conditional distribution of X given $Y \in B$, where $P(Y \in B) > 0$, is defined as

$$F_{X|Y \in B}(a) = P_{X|Y}(X \leq a | Y \in B) = \frac{P_{XY}(X \leq A, Y \in B)}{P_Y(Y \in B)}$$

Two random variables X and Y are **independent** ($X \perp\!\!\!\perp Y$) if

- $P_{X|Y}(X \leq a | Y \leq b) = P_X(X \leq a)$ for $a \in \mathbb{R}$, and for all b such that $P_Y(Y \leq b) > 0$, or, equivalently,
- $p_{XY}(x, y) = p_X(x) \cdot p_Y(y)$ (if discrete) or $f_{XY}(x, y) = f_X(x) \cdot f_Y(y)$ (if continuous).

Two random variables X and Y are said **identically distributed** ($X \sim Y$) if $F_X = F_Y$, i.e., $F_X(a) = F_Y(a)$ for $a \in \mathbb{R}$. If two random variables are both independent and identically distributed, they are said to be **independent and identically distributed** (i.i.d.).

Quantiles (percentiles)

Let X be a continuous random variable, and let p be a number in the interval $[0, 1]$. The p^{th} quantile (or $100p^{th}$ percentile) of the distribution of X is the smallest number q_p such that

$$F(q_p) = P(X \leq q_p) = p$$

The **median** of a distribution is the 50^{th} percentile. The **interquartile range** (IQR) is the difference between the 75^{th} and the 25^{th} percentiles. A more general definition, which holds also for discrete random variables, is

$$q_p = \inf_x \{P(X \leq x) \geq p\}$$

3 Probability distributions

3.1 Discrete distributions

Uniform distribution

$X \sim U(m, M)$

Models some experiment with $M - m + 1$ outcomes with the same probability of occurring. A random variable has uniform distribution if its PMF is given by

$$p(a) = P(X = a) = \frac{1}{M - m + 1} \qquad \text{for } a = m, m+1, \ldots, M$$

$$F(a) = \frac{\lfloor a \rfloor - m + 1}{M - m + 1} \qquad \text{for } m \leq a \leq M$$

$$\mathbb{E}[X] = \frac{m + M}{2} \qquad Var(X) = \frac{(M - m + 1)^2 - 1}{12}$$

Bernoulli distribution

$X \sim Ber(p)$

Models an experiment with two outcomes, success and failure, with probability $0 \leq p \leq 1$ of success. A random variable has the Bernoulli distribution if its PMF is given by

$$p(a) = P(X = a) = p^a(1 - p)^{1-a} \qquad \text{for } a = 0, 1$$

$$\mathbb{E}[X] = p \qquad Var(X) = p(1 - p)$$

Binomial distribution

$X \sim Bin(n, p)$

Models the number of successes in a sequence of n independent Bernoulli trials, each with probability $0 \leq p \leq 1$ of success. A random variable has the Binomial distribution if its PMF is given by

$$p(a) = P(X = a) = \binom{n}{a} p^a(1 - p)^{n-a} \qquad \text{for } a = 0, 1, \ldots, n$$

The sum of n independent Bernoulli r.v.s with parameter p is a Binomial r.v. with parameters n and p:

$$X = \sum_{i}^{n} X_i \sim Bin(n, p) \qquad \text{where } X_1, X_2, \ldots, X_n \sim Ber(p)$$

$$\mathbb{E}[X] = n \cdot p \qquad Var(X) = n \cdot p(1 - p)$$

Benford's law

$X \sim Ben$

Models the distribution of the leading digits in many real-life numerical datasets. A random variable has the Benford's law distribution if its PMF is given by

$$p(a) = P(X = a) = \log_{10}(1 + \frac{1}{a}) - \log_{10}(1 + \frac{1}{a+1}) \qquad \text{for } a = 1, 2, \ldots, 9$$

Geometric distribution

$X \sim Geo(p)$

Models the number of attempts needed to get the first success in a sequence of independent Bernoulli trials, each with probability $0 \leq p \leq 1$ of success. A random variable has the Geometric distribution if its PMF is given by

$$p(a) = P(X = a) = (1-p)^{a-1}p \qquad \text{for } a = 1, 2, \ldots$$

$$F(a) = 1 - (1-p)^a \qquad \text{for } a = 1, 2, \ldots$$

Given an infinite sequence of independent Bernoulli r.v.s with parameter p, the minimum number of trials needed to get a success is a Geometric r.v. with parameter p:

$$X = \min\{i : X_i = 1\} \sim Geo(p) \qquad \text{where } X_1, X_2, \cdots \sim Ber(p)$$

$$\mathbb{E}[X] = \frac{1}{p} \qquad Var(X) = \frac{1-p}{p^2}$$

Negative binomial (Pascal) distribution

$X \sim NBin(n, p)$

Models the number of failures before the n-th success in a sequence of independent Bernoulli trials, each with probability $0 \leq p \leq 1$ of success. A random variable has the Negative binomial distribution if its PMF is given by

$$p(a) = P(X = a) = \binom{a+n-1}{a} p^n (1-p)^a \qquad \text{for } a = 0, 1, \ldots$$

Given n i.i.d. Geometric r.v.s, we can obtain a Negative binomial r.v. with parameters n and p as follows:

$$X = \sum_i^n X_i - n \sim NBin(n, p) \qquad \text{where } X_1, X_2, \ldots, X_n \sim Geo(p)$$

$$\mathbb{E}[X] = \frac{n \cdot p}{(1-p)} \qquad Var(X) = n\frac{1-p}{p^2}$$

Poisson distribution

$X \sim Poi(\mu)$

Models the number of events occurring within some time interval, knowing the average rate of occurrence in that interval is μ. A random variable has the Poisson distribution if its PMF is given by

$$p(a) = P(X = a) = \frac{\mu^a}{a!} e^{-\mu} \qquad \text{for } a = 0, 1, 2, \ldots$$

The Poisson distribution can be approximated from the Binomial distribution:

$$Bin(n, p) \xrightarrow[n \to \infty]{} Poi(p \cdot n)$$

The approximation works for an experiment with an infinite number of Bernoulli trials, making it so that the mean rate of success is $\mu = p \cdot n$.

$$\mathbb{E}[X] = \mu \qquad Var(X) = \mu$$

Categorical distribution

$X \sim Cat(\vec{p})$

A generalization of the Bernoulli distribution to 3 or more possible outcomes, each with its own probability of occurring. A random variable has the Categorical distribution if its PMF is given by

$$p(i) = P(X = i) = p_i \qquad i = 1, 2, \ldots, n_C - 1$$

The parameter $\vec{p}$ is a vector of probabilities, such that $\sum_i p_i = 1$.

Multinomial distribution

$X \sim Mult(n, \vec{p})$

A generalization of the Binomial distribution to 3 or more possible outcomes, each with its own probability of occurring. A random variable has the Multinomial distribution if its PMF is given by

$$p(i_0, i_1, \ldots, i_{n_C-1}) = P(X = (i_0, i_1, \ldots, i_{n_C-1})) = \frac{n!}{i_0! \ldots i_{n_C-1}!} p_0^{i_0} p_1^{i_1} \ldots p_{n_C-1}^{i_{n_C-1}}$$

The sum of n independent Categorical r.v.s with parameter $\vec{p}$ is a Multinomial r.v. with parameters n and $\vec{p}$:

$$X = \sum_i^n X_i \sim Mult(n, \vec{p}) \qquad \text{where } X_1, X_2, \ldots, X_n \sim Cat(\vec{p})$$

3.2 Continuous distributions

Uniform distribution

$X \sim U(\alpha, \beta)$

Models some experiment with arbitrary outcomes in the interval $[\alpha, \beta]$. A random variable has the Uniform distribution if its PDF is given by

$$f(x) = \frac{1}{\beta - \alpha} \qquad \text{for } \alpha \leq x \leq \beta$$

$$F(x) = \frac{x - \alpha}{\beta - \alpha} \qquad \text{for } \alpha \leq x \leq \beta$$

$$\mathbb{E}[X] = \frac{\alpha + \beta}{2} \qquad Var(X) = \frac{(\beta - \alpha)^2}{12}$$

Exponential distribution

$X \sim Exp(\lambda)$

Models the time between subsequent events in a Poisson point process, with average rate of occurrence λ. A random variable has the Exponential distribution if its PDF is given by

$$f(x) = \lambda e^{-\lambda x} \qquad \text{for } x \geq 0$$

$$F(x) = 1 - e^{\lambda x}$$

$$\mathbb{E}[X] = \frac{1}{\lambda} \qquad Var(X) = \frac{1}{\lambda^2}$$

Normal (Gaussian) distribution

$X \sim N(\mu, \sigma^2)$

A random variable has a Normal distribution if its PDF is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \qquad \text{for } -\infty < x < \infty$$

The standard Normal distribution has $\mu = 0$ and $\sigma = 1$.
The Normal distribution can be approximated from the Binomial distribution:

$$Bin(n,p) \approx N(n \cdot p, n \cdot p(1-p)) \qquad \text{for } n \to \infty \text{ and } 0 \ll p \ll 1$$

There is no closed form of the CDF of the Normal distribution, but any variable can be turned into a standard Normal variable and its probability can be estimated using the right tail probability table of $N(0,1)$.

$$\mathbb{E}[X] = \mu \qquad Var(X) = \sigma^2$$

Erlang distribution

$X \sim Erl(n, \lambda)$

Models the time until n events occur in a Poisson point process, with average rate of occurrence λ. A random variable has the Erlang distribution if its PDF is given by

$$f(x) = \frac{\lambda(\lambda x)^{n-1} e^{\lambda x}}{\Gamma(\alpha)} \qquad \text{for } x \geq 0$$

$\Gamma(\alpha) = (\alpha - 1)!$ is called Gamma function, and is a normalization factor ensuring that the integral of the PDF is equal to 1.

$$\mathbb{E}[X] = \frac{n}{\lambda} \qquad Var(X) = \frac{n}{\lambda^2}$$

Gamma distribution

$X \sim Gam(\alpha, \lambda)$

Models the time until α quantities of something occur in a Poisson point process, with average rate of occurrence λ. It is a generalization of the Erlang distributions that also allows the first parameter to be any postive real number instead of a positive integer. A random variable has the Gamma distribution if its PDF is given by

$$f(x) = \frac{\lambda(\lambda x)^{\alpha-1} e^{\lambda x}}{\Gamma(\alpha)} \qquad \text{for } x \geq 0$$

The sum of n i.i.d. Exponential r.v.s. with parameter λ is Gamma distributed, with parameters n and λ:

$$X = \sum_{i}^{n} X_i \sim Gam(n, \lambda) \qquad \text{where } X_1, X_2, \ldots, X_n \sim Exp(\lambda)$$

$$\mathbb{E}[X] = \frac{n}{\lambda} \qquad Var(X) = \frac{n}{\lambda^2}$$

Cauchy distribution

$X \sim Cau(\alpha, \beta)$

A random variable has the Cauchy distribution if its PDF is given by

$$f(x) = \frac{\beta}{\pi(\beta^2 + (x-\alpha)^2)} \qquad \text{for } -\infty < x < \infty$$

A special case of the Cauchy distribution is the standard Cauchy distribution, with $\alpha = 0$ and $\beta = 1$. This distribution is also the same as the ratio between two standard Normal r.v.s.

$$\mathbb{E}[X] = \text{undefined} \qquad Var(X) = \text{undefined}$$

---

4 Expectation

The expectation (or expected value, mean, center of gravity) of a random variable is a number that summarizes the most central value in that variable's distribution.

Expectation

The expectation of a discrete random variable X is calculated as

$$\mathbb{E}[X] = \sum_i x_i \cdot P(X = x_i) = \sum_i x_i \cdot p(x_i)$$

The expectation of a continuous random variable X is calculated as

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

Expected value may be infinite or not exist for certain distributions. Consider the case of a continuous random variable. Its expected value, which is calculated as an integral I over $(-\infty, \infty)$ can be split into two terms, $I = I^- + I^+$, defined as follows:

$$I^- = \int_{-\infty}^{0} x \cdot f(x)$$

$$I^+ = \int_{0}^{\infty} x \cdot f(x)$$

Since $f(x)$ cannot take negative values, I^- is negative, and I^+ is positive. If I^- and I^+ are both finite, then the expected value exists and is finite. If one of them is infinite, the expected value is infinite. If both are infinite, the expected value does not exist. This can be generalized to discrete random variables, where the expectation is expressed as a sum instead of an integral (but can still similarly diverge or converge).

An example of distribution for which the expected value does not exist is the Cauchy distribution. An example of distribution for which the expected value is infinite is the Pareto distribution.

Change of variable formula (a.k.a. law of the unconscious/lazy statistician)

Let X be a random variable, and $g : \mathbb{R} \to \mathbb{R}$ be a function. If X is discrete, then

$$\mathbb{E}[g(X)] = \sum_i g(x_i) \cdot P(X = x_i)$$

If X is continuous, then

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot f(x)\ dx$$

Change of units theorem (for the expectation)

$\mathbb{E}[rX + s] = r\mathbb{E}[X] + s$

The expected value is **linear**. This means that $\mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c$ for any constants a, b, c. More in general, $\mathbb{E}[a_0 + \sum_i^n a_i \cdot X_i] = a_0 + \sum_i^n a_i \mathbb{E}[X_i]$

Jensen's inequality

Let g be a convex function, and let X be a random variable. Then

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$$

If g is concave, the inequality is reversed. If g is linear, the inequality becomes an equality.

Two-dimensional change of variable formula

Let X and Y be random variables, and let $g : \mathbb{R}^2 \to \mathbb{R}$ be a function. If X and Y are discrete, Then

$$\mathbb{E}[g(X, Y)] = \sum_i \sum_j g(a_i, b_i) P(X = a_i, Y = b_j)$$

If X and Y are continuous, then

$$\mathbb{E}[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y)\ dxdy$$

where $f(x, y)$ is their joint PDF.

If two variables are independent, then $\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}Y$. This holds for any set of independent random variables. More in general, given $X_1, X_2, \ldots, X_n$ independent random variables, and let $h_i : \mathbb{R} \to \mathbb{R}$ be a function; define the random variable $Y = h_i(X_i)$. Then, $Y_1, Y_2, \ldots, Y_n$ are also independent.
If we take two random variables, $X \perp\!\!\!\perp Y$ such that $Y > 0$, we have $\mathbb{E}[X/Y] \geq \mathbb{E}[X]/\mathbb{E}[Y]$. Let $g(y) = \frac{1}{y}$, the inequality follows from Jensen's inequality and the linearity of expectation.

Conditional expectation

$$\mathbb{E}[X|Y = b] = \sum_i a_i p(a_i|b) \qquad \mathbb{E}[X|Y = y] = \int_{-\infty}^{\infty} x f(x|y)\, dx$$

Also, the following theorem holds.

Law of iterated/total expectation

$$\mathbb{E}_Y[\mathbb{E}[X|Y]] = \mathbb{E}[X]$$

Proof:

$$\mathbb{E}_Y[\mathbb{E}[X|Y]] = \sum_j \sum_i a_i p_{X|Y}(a_i|b_j) \cdot p_Y(b_j) = \sum_j \sum_i a_i p_{X,Y}(a_i, b_j) = \sum_i a_i p_X(a_i) = \mathbb{E}[X]$$

5 Variance

The variance of a random variable is a measure of how much the values of that variable spread around the mean. A low variance means that most values are close to the mean, while a high variance means that the values are more spread out.

Variance

The variance of a random variable X is defined as

$$Var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

Often, the **standard deviation** ($\sigma = \sqrt{Var(X)}$) is used instead. This is because the variance is in squared units, so the standard deviation is on the same scale as the expectation and is easier to interpret.

Just like expectation, variance may be infinite or not exist. Variance does not exist if the expectation does not exist, but there may be distributions where the expectation exists while the variance does not: an example of such distribution are the Power Laws.

Change of units theorem (for the variance)

$Var(rX + s) = r^2 Var(X)$

The variance is **not linear**. This means that $Var(aX + bY + c) \neq aVar(X) + bVar(Y) + c$ in general. However, if X and Y are independent, then $Var(X + Y) = Var(X) + Var(Y)$.

6 Covariance

Covariance

The covariance of two random variables X and Y is the number:

$$Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

Given two random variables X and Y, the variance of their sum is:

$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$$

If the random variables are independent, their covariance is 0 (and so the variance of the sum is the sum of the variances).

Given X and Y two random variables, and $r, s, t, u \in \mathbb{R}$, then

$$Cov(rX + s, tY + u) = rtCov(X, Y)$$

Hence, $Var(rX + sY + t) = r^2 Var(X) + s^2 Var(X) + 2rsCov(X, Y)$.

---

7 Power laws and Zipf's law

Power laws are a family of "scale free" distributions. Most distributions have a typical size or scale, so they have some value around which measurements are centered. In contrast, power laws vary over a very large range where it's not possible to identify a typical value around which the distribution peaks.

Power law distribution

$X \sim Pow(x_{min}, \alpha)$

A random variable has the power law distribution if for some $\alpha > 1$ its PDF is given by

$$f(x) = C \cdot x^{-\alpha} \qquad x \geq x_{min}$$

C is called **intercept**, while α is called **exponent**. If the function is expressed in logarithmic scale, we have

$$\log(f(x)) = -\alpha \cdot \log(x) + \log(C)$$

i.e., there is a linear relationship between $\log(f(x))$ and $\log(x)$. Graphically, this means that the distribution is a straight line in a log-log plot. The reason parameter x_{min} is included is to specify what is the exact lower bound after which a distribution shows a power law behaviour.

Being scale-free, we can identify some constant b such that $p(bx) = g(b)p(x)$, meaning that even if we multiply the variable by this scaling factor, the form of the distribution remains the same. In this case, we write

$$p(bx) = b^{-\alpha}C \cdot x^{-\alpha}.$$

Notice how the value of the intercept is not specified in the definition above. This is because after fixing x_{min} and α, C is uniquely determined by the condition that the integral of the PDF over the entire range must be 1:

$$1 = \int_{x_{min}}^{\infty} C \cdot x^{-\alpha}\ dx = \frac{C}{-\alpha+1}[x^{-\alpha+1}]_{x_{min}}^{\infty} = \frac{C}{\alpha-1}x_{min}^{-\alpha+1} \iff \boxed{C = \frac{(\alpha-1)}{x_{min}^{-\alpha+1}}}$$

This integral is finite only if $\alpha > 1$. If $\alpha < 1$, then it simply diverges. If $\alpha = 1$, the denominator becomes 0, and the integral is not defined. By substituting this value in the formula of the PDF, we get

$$f(x) = \frac{\alpha-1}{x_{min}}\left(\frac{x}{x_{min}}\right)^{-\alpha}.$$

Using the same calculations we can find a closed formula for the CCDF:

$$P(X > x) = \int_{x}^{\infty} C \cdot x^{-\alpha}\ dx = \frac{C}{-\alpha+1}[x^{-\alpha+1}]_{x_{min}}^{\infty} = \frac{C}{\alpha-1}x^{-\alpha+1}.$$

Since we calculated C we can substitute it back in the formula to get

$$P(X > x) = \left(\frac{x}{x_{min}}\right)^{-\alpha+1}$$

Both the PDF and the CCDF have the same form, but with a different exponent. The CCDF also looks linear when plotted in a log-log scale. As for the expectation, we have

$$\mathbb{E}[X] = \int_{x_{min}}^{\infty} x \cdot C \cdot x^{-\alpha}\ dx = C\int_{x_{min}}^{\infty} x^{alpha+1}\ dx = \frac{C}{-\alpha+2}\left[x^{-\alpha+2}\right]_{x_{min}}^{\infty} = \frac{C}{\alpha-2}x_{min}^{-\alpha+2}.$$

Similarly to the calculations done to find C, we can observe how this integral is finite only for $\alpha > 2$: if $\alpha < 2$, the integral diverges, while if $\alpha = 2$, the denominator becomes 0 and the integral is not defined. Substituting the value of C back in the formula, we get

$$\mathbb{E}[X] = \frac{\alpha-1}{\alpha-2}x_{min}$$

Also for the variance, it is finite only for $\alpha > 3$.

Pareto distribution

$X \sim Par(x_{min}, \beta)$

A random variable has the Pareto distribution if for some $\beta > 0$ its density function is given by

$$f(x) = C \cdot x^{-(\beta+1)} \qquad x \geq x_{min}$$

A Pareto distribution is actually just a power law, but expressed differently:
$Par(x_{min}, \beta) = Pow(x_{min}, \beta + 1)$.

Discrete power law distribution

$X \sim Pow(\alpha, k_{min})$

A random variable has the discrete power law distribution if for some $\alpha > 1$ its PMF is given by

$$p(k) = C \cdot k^{-\alpha} \qquad k = k_{min}, k_{min} + 1, \dots$$

Since the sum of probabilities must be 1, C is determined as

$$C = \frac{1}{sum_{k=k_{min}}^{\infty} k^{-\alpha}} = \frac{1}{\zeta(\alpha, k_{min})}$$

$\zeta(\alpha, k_{min})$ is the **Hurwitz zeta function**. A special case of it is the **Riemann zeta function**, which is $\zeta(\alpha) = \zeta(\alpha, 1)$

When we are studying a data sample and want to check if it follows a power law, we can plot the frequency of its values in log-log scale and verify if the points are aligned in a straight line. However, since the values in the tail are rarer, the data sample will have few of them. The resulting plot will likely present a lot of noise around the tail of the distribution, and it may not be obvious whether it is a power law or some similar distribution (such as exponential or log-normal). To fix this issue, we can follow two approaches:

- We estimate and plot the CCDF in log-log scale. As mentioned above, the CCDF of a power law also appears linear when plotted this way, with the advantage of being much more stable in the tail.
- We construct an histogram using logarithmic binning. This means that the bins are not equally spaced, but they grow exponentially. For example, the first bin goes from 1 to 10, the second from 10 to 100, the third from 100 to 1000, and so on. Since bins aggregate values, the effect of noise is reduced.

Zipf's law distribution

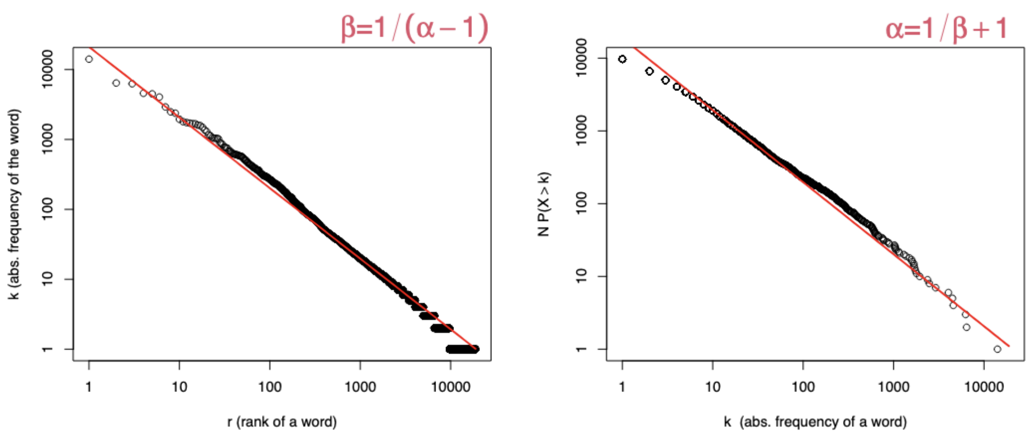
$X \sim Zipf(\alpha)$

A random variable has the Zipf's law distribution if for some $\alpha > 1$ its PMF is given by

$$p(r) = C \cdot r^{-\alpha} \qquad\qquad r = 1, 2, \ldots, N$$

In a Zipf's law distribution, probabilities are assigned to the **ranks** of events; the higher the rank (i.e. closer to 1), the higher the probability. This is different than a power law distribution, where probabilities directly depend on the frequencies. For example: a discrete power law may model the probability of a word with a certain number of accurrences in a text, while Zipf's law may model the probability of a word with a certain rank in a list of words sorted by frequency.

We can try to convert a power-law into a Zipf's law and vice-versa. By comparing the PMF of a Zipf's law and the CCDF of a power law, they have the same form, and are actually representing the same information but with the axes inverted:



The rank r of a word with frequency k is equal to the number of words with frequency larger than k plus 1. In other words, $r = 1 + N \cdot P(X > k)$. If $X \sim Pow(1, \alpha)$, then $r = 1 + N \cdot P(X > k) \propto k^{-(\alpha-1)}$. By inverting, we get that $k \propto r^{-\frac{1}{\alpha-1}}$, i.e., the frequencies are Zipf's law distributed with parameter $\frac{1}{\alpha-1}$.

$$X \sim Pow(1, \alpha) \iff R \sim Zipf\left(\frac{1}{\alpha - 1}\right)$$

(R is a r.v. that models the ranks).
For this distribution, C is calculated as

$$C = \frac{1}{\sum_{r=1}^{N} r^{-\alpha}} = \frac{1}{\zeta(\alpha) - \zeta(\alpha, N+1)}$$

8 Computations with random variables

Consider a random variable X with a given distribution. Let $Y = g(X)$ be another random variable obtained as a function of the first. Then, the following theorems hold:

- If X is a discrete random variable, the PMF of $Y = g(X)$ is

$$P_Y(Y = y) = \sum_{g(x)=y} P_X(X = x) = \sum_{x \in g^{-1}(y)} P_X(X = x)$$

- If X is a continuous random variable, the CDF and PDF of $Y = g(X)$ when g is invertible is

$$F_y(y) = F_X(g^{-1}(y)) \qquad\qquad f_Y(y) = f_X(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right|$$

Change of units transformation

Let X be a continuous random variable. If we change units to $Y = rX + s$ for $r, s \in \mathbb{R}, r > 0$ then

$$F_Y(y) = F_X\left(\frac{y-s}{r}\right) \qquad\qquad f_Y(y) = \frac{1}{r} f_X\left(\frac{y-s}{r}\right)$$

Convolution of random variables

Let X adn Y be two independent random variables. If they are discrete with PMFs $p_X(x)$ and $p_Y(y)$, then the PMF of $Z = X + Y$ is

$$p_Z(z) = \sum_y p_X(c - x) \cdot p_Y(y)$$

If X and Y are continuous with PDFs $f_X(x)$ and $f_Y(y)$, then the PDF of $Z = X + Y$ is

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(c - x) \cdot f_Y(y)\, dx$$

Maximum of random variables

Let $X_1, X_2, \ldots, X_n$ be n independent random variables with the same distribution function F, and let $Z = \max\{X_1, X_2, \ldots, X_n\}$. Then

$$F_Z(a) = (F(a))^n.$$

This is because $F_Z(a) = P(Z \leq a) = \prod_i^n P(X_i \leq a) = P(X_1 \leq a) \cdot P(X_2 \leq a) \cdot \ldots \cdot P(X_n \leq a) = (F(a))^n$.

Minimum of random variables

Let $X_1, X_2, \ldots, X_n$ be n independent random variables with the same distribution function F, and let $Z = \min\{X_1, X_2, \ldots, X_n\}$. Then

$$F_Z(a) = 1 - (1 - F(a))^n.$$

This is because $F_Z(a) = P(Z \leq a) = 1 - \prod_i^n P(X_i > a) = 1 - (1 - F(a))^n$.

Product of independent random variables

Let X and Y be two independent continuous random variables with PDFs f_X and f_Y. Then the PDF of $Z = XY$ is

$$f_Z(z) = \int_{-\infty}^{\infty} f_Y\left(\frac{z}{x}\right) f_X(x) \frac{1}{|x|}\, dx \qquad\qquad -\infty < z < \infty$$

Quotient of independent random variables

Let X and Y be two independent continuous random variables with PDFs f_X and f_Y. Then the PDF of $Z = X/Y$ is

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(zx) f_Y(x) |x|\, dx \qquad\qquad -\infty < z < \infty$$

Propagation of independence

Let $X_1, X_2, \ldots, X_n$ be independent random variables. For each i, let $h_i : \mathbb{R} \to \mathbb{R}$ be a function, and define the r.v.s

$$Y_i = h_i(X_i)$$

Then $Y_1, Y_2, \ldots, Y_n$ are also independent.

---

9 Moments

Moment

Let X be a continuous random variable with PDF $f(x)$. The k^{th} moment of X (if it exists) is

$$\mathbb{E}[X^k] = \int_{-\infty}^{\infty} x^k \cdot f(x)\, dx$$

Th expected value of a random variable is its first moment.

Central moment

Let X be a continuous random variable with PDF $f(x)$. The k^{th} central moment of X (if it exists) is

$$\mu_k = \mathbb{E}[(X-\mu)^k] = \int_{-\infty}^{\infty} (x-\mu)^k f(x)\, dx$$

The variance of a random variable is its second central moment.
Another related concept is the k^{th} standardized moment, which is the k^{th} central moment divided by the standard deviation raised to the k^{th} power:

$$\tilde{\mu}_k = \frac{\mu_k}{\sigma^k} = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^k\right]$$

- $\tilde{\mu}_1 = 0$ (since $\mathbb{E}[X-\mu] = 0$ for any random variable);
- $\tilde{\mu}_2 = 1$ (since $\mathbb{E}[(X-\mu)^2] = \sigma^2$);
- $\tilde{\mu}_3$ is called **skewness** and measures the magnitude and direction of the asymmetry of the distribution. If it is 0, the distribution is symmetric, and mean, median, and mode coincide. If it is positive, the distribution is **right-skewed** (the mean is greater than mode and median), while if it is negative, the distribution is **left-skewed** (the mean is less than mode and median).
- $\tilde{\mu}_4$ is called **kurtosis** and measures the dispersion of the random variable around the values $\mu + \sigma$ and $\mu - \sigma$. Specifically, the kurtosis of a distribution is compared to that of a Normal distribution, which is always 3. Then, if the kurtosis is equal to 3, the distribuion is **mesokurtic** (similar to a Normal); if it is greater than 3, it is **leptokurtic** (the tails are fatter, while the middle is thinner); if it is less than 3, it is **platykurtic** (the tails are thinner, but the middle is larger).

10 Distances between distributions

Distances and metrics

A distance over a set $\mathcal{A}$ is a function $d : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ such that:

- $d(x,y) \geq 0$ (non-negativity)
- $d(x,y) = 0$ iff $x = y$ (identity of indiscernibles)
- $d(x,y) = d(y,x)$ (symmetry)

Also, d is a metric if it also satisfies the triangle inequality: $d(x,z) \leq d(x,y) + d(y,z)$.

Distances and metrics over probability distributions are used to measure how far apart two distributions are. Calculating distances is very useful in statistics and machine learning: for example, after a dataset has been split into training and test sets, we can compare the distribution of the two to make sure they are similar (or, alternatively, to make sure they are different and study how well the model generalizes). This section will overview the most common distances used in statistics.

Total Variation (TV) distance

$$d_{TV}(X,Y) = \frac{1}{2}\sum_i |p_X(a_i) - p_y(a_i)| \qquad \text{(discrete case)}$$

$$d_{TV}(X,Y) = \frac{1}{2}\int |f_X(x) - f_Y(x)| dx \qquad \text{(continuous case)}$$

Kolmogorov-Smirnof (KS) distance

$$d_{KS}(X,Y) = \sup_x |F_X(x) - F_Y(x)|$$

Both are metrics. They have no closed forms, but they can be approximated from samples of the distributions.

Shannon's information entropy (H)

$$H(X) = \mathbb{E}[-\log(p(X))] = -\sum_i p(a_i)\log(p(a_i)) \qquad \text{(discrete case)}$$

$$H(X) = \mathbb{E}[-\log(f(X))] = -\int_{-\infty}^{\infty} f(x)\log(f(x))dx \qquad \text{(continuous case)}$$

Entropy measures the average level of information (or "surprise", "uncertainty") of a random variable. Information is inversely proportional to probability: the more unlikely an event is, the more information it carries. So, for example, a random variable that only takes a single value has zero entropy, because there is no uncertainty about its value. In contrast, a random variable that takes many values with equal probability has high entropy, because there is a lot of uncertainty about its value.

Let X be a discrete random variable with a finite domain of n elements. Per corollary of Jensen's inequality, since $log(x)$ is a concave function, we have:

$$H(X) = \mathbb{E}[-\log(p(X))] = \mathbb{E}\left[\log\left(\frac{1}{p(X)}\right)\right] \leq \log\left(\mathbb{E}\left[\frac{1}{p(X)}\right]\right)$$

Then, by change of variable:

$$\mathbb{E}\left[\frac{1}{p(X)}\right] = \sum_i \frac{p(a_i)}{p(a_i)} = n$$

So we can derive the following bound for the entropy:

$$H(X) \leq \log(n)$$

Cross entropy (H)

$$H(X;Y) = \mathbb{E}_X[-\log p_Y(Y)] = -\sum_i p_X(a_i)\log(p_Y(a_i)) \qquad \text{(discrete case)}$$

$$H(X;Y) = \mathbb{E}_X[-\log f_Y(Y)] = -\int_{-\infty}^{\infty} f_X(x)\log(f_Y(x))dx \qquad \text{(continuous case)}$$

Cross entropy measures the number of bits needed to encode values from X using a code based on Y. If the two have exactly the same distribution, the cross entropy is minimal: it is exactly equal to the entropy of X. The more the two are different, the more extra bits will be needed to encode the differences between the two.

Joint entropy (H)

$$H(X,Y) = \mathbb{E}[-\log(p(X,Y))] = -\sum_{i,j} p(a_i,a_j)\log(p(a_i,a_j)) \qquad \text{(discrete case)}$$

$$H(X,Y) = \mathbb{E}[-\log(f(X,Y))] = -\int_{-\infty}^{\infty} f(x,y)\log(f(x,y))dxdy \qquad \text{(continuous case)}$$

Joint entropy is simply the entropy of the joint distribution of two random variables X and Y. If the two are independent, then $p_{XY}(x,y) = p_X(x) \cdot p_Y(y)$ (and similarly for PDFs), so the above sum/integral can be split, making it so that $H(X,Y) = H(X) + H(Y)$.

Kullback-Leibler (KL) divergence

$$D_{KL}(X||Y) = \sum_i P_X(a_i)\log\left(\frac{p_X(a_i)}{p_y(a_i)}\right) = H(X;Y) - H(X) \qquad \text{(discrete case)}$$

$$D_{KL}(X||Y) = \int_{-\infty}^{\infty} f_X(x)\log\left(\frac{f_X(x)}{f_Y(x)}\right)dx = H(X;Y) - H(X) \qquad \text{(continuous case)}$$

KL divergence is also sometimes called relative entropy of X w.r.t. Y; it measures how well the distribution of the model Y can reconstruct the distribution of the data X. Since it can be expressed in terms of cross-entropy and entropy, it is easy to see that

- It is always non-negative, since the cross-entropy between two distributions can only be greater than or equal to the entropy of one of them.
- It is exactly 0 if the two distributions are the same.
- It is asymmetric.

Note that since it is not symmetric, it is not an actual distance.

Mutual information

Discrete case:

$$I(X,Y) = D_{KL}(p_{XY}||p_X p_Y) = \sum_{i,j} p_{XY}(a_i, a_j) \log\left(\frac{p_{XY}(a_i, a_j)}{p_X(a_i)p_y(a_j)}\right) =$$

$$= H(X) + H(Y) - H(X;Y)$$

Continuous case:

$$I(X,Y) = D_{KL}(f_{XY}||f_X f_Y) = \int_{-\infty}^{\infty} f_{XY}(x,y) \log\left(\frac{f_{XY}(x,y)}{f_X(x)f_Y(y)}\right) dxdy =$$

$$= H(X) + H(Y) - H(X;Y)$$

Mutual information measures how dependent the two distributions are. It quantifies how much the product of the marginals can reconstruct the joint distribution.

- It is always non-negative, since the sum of the individual entropies is always greater than or equal to the joint entropy.
- It is exactly 0 if $X \perp\!\!\!\perp Y$.
- It is symmetric.

In some cases, it may be useful to have a normalized measure of dependence. The **normalized mutual information** is defined as:

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

Suppose we have an unknown variable X, and we observe a noisy function of it, called Y. Let $Z = f(Y)$, i.e., a processing of the noisy observations. Intuitively, Z cannot contain more information about X than Y does. This is known as the **data processing inequality**:

$$I(X,Y) \geq I(X,Z)$$

If they happen to be equal, and Z is a summary of Y, then Z is a sufficient statistic for X: it means that we can reconstruct X from Z with the same accuracy as from Y.

Earth mover's distance (Wasserstein metric)

$$EMD(X,Y) = \frac{\sum_{i,j} F_{i,j} \cdot |a_i - a_j|}{\sum_{i,j} F_{i,j}}$$

Earth's mover distance measures the minimum cost required to transform one distribution into another; the F in the formulas is the **flow** which minimizes the numerator (the cost). In practice, for pairs of univariate random vairables X and Y, is is calculated as follows:

$$EMD(X,Y) = \sum_i |F_X(a_i) - F_Y(a_i)| \qquad \text{(discrete case)}$$

$$EMD(X,Y) = \int_{-\infty}^{\infty} |F_X(x) - F_Y(x)|\, dx \qquad \text{(continuous case)}$$

For empirical distributions, assuming the samples are sorted in increasing order:

$$EMD(X,Y) = \frac{1}{n} \sum_i |x_i - y_i|$$

11 The law of large numbers

For many experiments that concern natural phenomena, different executions of the same experiment will likely yield different results. The variation in the outcome is due to randomness caused by uncontrollable factors. To mitigate the effect of this randomness, the same experiment can be repeated a number of times and the **average** of the results is studied instead. Formally, given $X_1, X_2, \ldots, X_n$ independent random variables, their average is

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

Expected value 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

$$\mathbb{E}[\bar{X}_n] = \mu \qquad Var(\bar{X}_n) = \frac{\sigma^2}{n}$$

The random variables do not need to be identically distributed. Note that the variance is inversely proportional to the number of random variables contributing to the average: the more variables we have, the more stable the average becomes.

Markov's inequality

Let $X \geq 0$ be a random variable, and let $a > 0$. Then

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

Corollary: Assume $X \geq 0$, $\mathbb{E}[X] > 0$ and $k > 0$. Then

$$P(X \geq k\mathbb{E}[X]) \leq \frac{1}{k}$$

The proof is as follows: let $\mathbb{1}_{X \geq \alpha}$ be an indicator variable that is 1 if $X \geq \alpha$ and 0 otherwise. Then

$$\alpha \mathbb{1}_{X \geq \alpha} \leq X$$

$$\mathbb{E}[\alpha \mathbb{1}_{X \geq \alpha}] \leq \mathbb{E}[X]$$

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

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

Chebyshev's inequality

Let X be a random variable, and $a > 0$. Then

$$P(|X - \mathbb{E}[X]| \geq a) \leq \frac{Var(X)}{a^2}$$

This inequality claims that most of the probability mass of a random variable is within a few standard deviations from the expected value. It is a consequence of Markov's inequality:

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

Now, we can apply Chebyshev's inequality to the average of n independent random variables, obtaining the following result:

The (weak) law of large numbers

Let $\bar{X}_n$ be the average of n independent random variables with the same expectation μ and variance σ^2. Then, for any $\varepsilon > 0$,

$$\lim_{n \to \infty} P(|\bar{X}_n - \mu| > \varepsilon) = 0$$

This law confirms what we previously observed with the variance of the average. As n goes to infinity, the probability that the value of the average significantly deviates from its expectation (that is also the same of the individual random variables in the average) becomes zero. This also holds if σ^2 is infinite, as long as the individual random variables have finite expectation.

The consequence of the law of large numbers is that by calculating the average of a reasonably large enough set of random variables we can recover not only the mean and the stardard deviation, but pretty much any feature of the distribution of the random variables. Next up are two application examples.

Recovering the probability of an event Assume we want to know the probability that the outcome of some experiment falls within a certain range, i.e., $p = P(a \leq X \leq b)$. We run n independent measurements of this same experiment, and we model those results with the r.v.s $X_1, X_2, \dots, X_n$. Then, we can define an indicator variable for each X_i:

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

The Y_i are also independent (per the propagation of independence seen previously). Since Y_i is an indicator variable, we know that

$$\mathbb{E}[Y_i] = p = P(a \le X_i \le b) \qquad\qquad Var(Y_i) = p(1-p)$$

Let $\bar{Y}_n$ be the average of the indicator variables. By the law of large numbers:

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

Informally, this means that if we perform the experiment n times, count the amount of times the outcome falls within the range $[a, b]$, and divide by n, we get an estimate of the real probability of that event. The larger n is, the better the estimate becomes.

Estimating conditional probability We want to estimate the conditional probability for two random variables: $p = P(C = c|A = a) = P(A = a, C = c)/P(A = a) = p_{ac}/p_a$. We run n independent measurements of the experiment, modeling each result as a pair (A_i, C_i). We define indicator variables as the example above:

$$Y_i = \mathbb{1}_{A_i = a \wedge C_i = c} = \begin{cases} 1 & \text{if } A_i = a \wedge C_i = c \\ 0 & \text{otherwise} \end{cases}$$

$$Z_i = \mathbb{1}_{A_i = a} = \begin{cases} 1 & \text{if } A_i = a \\ 0 & \text{otherwise} \end{cases}$$

By applying the (strong) law of large numbers, we get that

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

$$\lim_{n\to\infty} P(\bar{Z}_n = p_a) = 1$$

The two limits can be condensed in a ratio to estimate the conditional probability:

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

Hoeffding bound

If $\bar{X}_n$ is the average of n independent random variables with the same expectation μ and $P(a \le X_i \le b) = 1$, then for any $\varepsilon > 0$:

$$P(|\bar{X}_n - \mu| \ge \varepsilon) \le 2e^{-2n\varepsilon^2/(b-a)^2}$$

This offers a tight bound on the probability that the average deviates from its expectation by an arbitrarily small amount, but requires that the random variables have a bounded support.

12 The central limit theorem

The central limit theorem

Let $X_1, X_2, \ldots, X_n$ ne any sequence of i.i.d. random variables with the same expectation μ and finite positive vairance σ^2. 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.

This theorem states that if we take the average of n random variables, remove its expectation, and divide by its standard deviation, the result is a random variable that converges to a standard normal distribution as n goes to infinity. But, in practice, how large should n be? A famous rule of thumb is to use $n \geq 30$ to get a good approximation, but it is mostly just a myth; the optimal value of n depends on the distribution of the random variables.

13 Summaries

Summaries are used to represent and describe the information contained in datasets. They csan be graphical summaries, which visually represent the data, or numerical summaries, which give a description of the data in terms of numbers.

13.1 Graphical summaries

Empirical CDF A random variable is completely characterized by its CDF. We can approximate the CDF with the empirical cumulative distribution function, which is defined as

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

where the x_i are the observations in the dataset. The **Glivenko-Cantelli theorem** states that the empirical CDF converges to the true CDF as n goes to infinity:

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

This approximation can be plugged into different formulas to estimate other quantities, such as the mean or the variance.

Bar plots and histograms A bar plot is used for discrete data. It provides a frequency count for the values in the dataset, and approximates the PMF (as a consequence of the law of large numbers, as seen in a previous example):

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

A histogram is used for continuous data. It provides frequency counts for ranges of values (instead of individual ones). The support of the random variable is first split into m intervals called **bins** (which can all have the same width or different widths), and the number of occurrences belonging to each bin is counted and normalized:

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

The bins can be plotted as rectangles so that their area is proportional to A_i; after fixing their width b_i, the height is found as $H_i = A_i/b_i$.

Bin width can be chosen in different ways, producing different results. It is common to choose the same width for all bins, such that, for a total number of bins m, the interval corresponding to the i^{th} bin is

$$B_i = (r + (i - 1)b, r + i * b)$$

where r is the minimum value taken by the random variable, and b is the bin width. The optimal width can be found using **Mean Integrated Squared Error** (**MISE**):

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

where $\hat{f}$ is the density estimation of the real PDF f. The minimum of the MISE for Normal distributions is represented by **Scott's normal reference rule**:

$$b = 3.49 \cdot s \cdot n^{-1/3}$$

wher s is the sample standard deviation.
Other options are:

- **Freedman-Diaconis' choice**:

  $$b = 2 \cdot \text{IQR} \cdot n^{-1/3}$$

  This choice is more robust to outliers than the previous.

- **Variable bin width** (such as logarithmic binning as seen in power-law distributions).

- **Fixing the number of bins, and derivaring the width from it**. Some common strategies are:

  $$m = \lceil \frac{\max x_i - \min x_i}{b} \rceil$$

  $$m = \lceil \sqrt{n} \rceil$$

  $$m = \lceil \log_2 n \rceil + 1 \text{ (Sturges' rule)}$$

  The latter assumes normal distribution for the true density. This distribution can be approximated by a $Bin(n, 0.5)$ distribution, so the absolute frequency of the i^{th} bin is $\binom{m-1}{i}$. The total frequency is $n = \sum_{i=0}^{m-1} \binom{m-1}{i} = 2^{m-1}$, from which m is derived.

Kernel density estimation A big disadvantage of histograms is that the result strictly depends on the number of bins/bin width chosen to visualize the data. Kernel density estimation is another popular method to summarize distributions which is not as sensitive to the choice of parameters.

The idea behind this method is to mix kernel functions (which can take different forms, see Fig 1) centered in each observation in the dataset. Since data is assumed to be of continuous nature, the presence of a certain value in the dataset also contributes to the density of the values around it. The kernel function models the way this density is distributed around that single observation, and by mixing together all the kernels, the result should be a good approximation of the true density.

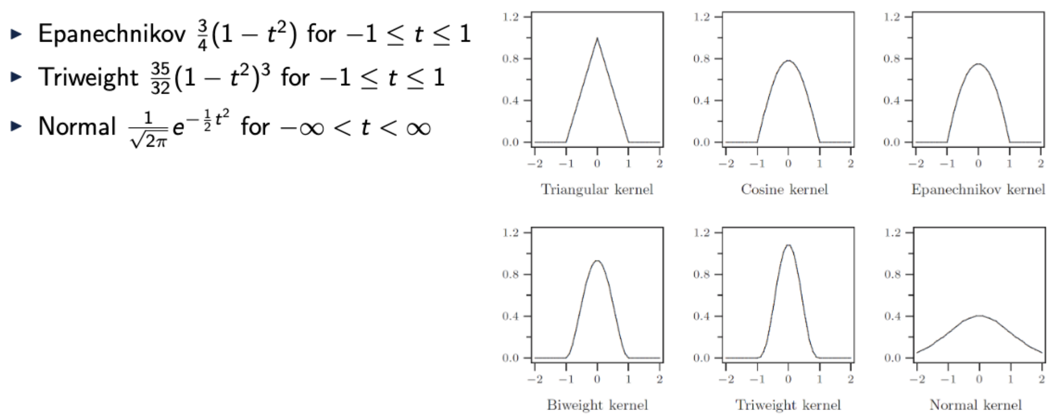


Figure 1: Examples of common kernels used in KDE.

Kernel

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

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

The last requirement is not strictly necessary, actually; for example, the Normal kernel has unbounded support.

Each kernel function is characterized by a center (the observation), and a **bandwidth** h, which is a scaling factor over the support of the kernel from $[-1, 1]$ to $[-h, h]$. In other words, the badwidth determines how tall-thin or short-wide the kernel is around its center. We can then write

$$X \sim K(t) \implies h \cdot X + x_i \sim \frac{1}{h} K\left(\frac{t - x_i}{h}\right)$$

because of the change-of-units transformation formulas. The final kernel density estimate is the result of the **mixture** of all the scaled and shifted kernel densities:

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

The $1/n$ in the formula is a normalization factor that ensures the density integrates to 1.

The choice of kernel is not critical to the final result; different kernels behave similarly. The key parameter is the bandwidth h. Also for KDE, MISE can be used to find the optimal value. Assuming the true density is Normal, the MISE is minimized for

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

For other distributions, the optimal bandwidth can be found using plug-in selectors or cross validation selectors.

Another issue that may arise is when the support of the random variable is bounded. If KDE is used as is, the result will present density event corresponding to values that are not possible. To avoid this, boundary correction techniques are used; some examples are

- Kernel truncation and renormalization (forced truncation of the kernel outside the support);
- Linear combination kernel;
- Beta boundary kernels;
- Reflective kernels.

13.2 Numerical summaries

Sample summaries Summaries of the empirical data can be used to estimate summaries of the true distribution. The following are some common ones:

- **Sample mean**:

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n}$$

- **Median**: Let $x_1, x_2, \ldots, x_n$ be the data in the sample, sorted.

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

  The median of a distriution corresponds to the 0.5^{th} quantile.

- **Sample variance and standard deviation**:

$$s_n^2 = \frac{\sum_i^n (x_i - \bar{x_n})^2}{n-1} \qquad s_n = \sqrt{s_n^2}$$

- **Median of absolute deviations**:

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

  If the distribution is symmetric, the MAD is exactly equal to the difference between the 0.75^{th} and 0.5^{th} quantiles.

Order statistics Let $x_1, x_2, \dots, x_n$ be the ordered sequence of values in a sample. $x_{\langle i \rangle}$ is the i^{th} order statistic. Order statistics are used to calculate empirical quantiles. Formally, the p^{th} quantile is the value q_p such that $q_p = \inf_x\{P(X \leq x) \geq p\} = \inf_x\{F(x) \geq p\}$ (to be read as: "the smallest number x such that the probability of X being less or equal than x is greater or equal than p"). To find the empirical quantiles, we use the empirical approximation of the CDF in place of the true CDF:

$$q_p = \inf_x\{F_n(x) \geq p\} = \inf_x \left\{ \frac{|\{i|x_i \leq x\}|}{n} \geq p \right\}$$

There are actually many ways to find the quantiles of a distribution. In `R`, there are 9 variants. The default one is type 7:

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

Another common choice is type 6:

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

The difference between the methods is irrelevant for big enough datasets.

What if the supposed index of the quantile is not an integer? In this case, the quantile is approximated using linear interpolation. Let $k = \lfloor p \cdot (n+1) \rfloor$ (or whatever other formula is used by the chosen method). Then,

$$q_p = x_{\langle k \rangle} + \alpha \cdot (x_{\langle k+1 \rangle} - x_{\langle k \rangle})$$

where $\alpha = p \cdot (n+1) - k$, i.e., the decimal part of the index.

Association and correlation **Association** measures how much information one variable provides on another. If two variables are independent, they are not associated. Association is maximum when one variable is a (invertible) function of the other. **Correlation** measures the presence and strenght of an increasing/decreasing trend between two variables. If two variables are independent, their correlation is always 0, but the opposite is not always true.

Correlation

Let X and Y be two random variables. The correlation coefficient ρ is defined to be 0 if $Var(X) = 0$ or $Var(Y) = 0$, and otherwise as

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X) \cdot Var(Y)}} \tag{1}$$

Some common correlation coefficients are:

- **Pearson's r**: it is obtained by substituting the sample covariance and the sample standard deviations of the random variables in the formula above:

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

  It is bounded in the interval $[-1, 1]$. The computational cost to calculate it is $O(n)$. A limitation of Pearson's r is that it can only detect linear relationships between random variables, and it can only be used when the variables are continuous.

- **Spearman's** ρ: this coefficient is calculated as the correlation between ranks of the observations. Let $rank(x)$ be the ranks of the values of the variable x. Then, Spearman's ρ is calculated as

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

  This coefficient assesses monotonic relationships of any kind (both linear and non-linear). The computational cost to calculate it is $O(n \cdot \log n)$, since it requires a sorting of the data to compute the ranks. It can be used when both variables are ordinal, continuous, or when one is ordinal and the other is continuous.

- **Kendall's** τ: this coefficient compares the sign of the differences between successive pairs of observations. It is calculated as

$$\tau_a = \frac{2\sum_{i<j} sign(x_i - x_j) \cdot sign(y_i - y_j)}{n \cdot (n-1)}$$

  It calculates the fraction of concordant pairs minus discordant pairs of values between the two variables, and it is bounded in the interval $[-1, 1]$. The computational cost to calculate it is $O(n^2)$. There is a variant, τ_b, which also takes into account the possibility of ties; instead of dividing by $n \cdot (n-1)$, it counts how many pairs do not present a tie in x or y. It can be used when both variables are ordinal, or when one is ordinal and the other is continuous.

- **Somer's D**: this coefficient is used when one variable is continuous and the other is binary. It can be seen as an asymmetric version of Kendall's τ:

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

  It calculates the fraction of concordant pairs minus discordant pairs over the number of unequal pairs of values in y. An example application can be seen with probabilistic classifiers: x is the confidence prediction of an example being positive, y is the true class, and D is the Gini index of the classifier.

- **Thiel's U**: it is used when both variables are nominal. It can be calculated in a symmetric and an asymmetric way:

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

  where $I(X,Y)$ is the mutual information between X and Y, and $H(X), H(Y)$ are the entropies of the two random variables. The asymmetrical version, specifically, indicates what fraction of X can be predicted by Y.

14 Estimators

A dataset is a st of repeated measurements of a specific phenomenon which we want to understand better. The phenomenon can be modeled using some probability distribuion, so the dataset (which is the realization of a random sample from that distribution) can be used to approximate its parameters.

Random sample

A random sample is a collection of i.i.d. random variables $X_1, X_2, \ldots, X_n \sim F(\alpha)$, where $F()$ is the distribution and α its parameter(s).

A classic example is the approximation of the speed of light by physicist A. A. Michelson, done in 1879. His dataset of measurements consisted of 100 different measurements, each which was in itself th average of repeated measurements on several variables (e.g., distance between the tools used). At the end, his estimate was off by about 150 km/s, likely due to him missing some source of error despite his meticulousness; still, he had the intuition of using the average of a dataset to estimate the parameter of a distribution (in this case, the one that describes the speed of light).

Estimand and estimate

An estimand θ is an unknown parameter of a distribution $F()$.
An estimate t of θ is a value that is obtained as a function $h()$ over a dataset:

$$t = h(x_1, \ldots, x_n) \tag{3}$$

Statistics and estimator

A statistics is a function $h(X_1, \ldots, X_n)$ of random variables.
An estimator of a parameter θ is a statistics $T_n = h(X_1, \ldots, X_n)$ intended to provide information about θ.

Using the speed of light example, we can model is as follows: the dataset of measurements is the realization of a random sample of random variables. Each random variable is defined as:

$$X_i = c + \epsilon_i$$

where c is the speed of light, and ϵ_i is a measurement error, assumed to be normally distributed with mean 0 ad variance σ^2. The **estimand** is the expected value of one of these variables: $\theta = \mathbb{E}[X_i]$. We can define an **estimator** as the average of the sample:

$$T_n = \bar{X}_n = \sum_{i=1}^{n} \frac{X_1 + X_2, \ldots, +X_n}{n}$$

Finally, the **estimate** is the actual value we get by plugging the values collected in the dataset to the estimator.

Unbiased estimator

An estimator $T_n = h(X_1, \ldots, X_n)$ of a parameter θ is unbiased if

$$\mathbb{E}[T_n] = \theta$$

If $\mathbb{E}[T_n] - \theta \neq 0$, the estimator is biased.

Bias, if present, can be either positive or negative. An estimator may be **asymptotically unbiased** if it its unbiased as the sample size n approaches infinity:

$$\lim_{n \to \infty} \mathbb{E}[T_n] = \theta$$

Sometimes, the estimator is indicated with the same symbol as the estimand, but with a hat on top: $\hat{\theta}$ (e.g., $\hat{\mu}$ is an estimator for the mean, $\hat{sigma}$ is an estimator for the standard deviation).

Bias can be thought of as a measure of how well the estimator can approsimate the parameter of interest. If it is unbiased, it means it has the capacity to estimate the parameter correctly. Estimators are also characterized by their variance. **Variance** is a measure of how much the estimate can sway from the true value of the parameter, regardless of bias. An estimator can have low bias, but high variance, meaning that it can approximate the parameter, but the response is not necessarily reliable. When the same estimand has multiple unbiased estimators, variance is a measure of their efficiency. Let T_1 and T_2 be unbiased estimators of θ; T_1 is **more efficient** than T_1 if $Var(T_2) < Var(T_1)$. The **relative efficiency** of T_2 w.r.t. T_1 is $Var(T_1)/Var(T_2)$. The standard deviation of the estimator is called **standard error** (**SE**).

Unbiased estimators for expectation and variance

Let $X_1, X_2, \dots, X_n$ be a random sample from a distribution with finite expectation μ and finite variance σ. Then

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

is an unbiased estimator for μ, and

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

is an unbiased estimator for σ^2.

We've already seen how the expected value of the mean is also the mean of the distribution. Also, per the central limit theorem, the variance of the mean goes to 0 as n goes to infinity.

Why is the estimator of the variance divided by $n-1$ instead of n? This is called Bessel's correction, and it is used to make sure the estimator is unbiased. The proof is explained below. First, for any random variable X_i, the following hold true:

$$(1) \quad \mathbb{E}[X_i - \bar{X}_n] = \mathbb{E}[X_i] - \mathbb{E}[\bar{X}_n] = \mu - \mu = 0$$

$$(2) \quad Var(X_i - \bar{X}_n) = \mathbb{E}[(X_i - \bar{X}_n)^2] - E[X_i - \bar{X}_n]^2 = \mathbb{E}[(X_i - \bar{X}_n)^2] = \sigma^2$$

X_i and $\bar{X}_n$ are not independent, because the latter is a function of the former. However, we can split the mean, removing X_i from it, getting two independent terms:

$$X_i - \bar{X}_n = X_i - \frac{1}{n} \sum_{j=1}^{n} X_j = X_i - \frac{1}{n} X_i - \frac{1}{n} \sum_{j=1, j \neq i} X_j = \frac{n-1}{n} X_i - \frac{1}{n} \sum_{j=1, j \neq i} X_j$$

We can now easily calculate the variance, since it is the sum of the variances of the two terms:

$$Var(X_i - \bar{X}_n) = Var\left(\frac{n-1}{n}X_i - \frac{1}{n}\sum_{j=1, j\neq i} X_j\right) = Var\left(\frac{n-1}{n}X_i\right) + Var\left(\frac{1}{n}\sum_{j=1, j\neq i} X_j\right) =$$

$$= \frac{(n-1)^2}{n^2}\sigma^2 - \frac{1}{n^2}(n-1)\sigma^2 = \frac{n-1}{n}\sigma^2$$

Finally, we apply the above calculations to find the expected value of the estimator for the variance:

$$\mathbb{E}[S_n^2] = \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right] = \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}[(X_i - \bar{X}_n)^2] =$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} Var(X_i - \bar{X}_n) = \frac{1}{n-1}n\frac{n-1}{n}\sigma^2 = \sigma^2$$

Additionally,

$$Var(S_n)^2 = \frac{1}{n}(\mu_4 - \frac{n-3}{n-1}\sigma^4)$$

which goes to 0 as n goes to infinity.

Intuitively, the need for this correction can be explained by the fact that the X_i and the mean $\bar{X}_n$ are not independent; if we know $n-1$ random variables and the mean, we can always deduce the n^{th} random variable. In this sense, the estimator has $n-1$ **degrees of freedom**: in general, the degrees of freedom of any estimator is the number of observations minus the number of parameters already estimated.

What about the standard deviation? Unfortunately, the square root of the estimator of the variance is not an unbiased estimator for the standard deviation. By Jensen's inequality, since the square root is a concave function, we have:

$$\mathbb{E}[\sqrt{S_n^2}] = \mathbb{E}[S_n] < \sqrt{\mathbb{E}[S_n^2]} = \sigma$$

This is true for most estimators: if T is unbiased for θ, $g(T)$ is not necessarily unbiased for $g(\theta)$; the only exception is when $g()$ is a linear transformation (since by Jensen's inequality the strict equality holds). A non-parametric unbiased estimator for the standard deviation does not exist: we need to know the distribution to estimate it unbiasedly.

Unbiased estimator for the median

Let $X_1, \ldots, X_n$ be a random sample from a distribution with PDF $f(x)$. Let $m = F^{-1}(0.5)$ be the true median of the distribution. Let

$$T = Med(X_1, \ldots, X_n).$$

Then

$$T \sim N\left(m, \frac{1}{4nf(m)^2}\right) \quad \text{as} \quad n \to \infty$$

and T is an unbiased estimator for the median as n goes to infinity.

Unbiased estimator for quantiles

Let $X_1, \dots, X_n$ be a random sample from a distribution with PDF $f(x)$. Let q_p the the true p^{th} quantile of the distribution. Let

$$T = q_{X_1,\dots,X_n}(p)$$

Then

$$T \sim N\left(q_p, \frac{p(1-p)}{nf(q_p)^2}\right) \quad \text{as} \quad n \to \infty$$

and T is an unbiased estimator for the p^{th} quantile as n goes to infinity.

Unbiased estimator for the Median of Absolute Deviations (MAD)

Let $X_1, \dots, X_n$ be a random sample from a distribution. Let Md be the true median of absolute deviations of the distribution. Let

$$T = MAD(X_1, \dots, X_n) = Med(|X_1 - Med(X_1, \dots, X_n)|, \dots, |X_n - Med(X_1, \dots, X_n)|)$$

Then

$$T \sim N\left(Md, \frac{\sigma_1^2}{n}\right)$$

(where σ_1^2 is defined in terms of Md, median, and CDF of the distribution) and T is an unbiased estimator for the MAD as n goes to infinity.

All the above estimators are found by applying the corresponding version of the central limit theorem (CLT for medians, CLT for quantiles, CLT for MAD).

For correlation, the various coefficients seen in the previous section are estimators of the true correlation coefficient between two random variables. Pearson's r is an estimator for ρ, but the distribution of r is highly skewed. To fix this issue, the **Fisher transformation** can be applied, defined as follows: $\mathit{FisherZ}(r) = \frac{1}{2}\log\frac{1+r}{1-r}$. The distribution of the Fisher-transformed coefficient is approximately normal:

$$\mathit{FisherZ}(r) \sim N\left(\mathit{FisherZ}(\rho), \frac{1}{n-3}\right)$$

Hence, if we apply the inverse transformation to its expected value we get

$$\mathit{FisherZ}^{-1}(\mathbb{E}[\mathit{FisherZ(r)}]) = \rho$$

This is also true for Spearman's ρ, sicne it is a special case of Pearson's r.
For Kendall's τ_a, we have, for $n > 10$, that

$$\tau_a(X, Y) \sim N\left(\theta, \frac{2(2n+5)}{9n(n-1)}\right)$$

where $\theta = \mathbb{E}_{X_1,X_2 \sim F_X, Y_1,Y_2 \sim F_Y}[sign(X_1 - X_2) \cdot sign(Y_1 - Y_2)]$. Hence τ_a is an unbiased estimator for θ.

Mean Squared Error (MSE)

The Mean Squared Error of an estimator T for a parameter θ is defined as

$$MSE(T) = \mathbb{E}[(T - \theta)^2]$$

The MSE can be used to compare different estimators by considering both their bias and variance. The lower the MSE, the better the estimator. The MSE can be decomposed into the sum of the variance and the square of the bias:

$$MSE(T) = \mathbb{E}[(T - \mathbb{E}[T] + \mathbb{E}[T] - \theta)^2] =$$

$$= \mathbb{E}[(T - \mathbb{E}[T])^2] + (\mathbb{E}[T] - \theta)^2 + 2\underbrace{\mathbb{E}[T - \mathbb{E}[T]]}_{(\textit{this is 0})}(E[T] - \theta) = \mathit{Var}(T) + \mathit{Bias}(T)^2$$

A biased estimator with low variance may be better than an unbiased estimator with high variance, since the latter may have a higher MSE.

Consistent estimator

An estimator T_n is a squared error consistent estimator if

$$\lim_{n\to\infty} MSE(T_n) = 0$$

A consistent estimator has both bias and variance go to 0 as n goes to 0. For example, $\bar{X}_n$ is a squared error consistent estimator of μ.

Minimum Variance Unbiased Estimator (MVUE)

An unbiased estimator T_n of θ is a Minimum Variance Unbiased Estimator if

$$Var(T_n) \leq Var(S_n)$$

for all unbiased estimators S_n of θ.

As a corollary, if T_n is a MVUE, then $MSE(T_n) \leq MSE(S_n)$. $\bar{X}_n$ is also a MVUE of μ when the random sample is normally distributed with parameters μ and σ^2.

---

15 Maximum Likelihood Estimation

The previous sections showed different ways to derive parameter estimators using the "plug-in method": knowing a sample, we use a formula of random variables, substitute the empirical data, and obtain an estimate. This section will introduce a more general parametric principle to derive estimators, called **Maximum Likelihood Estimation** (**MLE**). The maximum likelihood principle states: "Given a dataset, choose the parameter(s) of interest that maximize the likelihood of observing that dataset".

Likelihood, Log-likelihood

Let $x_1, \ldots, x_n$ be a dataset, realization of a random sample $X_1, \ldots, X_n$, where the PMF/PDF of X_i is $f_\theta()$, parametric on θ. The likelihood function is defined as

$$L(\theta) = \prod_{i=1}^{n} f_\theta(x_i)$$

and the log likelihood function is defined as

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f_\theta(x_i)$$

The goal of MLE is to find the parameter value(s) that maximize the likelihood of the data. Since finding the maximum (in practice, derivating) a product is difficult, the logarithm of it is used instead, since it is a monotonic function (so the maximum remains the same) and the original likelihood can be converted to a sum of terms. Another issue to consider when using products is that when it is computed, the result can become very small, leading to numerical instability.

Maximum Likelihood estimates

The maximum likelihood estimates of θ is the value $t = \arg\max_\theta L(\theta) = \arg\max_\theta l(\theta)$. The statistics over the random sample

$$\hat{\theta}_{ML} = \arg\max_\theta L(\theta) = \arg\max_\theta l(\theta)$$

is called maximum likelihood estimator for θ.

Often, loss functions are used to evaluate the quality of the estimator. Some examples are:

- **Negative log-likelihood**: $nLL(\theta) = -l(\theta)$
- **Akaike Information Criterion** (**AIC**): $AIC = 2|\theta| - 2l(\theta)$. This function balances model fit (in the second term) with model simplicity (in the first term).
- **Bayesian Information Criterion** (**BIC**): $BIC = |\theta| \log(\theta) - 2l(\theta)$. It has a stronger penalty for model complexity compared to AIC.

The last two can be used to compare models with different number of parameters since they consider complexity as well as fit.

MLE estimators can be biased, but under mild assumptions, they are always asymtotically unbiased, and (also asymptotically) they always have the smallest variance among unbiased estimators. Additionally, if $\hat{\theta}$ is the MLE estimator of θ and $g()$ is an invertible function, $g(\hat{\theta})$ is the MLE estimator of $g(\theta)$.

Cross entropy and negative log likelihood There is a connection between the cross entropy and the negative log likelihood functions. Assume we have the random variables X, Y with PMF p_x and p_y. The cross entropy of X with respect to Y is defined as

$$H(X;Y) = -\sum_i p_x(a_i) \log p_y(a_i).$$

The negative log-likelihood is

$$nLL(\theta) = -\sum_i \underbrace{\log(f_\theta(x_i))}_{This\ is\ \log p_y(a_i)} = H(X;Y)$$

where $X \sim F_n$ and $Y \sim F_\theta$. Minimizing the cross entropy (or the KL-divergence) between empirical and theoretical distributions is equivalent to maximizing the likelihood of the data.

Score function and Fisher Information The entropy of a random variable is the mean information carried by it. Then, the partial derivative (w.r.t. θ) of the logarithm of f_θ represents the change in information as θ varies. It turns out that

$$\mathbb{E}\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right] = 0$$

for any X. Instead of studying the expectation, we consider the variance. Formally, we define the score function and the Fisher information as follows:

Score function

The score function is the random variable

$$S(\theta) = \frac{\partial}{\partial\theta} l(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f_\theta(X_i)$$

Fisher information

The Fisher information is the variance of th score function:

$$I(\theta) = Var(S(\theta)) = \mathbb{E}[S(\theta^2)]$$

The Fisher information measures the sensitivity of X with respect to θ. If small changes in θ cause large changes in the density function, the Fisher information is high: this means that the data provides information on the correct θ.

For unbiased estimators, the **Cramér-Rao bound** holds, stating that for any unbiased estimator T:

$$Var(T) \geq \frac{1}{I(\theta)}$$

An unbiased estimator for which the equality holds is a Minimum Variance Unbiased Estimator. The **absolute efficiency** of an estimator can be then calculated as

$$e(T) = \frac{1}{I(\theta) \cdot Var(T)} \in [0, 1]$$

Recall that a MLE estimator is always asymptotically unbiased and has asymptotic minimum variance. Also by Cramér-Rao, asymptotically we have that

$$se(\hat{\theta}_{ML}) = \sqrt{Var(\hat{\theta}_{ML})} = \frac{1}{\sqrt{n \cdot I(\theta)}}$$

where se is the standard error of the estimator.

16 Regression

Regression analysis is a set of statistical processes used to estimate the relationship between one or more dependent variables, and one or more independent variables. This relationship may be linear or non-linear, and regression analysis can be used for both cases. This section will start with the simplest case (univariate simple linear regression), and gradually introduce more complex cases.

16.1 Simple linear regression

In a simple linear regression model for a bivariate dataset $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, we assume the x_i are non-random, and the y_i are realizations of random variables Y_i satisfying the following equation:

$$Y_i = \alpha + \beta x_i + U_i \qquad\qquad i = 1, 2, \ldots, n$$

where $U_1, U_2, \ldots, U_n$ are independent random variables with 0 expectation and constant variance σ^2. The equalities above describe a regression line, $y = \alpha + \beta x$, where α is the **intercept**, and β is the **slope** of the line. x is called **independent** (or *explanatory*) variable, while y is the **dependent** (or *response*) variable. Since the U_i are independent, and each Y_i is a function of the respective U_i, the Y_i are also independent (per the propagation of indipendence rule). They are not identically distributed, however, since each Y_i depends on a different x_i. All the Y_i have the same variance σ^2 (same as the U_i). This property is called **homoscedasticity**.

Least Squares Estimation To estimate the values of α and β, an initial idea is to use MLE. However, MLE is a parametric method, meaning that we need to know the distribution of the U_i beforehand. The alternative is to use the **Least Squares** method: let

$$y_i - \alpha - \beta x_i$$

be the **residual** of the i^{th} observation, which is the realization of $U_i = Y_i - \alpha - \beta x_i$. The least squares method aims to minimize the sum of squares of residuals across all observations:

$$\hat{\alpha}, \hat{\beta} = \arg\min_{\alpha,\beta} S(\alpha, \beta) = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

$S(\alpha, \beta)$ is called **Sum of Squares of Errors** (**SSE**) or Residual Sum of Squares (RSS). To minimize the function, we need to calculate the partial derivatives with respect to α and β, and set them to 0. The partial derivatives are:

$$\frac{\partial}{\partial \alpha} S(\alpha, \beta) = -2 \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) \qquad \frac{\partial}{\partial \beta} S(\alpha, \beta) = -2 \sum_{i_1}^{n} x_i (y_i - \alpha - \beta x_i)$$

By setting both to 0, we get the estimates

$$\hat{\alpha} = \bar{y}_n - \hat{\beta}\bar{x}_n \qquad \hat{\beta} = \frac{n \sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

The $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$ are called **fitted values**. The difference between the fitted value and the observed value is called residual: $y_i - \hat{y}_i$.
An equivalent form of $\hat{\beta}$ is the following:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{SXX} = r_{xy} \frac{s_y}{s_x}$$

where:

- $SXX = \sum_{i=1}^{n} (x_i - \bar{x}_n)^2$;
- r_{xy} is the Pearson's correlation coefficient between x and y;
- s_x and s_y are the sample standard deviations of x and y, respectively.

The line described by the equation $y = \hat{\alpha} + \hat{\beta} x$ always passes through the center of gravity $(\bar{x}_n, \bar{y}_n)$:

$$\hat{\alpha} = \bar{y}_n - \hat{\beta}\bar{x}_n \implies \hat{\alpha} + \hat{\beta}\bar{x}_n = \bar{y}_n - \hat{\beta}\bar{x}_n + \hat{\beta}\bar{x}_n = \bar{y}_n$$

Unbiasedness of LS estimators Both LS estimators of slope and interept are unbiased, and the following is the proof. Starting with the estimator of $\hat{\beta}$, it can be rewritten as:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n)(Y_i - \bar{Y}_n)}{SXX} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n) Y_i - \sum_{i_1}^{n} (x_i - \bar{x}_n) \bar{Y}_n}{SXX} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n) Y_i}{SXX}$$

(this is because $\sum_{i=1}^{n} (x_i - \bar{x}_n) = 0$). We can now calculate its expectation, which is:

$$\mathbb{E}[\hat{\beta}] = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n) \mathbb{E}[Y_i]}{SXX} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n)(\alpha + \beta x_i)}{SXX} =$$

$$\underbrace{\alpha \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n)}{SXX}}_{\textit{this it 0 as above}} + \beta \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n) x_i}{SXX} = \beta \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n) x_i}{SXX} = \beta$$

Moreover, its variance is:

$$Var(\hat{\beta}) = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2 Var(Y_i)}{SXX^2} = \sigma^2 \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2}{SXX^2} = \frac{\sigma^2}{SXX}$$

Now, we calculate the expectation of $\hat{\alpha}$:

$$\mathbb{E}[\hat{\alpha}] = \mathbb{E}[\bar{Y}_n] - \bar{x}_n \mathbb{E}[\hat{\beta}] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[Y_i] - \bar{x}_n \beta =$$

$$= \frac{1}{n} \sum_{i=1}^{n} (\alpha + \beta x_i) - \bar{x}_n \beta = \frac{n\alpha}{n} + \frac{\beta}{n} \sum_{i=1}^{n} x_i - \bar{x}_n \beta =$$

$$= \alpha + \bar{x}_n \beta - \bar{x}_n \beta = \alpha$$

Moreover, its variance is:

$$Var(\hat{\alpha}) = Var(\bar{Y}_n - \hat{\beta}\bar{x}_n) = Var(\bar{Y}_n) + \bar{x}_n^2 Var(\hat{\beta}) - 2Cov(\bar{Y}_n, \beta\bar{x}_n) =$$

$$= \frac{\sigma^2}{n} + \bar{x}_n^2 \frac{\sigma^2}{SXX} - 2\bar{x}_n \underbrace{Cov(\bar{Y}_n, \hat{\beta})}_{\textit{this is 0}} = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}_n^2}{SXX} \right)$$

Both variances of the two estimators use σ^2, which is unknown. We cannot estimate it using the unbiased estimator $\hat{\sigma}^2 = \sum_{i=1}^n (Y_i - \bar{Y}_n)^2/(n-1)$, because the Y_i all have a different expectation. In this case, an unbiased estimator for the variance is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2}{n-2}$$

$\hat{\sigma}$ is called **residual standard error**. The standard errors of the coefficient estimators are defined as estimates of their respective standard deviations:

$$se(\hat{\alpha}) = \hat{\sigma}\sqrt{\left(\frac{1}{n} + \frac{\bar{x}_n^2}{SXX}\right)} \qquad\qquad se(\hat{\beta}) = \frac{\sigma}{\sqrt{SXX}}$$

A measure close to the residual standard error is the **Root Mean Squared Error**, defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^n (y_i - \hat{a} - \hat{\beta}x_i)^2}$$

LSE and MLE MLE and LSE are equivalent in a special case: when the U_i random variables are normally distributed with mean 0 and variance σ^2. The Y_i then are also normally distributed; specifically, $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$.
The log-likelihood function is:

$$l(\alpha, \beta) = \sum_{i=1}^n \log\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_i - \alpha - \beta x_i}{\sigma}\right)^2} \right) = -n\log(\sigma\sqrt{2\pi}) - \frac{1}{2}\left(\frac{\sum_{i=1}^n (y_i - \alpha - \beta x_i)}{\sigma} \right)^2$$

It turns out that $\arg\max_{\alpha,\beta} l(\alpha, \beta) = \hat{\alpha}, \hat{\beta}$ as found for LSE.

Total variability and R^2 The total variability of the observed data is calculated as the **Sum of Squares Total** (**SST**):

$$SST = \sum_{i=1}^n (y_i - \bar{y}_n)^2$$

It is the sum of the squared differences between each observation y_i and their mean.
The total variability of the fitted data is instead calculated as the **Sum of Squares of Regression** (**SSR**):

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}}_n)^2$$

It is identical to the above, but uses the fitted values instead of those in the dataset.
The total variablity of the residuals, which is the unexplained variability, is calculatd as the aforementioned **Sum of Squares of Errors** (**SSE**):

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

The three quantities are related by the following equation:

$$SST = SSR + SSE$$

Also, $1 - SSE/SST$ (or SSR/SST) is the *fraction of explained variability over total variability.* We can now express the variances of the observations, of the residuals, and of the fitted values using the three quantities above:

$$\sigma_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y}_n)^2}{n-1} = \frac{SST}{n-1} \qquad \text{(Variance of the } y\text{)}$$

$$\sigma_{res}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{n-1} = \frac{SSE}{n-1} \qquad \text{(Variance of the residuals)}$$

$$\sigma_{\hat{y}}^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}}_n)^2}{n-1} = \frac{SSR}{n-1} \qquad \text{(Variance of the fitted values)}$$

These quantities are used to define the **coefficient of determination** R^2:

$$R^2 = 1 - \frac{\sigma_{res}^2}{\sigma_y^2} = \frac{\sigma_{\hat{y}}^2}{\sigma_y^2}$$

It is a measure of how well the regression line fits the data. For simple linear regression, it is equal to the square of the Pearson's correlation coefficient between y and $\hat{y}$:

$$R^2 = r_{y\hat{y}}^2 = \frac{[\sum_{i=1}^{n}(y_i - \bar{y}_n) \cdot (\hat{y}_i - \bar{\hat{y}}_n)]^2}{\sum_{i=1}^{n}(y_i - \bar{y}_n)^2 \cdot \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}}_n)^2}$$

When we take the adjusted sample variance of the residuals:

$$\hat{\sigma}_{res}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{SSE}{n-2}$$

we can then define the **adjusted coefficient of determination** $adjR^2$:

$$adjR^2 = 1 - \frac{\hat{\sigma}_{res}^2}{\sigma_y^2} = 1 - \frac{SSE/(n-2)}{SST/(n-1)} = 1 - \frac{SSE}{SST} \cdot \frac{n-1}{n-2}$$

16.2 Non-linear simple regression

16.3 Multiple linear regression

16.4 Multivariate multiple linear regression