Statistics with R

Chapter 1: Introduction to statistics

Tabea Rebafka

October 2018

Master AIMS 2018–19

Outline

What is statistics? I

What is the aim of statistics?

Analysis and interpretation of **data** (or observations, measurements)

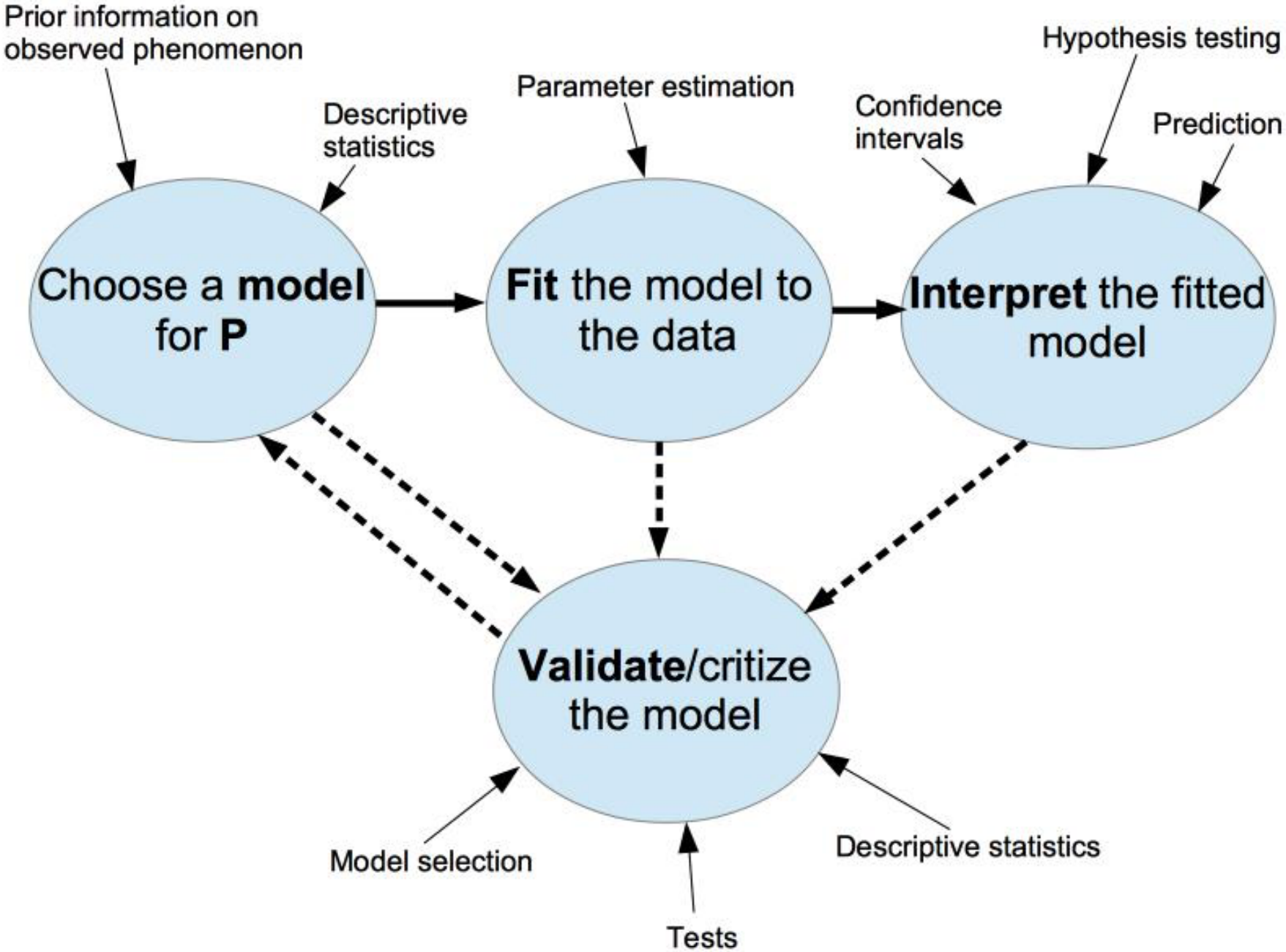
- understand an observed phenomenon by **statistical inference** (i.e. **modelling, estimation and testing**)
- recover unobserved features (**prediction**)

What is statistics? II

Statistical approach

- Use a **probabilistic model** to explain the nature of the data (in opposition to **data analysis**)
- Let $x_1, \ldots, x_n$ be the data. A statistician assumes that $(x_1, \ldots, x_n)$ is the realization of a random variable $\mathbf{X}$ with distribution $\mathbb{P}$.
- The distribution $\mathbb{P}$ is **unknown** (in opposition to probability theory).

What is statistics? III



Example: Coin tossing I

Data

- Observations: the outcome of n tosses of the same coin
- *Head* is encoded by 1, *tail* by 0.
- Data: $x_1, \ldots, x_n$ with $x_i \in \{0, 1\}$. The number n is called the **sample size**.

Probabilistic model

- Consider x_i as independent realizations of a Bernoulli distribution
- More precisely, let X_i be i.i.d. (independent and identically distributed) random variables with **Bernoulli distribution** $B(p)$ with parameter $p \in (0, 1)$, i.e.

$$\mathbb{P}(X_i = 1) = p = 1 - \mathbb{P}(X_i = 0)$$

- Bernoulli parameter p is **unknown**.

Example: Coin tossing II

Fit the model

- Estimate the Bernoulli parameter p from the data $x_1, \ldots, x_n$.
- Simple idea: we know that for $X_i \sim B(p)$ i.i.d., we have

$$\mathbb{E}[X_1] = p \quad \text{and} \quad \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} p \quad (n \to \infty).$$

- Use the **sample mean** $\bar{X}_n$ as an **estimate** of p:

$$\hat{p}_n = \bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Example: Coin tossing III

Properties of the estimator $\hat{p}_n$ of p

- $\hat{p}_n = \bar{X}_n \xrightarrow{P} p$ as $n \to \infty$, i.e. when the sample size n is large, $\hat{p}_n$ tends to be close to p (**consistency**).
- $\mathbb{E}[\hat{p}_n] = p$, i.e. in average $\hat{p}_n$ takes the target value p (**unbiased**)
- **Mean squared error** (MSE)

$$\mathbb{E}[(\hat{p}_n - p)^2] = \frac{p(1-p)}{n} \longrightarrow 0 \quad (n \to \infty).$$

- **Limit distribution and rate of convergence**

$$\sqrt{n}(\hat{p}_n - p) \xrightarrow{d} \mathcal{N}(0, p(1-p)) \quad (n \to \infty).$$

Example: Coin tossing IV

Quantify uncertainty of the estimate

Instead of a point estimator $\hat{p}_n$ compute an **interval** $\mathcal{I}$ that depends on the data (i.e. $\mathcal{I} = \mathcal{I}(x_1, \ldots, x_n)$) and that contains the target p with given probability γ (**confidence interval**):

$$\mathbb{P}(p \in \mathcal{I}) \geq \gamma.$$

The length of the interval $\mathcal{I}$ indicates the uncertainty about our estimation of p.

Example: Coin tossing V

Confidence interval for p in the Bernoulli model

- An asymptotic confidence interval is given by

$$\mathcal{I}_n = \left[\hat{p}_n + q^N_{\gamma_1}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}, \hat{p}_n + q^N_{\gamma_2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right]$$

with $\gamma_1 = (1-\gamma)/2$ and $\gamma_2 = (1+\gamma)/2$ and where q^N_α denotes the α-quantile of the standard normal distribution $\mathcal{N}(0,1)$ defined by

$$\mathbb{P}(Z \leq q^N_\alpha) = \alpha \quad \text{for } Z \sim \mathcal{N}(0,1).$$

- Interval length:

$$2q^N_{\gamma_2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}$$

by using $q^N_{\gamma_1} = -q^N_{\gamma_2}$.

Example: Coin tossing VI

Statistical testing

Answer questions as: Is the coin a fair coin?

- Mathematically speaking: Is $p = 1/2$ or $p \neq 1/2$?
- Estimate p and evaluate the uncertainty of the estimate
- If the estimate is too far away from $1/2$, then decide that $p \neq 1/2$. Otherwise conserve the hypothesis that $p = 1/2$.

Refresher on probability theory

Definition

- A **sample space** is any finite or infinite set Ω (it is thought as the set of all possible outcomes of a random experiment).
- Any subset $A \subset \Omega$ is called an **event**, including Ω and the empty set $\emptyset$.

Example: Dice

- Sample space of rolling a dice: $\Omega = \{1, \ldots, 6\}$.
- Some events:

$$A = \{2\},$$
$$B = \{2, 4, 6\} = \{\text{the result is even}\},$$
$$C = \emptyset,$$
$$D = \Omega.$$

Probability measures I

Basically, a probability measure assigns a probability to every event.

Definition

Let Ω be a sample space, a **probability measure** $\mathbb{P}$ on Ω is an application

$$\mathbb{P} : \{\text{Events}\} \to [0, 1]$$

such that

- $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\Omega) = 1$.
- **(Countable additivity)** For every sequence of disjoint events $A_1, A_2, \ldots$

$$\mathbb{P}\left(\bigcup_{n\geq 1} A_n\right) = \sum_{n\geq 1} \mathbb{P}(A_n).$$

A pair $(\Omega, \mathbb{P})$ is called a **probability space**.

Probability measures II

Examples

- The **uniform measure** on a finite set Ω is defined by

$$\mu(A) = \frac{\mathrm{card}(A)}{\mathrm{card}(\Omega)}.$$

- The **Dirac measure** (or *Dirac mass*) at some point a, denoted by δ_a, puts all the *mass* on a:

$$\delta_a(A) = \begin{cases} 1 & \text{if } a \in A, \\ 0 & \text{otherwise.} \end{cases}$$

The **Lebesgue measure** on $\mathbb{R}$ is the measure λ that assigns the length to each interval $[a, b]$:

$$\lambda\big([a, b]\big) = b - a.$$

It is not a probability measure as its values are not restricted to $[0, 1]$.

Probability measures III

Proposition

Let $(\Omega, \mathbb{P})$ be a probability space.

(i) If $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.

(ii) For any event A, $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

(iii) For any events A, B,

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B),$$

in particular $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$.

Probability measures IV

Proposition

(iv) **(Union bound)** More generally, let $(A_n)_{n\geq 1}$ be any sequence of sets (not necessarily disjoint),

$$\mathbb{P}\left(\bigcup_{n\geq 1} A_n\right) \leq \sum_{n\geq 1} \mathbb{P}(A_n).$$

(v) **(Law of total probability)** Let A be an event and $B_1, B_2, \ldots$ be a sequence of disjoint sets such that $\cup_{n\geq 1} B_n = \Omega$,

$$\mathbb{P}(A) = \sum_{n\geq 1} \mathbb{P}(A \cap B_n).$$

Random variables I

- From now on, we work on a fixed probability space $(\Omega, \mathbb{P})$ where $\mathbb{P}$ is a probability measure.
- Elements of Ω are often denoted by ω.

Definition

Any function $X : \Omega \to \mathbb{R}$ is a **random variable**.

Random variables II

Example: Indicator function

For a given event A, the **indicator function** of A is denoted by $\mathbb{1}_A$ and defined as

$$\mathbb{1}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{otherwise.} \end{cases}$$

Indicator functions are similar to Dirac measures as $\mathbb{1}_A(\omega) = \delta_\omega(A)$.

Random variables III

Definition

- The **distribution** or **law** of X, denoted by $\mathbb{P}_X$, is the probability measure on $\mathbb{R}$ such that for any event A

$$\mathbb{P}_X(A) = \mathbb{P}\left(\{\omega \text{ such that } X(\omega) \in A\}\right) = \mathbb{P}(X \in A).$$

  We write $X \sim \mathbb{P}_X$.

- The **cumulative distribution function** (or just distribution function) of X is the function $F_X : \mathbb{R} \mapsto [0, 1]$ defined by

$$F_X(t) = \mathbb{P}(X \leq t) \quad \text{for every } t.$$

Theorem

X and Y have the same law $\iff F_X(t) = F_Y(t)$ for every t.

Random variables IV

Properties of the distribution function

(i) F_X is non-decreasing.

(ii) F_X is right-continuous.

(iii) $\lim_{t \to -\infty} F_X(t) = 0, \quad \lim_{t \to +\infty} F_X(t) = 1.$

Theorem

Any function F with properties (i), (ii) and (iii) above, is the distribution function of some random variable.

Discrete distribution I

Definition

- We say that X has a **discrete distribution** if X takes its values in a finite or countable set $\{x_1, x_2, \dots\}$.
- Discrete distributions are entirely described by their **probability mass function** $p(x) = \mathbb{P}(X = x)$ for $x \in \{x_1, x_2, \dots\}$.

Discrete distribution II

Examples of discrete distributions

- **Bernoulli distribution** $B(p)$ with parameter $p \in [0,1]$ with values in $\{0,1\}$:

$$\mathbb{P}(X=1) = p, \qquad \mathbb{P}(X=0) = 1-p.$$

  Model of the success or failure of an experiment.

- **Binomial distribution** $B(n,p)$ with parameters $n \geq 1$ and $p \in [0,1]$:

$$\mathbb{P}(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k = 0,1,\ldots,n.$$

  Model of the number of successes in n Bernoulli trials.

Discrete distribution III

Examples of discrete distributions

- **Geometric distribution** with parameter $p \in [0, 1]$:

$$\mathbb{P}(X = k) = (1 - p)^{k-1} p \quad \text{for } k = 1, 2, \ldots$$

  Model of the number of Bernoulli trials until the first success.

- **Poisson distribution** with parameter $\lambda > 0$:

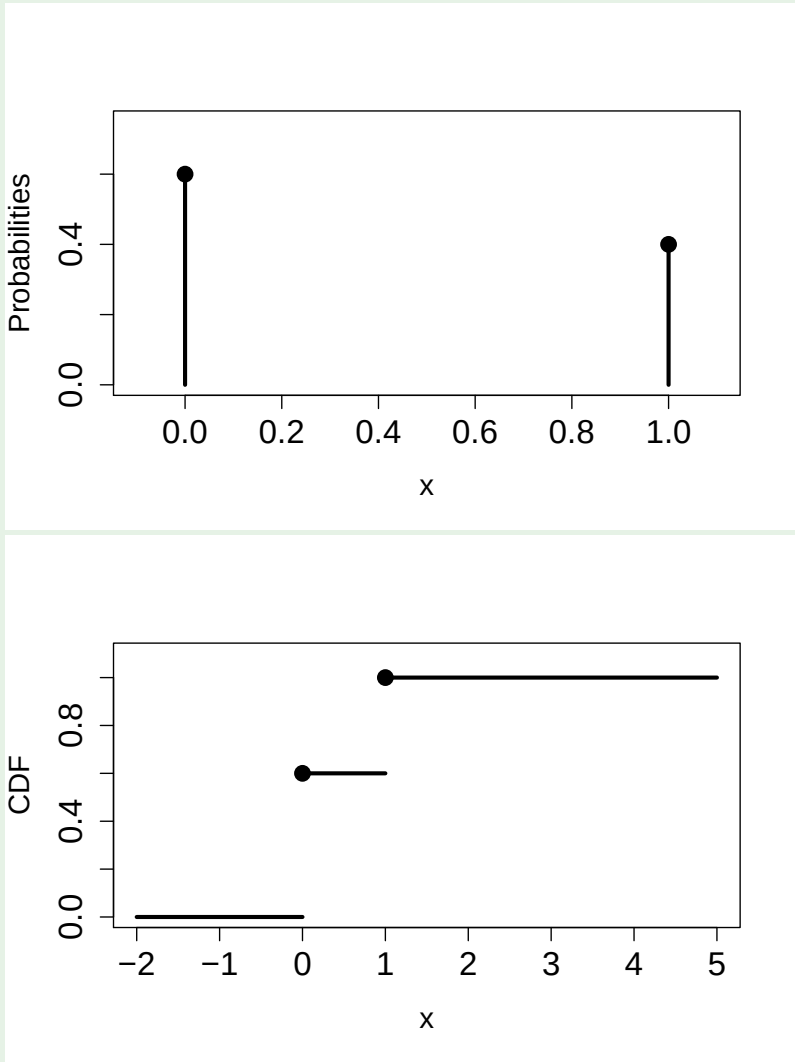
$$\mathbb{P}(X = k) = e^{-\lambda} \frac{\lambda^k}{k!} \quad \text{for } k = 0, 1, 2, \ldots$$

- **Discrete uniform distribution** on a finite set of values $\{x_1, \ldots, x_m\}$:

$$\mathbb{P}(X = x_k) = \frac{1}{m} \quad \text{for } k = 1, \ldots, m.$$

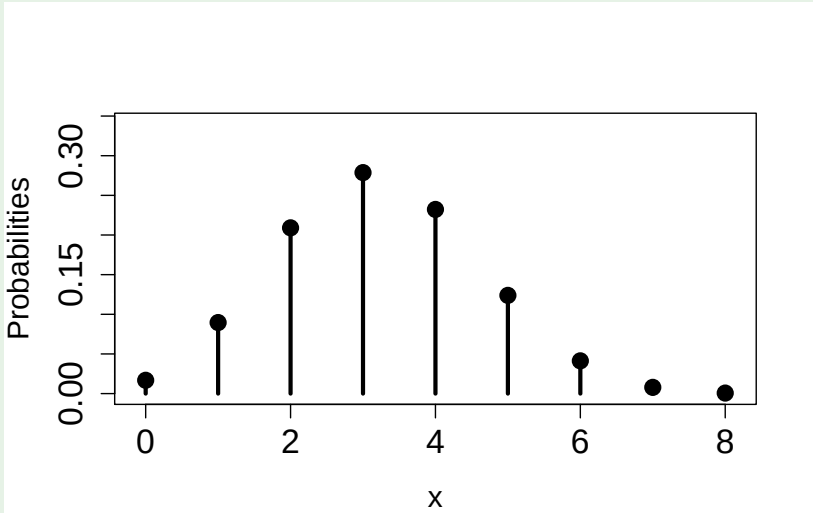
Discrete distribution IV

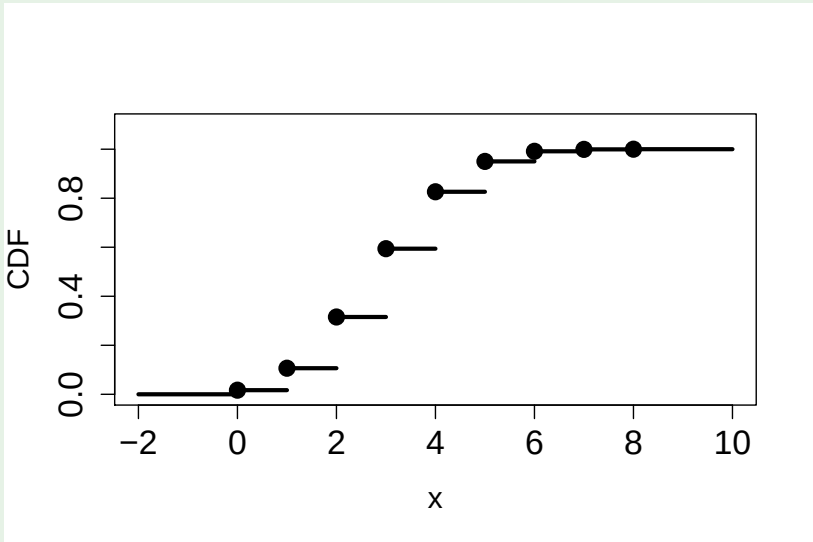
Bernoulli distribution with parameter $p = 0.4$



Discrete distribution V

Binomial distribution with parameters $n = 8$ and $p = 0.4$





Discrete distribution VI

- The cumulative distribution function of any **discrete** distribution is a step function.
- The jumps indicate the values taken by the random variable and
- the height of the jump indicates the associated probability.

Continuous distribution I

Definition

We say that X has **continuous distribution** if X takes its values in $\mathbb{R}$ (or in an interval of $\mathbb{R}$) and if there is a non-negative function f such that for any event A

$$\mathbb{P}(X \in A) = \int_A f(x)\mathrm{d}x.$$

The function f is called the **density** of X.

- Any density function f is non-negative and $\int_{\mathbb{R}} f(x)\mathrm{d}x = 1$.
- The density entirely describes the distribution of the random variable.

Continuous distribution II

Examples continuous distributions

- **Uniform distribution** $U[a, b]$ on $[a, b]$:

$$f(x) = \frac{1}{b-a} \mathbb{1}_{[a,b]}(x).$$

- **Exponential distribution** $\mathcal{E}(\lambda)$ with parameter $\lambda > 0$:

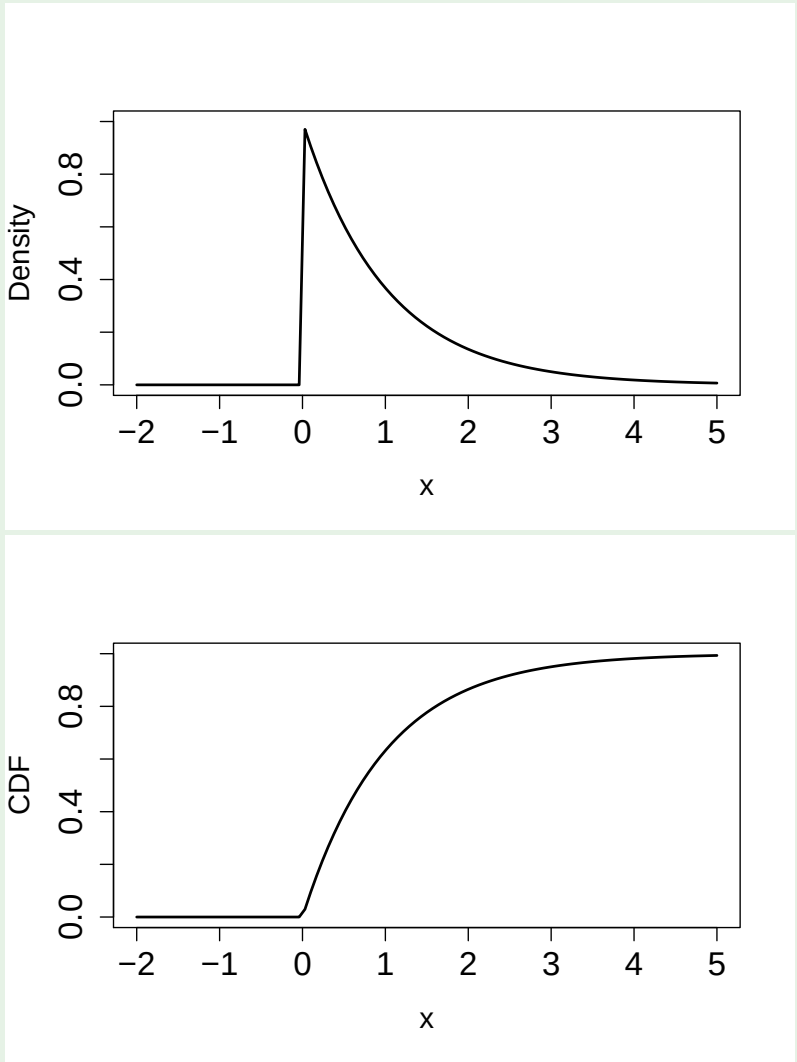
$$f(x) = \lambda \exp(-\lambda x) \mathbb{1}_{x \geq 0}.$$

- **Normal distribution** or gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ with parameters $\mu \in \mathbb{R}$, $\sigma^2 > 0$:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

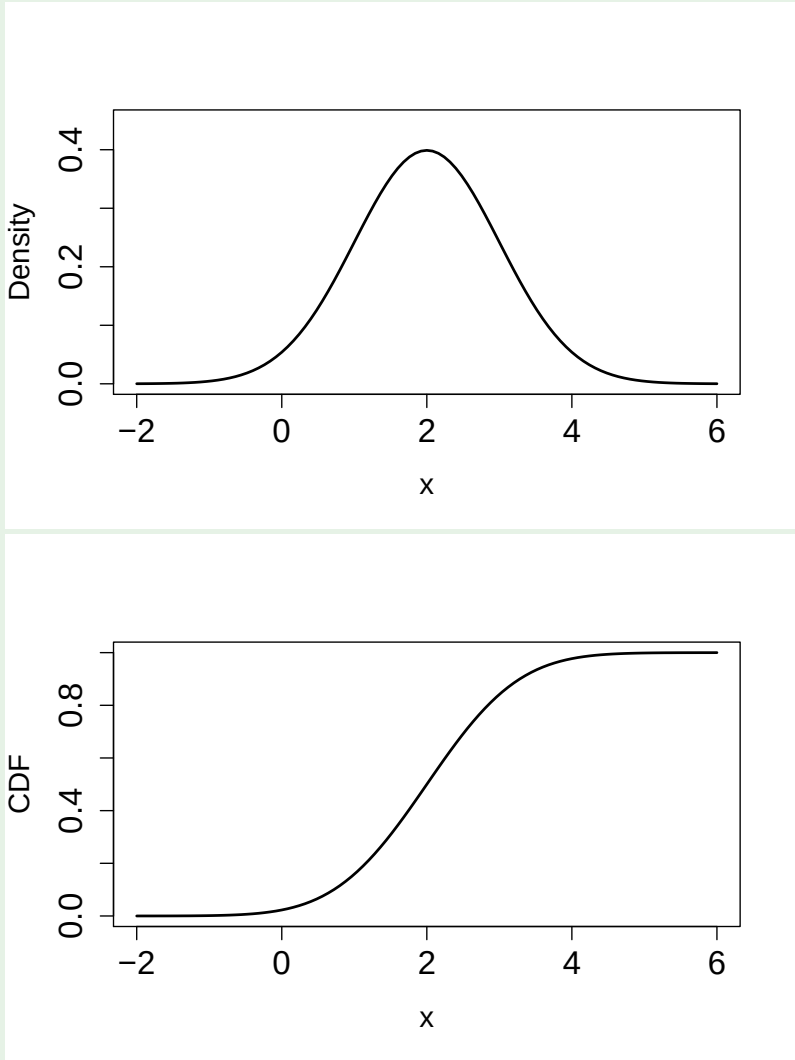
Continuous distribution III

Exponential distribution $\mathcal{E}(1)$ with parameter $\lambda = 1$



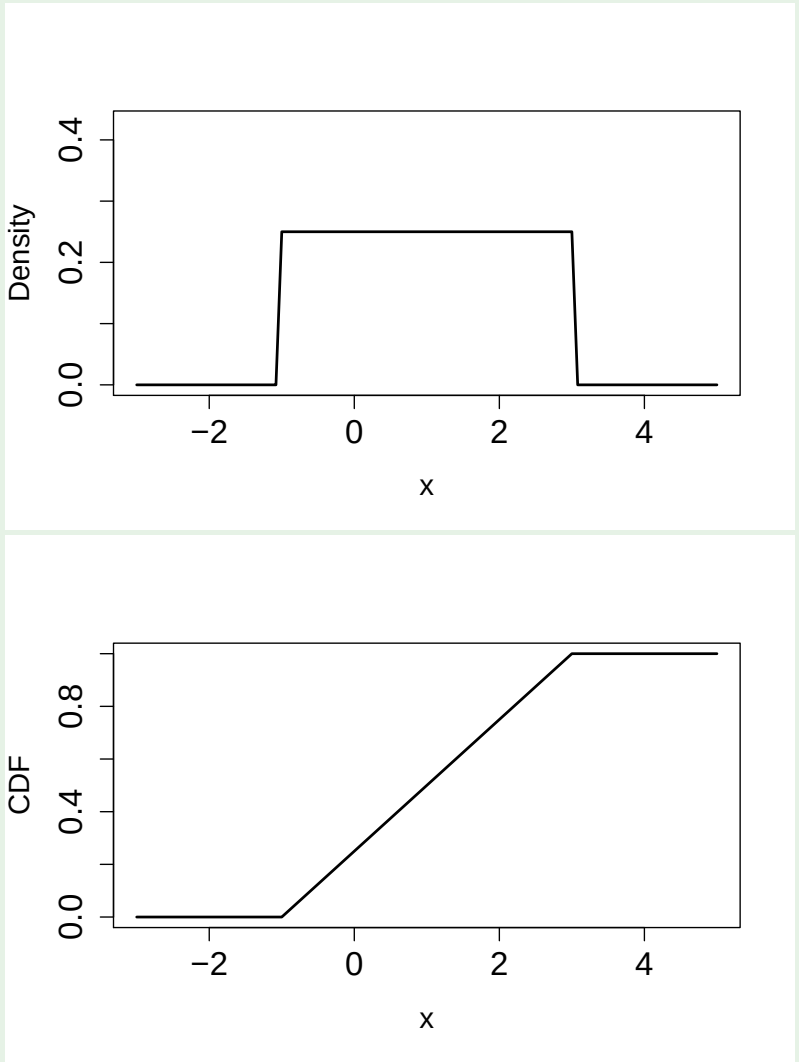
Continuous distribution IV

Normal distribution $\mathcal{N}(2, 1)$ with parameters $\mu = 2$ and $\sigma^2 = 1$



Continuous distribution V

Uniform distribution $U[-1,3]$ on $[-1,3]$



Continuous distribution VI

- The cumulative distribution function of any **continuous** distribution is continuous.
- We have $F_x(t) = \int_{-\infty}^{t} f(x)\mathrm{d}x$ for all t and
- $f(t) = F'(t)$ for almost all t.

Continuous distribution VII

- There exist random variables which are **neither discrete nor continuous!**
- For instance $X = \min\{1, Y\}$ where $Y \sim \mathcal{E}(1)$ (censored distribution).

Statistical modelling I

- In statistics, data $\mathbf{x} = (x_1, \ldots, x_n)$ are considered as a realization of a random vector $\mathbf{X} = (X_1, \ldots, X_n)$ with distribution $\mathbb{P}$.
- The distribution $\mathbb{P}$ is **unknown**.

Statistical model

- We introduce a family $\mathcal{P}$ of (known) probability distributions and **suppose** that $\mathbb{P}$ belongs to this family $\mathcal{P}$, i.e.

$$\mathbb{P} \in \mathcal{P}.$$

- $\mathcal{P}$ is called a **statistical model** and it is indeed a set of candidate distributions for $\mathbb{P}$.

Statistical modelling II

- A statistical model $\mathcal{P}$ is determined by using
  - our prior knowledge on the observed phenomenon and
  - tools from descriptive statistics.
- Any model is false. A model is only an approximation of reality.
- A model is always a **trade-off** between a precise description of a complex reality and mathematical convenience.

Statistical modelling III

Model parameter

- In general, we write $\mathcal{P} = \{\mathbb{P}_\theta, \theta \in \Theta\}$ where θ is the **model parameter** and Θ the **parameter set**.
- Denote $\theta_0 \in \Theta$ the "true value" of the parameter such that $\mathbb{P} = \mathbb{P}_{\theta_0}$.
- The problem of estimating $\mathbb{P}$ becomes the problem of estimating the parameter θ_0 from the data.

Identifiability

The model $\mathcal{P}$ is said to be **identifiable** if and only if

$$\forall \theta, \theta' \in \Theta, \mathbb{P}_\theta = \mathbb{P}_{\theta'} \Longrightarrow \theta = \theta'.$$

Statistical modelling IV

Example: Coin tossing

- The data $\mathbf{x} = (x_1, \ldots, x_n)$ are considered as a realization the random vector $\mathbf{X} = (X_1, \ldots, X_n)$ with $X_i \sim B(p)$ i.i.d. and unknown parameter $p \in (0, 1)$.
- In other words, we suppose that the distribution $\mathbb{P}$ of $\mathbf{X}$ belongs to the family

$$\mathcal{P} = \{B(p)^{\otimes n}, p \in (0, 1)\}.$$

Here, p is the model parameter.

Parameter estimation I

How to estimate θ_0 from the data $\mathbf{x} = (x_1, \ldots, x_n)$?

Definition

- Any function $S = S(\mathbf{x})$ defined on the data $\mathbf{x}$ is called a **statistic**.
- Examples: $S_1(\mathbf{x}) = 0, \forall \mathbf{x}$; $S_2(\mathbf{x}) = \bar{x}_n$.
- A statistic is called an **estimator** of θ_0 if the statistic is supposed to approach θ_0.

Parameter estimation II

There are different estimation approaches depending on the **size** of the parameter set Θ.

- If $\Theta \subset \mathbb{R}^d$ (i.e. if θ is a d-vector) for some $d < \infty$, the model is called **parametric**.
- If no parametrization of $\mathcal{P}$ exists such that Θ is of finite dimension, the model is called **non parametric**.

Examples of non parametric models:

- $\mathcal{P}$ = the set of all probability measures
- $\mathcal{P}$ = the set of all absolutely continuous probability measures
- $\mathcal{P}$ = the set of all absolutely continuous probability measures with continuous density