Statistics with R

Chapter 2: Descriptive statistics

Tabea Rebafka

October 2018

Master AIMS 2018–19

Outline

Descriptive statistics

Descriptive statistics or explanatory data analysis

- provides **simple** tools to analyze data:
  - data visualization,
  - graphical tools,
  - summary statistics and numerical indicators.
- is useful to determine a statistical model $\mathcal{P}$ for the distribution of our data.

Univariate observations

- Consider data $\mathbf{x} = (x_1, \ldots, x_n)$ with $x_i \in \mathbb{R}$ for all $i = 1, \ldots, n$ (**univariate observations**).
- Let's study two examples.

Example I: Deaths from lung diseases

Number of monthly deaths from lung diseases

	Jan	Feb	March	April	May	June	July	Aug	Sep	Oct	Nov	Dec
1974	3.04	2.55	2.70	2.55	2.01	1.66	1.72	1.52	1.60	2.07	2.20	2.51
1975	2.93	2.89	2.94	2.50	1.87	1.73	1.61	1.54	1.40	1.79	2.08	2.84
1976	2.79	3.89	3.18	2.01	1.64	1.58	1.49	1.30	1.36	1.65	2.01	2.82
1977	3.10	2.29	2.38	2.44	1.75	1.55	1.50	1.36	1.35	1.56	1.64	2.29
1978	2.82	3.14	2.68	1.97	1.87	1.63	1.53	1.37	1.36	1.57	1.54	2.49
1979	3.08	2.60	2.57	2.14	1.69	1.50	1.46	1.35	1.33	1.49	1.78	1.92

Number of monthly deaths from lung diseases in the UK from 1974 to 1979 (in thousands).

Example II: Scientific discoveries

Number of scientific discoveries									
1860	1861	1862	1863	1864	1865	1866	1867	1868	1869
5	3	0	2	0	3	2	3	6	1
1870	1871	1872	1873	1874	1875	1876	1877	1878	1879
2	1	2	1	3	3	3	5	2	4
1880	1881	1882	1883	1884	1885	1886	1887	1888	1889
4	0	2	3	7	12	3	10	9	2
1890	1891	1892	1893	1894	1895	1896	1899	1898	1899
3	7	7	2	3	3	6	2	4	3
1900	1901	1902	1903	1904	1905	1906	1907	1908	1909
5	2	2	4	0	4	2	5	2	3
1910	1911	1912	1913	1914	1915	1916	1917	1918	1919
3	6	5	8	3	6	6	0	5	2
1920	1921	1922	1923	1924	1925	1926	1927	1928	1929
2	2	6	3	4	4	2	2	4	7
1930	1931	1932	1933	1934	1935	1936	1937	1938	1939
5	3	3	0	2	2	2	1	3	4
1940	1941	1942	1943	1944	1945	1946	1947	1948	1949
2	2	1	1	1	2	1	4	4	3
1950	1951	1952	1953	1954	1955	1956	1957	1958	1959
2	1	4	1	1	1	0	0	2	0

Number of major scientific discoveries or important inventions per year from 1860 to 1959.

Type of distribution I

- Suppose that the observations x_i are i.i.d. realizations of some random variable X with distribution function F.
- Of what type is the distribution of X? **Discrete** or **continuous** ? Or neither of them?

Type of distribution II

Proposition

If F is a continuous distribution and $X_i \sim F, i = 1, 2, \ldots$ i.i.d., then

$$\mathbb{P}(X_i = X_j) = 0, \quad \forall i \neq j.$$

Thus, if F is a continuous distribution, the probability of multiple identical observations in the sample is 0.

- If the number of different values in the sample is of the order of n, then use a continuous distribution.
- Otherwise, when the sample contains values whose frequency is much larger than $1/n$, then use a discrete distribution.

Type of distribution III

Two types of discrete variables

Discrete random variables can be

- **quantitative**: number of siblings, grades, number of car accidents, number of fruits/vegetables per day...
- **qualitative**: sex, nationality, type of fruit, colors...

Barchart I

Barcharts are **only for observations from a discrete distribution!**

- Denote $\mathcal{V} = \{v_k, k = 1, \ldots, m\}$ the set of distinct values in the sample $x_1, \ldots, x_n$. (We have $m < n$).
- Denote $\hat{p}_k$ the proportion of the values v_k in the sample, i.e.

$$\hat{p}_k = \frac{\#\{i : x_i = v_k\}}{n}, \quad k = 1, \ldots, m.$$

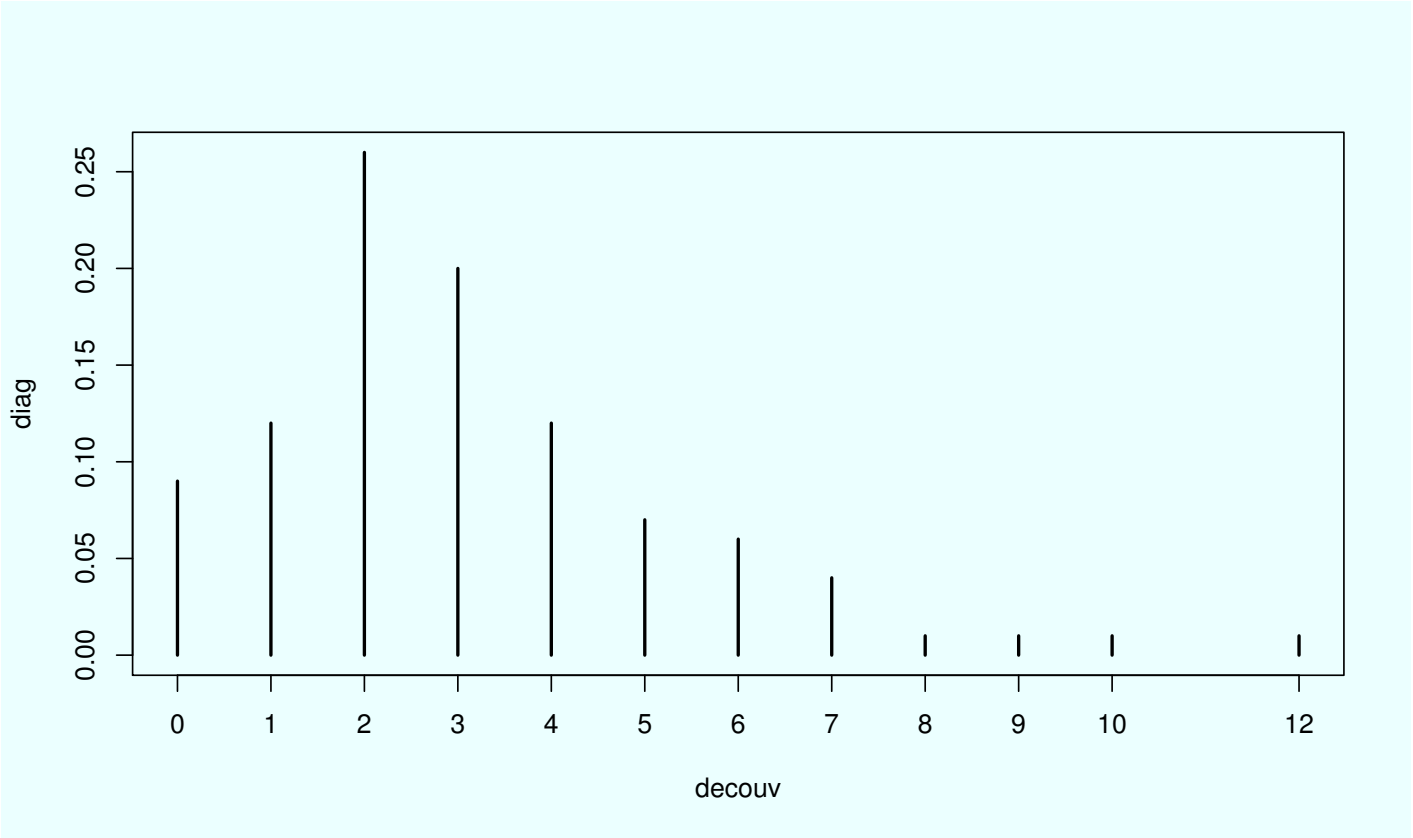
- For the barchart we draw vertical lines at v_k of length $\hat{p}_k$.

$\hat{p}_k$ is an approximation of $\mathbb{P}(X = v_k)$. Indeed,

$$\hat{p}_k \xrightarrow{P} \mathbb{P}(X = v_k), \quad (n \to \infty),$$

where $X \sim F$.

Barchart II



Barchart of scientific discoveries.

Histogram I

Histograms are **only for observations from a continuous distribution!**

1. Choose an interval $A = [a, b]$ such that $x_i \in A$ for all $i = 1, \ldots, n$.
2. Choose a partition size $m \in \mathbb{N}$ and define

$$A_j = [a + (j-1)h, a + jh], \quad j = 1, \ldots, m \quad \text{with } h = \frac{b-a}{m}.$$

3. Count the number of data points per subinterval:

$$N_j = \#\{i : x_i \in A_j\} = \sum_{i=1}^{n} \mathbb{1}\{x_i \in A_j\}.$$

Histogram II

Definition

The **histogram** $\hat{f}^H$ is defined by

$$\hat{f}^H(x) = \begin{cases} \frac{N_j}{n}, & \text{if } x \in A_j \\ 0, & \text{otherwise} \end{cases}$$

$$= \frac{1}{nh} \sum_{j=1}^{m} N_j \mathbb{1}\{x \in A_j\}, \quad x \in \mathbb{R}.$$

Histogram III

Properties

The histogram function $\hat{f}^H$

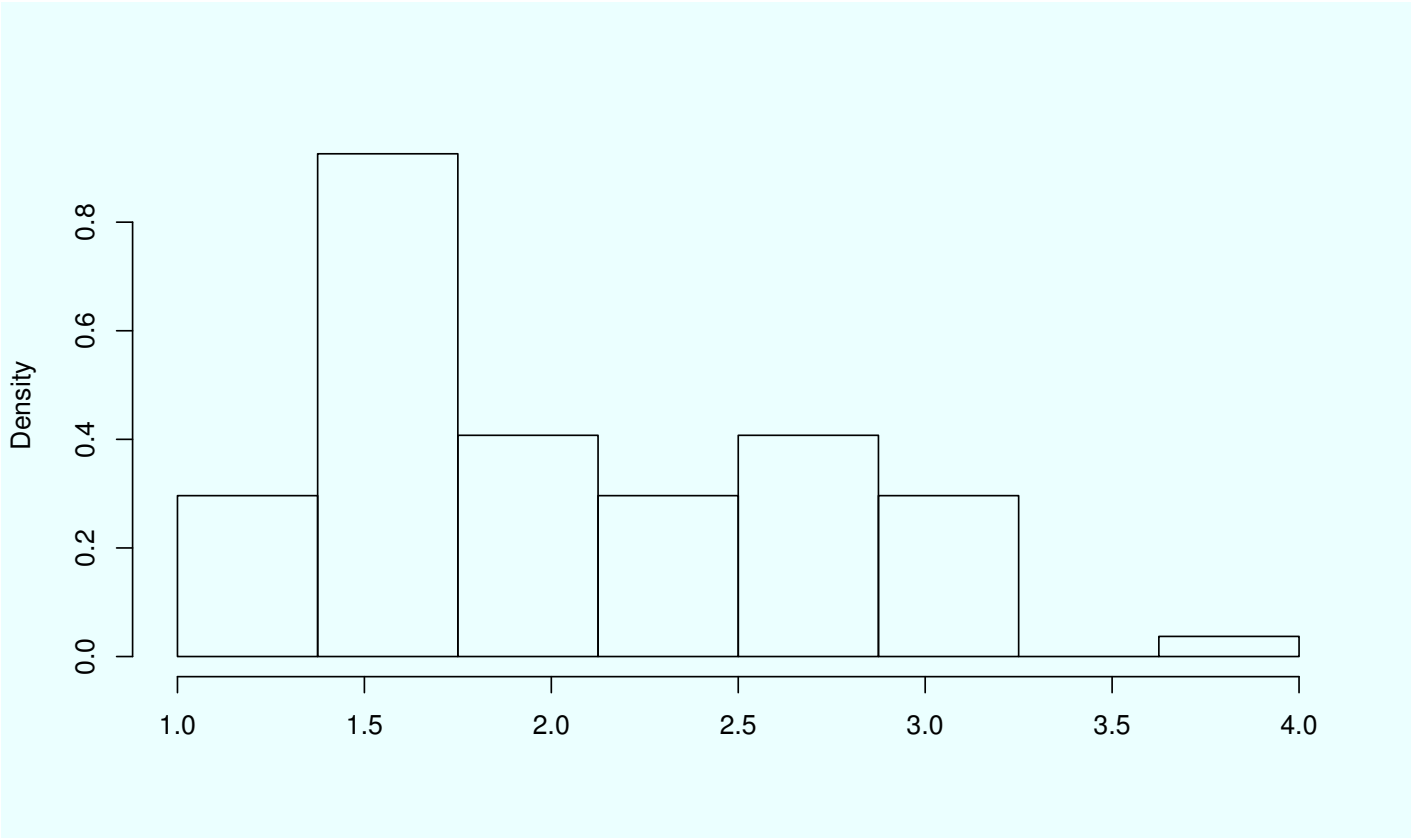
- is a piecewise constant function
- is a probability density function
- Under some mild regularity assumptions, we have pointwise convergence of the histogram, i.e. for all $x_0 \in \mathbb{R}$, we have

$$\hat{f}_n^H(x_0) \xrightarrow{P} f(x_0) \quad (n \to \infty).$$

The histogram $\hat{f}^H$ is an approximation of the density f of the distribution F of the data.

The histogram $\hat{f}^H$ is a nonparametric estimator of the density f.

Histogram IV



Histogram of the number of lung deaths.

Different types of convergence I

- Let X and $X_n, n \geq 1$ be random variables defined on $(\Omega, \mathbb{P})$.
- What does it mean that a sequence of random variables $(X_n)_{n \geq 1}$ converges to some random variable X?

Different types of convergence II

Definition

- $(X_n)_{n\geq 1}$ **converges** to X **in probability** ($X_n \xrightarrow{P} X$) if for all $\varepsilon > 0$

$$\lim_{n\to+\infty} \mathbb{P}\left(|X_n - X| > \varepsilon\right) = 0.$$

- $(X_n)_{n\geq 1}$ **converges** to X **almost surely** ($X_n \xrightarrow{\text{a.s.}} X$) if

$$\mathbb{P}\left(\left\{\omega \text{ such that } \lim_{n\to+\infty} X_n(\omega) = X(\omega)\right\}\right) = 1.$$

- $(X_n)_{n\geq 1}$ **converges in distribution** (or **in law**) ($X_n \xrightarrow{d} X$), if

$$\lim_{n\to\infty} F_{X_n}(t) = F_X(t),$$

for every $t \in \mathbb{R}$ such that F_X is continuous at t.

Different types of convergence III

Theorem

$$a.s. \quad \Longrightarrow \quad \mathbb{P} \quad \Longrightarrow \quad d$$

Different types of convergence IV

Strong law of large numbers

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. integrable random variables. Then

$$\bar{X}_n \longrightarrow \mathbb{E}[X_1] \ a.s. \quad \text{as } n \to \infty.$$

Different types of convergence V

Central limit theorem

Let $(X_n)_{n\geq 1}$ be i.i.d. random variables with finite variance. Denote $\mu = \mathbb{E}[X_1]$ and $\sigma^2 = \mathrm{Var}(X_1)$. Then

$$\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2) \quad \text{as } n \to \infty.$$

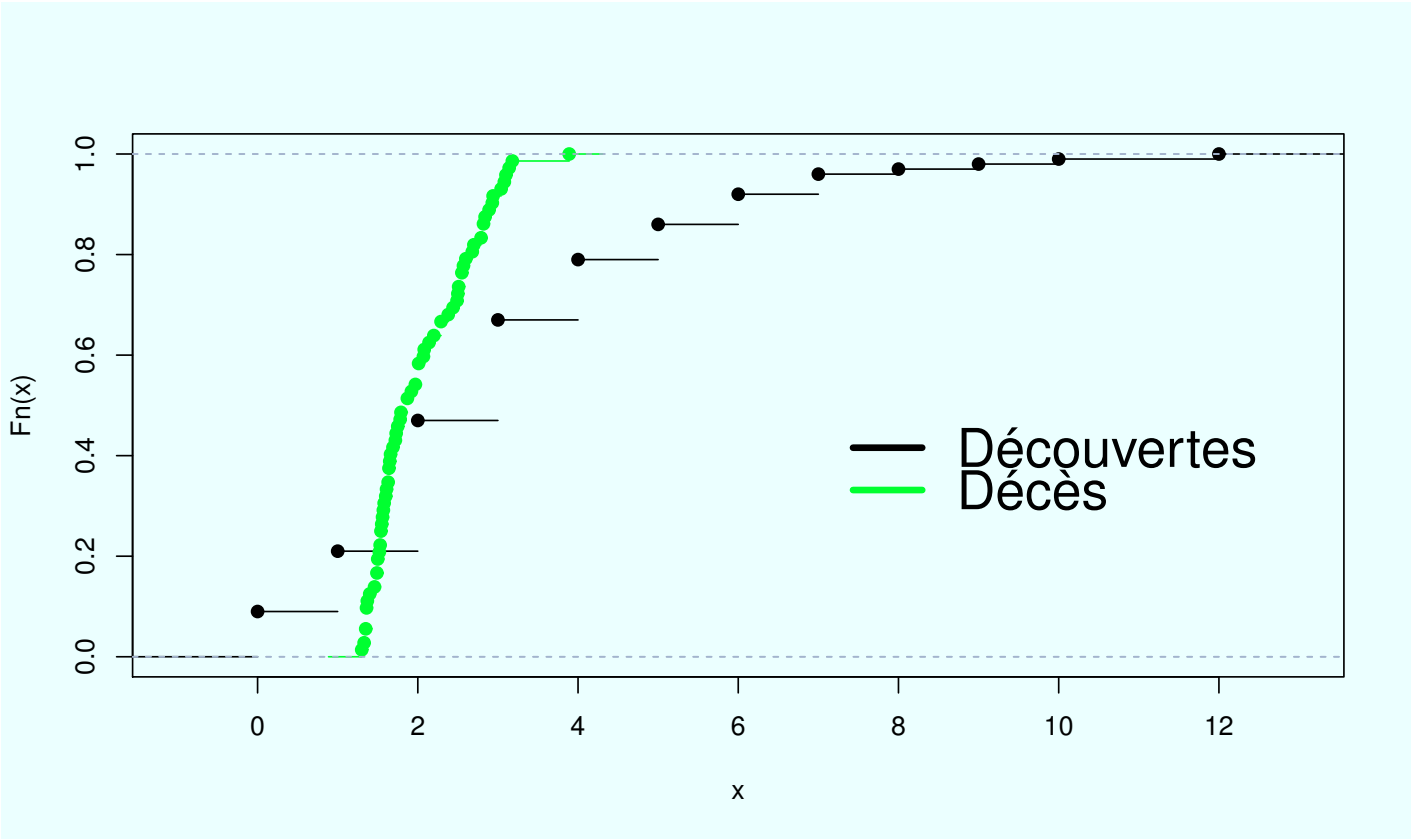
Empirical distribution function I

Definition

The **empirical cumulated distribution function** (ecdf) $\hat{F}$ associated with the sample $(x_1, \ldots, x_n)$ is defined by

$$\hat{F}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{x_i \leq t\} = \frac{\#\{i : x_i \leq t\}}{n}, \quad t \in \mathbb{R}.$$

Empirical distribution function II



Empirical distribution function $\hat{F}$ of the number of scientific discoveries (black line) and of the number of lung deaths (green line).

Empirical distribution function III

Properties

- $\hat{F}$ is a step function (i.e. non decreasing and piecewise constant) with jumps at x_i and jump height $h_i = \#\{j : x_j = x_i\}/n$ at x_i
- $\hat{F}$ is the cumulated distribution function of a discrete distribution, called the **empirical distribution** associated with $(x_1, \ldots, x_n)$. Indeed, if $Z \sim \hat{F}$, then $\mathbb{P}(Z = x_i) = h_i$ for $i = 1, \ldots, n$. That is, Z takes its values in $\{x_1, \ldots, x_n\}$.
- $\hat{F}$ is a (nonparametric) approximation of F.

Empirical distribution function IV

Theorem

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with distribution F and denote by $\hat{F}_n$ the empirical distribution function associated with $(X_1, \ldots, X_n)$.

(i) $n\hat{F}_n(t) \sim \mathrm{Bin}(n, F(t))$.

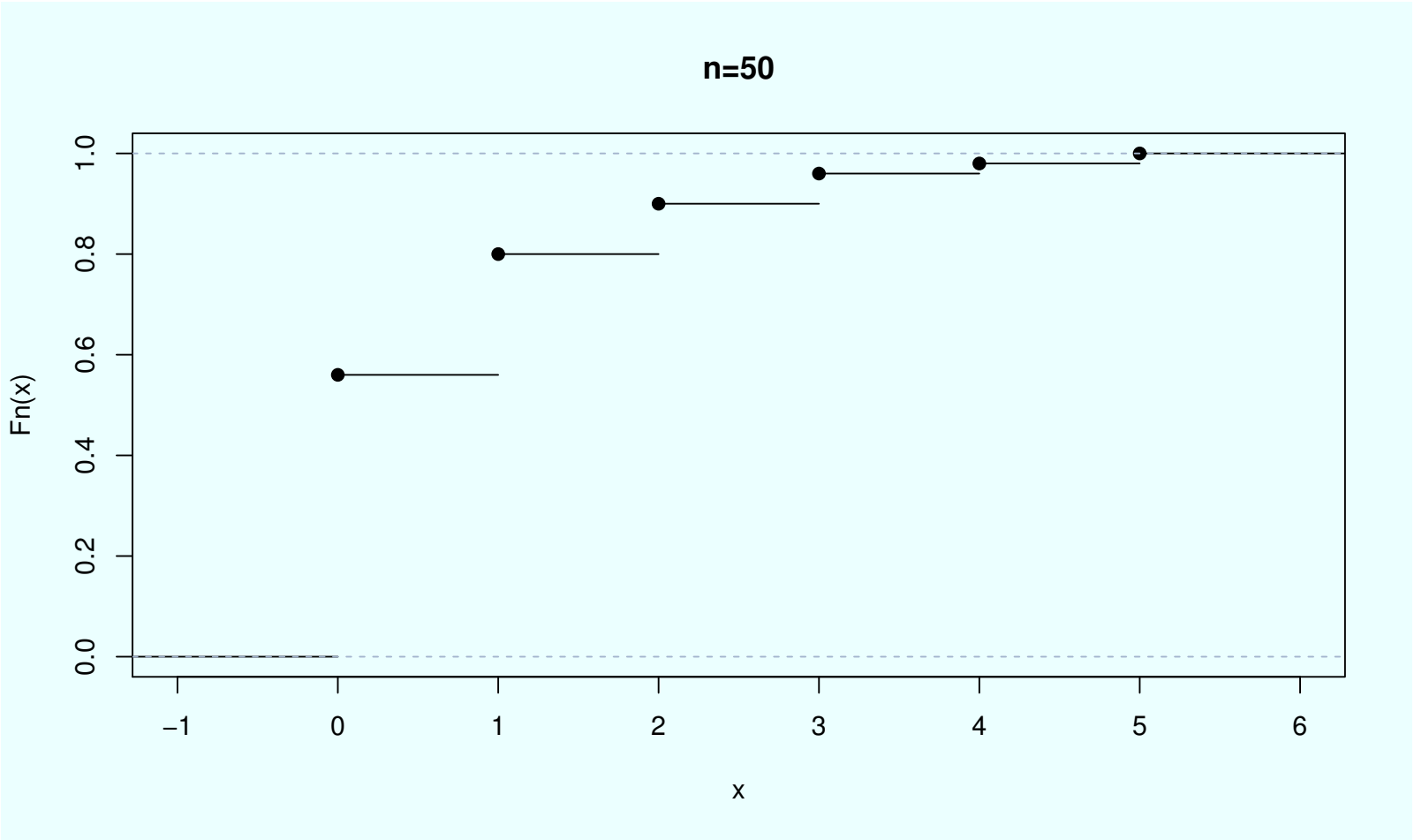
(ii) $\hat{F}_n(t) \longrightarrow F(t)$ *a.s.* when $n \to \infty$ for all $t \in \mathbb{R}$.

(iii) $\sqrt{n}(\hat{F}_n(t) - F(t)) \xrightarrow{d} \mathcal{N}(0, F(t)(1 - F(t)))$ when $n \to \infty$.

(iv) **(Glivenko-Cantelli Theorem)** $\hat{F}_n$ converges almost surely uniformly to F, that is

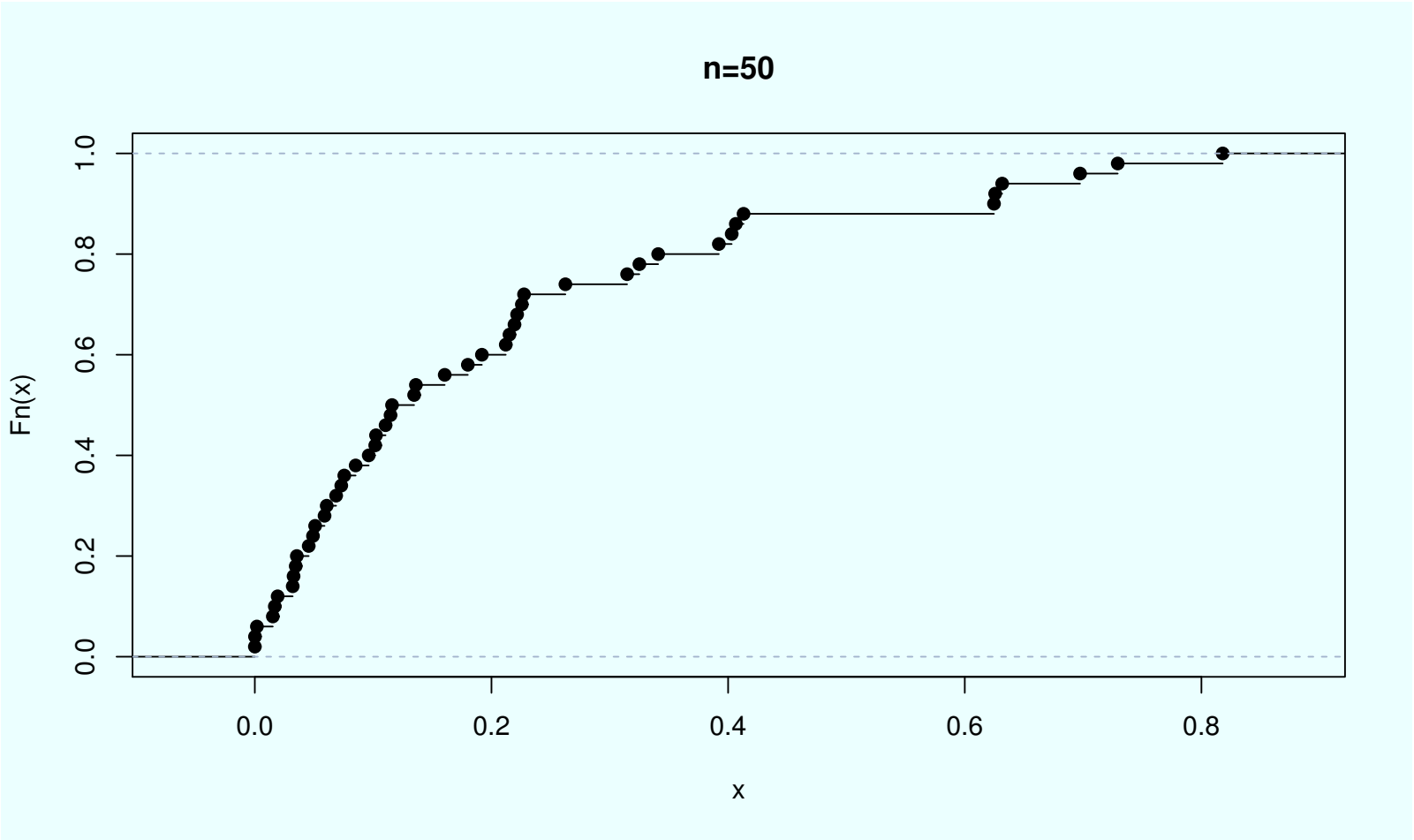
$$\|\hat{F}_n - F\|_\infty := \sup\left\{|\hat{F}_n(t) - F(t)|, t \in \mathbb{R}\right\} \longrightarrow 0 \text{ a.s.}, \quad n \to \infty .$$

Empirical distribution function V



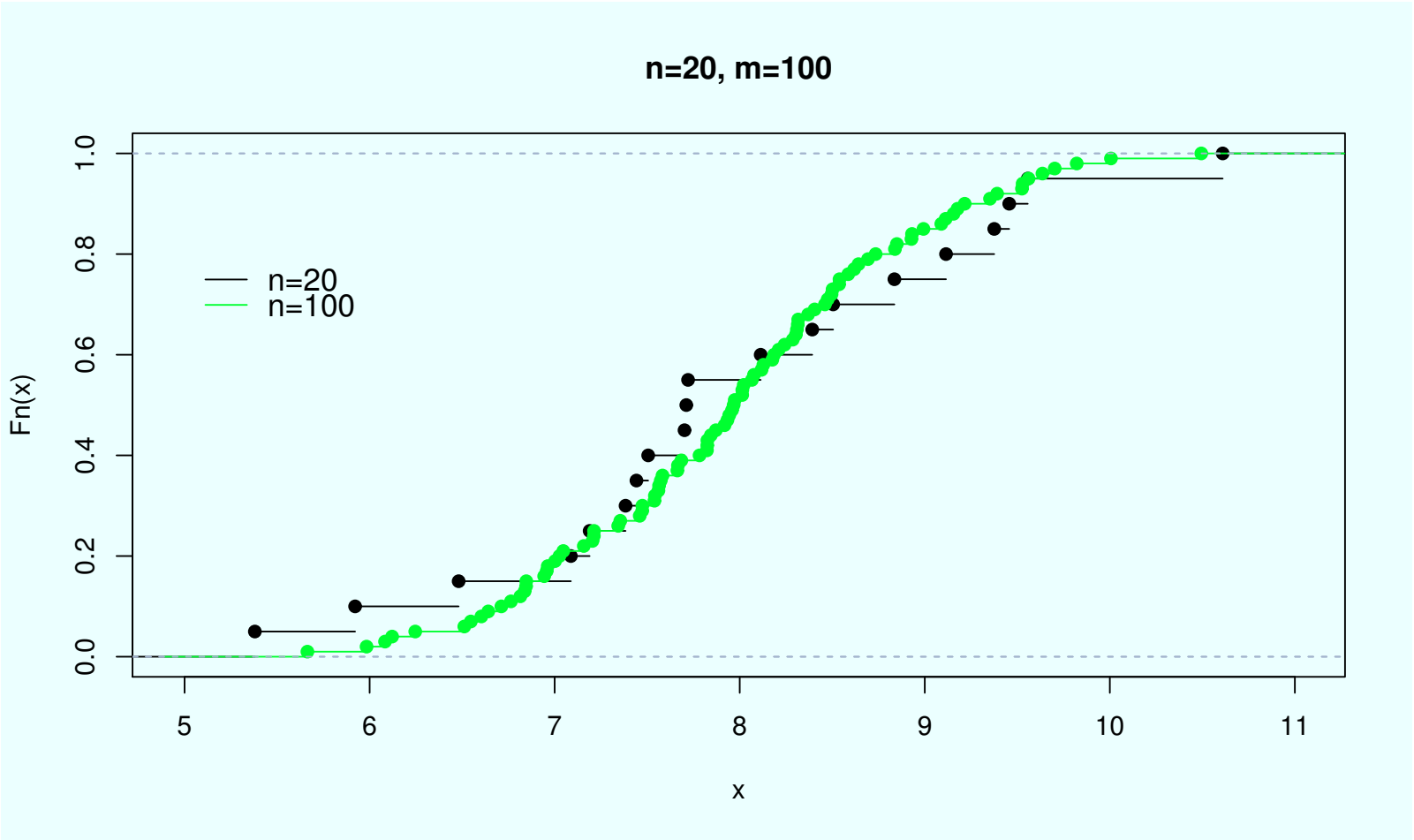
a)

Empirical distribution function VI



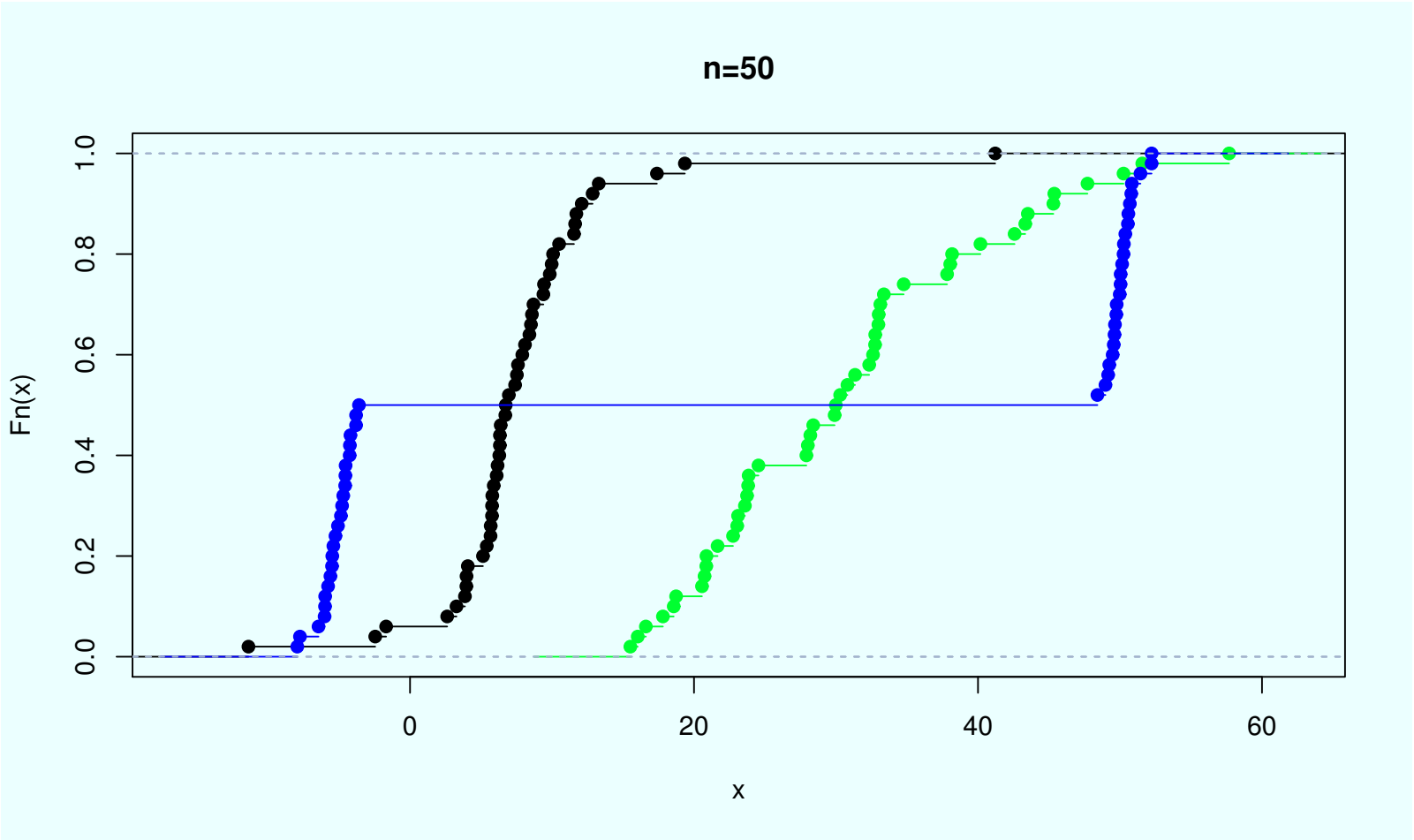
b)

Empirical distribution function VII



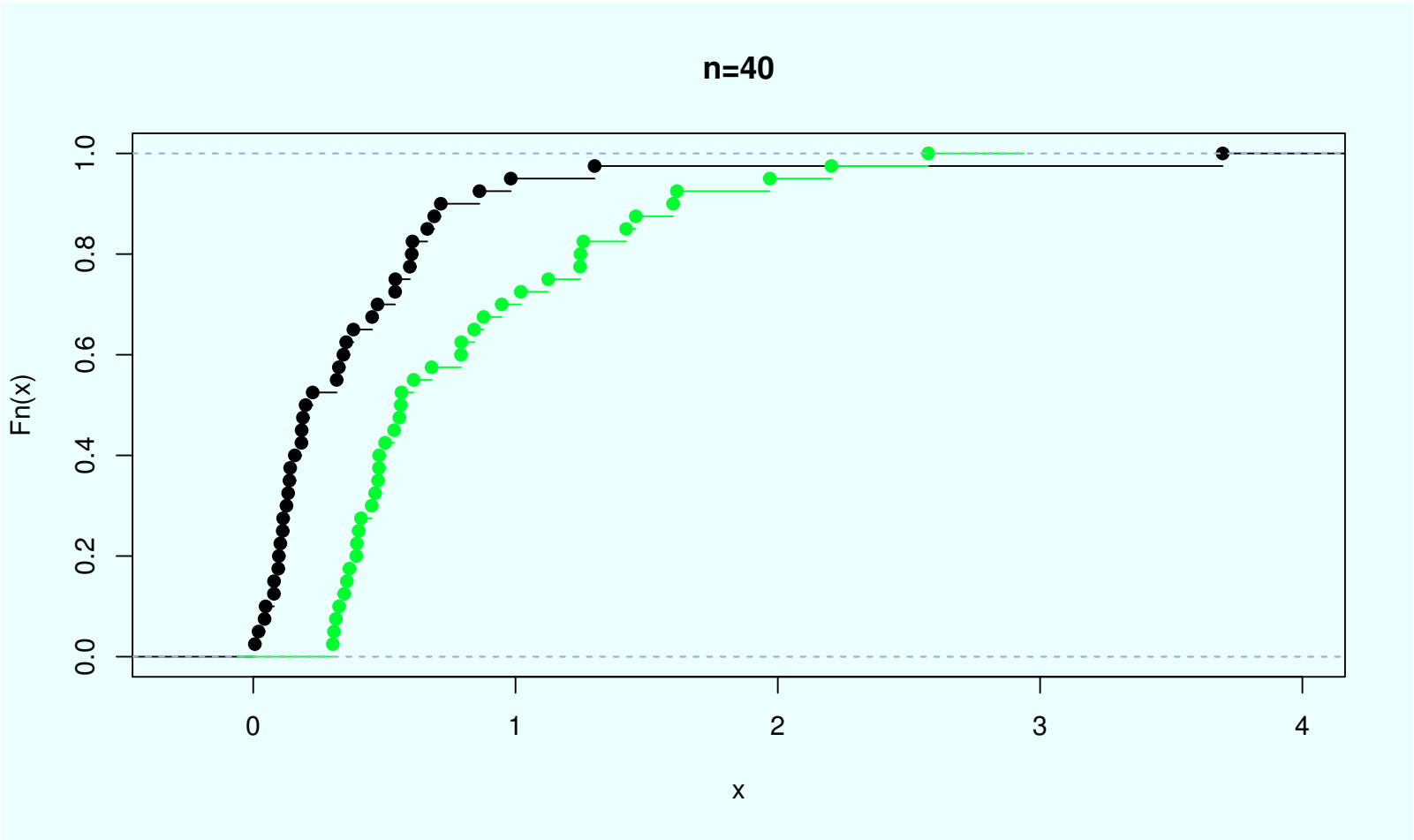
c)

Empirical distribution function VIII



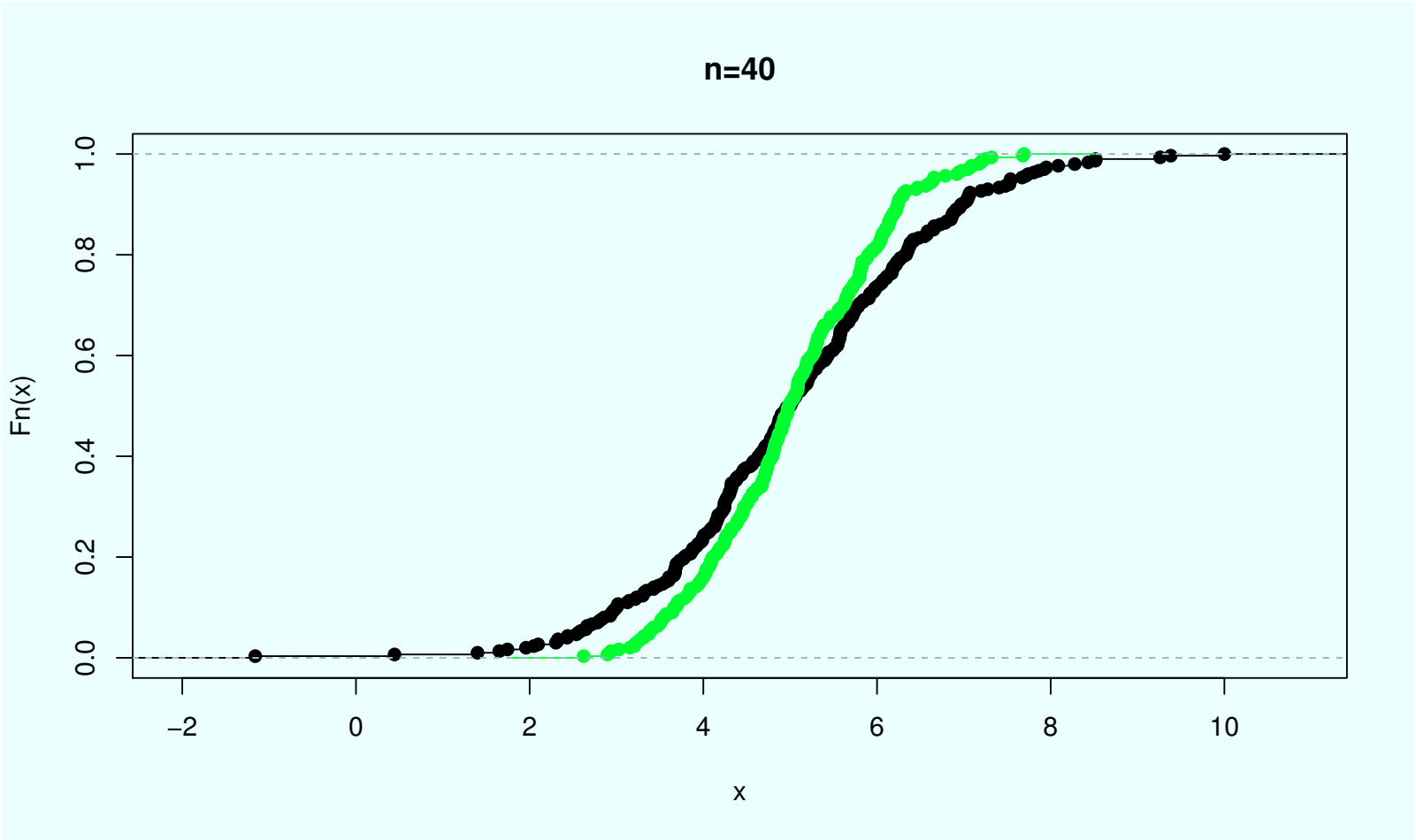
d)

Empirical distribution function IX



e)

Empirical distribution function X



f)

Empirical distribution function XI

- The ecdf may not be so useful to define a statistical model for the data (maybe only for deciding whether data come from a discrete or a continuous distribution).
- However, it can be useful to **compare** distributions (the ecdfs of two samples or one sample to a theoretical probability distribution)
- The empirical distribution is often used to define estimators (**plug-in method, method of moments**).
  Example: Approach the expectation $\mathbb{E}[X]$ of $X \sim F$ by the expectation of the empirical distribution, that is, by $\mathbb{E}[Z]$ where $Z \sim \hat{F}$, which is the sample mean, i.e. $\mathbb{E}[Z] = \bar{x}$.

Expectation I

The expectation of a distribution is the average value taken by a random variable having this distribution.

Definition

- Let X have a discrete distribution taking its values in $\{x_1, x_2, \dots\}$. If $\sum_{k\geq 1} |x_k| \mathbb{P}(X = x_k) < \infty$, the **expectation** or **mean** $\mathbb{E}[X]$ of X exists and is given by

$$\mathbb{E}[X] = \sum_{k\geq 1} x_k \mathbb{P}(X = x_k).$$

- Let X have a continuous distribution with density f. If $\int_{-\infty}^{\infty} |x| f(x) \mathrm{d}x < \infty$, the expectation $\mathbb{E}[X]$ of X exists and is given by

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x) \mathrm{d}x.$$

Expectation II

Counter-example

- The mean does not exist for all distributions.
- For instance, the **Cauchy distribution** is not integrable. Its density is given by

$$f(x) = \frac{1}{\pi(1+x^2)}.$$

Expectation III

Proposition

The expectation $\mathbb{E}[X]$ of X (if it exists) is a real number with properties

- If $X = \mathbb{1}_A$, then
$$\mathbb{E}[X] = \mathbb{E}[\mathbb{1}_A] = \mathbb{P}(A).$$
- **(Linearity)** For any real numbers a, b and any random variables X, Y
$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y].$$
- **(Monotonicity)** If $X \leq Y$ *a.s.*, i.e. $\mathbb{P}(\{\omega \text{ such that } X(\omega) \leq Y(\omega)\}) = 1$, then
$$\mathbb{E}[X] \leq \mathbb{E}[Y].$$

Expectation IV

Proposition

Let X be a random variable and $\phi : \mathbb{R} \to \mathbb{R}$ be a function. Set $Y = g(X)$.

- If X is discrete and $\mathbb{E}[Y]$ exists, then

$$\mathbb{E}[Y] = \mathbb{E}[\phi(X)] = \sum_{k \geq 1} \phi(x_k)\mathbb{P}(X = x_k).$$

- If X is continuous with density f and $\mathbb{E}[Y]$ exists, then

$$\mathbb{E}[Y] = \mathbb{E}[\phi(X)] = \int_{\mathbb{R}} \phi(x) f(x) \mathrm{d}x.$$

Variance I

The variance of a random variable is a measure of its dispersion around its mean.

Definition

Let X be a random variable such that $\mathbb{E}[X^2] < +\infty$. The **variance** of X is defined by

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

The **standard deviation** is defined as $\sigma = \sqrt{\mathrm{Var}(X)}$.

Variance II

Proposition

Let X be a random variable such that $\mathbb{E}[X^2] < +\infty$. Then

(i) $0 \leq \mathrm{Var}(X) < \infty$

(ii) $\mathrm{Var}(X) = 0 \iff \mathbb{P}(X = c) = 1$ for some constant c.

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

(iv) For any constants a, b, $\mathrm{Var}(aX + b) = \mathrm{Var}(aX) = a^2\mathrm{Var}(X)$.

The quantity $\mathbb{E}[X^2]$ is called the second moment of X, and $\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$ the second central moment.

Higher-order moments

Definition

Let X be a random variable such that $\mathbb{E}[|X|^k] < +\infty$ for some $k \in \mathbb{N}$.

- The ***k*-th moment** of X exists and is given by $\mathbb{E}[X^k]$.
- The ***k*-th central moment** of X exists and is given by

$$\mathbb{E}\left[(X - \mathbb{E}[X])^k\right].$$

Moments play an important role in statistics.

Characteristic function I

Definition

The **characteristic function** of a random variable X is the function $\Phi_X : \mathbb{R} \to \mathbb{C}$ defined by

$$\Phi_X(t) = \mathbb{E}[e^{itX}] \quad t \in \mathbb{R}.$$

Theorem

X and Y have the same law $\iff \Phi_X(t) = \Phi_Y(t)$ for all t.

Characteristic function II

Proposition

Let the distribution of X be such that $\mathbb{E}[|X|] < +\infty$. Then

$$\Phi'_X(t) = \frac{\partial}{\partial t}\mathbb{E}\left[e^{itX}\right] = \mathbb{E}\left[\frac{\partial}{\partial t}e^{itX}\right] = \mathbb{E}[iXe^{itX}].$$

In particular, $\Phi'_X(0) = i\mathbb{E}[X]$.

Quantiles I

Definition

- The **quantile function** $F^{-1} : (0,1) \to \mathbb{R}$ associated with some cdf F is defined as

$$F^{-1}(\alpha) = \inf\{t : F(t) \geq \alpha\}, \quad \alpha \in (0,1).$$

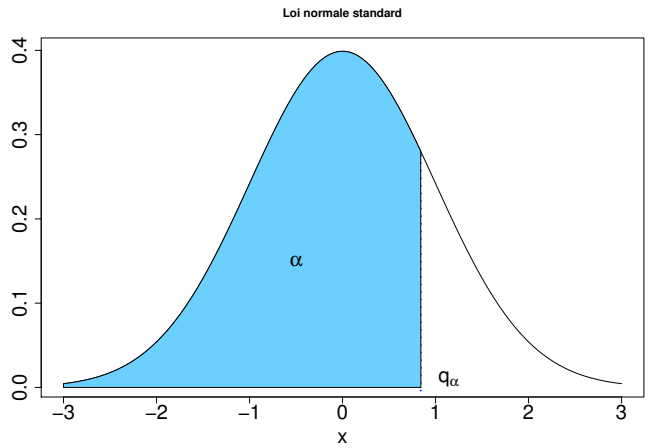
- The α-**quantile** of F is the value

$$q_\alpha = q_\alpha^F = F^{-1}(\alpha).$$

Quantiles II

- If $F : \mathbb{R} \to (0, 1)$ is a bijection, then F^{-1} is the traditional inverse function of F. However, in general $F \circ F^{-1} \neq \mathrm{Id}$.
- If F has a density f, then quantiles are characterised by the area under the density, since

$$\int_{-\infty}^{q_\alpha} f(x)\mathrm{d}x = F(q_\alpha) = F(F^{-1}(\alpha)) = \alpha.$$



Quantiles III

Definition

Let $X_1, \ldots, X_n$ be an i.i.d. sample from distribution F. Denote $\hat{F}$ the associated ecdf. Then the **α-sample quantile** $\hat{q}_\alpha$ is defined as

$$\hat{q}_\alpha = \hat{F}^{-1}(\alpha).$$

Quantiles IV

Definition

For a sample $X_1, \ldots, X_n$ we define the **order statistics** $X_{(1)}, \ldots, X_{(n)}$ by ordering the observations $X_1, \ldots, X_n$ such that

$$X_{(1)} \leq \cdots \leq X_{(n)} \quad \text{and} \quad X_{(i)} \in \{X_1, \ldots, X_n\}.$$

That is, $X_{(1)} = \min\{X_1, \ldots, X_n\}$ and $X_{(n)} = \max\{X_1, \ldots, X_n\}$

Theorem

We have

$$\hat{q}_\alpha = X_{(\lceil \alpha n \rceil)},$$

where $\lceil a \rceil$ denotes the smallest integer that is larger or equal a.

Quantiles V

Theorem

Let $X_i \sim F$ be i.i.d.. If F is strictly increasing in q_α^F, then

$$\hat{q}_\alpha \xrightarrow{P} q_\alpha^F \quad (n \to \infty).$$

Thus, the sample quantiles are consistent estimators of the theoretical quantiles.

Central tendency

The **central tendency** or **location** of a probability distribution F may be measured by

- the mean $\mathbb{E}[X]$ with $X \sim F$
- the median $q^F_{1/2}$.

From an i.i.d. sample $X_1, \ldots, X_n$ with distribution F, the central tendency of F may estimated by

- the **sample mean** $\bar{X}$ ($\xrightarrow{P} \mathbb{E}[X]$ if $\mathbb{E}[|X|] < \infty$)
- the **sample median** $q^{\hat{F}}_{1/2} = X_{(\lceil n/2 \rceil)}$ ($\xrightarrow{P} q^F_{1/2}$ if F is strictly increasing at $q^F_{1/2}$)

If F is symmetric, then $\mathbb{E}[X] = q^{\hat{F}}_{1/2}$. However, in general $\mathbb{E}[X] \neq q^{\hat{F}}_{1/2}$.

Dispersion

The **dispersion** or **variability** of a a probability distribution F may be measured by

- the variance $\mathrm{Var}(X)$
- the standard deviation $\sigma = \sqrt{\mathrm{Var}(X)}$
- the interquartile range $IQR = q^F_{3/4} - q^F_{1/4}$

Based on $X_1, \ldots, X_n \overset{i.i.d.}{\sim} F$, the dispersion of F may estimated by

- the **sample variance** $s_X^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$ ($\xrightarrow{P} \mathrm{Var}(X)$ if $\mathbb{E}[X^2] < \infty$)
- the **sample standard deviation** $\sigma = s_X$ ($\xrightarrow{P} \sigma$ if $\mathbb{E}[X^2] < \infty$)
- the **sample interquartile range** $IQR = q^{\hat{F}_n}_{3/4} - q^{\hat{F}_n}_{1/4}$ ($\xrightarrow{P} q^F_{3/4} - q^F_{1/4}$ if F is strictly increasing at $q^F_{1/4}$ and $q^F_{3/4}$)
- the **range** $X_{(n)} - X_{(1)}$ ($\xrightarrow{P} b - a$ if the support of F is $[a, b]$; $\xrightarrow{P} \infty$ if the support of F is not bounded)

Asymmetry I

Definition

- F is called **symmetric** (with respect to 0) if and only if $F(x) = 1 - F(x)$ for all $x \in \mathbb{R}$.
- F is called **symmetric with respect to** μ if and only if $F(\mu + x) = 1 - F(\mu - x)$ for all $x \in \mathbb{R}$.

Proposition

If F is symmetric and $\mathbb{E}[|X|^m] < \infty$ for $X \sim F$, then for all **odd** integer $r \leq m$ we have

$$\mathbb{E}[X^r] = 0.$$

Asymmetry II

Definition

Let $E[|X|^3] < \infty$ for $X \sim F$. We define the **skewness** of F by

$$\alpha_X = \frac{\mathbb{E}[(X - \mathbb{E}[X])^3]}{(\mathrm{Var}(X))^{3/2}}.$$

Proposition

If $\alpha_X \neq 0$, then F is not symmetric.

The skewness is a measure of asymmetry.

Asymmetry III

Proposition (Invariance to affine tranformations)

For all $a > 0$ and $b \in \mathbb{R}$,

$$\alpha_{aX+b} = \alpha_X.$$

For all $a < 0$ and $b \in \mathbb{R}$,

$$\alpha_{aX+b} = -\alpha_X.$$

Asymmetry IV

Definition

The **empirical skewness** associated to an i.i.d. sample $X_1, \ldots, X_n \sim F$, is defined by

$$\hat{\alpha}_n = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3}{s_X^3}.$$

$\hat{\alpha}_n$ is the skewness of the empirical distribution $\hat{F}$.

If $\mathbb{E}[|X|^3] < \infty$, then

$$\hat{\alpha}_n \xrightarrow{P} \alpha_X.$$

Kurtosis I

Definition

Let $E[X^4] < \infty$ for $X \sim F$. We define the **kurtosis** of F by

$$\beta_X = \frac{\mathbb{E}[(X - \mathbb{E}[X])^4}{(\mathrm{Var}(X))^2} - 3.$$

Proposition

1. If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\beta_X = 0$.
2. $\beta_X \geq -2$ for any distribution F of X.
3. (Invariance to affine tranformations) For all $a, b \in \mathbb{R}$,

$$\beta_{aX+b} = \beta_X.$$

Kurtosis II

Interpretation of the kurtosis

- $\beta_X > 0$: The distribution F has **heavy tails**, i.e. the probability $\mathbb{P}(|X| > b)$ decreases when $b \to \infty$ more slowly than for a normal distribution. (The probability to observe extreme values is quite elevated).
  Example: Student's t-distribution.
- $\beta_X < 0$: F has **light tails**, i.e. the probability $\mathbb{P}(|X| > b)$ decreases when $b \to \infty$ faster than for a normal distribution.
  Example: uniform distribution.

Kurtosis III

Definition

The **empirical kurtosis** associated to an i.i.d. sample $X_1, \ldots, X_n \sim F$, is defined by

$$\hat{\beta}_n = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^4}{s_X^4} - 3.$$

β_n is the kurtosis of the empirical distribution $\hat{F}$.

If $\mathbb{E}[X^4] < \infty$, then

$$\hat{\beta}_n \xrightarrow{P} \beta_X.$$

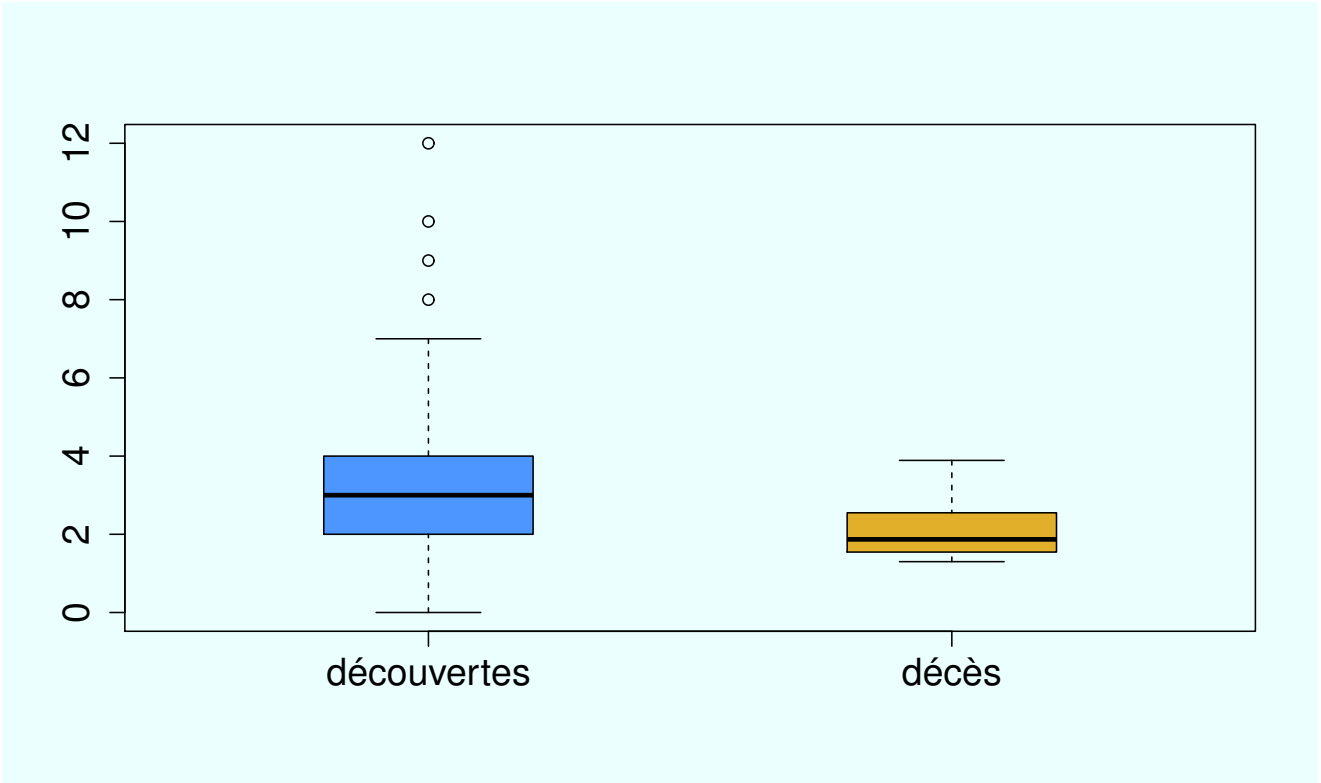
Boxplot I

Boxplot

The **boxplot** is a graphical representation of several summary statistics of a sample $x_1, \ldots, x_n$, namely the sample median, sample quartiles, sample interquartile range, outliers...

- The length of the whiskers shall not exceed $\frac{3}{2}IQR$.
- If there are data points beyond that distance, they are represented by isolated points and they are called **outliers**.
- Boxplot are useful to compare several datasets.

Boxplot II



Boxplot for the two data examples on scientific discoveries and the lung deaths.

Comparison of distributions and QQ-plot I

- Let $(X_1, \ldots, X_n)$ be an i.i.d. sample with distribution F.
- Let F_0 be a given distribution.
- Question: Does $F = F_0$ hold?
- Draw the QQ-plot to answer.

QQ-plot for a sample and a distribution F_0

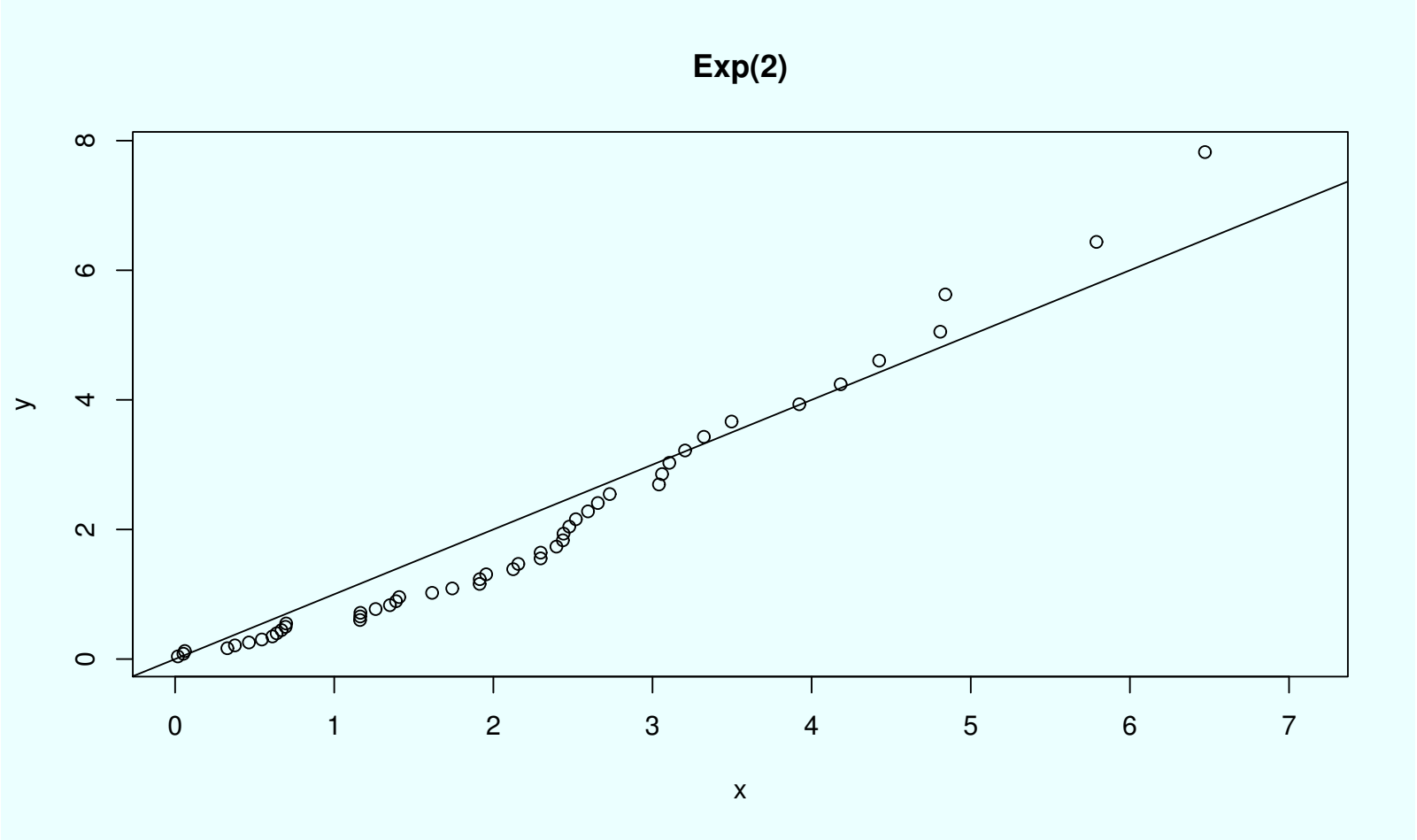
The **quantile-quantile diagram** or **QQ-plot** is defined as the scatter plot of the points

$$(\hat{q}_{j/n}, q_{j/n}^{F_0}), \qquad j = 1, \ldots, n,$$

where $q_\alpha^{F_0}$ denotes the (theoretical) α-quantile of F_0 and $\hat{q}_\alpha$ the corresponding sample quantile associated with $(X_1, \ldots, X_n)$.

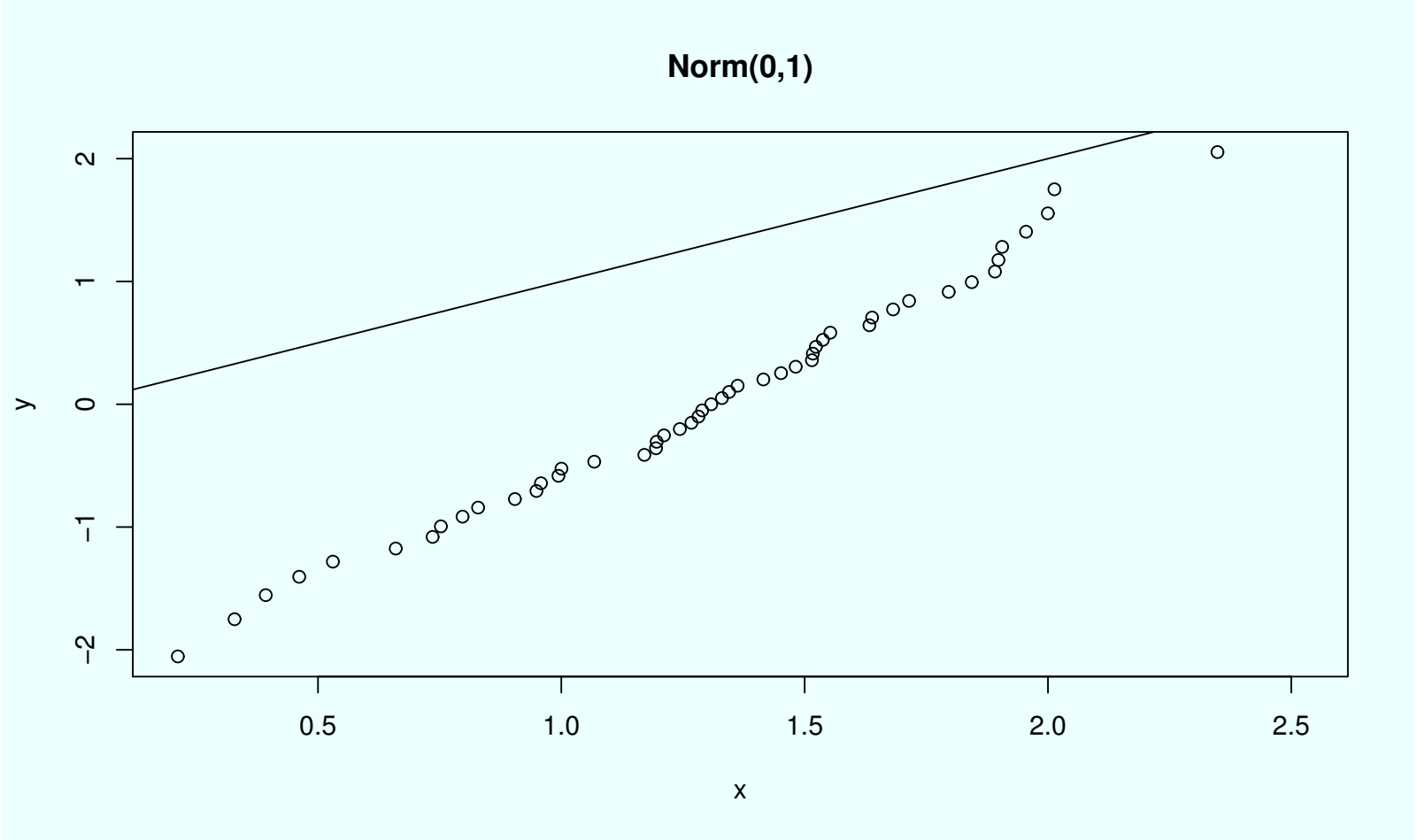
Notice that $\hat{q}_{j/n} = X_{(j)}, j = 1, \ldots, n$.

Comparison of distributions and QQ-plot II



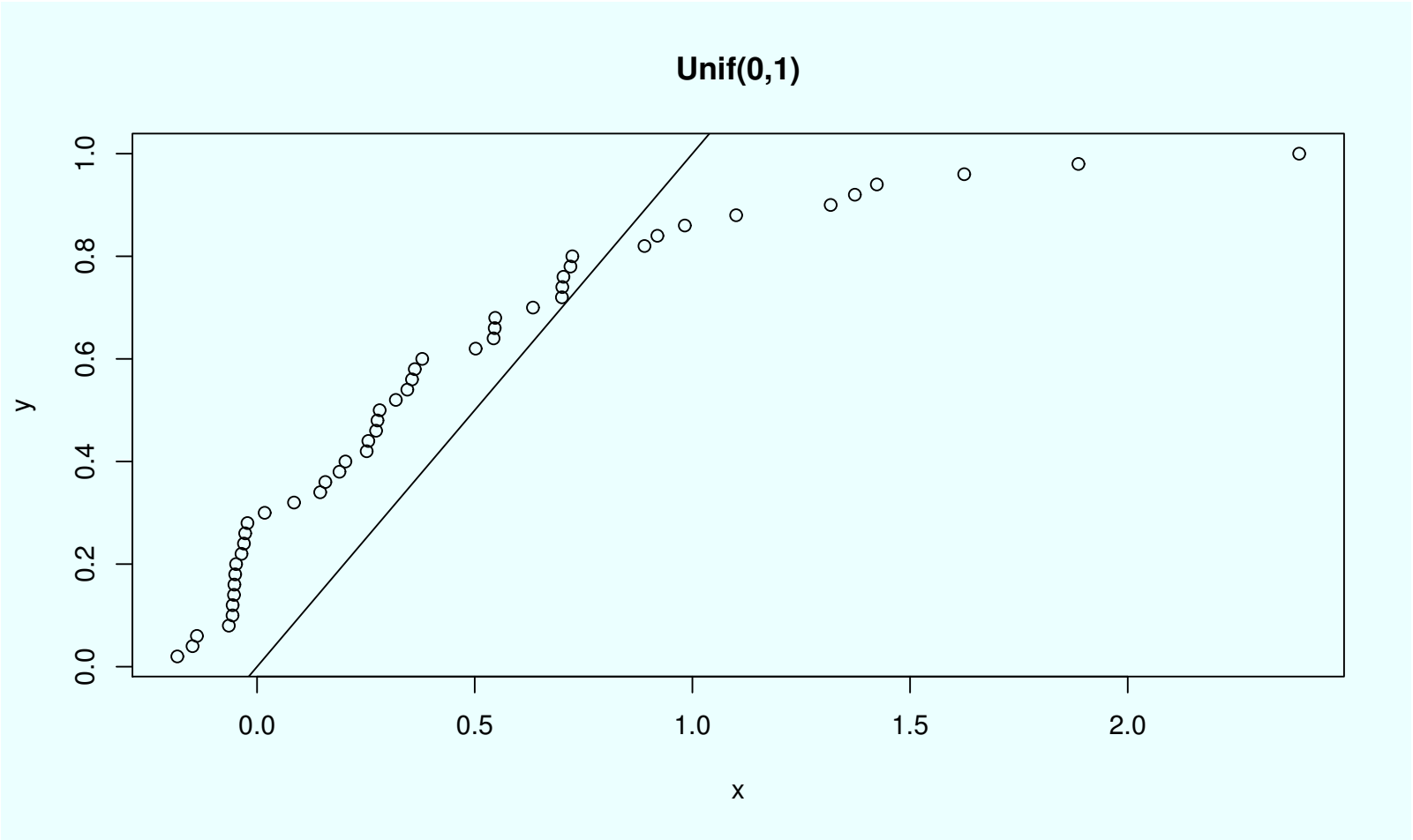
a)

Comparison of distributions and QQ-plot III



b)

Comparison of distributions and QQ-plot IV



c)

Comparison of distributions and QQ-plot V

Interpretation of the QQ-plot

- If the points of the QQ-plot are aligned on the line $x = y$, then F may be equal to F_0.
- If the points of the QQ-plot are on a straight line different from $x = y$, then F be may obtained by a translation-dilatation of F_0, i.e. $F = F_0((\cdot - \delta)/\sigma)$ for some constants $\delta \in \mathbb{R}$ and $\sigma > 0$

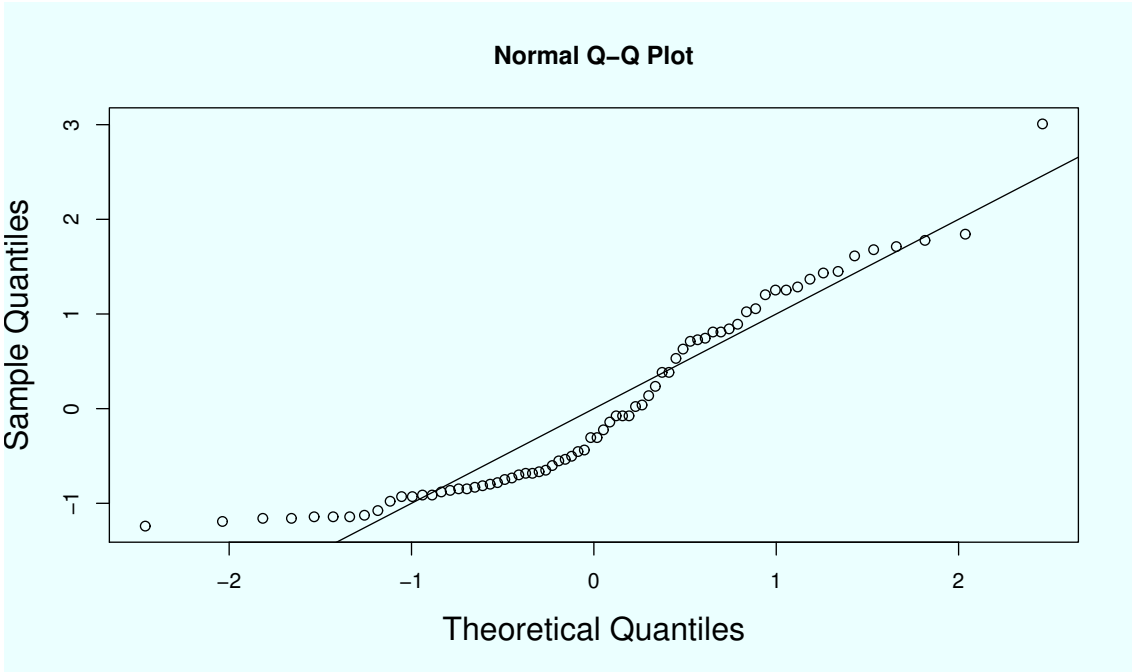
Comparison of distributions and QQ-plot VI

- In many applications we want to know whether the data come from a normal distribution $\mathcal{N}(\mu, \sigma^2)$ or not.
- To draw the QQ-plot, it is common to **standardize** the data by

$$\tilde{X}_i = \frac{X_i - \bar{X}_n}{s_{\mathbf{x}}}, \qquad i = 1, \ldots, n,$$

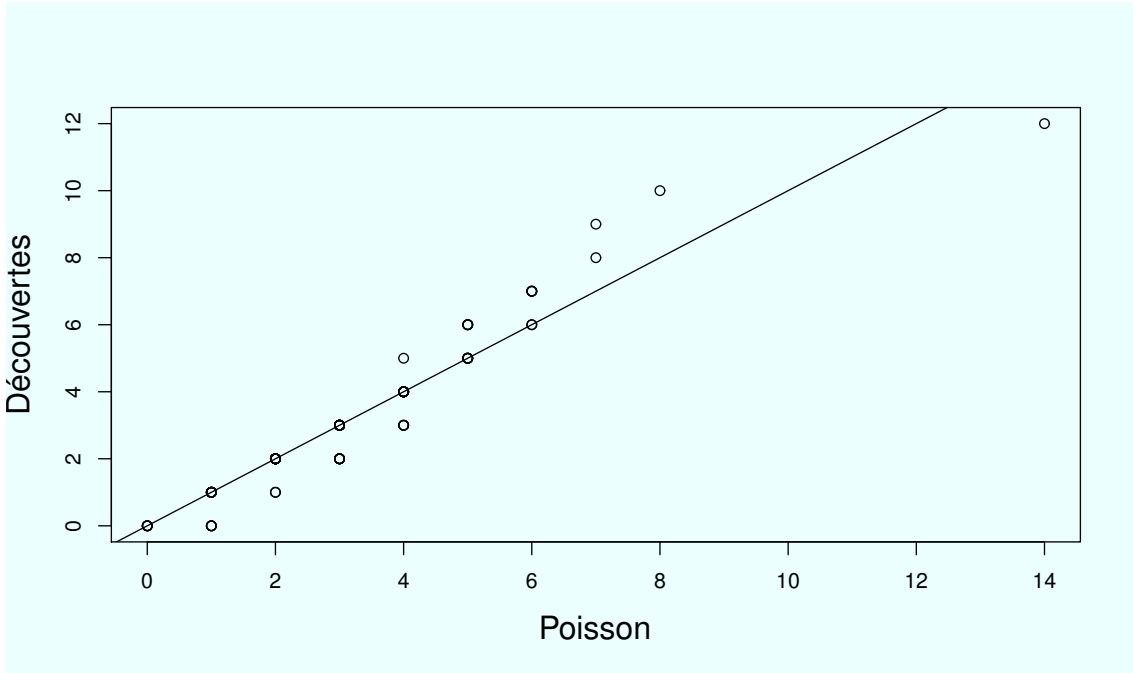
  and set $F_0 = \mathcal{N}(0, 1)$.
- Interpretation of the QQ-plot: if the points align on the line $x = y$, then F may be a normal distribution $\mathcal{N}(\mu, \sigma^2)$.

Comparison of distributions and QQ-plot VII



QQ-plot to compare the standardised data of the lung deaths to the standard normal distribution.

Comparison of distributions and QQ-plot VIII



QQ-plot to compare the scientific discovery data to the Poisson distribution Poi(3,1).

Comparison of distributions and QQ-plot IX

- Let $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_m)$ be to independant i.i.d. samples wit distribution F and G, respectively.
- To analyze the relation of F and G, draw the QQ-plot of the sample quantiles associated with both samples.

QQ-plot for two samples

The **QQ-plot** is defined as the scatter plot of the points

$$(\hat{q}^x_{k/r}, \hat{q}^y_{k/r}), \qquad k = 1, \ldots, r,$$

where $r = \min\{n, m\}$.

If $n = m$, then

$$(\hat{q}^x_{k/n}, \hat{q}^y_{k/n}) = (X_{(k)}, Y_{(k)}), \qquad k = 1, \ldots, n.$$

Comparison of distributions and QQ-plot X

Interpretation of the QQ-plot

- If the points are aligned on the line $x = y$, then $F \approx G$.
- If the points are aligned on a straight line different from $x = y$, then F is obtained from G be a linear transformation, i.e. $F \approx G((\cdot - \delta)/\sigma)$ for some constants $\delta \in \mathbb{R}$ and $\sigma > 0$.