Chapter 2: Descriptive statistics

Tabea Rebafka

October 2018

Master AIMS 2018–19
Outline

1. Type of distribution
2. Bar chart
3. Histogram
4. Convergence of sequences of random variables
5. Empirical distribution function
6. Expectation and moments
7. Quantiles
8. Summary statistics
9. Boxplot
10. QQ-plot
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 in the UK from 1974 to 1979 (in thousands).

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

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

<table>
<thead>
<tr>
<th>Year</th>
<th>1860</th>
<th>1861</th>
<th>1862</th>
<th>1863</th>
<th>1864</th>
<th>1865</th>
<th>1866</th>
<th>1867</th>
<th>1868</th>
<th>1869</th>
</tr>
</thead>
<tbody>
<tr>
<td>1860</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>1861</td>
<td>1870</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>1862</td>
<td>1871</td>
<td>1872</td>
<td>1873</td>
<td>1874</td>
<td>1875</td>
<td>1876</td>
<td>1877</td>
<td>1878</td>
<td>1879</td>
<td></td>
</tr>
<tr>
<td>1863</td>
<td>1880</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>1864</td>
<td>1881</td>
<td>1882</td>
<td>1883</td>
<td>1884</td>
<td>1885</td>
<td>1886</td>
<td>1887</td>
<td>1888</td>
<td>1889</td>
<td></td>
</tr>
<tr>
<td>1865</td>
<td>1890</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>7</td>
<td>12</td>
<td>3</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>1866</td>
<td>1891</td>
<td>1892</td>
<td>1893</td>
<td>1894</td>
<td>1895</td>
<td>1896</td>
<td>1897</td>
<td>1898</td>
<td>1899</td>
<td></td>
</tr>
<tr>
<td>1867</td>
<td>1900</td>
<td>3</td>
<td>7</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>2</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>1868</td>
<td>1901</td>
<td>1902</td>
<td>1903</td>
<td>1904</td>
<td>1905</td>
<td>1906</td>
<td>1907</td>
<td>1908</td>
<td>1909</td>
<td></td>
</tr>
<tr>
<td>1869</td>
<td>1910</td>
<td>5</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>1870</td>
<td>1911</td>
<td>1912</td>
<td>1913</td>
<td>1914</td>
<td>1915</td>
<td>1916</td>
<td>1917</td>
<td>1918</td>
<td>1919</td>
<td></td>
</tr>
<tr>
<td>1871</td>
<td>1920</td>
<td>3</td>
<td>6</td>
<td>5</td>
<td>8</td>
<td>3</td>
<td>6</td>
<td>0</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>1872</td>
<td>1921</td>
<td>1922</td>
<td>1923</td>
<td>1924</td>
<td>1925</td>
<td>1926</td>
<td>1927</td>
<td>1928</td>
<td>1929</td>
<td></td>
</tr>
<tr>
<td>1873</td>
<td>1930</td>
<td>2</td>
<td>6</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>1874</td>
<td>1931</td>
<td>1932</td>
<td>1933</td>
<td>1934</td>
<td>1935</td>
<td>1936</td>
<td>1937</td>
<td>1938</td>
<td>1939</td>
<td></td>
</tr>
<tr>
<td>1875</td>
<td>1940</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>1876</td>
<td>1941</td>
<td>1942</td>
<td>1943</td>
<td>1944</td>
<td>1945</td>
<td>1946</td>
<td>1947</td>
<td>1948</td>
<td>1949</td>
<td></td>
</tr>
<tr>
<td>1877</td>
<td>1950</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>1879</td>
<td>1952</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>
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?
Proposition

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

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

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 $X \sim F$. 

Tabea Rebafka
Statistics with R
Descriptive statistics 10 / 66
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

$$A_j = [a + (j - 1)h, a + jh], \quad j = 1, \ldots, m \quad \text{with} \quad h = \frac{b - a}{m}.$$ 
3. Count the number of data points per subinterval:

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

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

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

$$= \frac{1}{nh} \sum_{j=1}^{m} N_j \mathbb{1} \{x \in A_j\}, \quad x \in \mathbb{R}.$$
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 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 \(\varepsilon > 0\)
  \[
  \lim_{n \to +\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0.
  \]

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

\[ a.s. \implies \mathbb{P} \implies d \]
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

\[
\bar{X}_n \xrightarrow{} \mathbb{E}[X_1] \text{ a.s. as } n \to \infty.
\]
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 \(\sigma^2 = \text{Var}(X_1)\). Then

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

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

Tabea Rebafka

Statistics with R

Descriptive statistics 21 / 66
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 \( F \).
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) $n\hat{F}_n(t) \sim \text{Bin}(n, F(t))$.

(ii) $\hat{F}_n(t) \xrightarrow{} F(t)$ 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

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

\( n = 50 \)

\( F_n(x) \)

\( x \)

a)
Empirical distribution function VI

\[ F_n(x) \]

\[ n=50 \]

\[ x \]

b)
Empirical distribution function VII

n=20, m=100

$c$)
Empirical distribution function VIII

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure}
\caption{Empirical distribution function for a dataset of size 50.}
\end{figure}

\textbf{d)}
Empirical distribution function IX

\[ \text{e) } \]

\[ n=40 \]

\[ \begin{array}{c}
\text{\( F_n(x) \)} \\
\hline
0.0 & 0.2 & 0.4 & 0.6 & 0.8 & 1.0 \\
\hline
\end{array} \]

\[ \text{\( x \)} \]
Empirical distribution function $X$
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 $\{x_1, x_2, \ldots \}$. 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 $\int_{-\infty}^{\infty} |x| f(x) \, dx < \infty$, the expectation $\mathbb{E}[X]$ of $X$ exists and is given by

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

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

\[ f(x) = \frac{1}{\pi(1 + x^2)}. \]
The expectation $\mathbb{E}[X]$ of $X$ (if it exists) is a real number with properties:

- If $X = 1_A$, then
  \[
  \mathbb{E}[X] = \mathbb{E}[1_A] = \mathbb{P}(A).
  \]

- **(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
  \[
  \mathbb{E}[X] \leq \mathbb{E}[Y].
  \]
Let $X$ be a random variable and $\phi : \mathbb{R} \rightarrow \mathbb{R}$ 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
  \[
  \mathbb{E}[Y] = \mathbb{E}[\phi(X)] = \int_{\mathbb{R}} \phi(x) f(x) \, dx.
  \]
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

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

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

Tabea Rebafka
Statistics with R
Descriptive statistics 36 / 66
**Proposition**

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

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

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

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

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

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

Moments play an important role in statistics.
Characteristics function I

**Definition**

The characteristic function of a random variable $X$ is the function $\Phi_X : \mathbb{R} \rightarrow \mathbb{C}$ 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

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

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

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

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

- The **$\alpha$-quantile** of $F$ is the value

  $$q_\alpha = q^F_\alpha = F^{-1}(\alpha).$$
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 $F \circ F^{-1} \neq \text{Id}$.

- If $F$ has a density $f$, then quantiles are characterised by the area under the density, since

$$\int_{-\infty}^{q_\alpha} f(x) \, dx = F(q_\alpha) = F(F^{-1}(\alpha)) = \alpha.$$
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 $\alpha$-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 $a$. 
Theorem

Let \( X_i \sim F \) be i.i.d.. If \( F \) is strictly increasing in \( q^{F}_{\alpha} \), then

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

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 \( q_{1/2}^F \).

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** \( \hat{q}_{1/2}^F = X(\lceil n/2 \rceil) \) \( \xrightarrow{P} q_{1/2}^F \) if \( F \) is strictly increasing at \( q_{1/2}^F \)

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

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

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

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

- the **sample variance** $s_X^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \xrightarrow{P} \text{Var}(X)$ 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** $IQR = \hat{q}_{3/4}^n - \hat{q}_{1/4}^n \xrightarrow{P} q_{3/4}^F - q_{1/4}^F$ if $F$ is strictly increasing at $q_{1/4}^F$ and $q_{3/4}^F$
- 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 $\mu$** 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

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

**Proposition**

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

The skewness is a measure of asymmetry.
Asymmetry III

Proposition (Invariance to affine transformations)

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

$$\alpha aX + b = \alpha X.$$  

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

$$\alpha aX + b = -\alpha X.$$
Definition

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

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

$\hat{\alpha}_n$ 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

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

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 transformations) For all $a, b \in \mathbb{R}$,

$$
\beta_{aX+b} = \beta_X.
$$
### Interpretation of the kurtosis

- $\beta_X > 0$: The distribution $F$ has **heavy tails**, i.e. the probability $P(|X| > b)$ 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 $P(|X| > b)$ 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

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

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

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

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

\[
(\hat{q}_{j/n}, q_{F_0}^{F_0}), \quad j = 1, \ldots, n,
\]

where \(q_{F_0}^{F_0}\) denotes the (theoretical) \(\alpha\)-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

Exp(2)

a)
Comparison of distributions and QQ-plot III

![Diagram of a QQ-plot for Norm(0,1)]

b)
Comparison of distributions and QQ-plot IV

c)
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 $\sigma > 0$.
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

$$\tilde{X}_i = \frac{X_i - \bar{X}_n}{s_x} , \quad i = 1, \ldots, n,$$

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 $\mathcal{N}(\mu, \sigma^2)$. 
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 $\text{Poi}(3,1)$. 
Comparison of distributions and QQ-plot IX

- Let \((X_1, \ldots, X_n)\) and \((Y_1, \ldots, Y_m)\) be independent i.i.d. samples with 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

\[
(\hat{q}_x^{k/r}, \hat{q}_y^{k/r}), \quad k = 1, \ldots, r,
\]

where \(r = \min\{n, m\}\).

If \(n = m\), then

\[
(\hat{q}_x^{k/n}, \hat{q}_y^{k/n}) = (X_{(k)}, Y_{(k)}), \quad k = 1, \ldots, n.
\]
Comparison of distributions and QQ-plot

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 \( \sigma > 0 \).