2

Empirical Statistical Methods

Statistics deals to a large extend with the process of transforming observed data from one or more experiments into a probabilistic model – that is a probability measure on the sample space E. In many cases we may want to obtain an understanding of the data in a more superficial way – if we just want to understand some aspects of the probability measure, say. And even if we want to identify a specific probability measure we may find a "quick and dirty" approximation sufficient. The methods developed in this chapter are called empirical methods or descriptive statistics, because the methods developed mostly serve the purpose of summarising and describing the empirical data in a convenient way. We will also discuss some of the theoretical properties of the empirical methods that we develop.

2.1 The Empirical Probability Measure

Throughout we assume that we have a dataset,

$$x_1, \ldots, x_n \in E,$$

given. This means that we imagine that we have conducted an experiment n times and obtained the observations $x_1, \ldots, x_n$. One should remember that this notation, though quite innocently looking, covers a very wide range of possibilities. We could have a simple sample space like $E = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ but n being in the order of millions, i.e. the dataset could be a DNA-sequence (a genome) being millions or even milliards of letters long. The dataset could also contain protein structures for n different proteins, each structure being represented by three-dimensional coordinates in space for each atom in the protein. Each observation in such a dataset is therefore a real valued vector typically containing several thousand coordinates. So looking at the raw sequence of data may not provide any insight whatsoever about what probability measure that generated the data.

The first thing we are going to do is to transform the dataset into a probability measure – the *empirical probability measure*, to be defined below. This does not seem to simplify things much from a practical point of view, but the rest of this chapter will almost exclusively deal with deriving different kinds of descriptive statistical methods based on the empirical measure. The viewpoint of this chapter is therefore that descriptive statistics is a two step procedure: (i) Decide upon a useful, quantitative property that can be computed for all (or most) probability measures and (ii) compute that property for the empirical probability measure.

Definition 2.1.1. *Given a dataset $x_1, \ldots, x_n$ we define the empirical probability measure, or simply the empirical measure, ε_n, on E by*

$$\varepsilon_n(A) = \frac{1}{n}\sum_{k=1}^{n} 1(x_k \in A) \tag{2.1}$$

for all events $A \subseteq E$.

The empirical measure is the collection of frequencies, $\varepsilon_n(A)$, for all events $A \subseteq E$, which we encountered when discussing the frequency interpretation in Section 1.2. It is also the frequency interpretation that provides the rationale for considering the empirical measure. We will throughout this chapter assume that the dataset has been generated as a realisation of n iid stochastic variables $X_1, \ldots, X_n$ with distribution $\mathbb{P}$. The frequency interpretation then states that

$$\varepsilon_n(A) \simeq \mathbb{P}(A)$$

for all events $A \subseteq E$ when n is sufficiently large. The empirical measure therefore seems to be a reasonable approximation of the true probability measure.

We will in the following also regard the empirical measure of any event $A \subseteq E$ as a stochastic variable taking values in $[0, 1]$ and defined by

$$\varepsilon_n(A) = \frac{1}{n}\sum_{k=1}^{n} 1(X_k \in A).$$

Formally this is a substitution on the r.h.s. in (2.1) of x_k with X_k. Thus the empirical measure ε_n has two meanings. It means a collection of stochastic variables,

$$\frac{1}{n}\sum_{k=1}^{n} 1(X_k \in A), \; A \subseteq E,$$

as well as a concrete realisation of these stochastic variables,

$$\frac{1}{n}\sum_{k=1}^{n} 1(x_k \in A),\ A \subseteq E,$$

for $x_1, \dots, x_n$ a concrete realisation of $X_1, \dots, X_n$. Usually this abuse of notation doesn't lead to problems, since it will we clear from the context whether we regard $\varepsilon_n(A)$ as a stochastic variable or a concrete realisation. The purpose of regarding $\varepsilon_n(A)$ as a stochastic variable is that we can then analyse its distribution, and we can discuss to what extend a concrete realisation of the empirical measure will be a good approximation of $\mathbb{P}$. In the following we will always assume unless otherwise stated that X denotes a stochastic variable with distribution $\mathbb{P}$.

2.2 Mean, Variance and Covariance

If X is a real valued stochastic variable, i.e. the sample space is $\mathbb{R}$, we will introduce the expectation (or mean) and the variance of X. They provides us with two numbers representing the the typical *location* of a realisation of X and the spread around this location.

2.2.1 Expectations

To define *the expectation operator* $\mathbb{E}$ we choose for each n a division of the real line into n disjoint intervals $I_1, I_2, \dots, I_n$ given as

$$\begin{array}{cccccccccc} I_1 & I_2 & \dots & I_k & I_{k+1} & \dots & I_{n-1} & I_n \\ \shortparallel & \shortparallel & \dots & \shortparallel & \shortparallel & \dots & \shortparallel & \shortparallel \\ (-\infty, s_1] & (s_1, s_2] & \dots & (s_{k-1}, 0] & (0, s_{k+2}] & \dots & (s_{n-1}, s_n] & (s_n, \infty) \end{array}$$

with $s_k = s_{k+1} = 0$. Then we can compute the average of the s_i's *weighted* with the the probabilities that $X \in I_i$:

$$\xi_n(X) := \sum_{i=1}^{n} s_i \mathbb{P}(X \in I_i).$$

If the size of all the intervals shrinks towards zero as $n \to \infty$, the differens between X and s_i when $X \in I_i$ will also shrink towards zero and in many cases the number ξ_n converges towards something that we will call the expectation of X. The limiting value, if it exists, is a suitable (generalised) weighted average of the values that X can take – weighted according to the distribution of X. If we had to summarise our *expectations* to the outcome into one single number this limit seems to be a reasonable choice.

If X is only taking positive values it is possible to show that $\xi_n(X)$ always converges to something as $n \to \infty$ but that this something may be ∞. In case the limit is finite we say that X has finite expectation and otherwise it has infinite expectation. If X is any real valued stochastic variable its absolute value, $|X|$, is a *positive* real valued stochastic variable, and it is then possible to show that if $|X|$ has finite expectation the sequence $\xi_n(X)$ will converge to some finite real number.

Definition 2.2.1. *If X is a real valued stochastic variable we say that it has finite expectation if*

$$\lim_{n\to\infty} \xi_n(|X|) < \infty.$$

In this case the expectation operator $\mathbb{E}$ is defined as

$$\mathbb{E}X = \lim_{n\to\infty} \xi_n(X).$$

The formal properties of the expectation operator is best studied within the framework of *measure and integration theory.* It is far beyond the scope of these notes to go into the mathematical details of this theory, and we will instead simply state how to do computations with the expectation operator $\mathbb{E}$.

Theorem 2.2.2. *The expectation operator is linear. If X and Y are two real valued stochastic variables with finite expectation then $X+Y$ has finite expectation and*

$$\mathbb{E}(X+Y) = \mathbb{E}X + \mathbb{E}Y.$$

Furthermore, if $c \in \mathbb{R}$ is a real valued constant then cX has finite expectation and

$$\mathbb{E}cX = c\mathbb{E}X$$

Theorem 2.2.3. *If X is a positive real valued stochastic variable with distribution function F then it has finite expectation if and only if $\int_0^\infty 1 - F(x)\mathrm{d}x < \infty$ in which case*

$$\mathbb{E}X = \int_0^\infty 1 - F(x)\mathrm{d}x.$$

The two preceding theorems are in principle sufficient for computing the expectation of any real valued stochastic variable. If we define

$$X^+ = \max\{X, 0\} \quad \text{and} \quad X^- = \max\{-X, 0\},$$

which are called the positive and negative part of X respectively, then

$$X = X^+ - X^- \quad \text{and} \quad |X| = X^+ + X^-.$$

Both X^+ and X^- are positive stochastic variables and with F^+ and F^- their respective distribution functions we get that if $\mathbb{E}(X^+) < \infty$ and $\mathbb{E}(X^-) < \infty$ then X has finite expectation and

$$\begin{aligned} \mathbb{E}X &= \mathbb{E}X^+ - \mathbb{E}X^- \\ &= \int_0^\infty 1 - F^+(x)dx - \int_0^\infty 1 - F^-(x)dx \\ &= \int_0^\infty F^-(x) - F^+(x)dx. \end{aligned}$$

Theorem 2.2.4. *If X is a real valued stochastic variable with density f and finite expectation then*

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

Theorem 2.2.5. *If X is a discrete stochastic variable taking values in $E \subseteq \mathbb{R}$ with point probabilities $(p(x))_{x\in E}$ and finite expectation then*

$$\mathbb{E}X = \sum_{x\in E} xp(x).$$

Example 2.2.6. If X is a Bernoulli variable with succes probability p we find that

$$\mathbb{E}X = 1 \times \mathbb{P}(X = 1) + 0 \times \mathbb{P}(X = 0) = p. \tag{2.2}$$

Example 2.2.7. Let X be uniformly distributed in $[a, b]$. Then the density for the distribution of X is

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

We find the expectation of X to be

$$\begin{aligned} \mathbb{E}X &= \int_a^b x\frac{1}{b-a}\mathrm{d}x = \frac{1}{2(b-a)}x^2\Big|_a^b \\ &= \frac{1}{2}\frac{b^2-a^2}{b-a} = \frac{1}{2}\frac{(b-a)(b+a)}{b-a} = \frac{1}{2}(a+b). \end{aligned}$$

We see that $\mathbb{E}X$ is the midpoint between a and b.

Example 2.2.8. Let X be an exponentially distributed random variable with intensity parameter $\lambda > 0$. The density is

$$f(x) = \lambda\exp(-\lambda x)$$

for $x \geq 0$ (and $f(x) = 0$ for $x \leq 0$). We find using Theorem 2.2.4 and partial integration that

$$\begin{aligned} \mathbb{E}X &= \int_0^\infty x\lambda \exp(-\lambda x) dx \\ &= x\exp(-\lambda x)\Big|_0^\infty - \int_0^\infty \exp(-\lambda x)\mathrm{d}x \\ &= \frac{1}{\lambda}\exp(-\lambda x)\Big|_0^\infty = \frac{1}{\lambda}. \end{aligned}$$

Example 2.2.9. Let $X \sim \mathrm{Bin}(n,p)$ then

$$X = X_1 + \ldots + X_n$$

where $X_1, \ldots, X_n$ are iid Bernoulli variables with succes probability p. We find using Theorem 2.2.5 that

$$\mathbb{E}X = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k}.$$

It requires a little work to compute this sum. It is much easier to use Theorem 2.2.2 together with (2.2) to obtain that

$$\mathbb{E}X = \mathbb{E}X_1 + \ldots + \mathbb{E}X_n = p + \ldots + p = np.$$

Theorem 2.2.10 (Transformations). *Let X be a stochastic variable taking values in a discrete set E with distribution $\mathbb{P}$ given by the point probabilities $(p(x))_{x\in E}$. If $h : E \to \mathbb{R}$ is any real valued function then if $h(X)$ has finite expectation*

$$\mathbb{E}h(X) = \sum_{x\in E} h(x)p(x).$$

Proof: The stochastic variable $h(X)$ takes values in the discrete subset $E' \subseteq \mathbb{R}$ given by

$$E' = \{h(x) \mid x \in E\}.$$

For each $z \in E'$ we let $A_z = \{x \in E \mid h(x) = z\}$ denote the set of all x's in E which h maps to z. Note that each $x \in E$ belongs to exactly one set A_z. We say that the sets A_z, $z \in E'$, form a *disjoint partion* of E. The distribution of $h(X)$ has point probabilities $(p(z))_{z\in E'}$ given by

$$p(z) = \mathbb{P}(A_z) = \sum_{x\in A_z} p(x)$$

by Definition 1.7.1, and using Theorem 2.2.5 the expectation of $h(X)$ can be written as

$$\mathbb{E}h(X) = \sum_{z\in E'} zp(z) = \sum_{z\in E'} z \sum_{x\in A_z} p(x).$$

Now the function h is constantly equal to z on A_z so we get

$$\mathbb{E}h(X) = \sum_{z \in E'} \sum_{x \in A_z} h(x)p(x).$$

Since the sets A_z form a disjoint partition of the probability space E the sum on the r.h.s. above is precisely a sum over all elements in E, hence

$$\mathbb{E}h(X) = \sum_{x \in E} h(x)p(x).$$

□

Math Box 2.2.11 (General Transformations). For a stochastic variable X taking values in $\mathbb{R}^n$ there is an analoge of Theorem 2.2.10. If $h : \mathbb{R}^n \to \mathbb{R}$ is a real valued function, if the distribution of X has density $f : \mathbb{R}^n \to \mathbb{R}$ and if $h(X)$ has finite expectation then

$$\begin{aligned} \mathbb{E}h(X) &= \int h(x)f(x)\mathrm{d}x \\ &= \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{n} h(x_1, \ldots, x_n)f(x_1, \ldots, x_n)\mathrm{d}x_1 \cdots \mathrm{d}x_n. \end{aligned}$$

As for the discrete case one can verify whether h has finite expectation by computing

$$\begin{aligned} \mathbb{E}|h(X)| &= \int |h(x)|f(x)\mathrm{d}x \\ &= \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{n} |h(x_1, \ldots, x_n)|f(x_1, \ldots, x_n)\mathrm{d}x_1 \cdots \mathrm{d}x_n. \end{aligned}$$

The result is very useful since we may not be able to find an explicit analytic expression for the density of the distribution of $h(X)$ – the distribution may not even have a density – but often we have the distribution of X specified in terms of the density f. The computation of the iterated integrals can, however, be a horrendous task.

Remark 2.2.12. It always holds that

$$\mathbb{E}|h(X)| = \sum_{x \in E} |h(x)|p(x)$$

which can be verified by computations similarly to those above. This expectation may be infinite and $h(X)$ has finite expectation if and only if

$$\sum_{x \in E} |h(x)| p(x) < \infty.$$

It is silently understood above that for a stochastic variable under consideration we have given a fixed distribution $\mathbb{P}$ which is always used in the computations. If it is not clear from the context what the distribution of X is when we are computing $\mathbb{E}X$ we write $\mathbb{E}_{\mathbb{P}}X$ instead of $\mathbb{E}X$ if the distribution of X is $\mathbb{P}$. If $x_1, \ldots, x_n$ is the outcome of n iid replications of our experiment, the corresponding empirical measure ε_n is an approximation of $\mathbb{P}$. If we substitute the true distribution $\mathbb{P}$ of X, which is unknown, with the empirical distribution we can compute the approximation

$$\hat{\mu}_n = \mathbb{E}_{\varepsilon_n} X$$

of the true mean $\mathbb{E}_{\mathbb{P}}X$. We call $\hat{\mu}_n$ the empirical mean. It follows that

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

2.2.2 Variance

Definition 2.2.13. *If X is a real valued stochastic variable with expectation $\mathbb{E}X$, then if X^2 has finite expectation we define the variance of X as*

$$\mathbb{V}X = \mathbb{E}(X - \mathbb{E}X)^2 \tag{2.3}$$

and the standard deviation is defined as $\sqrt{\mathbb{V}X}$.

The variance is the expectation of the squared difference between X and its expectation $\mathbb{E}X$. This is a natural way of measuring how variable X is.

Remark 2.2.14. Writing out $(X - \mathbb{E}X)^2 = X^2 - 2X\mathbb{E}X + (\mathbb{E}X)^2$ and using Theorem 2.2.2 we obtain

$$\mathbb{V}X = \mathbb{E}X^2 - 2\mathbb{E}X\mathbb{E}X + (\mathbb{E}X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2, \tag{2.4}$$

which is a useful alternative way of computing the variance. The expectation of X^2, $\mathbb{E}X^2$, is called the *second moment* of the distribution of X.

Remark 2.2.15. For any $\mu \in \mathbb{R}$ we can write

$$\begin{aligned}(X-\mu)^2 &= (X-\mathbb{E}X+\mathbb{E}X-\mu)^2 \\ &= (X-\mathbb{E}X)^2+2(X-\mathbb{E}X)(\mathbb{E}X-\mu)+(\mathbb{E}X-\mu)^2,\end{aligned}$$

from which

$$\begin{aligned}\mathbb{E}(X-\mu)^2 &= \mathbb{E}(X-\mathbb{E}X)^2+2(\mathbb{E}X-\mathbb{E}X)(\mathbb{E}X-\mu)+(\mathbb{E}X-\mu)^2 \\ &= \mathbb{V}X+(\mathbb{E}X-\mu)^2 \geq \mathbb{V}X \qquad (2.5)\end{aligned}$$

with equality if and only if $\mathbb{E}X = \mu$. The number $\mathbb{E}(X-\mu)^2$ is the expected squared difference between μ and X, thus a measure of how much the outcome deviates from μ on average. We see that the expectation $\mathbb{E}X$ is the unique value of μ that minimises this measure of deviation. The expectation is therefore in this sense the best constant approximation to any outcome of our experiment.

Example 2.2.16. If X is a Bernoulli stochastic variable with succes probability p we known from Example 2.2.6 that $\mathbb{E}X = p$. We find that

$$\begin{aligned}\mathbb{V}X &= \mathbb{E}(X-p)^2 = (1-p)^2\mathbb{P}(X=1)+p^2\mathbb{P}(X=0) \\ &= (1-p)^2p+p^2(1-p) = (1-p)p(1-p+p) = (1-p)p.\end{aligned}$$

Example 2.2.17. If X is a stochastic variable with mean 0 and variance 1 then

$$\mathbb{E}(\sigma X+\mu) = \sigma\mathbb{E}(X)+\mu = \mu$$

and

$$\mathbb{V}(\sigma X+\mu) = \mathbb{E}(\sigma X+\mu-\mu)^2 = \mathbb{E}(\sigma^2X^2) = \sigma^2\mathbb{V}X = \sigma^2.$$

This shows that if we make a position-scale transformation of a mean 0 and variance 1 stochastic variable X, with μ and σ the position and scale parameters, then the mean and variance of the position-scaled transformed variable becomes μ and σ^2 respectively. In the other direction we find that if $\mathbb{E}X = \mu$ and $\mathbb{V}X = \sigma^2$ then

$$\mathbb{E}\left(\frac{X-\mu}{\sigma}\right) = 0 \quad \text{and} \quad \mathbb{V}\left(\frac{X-\mu}{\sigma}\right) = 1.$$

Example 2.2.18 (The Normal distribution). If $X \sim N(0,1)$, that is the distribution of X has density

$$f(x) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2}),$$

then

$$\mathbb{E}X = 0 \quad \text{and} \quad \mathbb{V}X = 1.$$

We will not show this (although the mean can be found by simple symmetri argument). Using the example above we find that if $X \sim N(\mu, \sigma^2)$ then

$$\mathbb{E}X = \mu \quad \text{and} \quad \mathbb{V}X = \sigma^2.$$

Computing the empirical variance (the variance under the empirical measure) amounts to

$$\begin{aligned}\tilde{\sigma}_n^2 &= \mathbb{V}_{\varepsilon_n} X = \mathbb{E}_{\varepsilon_n}(X - \mathbb{E}_{\varepsilon_n} X)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_n)^2 \qquad (2.6)\end{aligned}$$

with $\hat{\mu}_n = \mathbb{E}_{\varepsilon_n}(X)$ the empirical mean. If we use (2.4) we obtain the alternative formula

$$\tilde{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \hat{\mu}_n^2 \qquad (2.7)$$

It should be remarked that whereas (2.4) can be quite useful for theoretical computations it is *not* recommended that one uses (2.7) for numerical computations. This is because both $\frac{1}{n}\sum_{i=1}^{n} x_i^2$ and $\hat{\mu}_n^2$ can attain very large numerical values, and subtracting numerically large numbers can lead to a serious loss of precision.

As for the empirical mean we regard the empirical variance $\tilde{\sigma}_n^2$ as an approximation of the true variance $\mathbb{V}(X)$ of X under $\mathbb{P}$. Likewise, the square root

$$\tilde{\sigma}_n = \sqrt{\tilde{\sigma}_n^2}$$

of the empirical variance is an approximation of the standard deviation.

Example 2.2.19 (Empirical normalisation). If we have the dataset $x_1, \ldots, x_n$ and let X be a stochastic variable with distribution ε_n (the empirical distribution). Then by definition

$$\mathbb{E}_{\varepsilon_n} X = \hat{\mu}_n \quad \text{and} \quad \mathbb{V}_{\varepsilon_n} X = \widetilde{\sigma}_n^2.$$

The distribution of

$$X' = \frac{X - \hat{\mu}_n}{\widetilde{\sigma}_n^2}$$

is the empirical measure defined by the *normalised dataset*

$$x_i' = \frac{x_i - \hat{\mu}_n}{\widetilde{\sigma}_n^2}.$$

That is, the distribution of X' is ε_n' given by

$$\varepsilon_n'(A) = \sum_{i=1}^{n} 1(x_i' \in A).$$

By Example 2.2.17 it follows that

$$\mathbb{E}_{\varepsilon_n'} X = \mathbb{E}_{\varepsilon_n} X' = 0 \quad \text{and} \quad \mathbb{V}_{\varepsilon_n'} X = \mathbb{V}_{\varepsilon_n} X' = 1.$$

This means that having normalised the dataset with the empirical mean and variance results in a dataset that has empirical mean 0 and empirical variance 1.

2.2.3 Multivariate Distributions

If we consider two real valued stochastic variables X and Y the bundled variable (X,Y) takes values in $\mathbb{R}^2$. The mean and variance of each of the variables X and Y rely exclusively on the marginal distributions of X and Y. Thus they tell us nothing about the simultaneous distribution of X and Y. We introduce the *co*variance as a measure of dependency between X and Y.

Definition 2.2.20. *If XY has finite expectation the covariance of the stochastic variables X and Y is defined as*

$$\mathbb{V}(X,Y) = \mathbb{E}(X-\mathbb{E}X)(Y-\mathbb{E}Y) \tag{2.8}$$

and the correlation is defined by

$$\mathrm{corr}(X,Y) = \frac{\mathbb{V}(X,Y)}{\sqrt{\mathbb{V}X\mathbb{V}Y}} \tag{2.9}$$

The covariance is a measure of the covariation, that is the dependency, between the two stochastic variables X and Y. It is possible to show that if X and Y are *independent* then

$$\mathbb{V}(X,Y) = 0,$$

but having covariation equal to zero does *not* guaranty independence. The correlation is a standardisation of the covariance by the variances of the coordinates X and Y, and it can be shown that

$$-1 \leq \mathrm{corr}(X,Y) \leq 1.$$

We should note that the covariance is symmetric in X and Y:

$$\mathbb{V}(X,Y) = \mathbb{V}(Y,X).$$

Furthermore, if $X = Y$ then

$$\mathbb{V}(X,X) = \mathbb{E}(X-\mathbb{E}X)^2 = \mathbb{V}X$$

that is the covariance of X with X is simply the variance of X. Finally observe that

$$\begin{aligned}
\mathbb{V}(X,Y) &= \mathbb{E}(X-\mathbb{E}X)(Y-\mathbb{E}Y) \\
&= \mathbb{E}(XY - X\mathbb{E}Y - Y\mathbb{E}X + \mathbb{E}X\mathbb{E}Y) \\
&= \mathbb{E}XY - \mathbb{E}X\mathbb{E}Y - \mathbb{E}X\mathbb{E}Y + \mathbb{E}X\mathbb{E}Y \\
&= \mathbb{E}XY - \mathbb{E}X\mathbb{E}Y, \qquad (2.10)
\end{aligned}$$

which gives an alternative formula for computing the covariance. It also provides us with the formula

$$\mathbb{E}XY = \mathbb{E}X\mathbb{E}Y + \mathbb{V}(X,Y)$$

for the expectation of the product of two stochastic variables.

Theorem 2.2.21. *If X and Y are two stochastic variables with finite variance then the sum $X+Y$ has finite variance and*

$$\mathbb{V}(X+Y) = \mathbb{V}X + \mathbb{V}Y + 2\mathbb{V}(X,Y). \tag{2.11}$$

If X is a stochastic variable with finite variance and $c \in \mathbb{R}$ is a constant then cX has finite variance and

$$\mathbb{V}cX = c^2\mathbb{V}X. \tag{2.12}$$

Proof: We use (2.4) to compute the variance of $X+Y$:

$$\begin{aligned}
\mathbb{V}(X+Y) &= \mathbb{E}(X+Y)^2 - (\mathbb{E}(X+Y))^2 \\
&= \mathbb{E}(X^2+Y^2+2XY) - ((\mathbb{E}X)^2 + (\mathbb{E}Y)^2 + 2\mathbb{E}X\mathbb{E}Y) \\
&= \mathbb{E}X^2 - (\mathbb{E}X)^2 + \mathbb{E}Y^2 - (\mathbb{E}Y)^2 + 2(\mathbb{E}XY - \mathbb{E}X\mathbb{E}Y) \\
&= \mathbb{V}X + \mathbb{V}Y + 2\mathbb{V}(X,Y)
\end{aligned}$$

using (2.4) again together with (2.10) for the last equality. □

Remark 2.2.22. We observe that the formula

$$\mathbb{V}(X+Y) = \mathbb{V}X + \mathbb{V}Y \tag{2.13}$$

holds if and only if $\mathbb{V}(X,Y) = 0$, which in particular is the case if X and Y are independent. Note also that it follows from the theorem that

$$\begin{aligned}
\mathbb{V}(X-Y) &= \mathbb{V}(X+(-1)Y) = \mathbb{V}X + \mathbb{V}(-1)Y + 2\mathbb{V}(X,-Y) \\
&= \mathbb{V}X + (-1)^2\mathbb{V}Y + 2\mathbb{V}(X,-Y) \\
&= \mathbb{V}X + \mathbb{V}Y - 2\mathbb{V}(X,Y)
\end{aligned}$$

If we consider not just two stochastic variables but a k-dimensional vector $X = (X_1,\ldots,X_k)$ of real valued stochastic variables we can compute the k^2 covariances for each pair of variables X_i and X_j. One usually arrange the covariances in a $k \times k$ matrix Σ given by

$$\Sigma_{ij} = \mathbb{V}(X_i, Y_j).$$

That is

$$\Sigma = \left\{ \begin{array}{cccc}
\mathbb{V}X_1 & \mathbb{V}(X_1,X_2) & \cdots & \mathbb{V}(X_1,X_n) \\
\mathbb{V}(X_2,X_1) & \mathbb{V}(X_2) & \cdots & \mathbb{V}(X_2,X_n) \\
\vdots & \vdots & \ddots & \vdots \\
\mathbb{V}(X_n,X_1) & \mathbb{V}(X_n,X_2) & \cdots & \mathbb{V}X_n
\end{array} \right\}$$

Note that due to the symmetry of the covariance we have that the covariance matrix Σ is symmetric:

$$\Sigma_{ij} = \Sigma_{ji}.$$

With $x_1, \ldots, x_n \in E$ a dataset where $x_l = (x_{1l}, \ldots, x_{kl}) \in \mathbb{R}^k$ the computation of the covariance under the empirical measure gives

$$\begin{aligned}\widetilde{\sigma}^2_{ij,n} &= \mathbb{V}_{\varepsilon_n}(X_i, X_j) = \mathbb{E}_{\varepsilon_n}(X_i - \mathbb{E}_{\varepsilon_n} X_i)(X_j - \mathbb{E}_{\varepsilon_n} X_j) \\ &= \frac{1}{n}\sum_{l=1}^{n}(x_{il} - \widehat{\mu}_{i,n})(x_{jl} - \widehat{\mu}_{j,n})\end{aligned}$$

where

$$\widehat{\mu}_{i,n} = \frac{1}{n}\sum_{l=1}^{n} x_{il}.$$

Using (2.10) instead we obtain that

$$\widetilde{\sigma}^2_{ij,n} = \frac{1}{n}\sum_{l=1}^{n} x_{il}x_{jl} - \widehat{\mu}_{i,n}\widehat{\mu}_{j,n}. \tag{2.14}$$

As for the variance this is not a recommended formula to use for the practical computation of the empirical covariance.

The empirical covariance matrix $\widetilde{\Sigma}_n$ is given by

$$\widetilde{\Sigma}_{ij,n} = \widetilde{\sigma}^2_{ij,n}$$

The empirical correlation becomes

$$\widetilde{\mathrm{corr}}_{ij,n} = \frac{\widetilde{\sigma}^2_{ij,n}}{\widetilde{\sigma}_{i,n}\widetilde{\sigma}_{j,n}} = \frac{\sum_{l=1}^{n}(x_{il} - \widehat{\mu}_{i,n})(x_{jl} - \widehat{\mu}_{j,n})}{\sqrt{\sum_{l=1}^{n}(x_{il} - \hat{\mu}_{i,n})^2 \sum_{l=1}^{n}(x_{jl} - \hat{\mu}_{j,n})^2}}.$$

2.2.4 Properties of the Empirical Approximations

We recall that the empirical measure can be regarded as a realisation of a stochastic variable. Thus with a different realisation we would get a different empirical mean $\hat{\mu}_n$ and a different empirical variance $\tilde{\sigma}^2_n$. To evaluate the performance of these empirical quantities as approximations of the expectation and variance respectively we can study their distributions when regarded as stochastic variables. In particular we can compute the expectation and variance of $\hat{\mu}_n$ and $\tilde{\sigma}^2_n$.

Theorem 2.2.23. *Regarding ε_n as a stochastic variable, the empirical mean $\hat{\mu}_n$ and the empirical variance $\tilde{\sigma}^2_n$ are also stochastic variables and*

$$\mathbb{E}\hat{\mu}_n = \mathbb{E}X \quad \textit{and} \quad \mathbb{V}\hat{\mu}_n = \frac{1}{n}\mathbb{V}X \tag{2.15}$$

together with

$$\mathbb{E}\tilde{\sigma}^2_n = \frac{n-1}{n}\mathbb{V}X. \tag{2.16}$$

Proof: The derivation of these results is an exercise in using the properties of the expectation operator and independence of the stochastic variables $X_1, \ldots, X_n$.

$$\mathbb{E}\hat{\mu}_n = \mathbb{E}\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}X_i = \mathbb{E}X,$$

$$\mathbb{V}\hat{\mu}_n = \mathbb{V}\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}X_i = \frac{1}{n}\mathbb{V}X,$$

and

$$\begin{aligned}\mathbb{E}\tilde{\sigma}_n^2 &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \hat{\mu}_n^2\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}X_i^2 - \mathbb{E}\hat{\mu}_n^2 \\ &= \mathbb{E}X^2 - \mathbb{V}\hat{\mu}_n - (\mathbb{E}\hat{\mu}_n)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2 - \frac{1}{n}\mathbb{V}X = \frac{n-1}{n}\mathbb{V}X.\end{aligned}$$

□

The theorem shows that the expected value of $\hat{\mu}_n$ equals the true expectation $\mathbb{E}X$ and that the variance of $\hat{\mu}_n$ decreases as $1/n$. Thus for large n the variance of $\hat{\mu}_n$ becomes negligible and $\hat{\mu}_n$ will always be a very close approximation to $\mathbb{E}X$. How large n should be depends on the size of $\mathbb{V}X$. Regarding the empirical variance its expectation does not equal the true variance $\mathbb{V}X$. The expected value is always a little smaller than $\mathbb{V}X$. The relative deviation is

$$\frac{\mathbb{V}X - \mathbb{E}\tilde{\sigma}_n^2}{\mathbb{V}X} = \frac{1}{n},$$

which becomes negligible when n becomes large. However, for $n = 5$, say, the empirical variance undershoots the true variance by 20% on average. For this reason the empirical variance is *not* the preferred way of computing an empirical approximation to the variance. Instead the standard choice of approximation is

$$\hat{\sigma}_n^2 = \frac{n}{n-1}\tilde{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu}_n)^2. \tag{2.17}$$

It follows from Theorem 2.2.23 and linearity of the expectation operator that

$$\mathbb{E}\hat{\sigma}_n^2 = \mathbb{V}(X).$$

The square root $\hat{\sigma}_n = \sqrt{\hat{\sigma}_n^2}$ naturally becomes the corresponding approximation of the standard deviation. Note, however, that the expectation argument doesn't carry over to the standard deviations. In fact, it is possible to show that

$$\mathbb{E}\hat{\sigma}_n < \sqrt{\mathbb{V}(X)}$$

so $\hat{\sigma}_n$ is still expected to undershoot the standard deviation.

It is also possible to compute the variance of $\tilde{\sigma}_n^2$ but the derivation is long and tedious so we will skip it. The result is

$$\mathbb{V}\tilde{\sigma}_n^2 = \frac{n-1}{n^3}\left((n-1)\mathbb{E}(X-\mathbb{E}X)^4 - (n-3)(\mathbb{V}X)^2\right), \tag{2.18}$$

which is not a particular nice formula either. One can observe though that the variance decreases approximately as $1/n$, which shows that also the empirical variance becomes a good approximation of the true variance when n becomes large. But regardless of whether we can compute the variance of the empirical variance, we can compare the variance of $\tilde{\sigma}_n^2$ with the variance of $\hat{\sigma}_n^2$ and find that

$$\mathbb{V}(\hat{\sigma}_n^2) = \left(\frac{n}{n-1}\right)^2 \mathbb{V}(\tilde{\sigma}_n^2).$$

Hence the variance of $\hat{\sigma}_n^2$ is *larger* than the variance of the empirical variance $\tilde{\sigma}_n^2$. This is not necessarily problematic, but it should be noticed that what we gain by correcting the empirical variance so that the expectation becomes correct is (partly) lost by the increased variance.

R Box 2.2.24 (Mean and variance). If `x` is a numeric vector one can compute the (empirical) mean of `x` simply by

```
> mean(x)
```

Likewise, the (empirical) variance can be computed by

```
> var(x)
```

Using `var` results in $\hat{\sigma}_n^2$ where we divide by $n-1$.

The normalised dataset where we subtract the mean and divide by the variance can be efficiently computed by

```
> y <- scale(x)
```

If we consider a k-dimensional stochastic variable $X = (X_1, \ldots, X_k)$ we can also derive a result about the expectation of the empirical covariance.

Theorem 2.2.25. *The expectation of the empirical covariance regarded as a stochastic variable is*

$$\mathbb{E}\tilde{\sigma}_{ij,n} = \frac{n-1}{n}\mathbb{V}(X_i, X_j).$$

Proof: Using (2.14) yields

$$\begin{aligned}\mathbb{E}\tilde{\sigma}_{ij,n} &= \frac{1}{n}\sum_{l=1}^{n}\mathbb{E}X_{il}X_{jl} - \mathbb{E}\widehat{\mu}_{i,n}\widehat{\mu}_{j,n} \\ &= \mathbb{E}X_iX_j - \frac{1}{n^2}\sum_{l=1}^{n}\sum_{m=1}^{n}\mathbb{E}X_{il}X_{jm}\end{aligned}$$

Observing then that due to independence of X_{il} and X_{jm} when $m \neq l$

$$\mathbb{E}X_{il}X_{jm} = \mathbb{E}X_i\mathbb{E}X_j.$$

There are $n(n-1)$ such terms in the last sum above. There are n terms equalling $\mathbb{E}X_iX_j$. This gives that

$$\begin{aligned}\mathbb{E}\tilde{\sigma}_{ij,n} &= \mathbb{E}X_iX_j - \frac{1}{n}\mathbb{E}X_iX_j - \frac{n-1}{n}\mathbb{E}X_i\mathbb{E}X_j \\ &= \frac{n-1}{n}\left(\mathbb{E}X_iX_j - \mathbb{E}X_i\mathbb{E}X_j\right) = \frac{n-1}{n}\mathbb{V}(X_i, X_j).\end{aligned}$$

□

As we can see the empirical covariance also generally undershoots the true covariance leading to the alternative approximation

$$\hat{\sigma}_{ij,n} = \frac{1}{n-1}\sum_{l=1}^{n}(x_{il} - \widehat{\mu}_{i,n})(x_{jl} - \widehat{\mu}_{j,n}) \tag{2.19}$$

of the true variance with $\mathbb{E}\hat{\sigma}_{ij,n} = \mathbb{V}X(X_i, X_j)$.

2.3 Frequencies and Tables

Recall that the empirical measure is a collection of frequencies $\varepsilon_n(A)$ for all events $A \subseteq E$. In the previous section we discussed how to derive approximations from the empirical measure, hence from the frequencies, of the two characteristics of $\mathbb{P}$ – the mean and variance. In this section we return to the frequencies themselves and study how we can organise and present this collection of frequencies in a reasonable manner. First we give a theorem that tells us something about the distribution of the empirical measure and hence why $\varepsilon_n(A)$ is a reasonable approximation of $\mathbb{P}(A)$.

Theorem 2.3.1. *If $X_1, \ldots, X_n$ are independent and identically distributed with distribution $\mathbb{P}$ on E and ε_n is the corresponding empirical probability measure,*

$$\varepsilon_n(A) = \sum_{k=1}^{n} 1(X_k \in A), \quad A \subseteq E,$$

then for any event A it holds that

$$\mathbb{E}\varepsilon_n(A) = \mathbb{P}(A) \tag{2.20}$$

and

$$\mathbb{V}\varepsilon_n(A) = \frac{1}{n}\mathbb{P}(A)(1-\mathbb{P}(A)). \tag{2.21}$$

Proof: Since $1(X_i \in A)$ is a Bernoulli variable we can use Example 2.2.6 to find that

$$\mathbb{E}1(X_i \in A) = \mathbb{P}(A)$$

so

$$\begin{aligned}
\mathbb{E}\varepsilon_n(A) &= \mathbb{E}\frac{1}{n}\sum_{i=1}^{n} 1(X_i \in A) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}1(X_i \in A) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}(A) = \mathbb{P}(A).
\end{aligned}$$

Example 2.2.16 gives that

$$\mathbb{V}1(X_i \in A) = \mathbb{P}(A)(1-\mathbb{P}(A)).$$

Hence by independence of the stochastic variables $X_1, \ldots, X_n$

$$\begin{aligned}
\mathbb{V}\varepsilon_n(A) &= \mathbb{V}\frac{1}{n}\sum_{i=1}^{n} 1(X_i \in A) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}1(X_i \in A) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{P}(A)(1-\mathbb{P}(A)) = \frac{1}{n}\mathbb{P}(A)(1-\mathbb{P}(A)).
\end{aligned}$$

□

As for all other probability measures the collection of numbers $\varepsilon_n(A)$ for all events $A \subseteq E$ is enormous even for a small, finite set E. If E is finite we will therefore prefer the smaller collection of frequencies

$$\varepsilon_n(z) = \frac{1}{n}\sum_{i=1}^{n} 1(x_i = z)$$

for $z \in E$ – which is also sufficient for completely determining the empirical measure just like for any other probability measure on a discrete set. If $\mathbb{P}$ is given by the point probabilities $(p(z))_{z\in E}$ Theorem 2.3.1 tells us that

$$\mathbb{E}\varepsilon_n(z) = p(z) \tag{2.22}$$

and

$$\mathbb{V}\varepsilon_n(z) = \frac{1}{n}p(z)(1 - p(z)). \tag{2.23}$$

It is common to arrange the frequencies in a table like the following

z_1	z_2	z_3	$\ldots$	z_N
$\varepsilon_n(z_1)$	$\varepsilon_n(z_2)$	$\varepsilon_n(z_3)$	$\ldots$	$\varepsilon_n(z_N)$

where $E = \{z_1, \ldots, z_N\}$, but usually a table of $n\varepsilon_n(z)$ is presented instead – this table simply being a table of the counts of each of the possible outcomes.

Example 2.3.2. If the dataset consists of the DNA sequence

```
AGACTTGACAGCAGTACCCAG
```

of length 21 the resulting frequency and count tables become

A	C	G	T
0.33	0.29	0.24	0.14

A	C	G	T
7	6	5	3

If $E = E_1 \times E_2$ is a product space, we will organise the counts into a two-dimensional $(N+1) \times (M+1)$ table in the following way:

	w_1	w_2	$\ldots$	w_M	
z_1	$n\varepsilon_n(z_1, w_1)$	$n\varepsilon_n(z_1, w_2)$	$\ldots$	$n\varepsilon_n(z_1, w_M)$	$n\varepsilon_n(z_1, \bullet)$
z_2	$n\varepsilon_n(z_2, w_1)$	$n\varepsilon_n(z_2, w_2)$	$\ldots$	$n\varepsilon_n(z_2, w_M)$	$n\varepsilon_n(z_2, \bullet)$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
z_N	$n\varepsilon_n(z_N, w_1)$	$n\varepsilon_n(z_N, w_2)$	$\ldots$	$n\varepsilon_n(z_N, w_M)$	$n\varepsilon_n(z_N, \bullet)$
	$n\varepsilon_n(\bullet, w_1)$	$n\varepsilon_n(\bullet, w_2)$	$\ldots$	$n\varepsilon_n(\bullet, w_M)$	n

Here we use the notation

$$(z, \bullet) = \{(z, w) \in E \mid w \in E_2\} \quad \text{and} \quad (\bullet, w) = \{(z, w) \in E \mid z \in E_1\}.$$

We place the *row sums*

$$n\varepsilon_n(z_i, \bullet) = n \sum_{k=1}^{M} \varepsilon_n(z_i, w_k), \quad i = 1, \ldots, N,$$

at the right hand side of each row and the *column sums*

$$n\varepsilon_n(\bullet, w_j) = n \sum_{k=1}^{N} \varepsilon_n(z_k, w_j), \quad j = 1, \ldots, M,$$

at the bottom of each column. The probability measures given by the point probabilities $\varepsilon_n(z, \bullet)$ and $\varepsilon_n(\bullet, w)$ are the marginals of the empirical measure ε_n on $E_1 \times E_2$.

Example 2.3.3. If our dataset consists of two (gaplessly) aligned DNA sequences

```
AGACTTGACAGCAGTACCCAG
AGCCGTTACAGCGGTAACCAA
```

both of length 21 the resulting 4×4 frequency and count tables become

	A	C	G	T	
A	0.238	0.048	0.048	0.000	0.33
C	0.048	0.238	0.000	0.000	0.29
G	0.048	0.000	0.143	0.048	0.24
T	0.000	0.000	0.048	0.095	0.14
	0.33	0.29	0.24	0.14	1

	A	C	G	T	
A	5	1	1	0	7
C	1	5	0	0	6
G	1	0	3	1	5
T	0	0	1	2	3
	7	6	5	3	21

If our sample space is the product of more than two spaces one can in principle presents tables of frequencies or counts that are three or higher dimensional. On a computer there is no problem in representing such higher dimensional data by tables but it is a little more problematic to visualise three or higher dimensional tables. One approach especially useful for three dimensional tables is to *stratify* the two-dimensional tables for two of the coordinates according to the third (and forth and ...) variable. A theory for systematically analysing multidimensional discrete data is known as analysis of contingency tables. We will not pursue this further in these notes, but if one wants to make probabilistic models of letters occurring in the alignment of three or more sequences, the theory of analysing contingency tables is a good starting point.

2.4 Histograms and Quantiles

If $E \subseteq \mathbb{R}$ is a subset of the real line there are several different methods for representing the empirical measure.

Definition 2.4.1. *A histogram with break points $q_1 < q_2 < \ldots < q_k$, chosen so that*

$$q_1 < \min_{i=1,\ldots,n} x_i \leq \max_{i=1,\ldots,n} x_i < q_k,$$

is the function h given by

$$h(x) = \frac{1}{q_{i+1} - q_i} \varepsilon_n \left((q_i, q_{i+1}] \right) \quad \text{for } q_i < x \leq q_{i+1}. \tag{2.24}$$

together with $h(x) = 0$ for $x \notin (q_1, q_n]$. Usually the plot of h with a box of height $h(q_{i+1})$ located over the interval $(q_i, q_{i+1}]$ is what most people associate with a histogram.

The function h is constructed so that

$$\begin{aligned}\int h(x) &= \sum_{i=1}^{k-1} \int_{q_i}^{q_{i+1}} \frac{1}{q_{i+1} - q_i} \varepsilon_n\left((q_i, q_{i+1}]\right) \mathrm{d}x \\ &= \sum_{i=1}^{k-1} \varepsilon_n\left((q_i, q_{i+1}]\right) \\ &= \varepsilon_n((q_1, q_n]) = 1\end{aligned}$$

where we use that all the data points are contained within the interval $(q_1, q_n]$. Since the function h integrates to 1 it is a probability density. The purpose of the histogram is to approximate the density of the true distribution of X – assuming that the distribution has a density.

Sometimes one encounters the *unnormalised* histogram, given by the function

$$\tilde{h}(x) = n\varepsilon_n(q_i, q_{i+1}] \quad \text{for } q_i < x \leq q_{i+1}.$$

Here $\tilde{h}(x)$ is constantly equal to the number of observations falling in the interval $(q_i, q_{i+1}]$. Since the function doesn't integrate to 1 it can not be compared directly with a density.

Example 2.4.2. We consider the histogram of 100 and 1000 simulated $N(0, 1)$ iid stochastic variables. We choose the breaks to be equidistant from -4 to 4 with a distance of 0.5, thus the break point are

$$-4 \quad -3.5 \quad -3 \quad -2.5 \quad \ldots \quad 2.5 \quad 3 \quad 3.5 \quad 4.$$

We find the histograms in Figure 2.1. Note how the histogram corresponding to the 1000 simulated stochastic variables approximates the density more closely.

Definition 2.4.4. *If $F : \mathbb{R} \to [0, 1]$ is a distribution function for a probability measure $\mathbb{P}$ on $\mathbb{R}$, then $Q : [0, 1] \to \mathbb{R}$ is a quantile function for $\mathbb{P}$ if*

$$F(Q(y) - \varepsilon) \leq y \leq F(Q(y)) \tag{2.25}$$

for all $y \in [0, 1]$ and all $\varepsilon > 0$.

Theorem 2.4.5. *The generalised inverse distribution function $F^{\leftarrow}$, cf. Section 1.8, is a quantile function.*

Proof: To see this, first observe that with $x = F^{\leftarrow}(y)$ then

$$F^{\leftarrow}(y) \leq x \Rightarrow y \leq F(x) = F(F^{\leftarrow}(y))$$

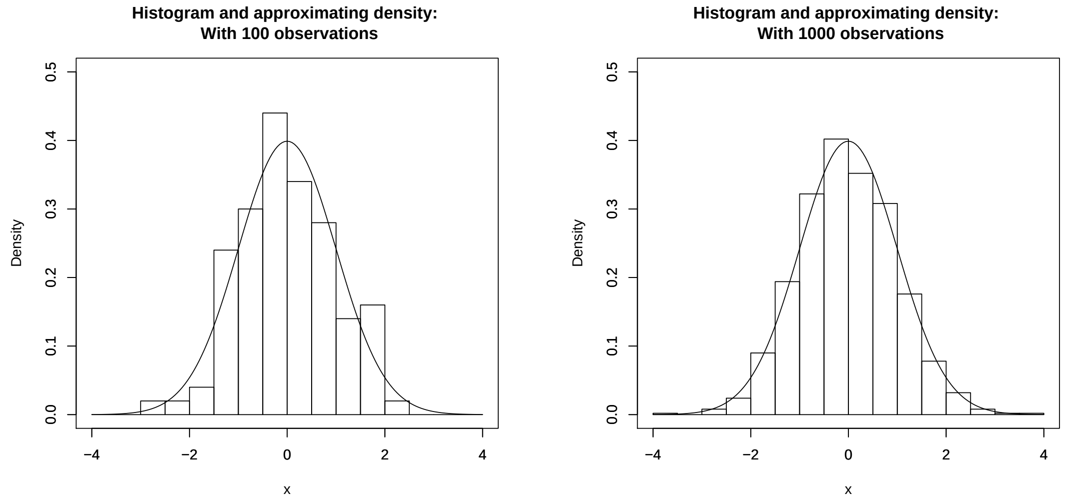

Figure 2.1: The histograms for the realisation of 100 (right) and 1000 (left) simulated iid $N(0,1)$ stochastic variables. For both histograms we compare the histogram with the corresponding density for the normal distribution.

by the definition of $F^{\leftarrow}$. On the other hand, suppose that there exists a $y \in [0,1]$ and an $\varepsilon > 0$ such that $F(F^{\leftarrow}(y) - \varepsilon) \geq y$ then again by the definition of $F^{\leftarrow}$ it follows that

$$F^{\leftarrow}(y) - \varepsilon \geq F^{\leftarrow}(y),$$

which can not be the case. Hence there exists no such $y \in [0,1]$ and $\varepsilon > 0$ and

$$F(F^{\leftarrow}(y) - \varepsilon) < y$$

for all $y \in [0,1]$ and $\varepsilon > 0$. This shows that $F^{\leftarrow}$ is a quantile function. □

There may exist other quantile functions besides the generalised inverse of the distribution function, which are preferred from time to time. However, if F has an inverse function then the inverse is the only quantile function.

Definition 2.4.6. *If F is a distribution function and Q a quantile function for F the median or second quartile of F is defined as*

$$q_2 = \text{median}(F) = Q(0.5).$$

In addition we call $q_1 = Q(0.25)$ and $q_3 = Q(0.75)$ the first end third quartiles of X. The difference

$$\text{IQR} = q_3 - q_1$$

is called the interquartile range.

R Box 2.4.3 (Histograms). A histogram of the data in the numeric vector `x` is produced in R by the command

```
> hist(x)
```

This automatically opens a graphics window and plots a histogram using default settings. The break points are by default chosen by R in a suitable way. It is possible to explicitly set the break points by hand, for instance

```
> hist(x,breaks=c(0,1,2,3,4,5))
```

produces a histogram with break points $0, 1, 2, 3, 4, 5$. Note that R will produce an error if the range of the break points does not contain all the data points in `x`. Note also that the default behaviour of `hist` is to plot the *unnormalised* histogram if the break points are equidistant. Otherwise it produces the normalised histogram. One can always make `hist` produce normalised histograms by

```
> hist(x,freq=FALSE)
```

Note that the definition of the median and the quartiles depend on the choice of quantile function. If the quantile function is not unique these numbers are not necessarily uniquely defined. The median represents like the mean a single number summarising the location of the probability measure given by F. The interquartile range is then comparable with the spread in the sense that it defines a measure of how spread out around the median the distribution is.

Definition 2.4.7. *If $E \subseteq \mathbb{R}$ we define the empirical distribution function F_n as the distribution function for the empirical measure ε_n:*

$$F_n(x) = \varepsilon_n((-\infty, x]). \tag{2.26}$$

An empirical quantile function is a quantile function for F_n.

If Q_n is an empirical quantile function for F_n then the quartiles for Q_n

$$\hat{q}_1 = Q_n(0.25), \quad \hat{q}_2 = Q_n(0.5), \quad \hat{q}_3 = Q_n(0.75),$$

are called the empirical quartiles (and $\hat{q}_2$ is also called the empirical median).

The empirical distribution function is notoriously ugly meaning that it never has an inverse and the quantile function is never unique. On the contrary there exists a

R Box 2.4.8 (Empirical distribution functions). If `x` is a numeric vector in R containing our data we can construct a `ecdf`-object (empirical cumulative distribution function). This requires the `stats` library:

```
> library(stats)
```

Then

```
> edf <- ecdf(x)
```

gives the empirical distribution function for the data in `x`. One can evaluate this function like any other function:

```
> edf(1.95)
```

gives the value of the empirical distribution function evaluated at 1.95. It is also easy to plot the distribution function:

```
> plot(edf)
```

produces a nice plot.

variety of quantile functions, and which one to choose can lead to much debate and confusion. In particular one can debate what the empirical median and the empirical quartiles should be. From a practical and pragmatic point of view it is likely to be completely irrelevant which quantile function one chooses, but it is good to keep in mind that different choices can lead to (small) discrepancies.

If we sort our dataset $x_1, \ldots, x_n$ in increasing order, writing

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$

for the sorted data so that $x_{(i)}$ is the i'th smallest observation, then we can observe that

$$F_n(x) = i/n \quad \text{for } x_{(i)} \leq x < x_{(i+1)}$$

for $i = 1, \ldots, n-1$. If we don't have ties, i.e. $x_{(i)} < x_{(i+1)}$ for $i = 1, \ldots, n$, then the empirical distribution function is constant in between the sorted observations and has jumps of size $1/n$ at each observation. If there are ties F_n has larger jumps.

By the definition of quantile functions one can see that

$$Q_n(y) = x_{(i)} \quad \text{for } \frac{i-1}{n} < y < \frac{i}{n}.$$

with $i \in 1, \ldots, n$. If there is no ties Q_n is in addition going to satisfy that

$$x_{(i)} \leq Q_n\left(\frac{i}{n}\right) \leq x_{(i+1)}.$$

The generalised inverse distribution function is given by

$$F_n^{\leftarrow}(y) = x_{(i)} \quad \text{for } \frac{i-1}{n} < y \leq \frac{i}{n},$$

that is, $F^{\leftarrow}$ is the quantile function that takes the value $x_{(i)}$ in i/n. We see that the ambiguity of which quantile function to choose is present only at the finite many numbers i/n, $i = 1, \ldots, n$. This is, however, quite annoying when computing the empirical median, say. If n is even, then 0.5 is one of the problematic points of ambiguity. If n is odd, however, there is not ambiguity and a uniquely defined empirical median.

One of the applications of quantiles and the empirical quantile function is to compare two distributions by comparing their quantiles.

Definition 2.4.9. *If F_1 and F_2 are two distribution functions with Q_1 and Q_2 their corresponding quantile functions a QQ-plot is a plot of Q_1 against Q_2.*

R Box 2.4.10 (QQ-plots). If `x` and `y` are numeric vectors then

```
> qqplot(x,y)
```

produces a QQ-plot of the empirical quantiles for `y` against those for `x`.

```
> qqnorm(x)
```

results in a QQ-plot of the empirical quantiles for `x` against the quantiles for the normal distribution.

Usually when making a QQ-plot one of the distributions, F_1, say, is empirical. It is then common only to plot

$$(Q_2(i/n), x_{(i)}), \quad i = 1, \ldots, n-1,$$

choosing the generalised inverse of F_1 as quantile function. If the empirical quantile function Q_1 is created from a realisation of n iid stochastic variables having distribution function F with quantile function Q_2 then the points in the QQ-plot should lie close to a straight line with slope 1 and intercept 0. It can be beneficial to plot the straight line to be able to visualise any discrepancies from the straight line.

We are often interested in comparing the empirical distribution with a distribution where we know the *form* of the distribution but not the location and scale. If X has distribution with quantile function Q_2 and our dataset is a realisation of n iid stochastic variables each having the same distribution as

$$\sigma X + \mu$$

for some unknown scale $\sigma > 0$ and position $\mu \in \mathbb{R}$, then if we make a QQ-plot of the empirical quantile function against Q_2 it will still result in points that lie close to a straight line, but with different slope and intercept. We typically normalise the dataset

$$x_i' = \frac{x_i - \hat{\mu}_n}{\hat{\sigma}_n}$$

and make the QQ-plot of the empirical quantile function Q_n' based on the normalised dataset against Q_2 in which case we should get points approximately on a straight line with slope 1 and intercept 0.

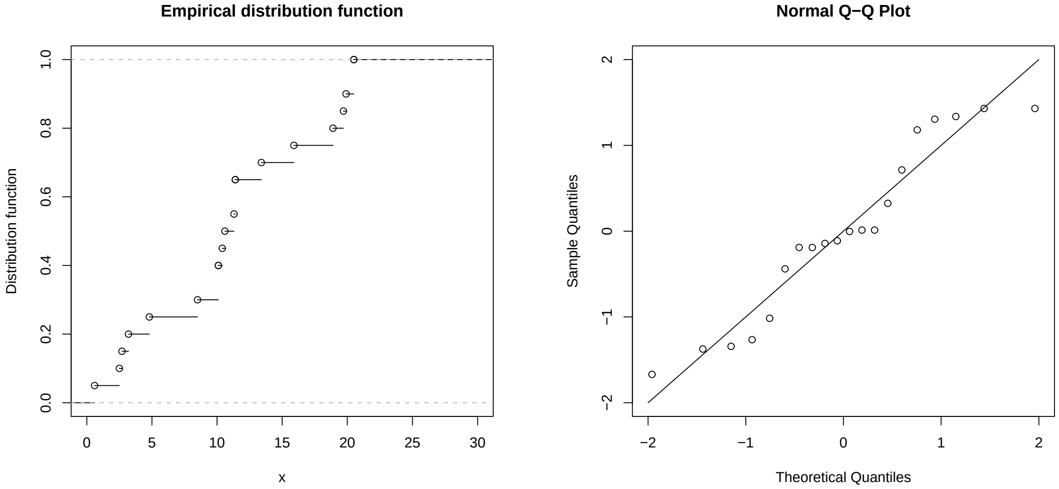

Figure 2.2: The right graph shows the empirical distribution function for the dataset in Example 2.4.11. The left graph shows a QQ-plot of the normalised dataset against the normal distribution with mean 0 and variance 1.

Example 2.4.11. Consider the dataset

2.5	4.8	11.4	0.6	13.4	10.6	10.1	10.4	3.2	11.4
20.5	11.3	19.7	18.9	19.9	8.5	10.1	2.7	15.9	20.5

consisting of 20 real values. The empirical distribution function is shown in Figure 2.2. The mean and variance are

$$\hat{\mu}_{20} = 11.3 \quad \text{and} \quad \hat{\sigma}_n^2 = 41.2.$$

To compare the empirical distribution with a normal distribution, say, we normalise the observations to have mean 0 and variance 1. On Figure 2.2 we also find the QQ-plot of the normalised data against the normal distribution.

One could also compare distribution functions directly instead of comparing quantile functions. It is, however, often more difficult to see the differences between two distribution functions. Especially if the differences are mostly occurring in the tails of the distribution functions. Then the differences will show up nicely on a QQ-plot but may be undetectable by comparing distribution functions directly.

Histograms are useful for representing a single empirical distribution and QQ-plots are valuable for comparing an empirical distribution with another empirical distribution or a theoretical distribution. The *box plot* is a useful tool for visualising and comparing three or more empirical distributions. It may also be useful for visualising just a single empirical distribution if all you want is a rough picture of location and scale.

R Box 2.4.12 (Box plots). For a numeric vector `x` we get a single box plot by

```
> boxplot(x)
```

If `x` is a dataframe the command will instead produce (in one figure) a box plot of each column. By specifying the `range` parameter (= whisker coefficient), which by default equals 1.5, we can change the length of the whiskers.

```
> boxplot(x,range=1)
```

produces a box plot with whisker coefficient 1.

Definition 2.4.13. *One defines a box plot using quantile function Q and whisker coefficient $c > 0$ in terms of a five-dimensional vector*

$$(w_1, q_1, q_2, q_2, w_2)$$

with $w_1 \leq q_1 \leq q_2 \leq q_3 \leq w_2$. Here

$$q_1 = Q(0.25), \quad q_2 = Q(0.5), \quad q_3 = Q(0.75)$$

are the three quartiles and

$$\begin{aligned} w_1 &= \min\{x_i \mid x_i \geq q_1 - c(q_3 - q_1)\} \\ w_2 &= \max\{x_i \mid x_i \leq q_3 + c(q_3 - q_1)\} \end{aligned}$$

are called the whiskers. The box plot is drawn as a vertical box from q_1 to q_3 with "whiskers" going out to w_1 and w_2. If datapoints lie outside the whiskers they are often plotted as points.

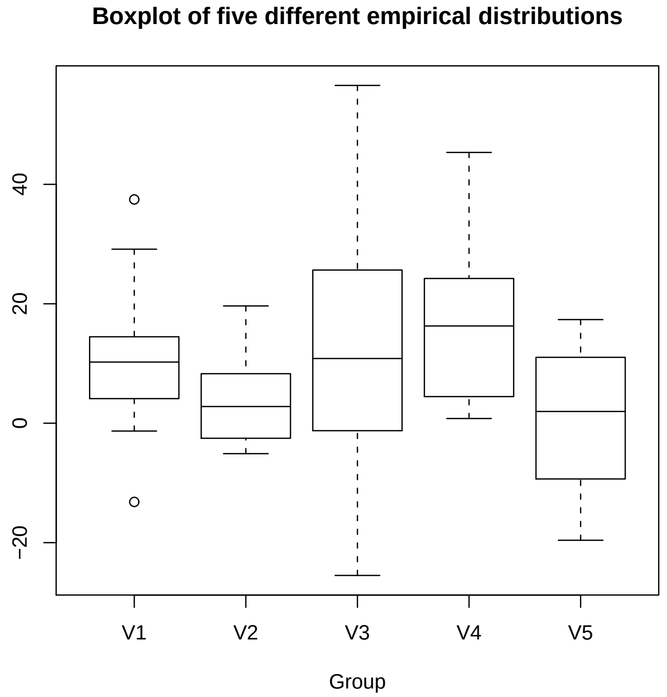

Figure 2.3: A box plot of the five datasets from Example 2.4.14. Note that in all cases except the first the whiskers stretch out to the most extreme observations.

Example 2.4.14. Consider the following five datasets

12.99	-0.28	-1.84	16.15	-15.85
6.09	-2.44	-9.64	3.57	3.81
14.36	5.46	2.24	1.50	5.66
5.10	-1.17	-21.74	26.87	-10.57
29.14	-4.81	-19.14	6.90	8.62
7.05	14.28	38.41	45.34	10.80
27.15	-2.60	56.56	11.69	0.13
-1.31	-2.77	31.15	0.87	8.59
10.68	1.33	-0.59	16.42	17.35
14.21	-4.68	33.58	31.18	-1.18
2.48	-1.93	9.04	37.60	-16.90
3.13	15.64	20.80	22.54	11.27
9.02	18.14	-25.49	5.34	-9.57
14.59	10.00	-0.67	25.93	-6.84
18.01	6.58	19.77	22.21	13.19
11.64	6.16	15.97	21.92	-9.10
37.47	5.75	15.08	17.56	12.58
-13.18	-5.10	7.39	2.30	-3.54
-0.49	4.26	12.63	14.43	-19.60
9.80	19.64	30.50	0.79	12.73

We want to compare these five datasets in one figure. We do so by drawing a box plot of each dataset in the same figure, cf. Figure 2.3. The figure shows that the five datasets all seems to differ from each other in one way or the other. The five datasets are simulations of 20 stochastic variables having distribution (in order) $N(10, 10)$, $N(5, 10)$, $N(10, 20)$, exponentially distributed with intensity parameter 1/15, and uniformly distributed on $[-20, 20]$.

2.5 Simulating from the empirical distribution

We may be interested in *simulating* from the unknown probability measure $\mathbb{P}$ that models the experiment with sample space E. All we have though is the observations $x_1, \ldots, x_n$ being the realisation of n iid stochastic variables $X_1, \ldots, X_n$ with distribution $\mathbb{P}$. Since the empirical measure is an approximation of $\mathbb{P}$ we can choose to simulate from ε_n instead as an approximation to simulating from $\mathbb{P}$. Being able to do so will play an important role a subsequent chapter. There are several possible approaches for simulating from ε_n – some more intelligent than others.

First of all one should observe that the empirical probability measure is always essentially a probability measure on a discrete set. This is because even if our sample space E is not discrete the empirical probability measure only assigns probabilities to the finite number of points $x_1, \ldots, x_n$ observed. That is, the sample space under the empirical measure can be restricted to

$$E_n = \{z \in E \mid z = x_i \text{ for some } i = 1, \ldots, n\} \subseteq E$$

and the empirical probability measure has point probabilities

$$\varepsilon_n(z) = \frac{1}{n} \sum_{i=1}^{n} 1(x_i = z)$$

for $z \in E_n$. If all the x_i's differ, which is the typical situation, then E_n simply consists of the n different observed values each with the point probability $1/n$. In this case the empirical distribution is the uniform distribution on E_n. Note, however, that the definition of E_n depends on the observed values $x_1, \ldots, x_n$.

Since ε_n is a probability measure on the discrete set E_n we can use Algorithm 1.8.3. This is certainly a possibility but applying the algorithm literally requires that we construct intervals $I(z)$ for each $z \in E_n$ with length $\varepsilon_n(z)$. Doing this directly for each new dataset is not necessary. The solution presented below in Theorem 2.5.1 is an adaption of Algorithm 1.8.3 suitable for simulating from any empirical measure.

If $E \subseteq \mathbb{R}$ it is also possible to use Algorithm 1.8.6 with the generalised inverse distribution function $F_n^{\leftarrow}$ of the empirical distribution function F_n. Although a possibility this approach is first of all restricted to sample spaces being subsets of $\mathbb{R}$ and secondly we need to construct the generalised inverse distribution function for each new dataset.

The recommended approach is to sample *indices* of the dataset uniformly. This approach is efficient and completely generic. The implementation requires no knowledge about the sample space whatsoever.

Theorem 2.5.1. *Let $x_1, \ldots, x_n$ be a dataset with values in the sample space E and corresponding empirical measure. If U is uniformly distributed on $\{1, \ldots, n\}$ then the distribution of*

$$X = x_U$$

is the empirical measure ε_n.

Proof: With $E_n = \{z \in E \mid x_i = z \text{ for some } i = 1, \ldots, n\}$ we find that for $z \in E_n$ and with $I_z = \{i \in \{1, \ldots, n\} \mid x_i = z\}$ then

$$\mathbb{P}(X = z) = \mathbb{P}(x_U = z) = \mathbb{P}(U \in I_z) = \frac{|I_z|}{n} = \frac{1}{n} \sum_{i=1}^{n} 1(x_i = z) = \varepsilon_n(z).$$

□

Remark 2.5.2. It follows from the theorem that if $U_1, \ldots, U_B$ are B iid uniformly distributed stochastic variables taking values in $\{1, \ldots, n\}$ then $X_1, \ldots, X_B$ defined by

$$X_i = x_{U_i}$$

for $i = 1, \ldots, B$ are iid with distribution ε_n. Taking $U_1, \ldots, U_B$ to be iid uniformly from $\{1, \ldots, n\}$ is known as *sampling with replacement* B times from $\{1, \ldots, n\}$. The stochastic variables $X_1, \ldots, X_B$ can therefore be regarded as B samples with replacement from the set $\{x_1, \ldots, x_n\}$. How we chose to perform the simulation of $U_1, \ldots, U_B$ is another issue. It could be done by Algorithm 1.8.3, but the details are not important. The implementation of that simulation can be done once and for all and optimised sufficiently.

R Box 2.5.3 (Simulation from the empirical measure). If `x` is a vector of length n containing the dataset $x_1, \ldots, x_n$ we can obtain a sample of size $B = 1000$ from the empirical measure by

```
> y <- sample(x,1000,replace=TRUE)
```

The vector `y` then contains 1000 simulations from the empirical measure. Note the parameter `replace` which by default is `FALSE`.