Basics of probability theory

Alexander Ritz

30. Juli 2020

1 Preface

The following introduction is meant to be a very basic overview of the most relevant theoretical concepts utilised in the field of statistics. For a more useful, thorough introduction and the more technical formalisations, the book **Measure Theory: Second Edition** by Donald L. Cohn is wholeheartedly recommended. More advanced texts can be recommended to enthusiastic students on request, since recommendations on an individual basis are likely to result in higher long-term satisfaction... or rather fewer regrets about money ill-spent.
This text is heavily inspired by the very insightful work of Dr. Klaus Schindler, especially his mathematically inspiring lecture „Mathematics D: Pricing of Financial Derivatives“.
Suggestions, corrections and questions regarding this script can be addressed at alexander.ritz@stud.uni-goettingen.de

2 Sample spaces

Let Ω denote the set of all potential outcomes or results ω of a random experiment. The set Ω is usually called sample space or possibility space, subsets of Ω are called events. Given an outcome ω, we say, „The event A was realised“, if $\omega \in A$ holds. In case $\omega \notin A$ we say, „The event A was not realised“. An outcome is called known, if it was either realised or not realised.

3 σ-algebras

The realisation of an outcome ω does not just contain information about the occurence of singular outcomes, but more complex composite outcomes [1]. Are A and B known outcomes, then set-theoretic reasons imply the knowledge of further outcomes. For example A^{C}, $A \cap B$ or $A \cup B$. A system $\mathscr{A}$ of observable results which adheres to these more or less intuitive set-theoretic properties is called σ-algebra. More formally:

A system $\mathscr{A}$ of subsets of Ω is called σ-Algebra in Ω, if it has the following properties:

- $\Omega \in \mathscr{A}$
- $A \in \mathscr{A} \implies A^{\mathrm{C}} \in \mathscr{A}$
- $A_1, A_2, A_3, ... \in \mathscr{A} \implies \bigcup_{l \in \mathbb{N}} A_l \in \mathscr{A}$

A tuple $(\Omega, \mathscr{A})$, consisting of a sample space Ω and a σ-algebra $A \subset \mathcal{P}(\Omega)$ is called measure space.[2]

[1] This information-theoretic interpretation of σ-algebras was established by the German mathematician Klaus Schindler.

[2] Please note that this „stability over unions“(the last property) only holds in case of a countable number of unions! Uncountable unions of elements of σ-algebra $\mathscr{A}$ do not generally lie in $\mathscr{A}$ as well! Appropriate caution is therefore recommended when working at the interface of discrete and continuous random variables.

4 Probability spaces

Although it is usually impossible to tell which outcomes are going to be realised, it is very often possible to give an opinion of the likelihood of an outcome in a specific σ-algebra, given its relative plausability. This can be formalised by assigning numeric values between 0 and 1,which are then called probability. If $A \subset \Omega$ is an event, then $P(A)$ denotes the probability of A being realised. The function P is called probability measure. Based on set-theoretic[3] considerations it is sensible to demand certain properties from this measure. Are A und B disjoint events, then $P(A \cup B) = P(A) + P(B)$ should hold. Additionally, the probabilities of all events in the given σ-algebra should be calculable. These considerations lead to the following definition.

Let $\mathscr{A}$ be a σ-algebra in the sample space Ω. A function $P : \mathscr{A} \to [0, 1]$ is called probability measure on $\mathscr{A}$, if

- $P(\Omega) = 1$
- P is σ-additive, that is, for all sequences of pairwise disjoint sets $A_1, A_2, ...$ holds:

$$P\left(\dot{\bigcup_{i\in\mathbb{N}}} A_i\right) = \sum_{i\in\mathbb{N}} P(A_i)$$

The triple $(\Omega, \mathscr{A}, P)$ is called probability space.

5 Random Variables and Measurability

As general and elegant as the concept of probability spaces is[4], it does not lend itself to applicability, since a complete determination of the entire sample space Ω is generally impossible or impractical due to the complexity involved. One is therefore going to restrict the model to the relevant data and outcomes that are of particular interest. These quantities, for example stock prices or temperatures, with values depending upon the random future outcomes are called random quantities.

[3] One could more correctly speak of *measure*-theoretic.

[4] This basic axiomatic model was established by the Russian mathematician Andrei Kolmogoroff.

A function on the sample space Ω

$$X : \Omega \to \mathbb{R}^d \quad \text{with} \quad \omega \mapsto X(\omega)$$

is called random quantity. In case of $d = 1$ X is called a random variable. In case of $d > 1$ X is a vector of random variables, that is $X = (X_1, ..., X_d)$. In that case, X is called d-dimensional random vector.[5]
Instead of modelling every single potential event one is going to restrict oneself to those which are related to a given random quantity X. Because of the unpredictable stochastic nature of our environment, it is sensible to examine the realisation of intervals of potential values of X instead of singular points. Of predominant interest are events in Ω, for which X assumes values within a specific interval, the preimage or inverse image

$$X^{-1}(]-\infty, z]) = \{\omega \in \Omega \mid -\infty < X(\omega) \leq z\} =: \{X \leq z\}$$

One can only call a random variable X „manageable“if these events are observable or „measurable“, i.e. if their probability is calculable. Formally, this means that they have to be elements of the domain of definition of the probability measure. This would make them elements of the related σ-algebra. Measurability is often seen as a necessary property of random quantities, and we will assume measurability in future applications involving random variables, unless otherwise stated.

A random quantity $X : \Omega \to \mathbb{R}^d$ is called measurable with respect to the σ-algebra $\mathscr{A}$, if:

$$\forall z \in \mathbb{R}^d : \{X \leq z\} \in \mathscr{A}$$

with $\{X \leq z\} \in \mathscr{A}$ being shorthand for the set of outcomes $\omega \in \Omega$, for which the function $X = (X_1, ..., X_d)$ results in values below $z = (z_1, ..., z_d)$, i.e.:

$$\{X \leq z\} = \{\omega \in \Omega \mid X(\omega) \leq z\} = \{\omega \in \Omega \mid X_1(\omega) \leq z_1, ..., X_d(\omega) \leq z_d\}$$

[5]Please note that this definition was restricted to the case of real valued random quantities, for reasons of ease of understanding. The concept is trivially generalised though.

6 Distribution of Random Variables

Let $X : \Omega \to \mathbb{R}^d$ be a measurable random variable on the measure space $(\Omega, \mathscr{A})$. The probabilities of *all* events belonging to X as a whole are called *probability distribution* or *cumulative distribution* of the random variable X. The entire distribution is determined by the function:

$$F_X : \mathbb{R}^d \to [0,1] \tag{1}$$

$$z \mapsto F_X(z) := P(\{X \leq z\}) = P(\{X_1 \leq z_1, \ldots, X_d \leq z_d\}), \tag{2}$$

leading to F_X being called *distribution function* of X. $F_X(z)$ denotes the probability of X assuming values lying within the d-dimensional interval $]-\infty, z_1] \times \cdots \times]-\infty, z_d]$. In case the distribution function is differentiable, F'_X is called *probability density* or just *density* of X.
Distribution functions offer an especially easy method for calculating probabilities of the structure $P(\{a < X \leq b\})$. Using *Riemann-Stieltjes integrals* (see Appendix), one can write:

$$P(\{a < X \leq b\}) = F_X(b) - F_X(a) = \int_a^b \mathrm{d}F_X(z)$$

Given that F_X is continuous.
In case F_X is differentiable, the probabilities can be given as:

$$\int_a^b \mathrm{d}F_X(z) = \int_a^b F'_X(z)\,\mathrm{d}z$$

Under the further restriction that F'_X be continuous, the following holds for infinitesimal changes $\mathrm{d}z$:

$$P(\{z < X \leq z + \mathrm{d}z\}) = F_X(z + \mathrm{d}z) - F_X(z) = F'_X(z)\,\mathrm{d}z = \mathrm{d}F_X(z)$$

Accordingly the density F'_X can be said to measure the chance of X assuming a value within a small environment of z. It can *not* be understood as the probability of X taking the value z. As for continuous F'_X, the fundamental theorem of calculus offers the insight that:

$$P(\{X = z\}) = \int_z^z F'_X(t)\,\mathrm{d}t = F_X(z) - F_X(z) = 0$$

A rather convincing argument against the interpretation as a probability can also be encountered when considering the density of normal or log-normal random variables, which can take values strictly greater than 1.
The distribution F_X of a random variable X is commonly called *pushforward measure*[6] of P under X. It is then denoted as P_X or $X(P)$, since X „transports "the probability measure P onto the real numbers. Meaning that

$$P_X([a,b]) := P(\{a < X \leq b\}) = \int_a^b \mathrm{d}F_X(z)$$

defines a probability measure on the σ-algebra generated by the real-valued intervals[7], i.e. on the *Borel sets*. All integrals related to P_X can be calculated with the help of the Riemann-Stieltjes integrals generated by the distribution function F_X. Therefore, the concept of distributions allows transferring the work with multiple random variables (on potentially differing sample spaces Ω) onto a common, shared space. If a density f exists, meaning $P_X(]a,b]) = \int_a^b \mathrm{d}P_X(z) = \int_a^b \mathrm{d}F_X(z) = \int_a^b f(z)\,\mathrm{d}z$, then we can note this in the form of

$$\mathrm{d}P_X(z) = f(z)\,\mathrm{d}z$$

or

$$\frac{\mathrm{d}P_X(z)}{\mathrm{d}z} = f(z)$$

The exact conditions under which a density f with $\mathrm{d}Q = f\,\mathrm{d}P$ exists for two measures Q and P can be found in the *Radon-Nikodym theorem*[8].

7 Moments of random variables

Moments are characteristic quantities of random variables. The most important moments routinely encountered in statistics are the *expected value* and *variance*. In general, expected value and variance are not sufficient to determine a distribution uniquely, but they are nonetheless capable of giving a first impression of a distribution.
Due to the obvious impossibility of determining the value a non-deterministic

[6] More elegantly called „Bildmaß“ in German.
[7] Or rather cuboids in the general case.
[8] The theorem can be found in the appendix, but the mathematics involved are too elaborate to be detailed here.

quantity is about to assume, the utilisation of a „mean“ value seems intuitive. For a deterministic function $F : \mathbb{D} \to \mathbb{R}$ with a finite domain $\mathbb{D} = \{x_1, \ldots, x_n\}$, this would be the sum of the function values weighted by their relative frequency[9]

$$\bar{f} := \frac{1}{n} \sum_{k=1}^{n} f(x_k)$$

In the case of an infinite domain $\mathbb{D} = [a, b]$ $\bar{f}$ assumes the form of an integral accordingly. The weighting „factor“ now being $\frac{\mathrm{d}x}{b-a}$:

$$\bar{f} := \frac{\mathrm{d}x}{b-a} \int_a^b f(x)\,\mathrm{d}x$$

Appropriating this concept in the stochastic case by replacing the relative frequency with the probability of a particular value being assumed, $P(\{z < X \leq z + \mathrm{d}z\}) = F_X(z + \mathrm{d}z) - F_X(z) = \mathrm{d}F_X(z)$, gives

$$\bar{X} := \int_{\mathbb{R}} z\,\mathrm{d}F_X(z)$$

Accordingly, the expected value $\mathbb{E}(X)$ of a random variable X is defined as its mean function value[10]

$$\mathbb{E}(X) = \int_{\mathbb{R}^d} z\,\mathrm{d}F_X(z)$$

The variance $\mathrm{Var}(X)$ of a random variable X is a measure of the strength of dispersion around the average value of X. It is defined as the quadratic deviation of X from its expected value, i.e.

$$\mathrm{Var}(X) = \mathbb{E}(\mid X - \mathbb{E}(X) \mid^2)$$

The variance, being defined as a quadratic quantity, differs in its unit of measurement from X [11]. A better intuition of the strength of dispersion can be had by examining the *standard deviation* $s(X)$

$$s(X) := \sqrt{\mathrm{Var}(X)}$$

[9] Being $\frac{1}{n}$ for all x_k in the case of a deterministic function.

[10] The more commonly associated form of the integral can be found within the appendix, under the definition of the Riemann-Stieltjes integral.

[11] The definition relying on the quadratic deviation instead of absolute deviation can be explained by the non-differntiability of the absolute value function. There are however drawbacks to this convention as well.

A Prerequisite Knowledge

The following two sections will give a very brief overview of the necessary concepts relating to set theory and formal logic[12], while the section following these gives a definition of continuity of functions. The set theory should mostly be familiar from school or earlier university experiences, while the formal logic is restricted to a summary of the usual notation required in order to be able to write mathematical statements with sufficient clarity and succinctness, i.e. defining quantifiers. Students confident in their mathematical literacy can therefore skip these sections without missing out on any mathematical pearls of wisdom.

A.1 Formal logic

A.1.1 Logical connectives

Logical connectives[13] are operators joining singular *propositions* into a compound proposition, with the truth of the compound proposition relying exclusively on the original component propositions.
A trivial example would be the connecting of the propositions „Earth is round.“ and „Earth is flat.“ , the first proposition being true and the second being false[14]. One can now join these propositions with e.g. the logical connective *and*. The resulting compound proposition being „Earth is round and Earth is flat“, which would intuitvely be seen as a false proposition, given its contradictory nature. The usually defined logical connectives are

- $\wedge$, being read as *and*.
- $\vee$, being read as *or*.
- $\neg$, being read as *not*.
- $\Longrightarrow$, being read as *implies*.
- $\Longleftrightarrow$, being read as *if and only if*.

[12] We will ignore the intricacies of different branches of logic and rely on a basic sketch of *First-order logic*, also known as *predicate logic*, *quantificational logic* or *first-order predicate calculus*.
[13] Called „Junktoren“ in German.
[14] Interestingly, not as uncontroversial an example as one would hope.

A.1.2 Quantifiers

In order to permit a short way of writing statements like „For all $x \dots$ holds“ or „There exist $x \dots$ such that“, one defines *quantifiers*, namely the *universal* quantifier $\forall$ and the *existential* quantifer $\exists$.

$$\forall x \in \mathbb{D} : p(x)$$

can be read as „For all x in $\mathbb{D}$, $p(x)$ holds.“.

$$\exists x \in \mathbb{D} : p(x)$$

can be read as „There exists an x in $\mathbb{D}$ such that $p(x)$ holds.“.
It is particularly important to familiarise oneself with quantifiers as many mathematical statements involve a combination of quantifiers.

$$\forall x \in \mathbb{D}\ \exists y \in \mathbb{W} : x = y^2$$

can be read as „For all x in $\mathbb{D}$ there exists at least one $y \in \mathbb{W}$ such that $x = y^2$ holds.“.

$$\exists x \in \mathbb{D}\ \forall y \in \mathbb{W} : x = y^2$$

can be read as „There exists at least one x in $\mathbb{D}$ for all $y \in \mathbb{W}$ such that $x = y^2$ holds.“.

A.2 Basic set theory

A *set* M is a collection of well-defined objects, which are called *elements* of the set. An element m of the set M is denoted as $m \in M$. If m is not an element of M, this is written as $m \notin M$. It is a widespread convention to denote sets with upper case letters and its elements with lower case letters.
Sets can be defined in several ways:

- Explicit construction of the set, i.e. listing of the contained elements seperated by commas and framed by curly brackets. The order of the elements holds no importance and multiple listings of elements remains without consequence. Such that the set of all letters contained in the word „trivial“ can

be written like so:

$$\{r, i, v, t, a, l\} = \{l, t, a, v, i, r\} = \{v, i, r, l, a, t, v\}$$

- Implicit construction of the set, i.e. giving properties $P_1, P_2, \ldots, P_n$ and constructing the set M of all objects satisfying these properties. Written like this:

  $$M := \{x \mid P_1(x) \wedge P_2(x) \wedge \cdots \wedge P_n(x)\},$$

  which can be read as, „The set M of all x with the properties $P_1, P_2, \ldots, P_n$“.

The *empty set*, which does not contain any elements, is denoted by $\emptyset$. It can be constructed implicitly by demanding properties which are not satisfied by any x, e.g. the set $\{x \in \mathbb{N} \mid 0 < x < 1\}$.

It is important to note that the set $\{a\}$, meaning the set containg nothing but the element a, is not the same as the element a. Just like a box containing a hat is not the same as the hat itself.

A non-empty set is called *finite*, if it contains only finitely many elements; $\mid M \mid$ denoting the number of elements.

In order to avoid running afoul of *Russel's paradox*[15], one usually defines a relative universal set G which contains all elements of the interest. A set A is then called *subset* of B, if:

$$\forall x \in G : (x \in A \implies x \in B),$$

which is then denoted by $A \subset B$. The equality of two sets is accordingly defined as

$$\forall x \in G : (x \in A \iff x \in B),$$

denoted by $A = B$ It is important to note that this definition of the subset relation does not exclude the possibility of A and B being equal. A multitude of texts therefore define the given symbol for the case of *strict* subsets, and define the symbol $\subseteq$ for the case of possible equality.

For sets A, B and C, it holds

- $A \subset A$

[15] In German called „Russelsche Antimonie“.

- $(A = B) \iff (A \subset B) \wedge (B \subset A)$
- $(A \subset B) \wedge (B \subset C) \implies (A \subset C)$
- $\emptyset \subset A$

The power set of a set A is given by

$$\mathcal{P}(A) := \{M \mid M \subset A\},$$

being the set of all subsets M of A. It holds that $\mid \mathcal{P}(A) \mid = 2^{|A|}$.

The *union* of sets $A_1, \ldots, A_n$ is defined as

$$\bigcup_{i=1}^{n} A_i := \{x \mid \exists i \in \{1, \ldots, n\} : x \in A_i\}$$

while the *intersection* of sets $A_1, \ldots, A_n$ is defined as

$$\bigcap_{i=1}^{n} A_i := \{x \mid \forall i \in \{1, \ldots, n\} : x \in A_i\}$$

Two sets are called *disjoint*, if their intersection is equal to the empty set. The *set difference* between tow sets A and B is defined as

$$A \setminus B := \{x \in A \mid x \notin B\}$$

A.3 Continuity of functions

Ignoring certain mathematical technicalities, a function $f : \mathbb{D} \to \mathbb{R}^M$ with $\mathbb{D} \subset \mathbb{R}^N$, $N, M \in \mathbb{N}$ is called continuous in x_0, if

$$x_0 \in \mathbb{D} \text{ and } \lim_{x \to x_0} f(x) = f(x_0) = f(\lim_{x \to x_0} x)$$

f is then called continuous on $\mathbb{D}$ if f is continuous in all points of $\mathbb{D}$.

B Integration theory and Radon-Nikodym theorem

Both of the following sections of the appendix are made available for the sake of completeness and out of mathematical fervor, it is not necessary for a sufficient

understanding of the main text. The most immediately useful parts should be the summary of properties of Riemann-Stieltjes integrals. While the Radon-Nikodym theorem is of utmost relevance, it might leave a rather abstract impression on readers unfamiliar with measure theory.

B.1 Riemann-Stieltjes integral

For two functions $f, \alpha : [a, b] \to \mathbb{R}$ and a partition $\mathscr{P} := \{t_0, t_1, \ldots, t_n\}$, one may choose $k = 0, 1, \ldots, n-1$ points $\tau_k \in [t_k, t_{k+1}]$ and construct:

$$I(\mathscr{P}, \tau_k) := \sum_{k=0}^{n-1} f(\tau_k) \cdot (\alpha(t_{k+1}) - \alpha(t_k))$$

If $I(\mathscr{P}, \tau_k)$ converges toward the limit I for $\max_k(t_{k+1} - t_k) \to 0$, independent of the choice of either $\mathscr{P}$ and τ_k, then I is called Riemann-Stieltjes integral of f. It is written as

$$I = \int_a^b f(t) \, \mathrm{d}\alpha(t)$$

When $\alpha(t) = t$, the Riemann integral is obtained as a special case.

If f is continuous on $[a, b]$ and α is differiantiable with bounded derivative, it holds that

$$\int_a^b f(t) \, \mathrm{d}\alpha(t) = \int_a^b f(t) \cdot \alpha'(t) \, \mathrm{d}t$$

, which leads to the usual formula given for the expected value:

$$\mathbb{E}(X) = \int_{\mathbb{R}^d} z \, \mathrm{d}F_X(z) = \int_{\mathbb{R}^d} z \cdot F_X'(z) \, \mathrm{d}z$$

B.2 Radon-Nikodym theorem

Let λ and μ be two positive measures on the measure space $(\Omega, \mathscr{A})$ with

- $0 < \mu(\Omega) < \infty$ and $0 < \lambda(\Omega) < \infty$
- λ being absolutely continuous[16] w.r.t μ, i.e. $\mu(A) = 0 \implies \lambda(A) = 0$ for all $A \in \mathscr{A}$.

[16] A stronger property than „simple“ continuity of a function.

Then a non-negative $\mathscr{A}$-measurable function h exists on Ω, such that

$$\forall A \in \mathscr{A} : \lambda(A) = \int_A h \, \mathrm{d}\mu$$

In particular, it holds for all measurable functions that

$$\int f \, \mathrm{d}\lambda = \int f \cdot h \, \mathrm{d}\mu$$

Oftentimes the theorem is stated with the shorter notation of $\lambda = h \cdot \mu$ and h being called the density of λ w.r.t. μ. Based on its construction, h is also called *Radon-Nikodym derivative*, which is then written as $h = \frac{\mathrm{d}\lambda}{\mathrm{d}\mu}$. This is only natural since the function can be easily related to a derivative as defined within calculus. It should be noted that the function h satisfying the property above is uniquely defined *up to a μ-null set*. Meaning that another function f satisfying the property will be equal to h μ-almost everywhere.