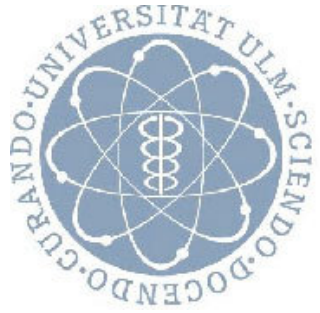
A Refresher in Probability Calculus



VERSION: April 28, 2011

Contents

1 Basic Definitions

This part introduces the concepts of measure theory and Lebesgue integration. While the theory is rather demanding, a general understanding of the subject matter is quite useful when dealing with probability theory. The following text gives an overview of basic concepts and definitions; detailed theory is taught in the Applied Analysis course. We assume most readers are familiar with basic set theory and Riemann integration.

1.1 Measure

The language of modeling financial markets involves that of probability, which in turn involves that of *measure theory*. This originated with Henri Lebesgue (1875-1941), in his thesis, 'Intégrale, longueur, aire'. Measure theory is the study of measures. It generalises notions such as "length", "area", and "volume", though not all of its applications have to do with physical sizes.

A measure on a set Ω is a function which assigns a real number to subsets of Ω; an intuitive way to think about this is that a measure assigns "size" or "volume" for sets.

Example 1.1. *Let C be a finite set. The counting measure of C is defined by $\mu(C) =$ number of elements in C*

Example 1.2. *Let I be the closed interval $[a, b]$ of real numbers. The measure (length) of the interval is $\mu(I) = b - a$. The open interval (a, b) has the same measure, since the points a and b have measure zero.*

Example 1.3. *The probability measure $\mathbb{P}$ is a special kind of measure that tell us the probability of an event to occur. Probability measures are discussed in detail in Section 2.*

When we define a measure we want to assign such a size to every subset of Ω but often this is not possible (i.e. some subsets of Ω are *not measurable*). Instead we focus on a specific collection of subsets of Ω, which are called measurable sets, and which are closed under operations that we would expect to preserve measurability. A σ-algebra is such a collection.

Definition 1.1. *Let $\mathcal{A}_0$ be a collection of subsets of Ω such that*

- $\emptyset \in \mathcal{A}_0$
- *Any union of countably many elements of $\mathcal{A}_0$ is an element of $\mathcal{A}_0$ (i.e. if A_1, A_2, A_3, ... are in $\mathcal{A}_0$, then so is $A = A_1 \cup A_2 \cup A_3 \cup ...$).*
- *The complement of any element of $\mathcal{A}_0$ in Ω is an element of $\mathcal{A}_0$ (i.e. if A is in $\mathcal{A}_0$, then so is its complement, Ω/A).*

Thus a σ-algebra on Ω is a family of subsets of Ω closed under any countable collection of set operations. It follows from the definition that any σ-algebra $\mathcal{A}_0$ in Ω also satisfies:

- $\Omega \in \mathcal{A}_0$.
- Any intersection of countably many elements of $\mathcal{A}_0$ is an element of $\mathcal{A}_0$.

Elements of the σ-algebra are called *measurable sets*. An ordered pair $(\Omega, \mathcal{A}_0)$, where Ω is a set and $\mathcal{A}_0$ is a σ-algebra over Ω, is called a *measurable space*.

Example 1.4. *(σ-algebras over Ω)*

- *The set $\{\emptyset, \Omega\}$*
- *The power set of Ω (i.e. the set of all subsets of Ω, see Definition 4.4)*

- *The Borel σ-algebra in $\mathbb{R}$, $\mathcal{B} = \mathcal{B}(\mathbb{R})$ is the σ-algebra, generated by the open intervals of $\mathbb{R}$ (see Definition 4.6). In other words, the Borel σ-algebra is equal to the intersection of all σ-algebras $\mathcal{A}$ of $\mathbb{R}$ having the property that every open set of $\mathbb{R}$ is an element of $\mathcal{A}$.*

As our aim is to define measures on collection of sets we now turn to set functions. In order to qualify as a measure, a function must satisfy a few conditions. One important condition is countable additivity. This condition states that the size of the union of disjoint subsets is equal to the sum of the sizes of the subsets. Intuitively, one may consider any two disjoint subsets $[a, b]$ and $[c, d]$ of $\mathbb{R}$ and their union $I = [a, b] \cup [c, d]$; the measure function should guarantee that the sum of the two subsets' sizes is the same as the size of I. To be more rigorous, we give the following definition:

Definition 1.2. *Let Ω be a set, $\mathcal{A}$ a σ-algebra on Ω and μ_0 a non-negative set function $\mu_0 : \mathcal{A} \to [0, \infty]$ such that $\mu_0(\emptyset) = 0$. μ_0 is called:*
(i) additive, if $A, B \in \mathcal{A}, A \cap B = \emptyset \Rightarrow \mu_0(A \cup B) = \mu_0(A) + \mu_0(B)$,
(ii) countably additive, if whenever $(A_n)_{n\in\mathbb{N}}$ is a sequence of pairwise disjoint sets in $\mathcal{A}$ with $\bigcup A_n \in \mathcal{A}$ then

$$\mu_0\left(\bigcup_{n=0}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu_0(A_n).$$

We are ready to define the measure function.

Important

Definition 1.3. *Let $(\Omega, \mathcal{A})$ be a measurable space. Let μ be a countably additive map given by*

$$\mu : \mathcal{A} \to [0, \infty]$$

Then μ is called a measure on $(\Omega, \mathcal{A})$. The triple $(\Omega, \mathcal{A}, \mu)$ is called a measure space.

In particular we are interested in a special class of measures - probability measures.

Definition 1.4. *A measure $\mathbb{P}$ on a measurable space $(\Omega, \mathcal{A})$ is called a probability measure if* **Important**

$$\mathbb{P}(\Omega) = 1.$$

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

Remark A probability measure $\mathbb{P}$ satisfies:

(i) $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\Omega) = 1$

(ii) $\mathbb{P}(A_i) \geq 0 \forall i$

(iii) If $A_1, A_2, ...$ are disjoint, $\mathbb{P}(\bigcup_i A_i) = \sum_i \mathbb{P}(A_i)$

Probability measure and probability spaces are discussed in detail in section 2.

In section 2, we will define the expected value - a fundamental concept in probability theory with many practical applications. But in order to define it, we need to introduce the notion of the Lebesgue integral. In this section we discuss Lebesgue integration, named after the French mathematician Henri Lebesgue (1875-1941).

1.2 Lebesgue Integral

The idea of the Lebesgue integral is to extend the class of integrable functions over those that are not Riemann integrable. For functions that are Riemann integrable, Lebesgue theory assigns the same numerical value to $\int_a^b f(x)dx$ as Riemann theory. On the other hand, functions that are not Riemann integrable might still be Lebesgue integrable. In this sense, Lebesgue theory can be thought of as a kind of completion of the Riemann integration theory.

Let $(\Omega, \mathcal{A})$ be a measurable space and let μ be a measure on $(\Omega, \mathcal{A})$. We want to define integration for a suitable class of real valued functions with respect to μ. In Lebesgue theory, integrals are defined for *measurable functions.*

Definition 1.5. *Let $(\Omega, \mathcal{A})$ be a measurable space and let $f : \Omega \to \mathbb{R}$. For $A \subset \mathbb{R}$ define $f^{-1}(X) = \{\omega \in \Omega : f(\omega) \in X\}$. f is called $\mathcal{A}$-measurable (or simply measurable) if*

$$f^{-1}(B) \in \mathcal{A} \text{ for all } B \in \mathcal{B}.$$

First we define the integral for simple functions, which are always measurable. Then we extend the definition for nonnegative measurable functions. In the following two definitions, the function f is always defined as in Definition 1.5.

Definition 1.6. *Let $f = \sum_{i=1}^{n} a_i \mathbf{1}_{\mathcal{A}}$ be a nonnegative simple $\mathcal{A}$-measurable function ($\mathbf{1}_{\mathcal{A}}$ is the indicator function, see Definition 4.8). Then the integral of f is defined as*

$$\int f d\mu := \sum_{i=1}^{n} a_i \mu(\mathcal{A})$$

We are restricted to nonnegative functions since we admit the case $\mu(\mathcal{A}) = \infty$. If we were dealing with a finite measure μ, the definition would work for all $\mathcal{A}$-measurable simple functions.

Example 1.5. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and let $X = \sum_{i=1}^{n} a_i \mathbf{1}_{\mathcal{A}}$ be a simple random variable. Than the expectation of X is given by $E(X) := \int X dP$. This will be useful later in the probability part, when we want to compute expectations.*

We will give precise definition of a random variable and of the expected value in section 2. For the sake of the example, one can think of random variable

as the possible outcomes of some random experiment, such as tossing a coin or a six sided die; the expected value can be regarded as the "average" outcome of this random experiment. Most importantly, we see that to use fundamental concepts such as expectation, we need the notion of the Lebesgue integral.

Definition 1.7. *Let f be a nonnegative measurable function (with possible values ∞ at some points) and let h be a simple function, such that $h(\omega) \leq f(\omega)$ for all $\omega \in \Omega$. Then the integral of f is defined as*

$$\int f d\mu := \sup_{\omega \in \Omega} \{ \int h d\mu \}$$

Now we have defined the integral for every nonnegative measurable function. The value of the integral may be ∞. In order to define the integral for measurable functions which may take both positive and negative values, we have to exclude infinite integrals.

Definition 1.8. *A measurable function f is μ-integrable if $\int f^+ d\mu < \infty$ and $\int f^- d\mu < \infty$, where $f^+ = \max(f, 0)$ and $f^- = \max(-f, 0)$. If f is μ-integrable then*

$$\int f d\mu := \int f^+ d\mu - \int f^- d\mu$$

2 Basic Probability Background

As we remarked in the introduction of this chapter, the mathematical theory of probability can be traced back to 1654, to correspondence between Pascal (1623–1662) and Fermat (1601–1665). However, the theory remained both incomplete and non-rigorous until the 20th century. It turns out that the Lebesgue theory of measure and integral sketched above is exactly the machinery needed to construct a rigorous theory of probability adequate for modeling reality (option pricing, etc.) . This was realised by Kolmogorov (1903-1987), whose classic

book of 1933, *Grundbegriffe der Wahrscheinlichkeitsrechnung* (Foundations of Probability Theory), inaugurated the modern era in probability.

2.1 Fundamentals

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space (recall Definition 1.4). We use probability spaces to model a random experiment. To do so, we use a sample space Ω to describe the set of all possible outcomes. An outcome is the result we get from running the experiment once, so depending on the experiment we run and on the set of possible outcomes, we choose different sample spaces .

The σ-algebra $\mathcal{F}$ is a set of subsets of Ω. Intuitively, this is the set of events, containing zero or more outcomes.

The probability measure $\mathbb{P}$ is a function, returning the probability of an event $A \in \mathcal{F}$. The probability is a number between 0 (event never happens) and 1 (event happens in every trial).

Example 2.1. *Consider tossing a fair coin once. The possible outcomes are only two: H(heads) and T(tails) and the sample space is* $\Omega = \{H, T\}$. *By definition, the* σ*-algebra contains the empty set and* Ω. *The other possible events are getting heads or tails. Therefore* $\mathcal{F} = \{\emptyset, H, T, \Omega\}$.
Since the coin is fair, both heads and tails come with 50% *probability, i.e.* $\mathbb{P}(H) = \mathbb{P}(T) = \frac{1}{2}$. *We toss the coin once, so we obtain either heads or tails but certainly something which is in* Ω, *thus* $\mathbb{P}(\Omega) = 1$. *Furthermore, since we obtain something, we cannot get the empty set, namely* $\mathbb{P}(\emptyset) = 0$.
Note that $\mathbb{P}(\Omega)$ *is the probability of getting either heads or tails.*

Example 2.2. *Consider the same experiment as above, but with the coin being tossed three times. There are 8 possible outcomes:* Ω = *{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT}. Note that for example HHT is different from HTH, due to the order of the outcomes.*
A σ*-algebra* $\mathcal{F}$ *on* Ω *can be the set of events, containing all the possible combinations of outcomes:*

$$\begin{aligned} \mathcal{F} \;=\; & \{\emptyset, HHH, HHT, ..., \{HHH, HHT\}, \{HHH, HTH\}, ..., \\ & ..., \{HHH, HHT, HHH\}, \{HHH, HHT, HHT\}, ..., \Omega\} \end{aligned}$$

This σ*-algebra contains all the subsets of* Ω *or, equivalently, the* σ*-algebra* $\mathcal{F}$ *is the power set of* Ω *(see Definition 4.4). Since the set* Ω *contains 8 elements, the power set* $\mathcal{F}$ *of* Ω *contains* $2^8 = 256$ *elements (events).*
Now consider the case where we know the number of tails after three tosses. There can be either 0, 1, 2 or 3 tails. The set of outcomes we obtain is $\Omega = \{HHH\} \cup \{HHT, HTH, THH\} \cup \{HTT, THT, TTH\} \cup \{TTT\}$; *accordingly the* σ*-algebra* $\mathcal{F}$ *has* $2^4 = 16$ *events.*

For a set of outcomes $A_1, A_2, ... \in \Omega$, we want the probability measure to satisfy:
(i) $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\Omega) = 1$
(ii) $\mathbb{P}(A_i) \geq 0 \forall i$
(iii) If $A_1, A_2, ...$ are disjoint, $\mathbb{P}(\bigcup_i A_i) = \sum_i \mathbb{P}(A_i)$

These three conditions are known as the Kolmogorov axioms. The reason we require that from $\mathbb{P}$ is the fact that it is a measure and as such it should satisfy all the conditions for a function to be a measure (recall Definition 1.2 and Definition 1.3). But even if we observe the axioms intuitively, their motivation should be obvious.

As mentioned in the first example above, when we run an experiment we always get some outcome and never the empty set, so we have (i). Furthermore, it does not make sense for an event to occur with negative probability, hence (ii). Finally, (iii) comes from the countable additivity property. For example, consider rolling a six sided die, where every side has the probability $\frac{1}{6}$. Than the probability assigned to $\{1, 2, 3\}$ is $\frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$ and to $\{5, 6\}$ is $\frac{1}{6} + \frac{1}{6} = \frac{1}{3}$. This means we get $1, 2$ or 3 half of the time and $5, 6$ only a third of the time - which is what one would expect.

Definition 2.1. *A probability space, or Kolmogorov triple, is a triple* $(\Omega, \mathcal{F}, \mathbb{P})$ *satisfying the Kolmogorov axioms (i), (ii) and (iii) above.*

2.1.1 Random Variables

Often we want to assign a value to each possible outcome of some random experiment. Such values can be used for the analysis of the experiment or to make predictions, based on the obtained data. We quantify outcomes ω by defining a real-valued function X on Ω, i.e. $X : \Omega \to \mathbb{R}$. If such a function is measurable it is called a random variable.

Using the definition of a measurable function, given in 1.5, we can provide a precise definition of the above statement.

Definition 2.2. *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space. A random variable (vector)* X *is a function* $X : \Omega \to \mathbb{R}$ *(*$X : \Omega \to \mathbb{R}^k$*) such that* $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}$ *for all Borel sets* $B \in \mathcal{B}(\mathbb{R})$ *(*$B \in \mathcal{B}(\mathbb{R}^k)$*).*

Example 2.3. *Consider the most basic experiment - tossing a fair coin (i.e. the coin has an equal probability of landing "heads" or "tails"). Let us say if it lands "Heads", we assign a value of 0 and if it lands "Tails" we assign value of 1. The random variable* X *is then given by*

$$X = \begin{cases} 0 & \text{if heads} \\ 1 & \text{if tails} \end{cases}$$

Here the values 0 and 1 might be the value of a bet, or any other numerical value used for quantifying the results of the experiment.

There are two types of random variables - discrete and continuous. Intuitively, discrete data is data we can count. Similarly, the discrete random variables takes values we can "count". In example 2.3, the random variable X takes only two values - 0 and 1. These are values we can count, therefore X is a discrete random variable.

Continuous random variables take infinite number of values (or values we cannot count). For example, the large hand of a clock can take any value between 0 and 60. Let Y be a random variable that takes values $y \in [0, 60]$, whenever the large hand of the clock is at y. We can have $Y = 10$, when it is ten minutes past, but also $Y = 10.1$, or even $Y = 10.01$. We cannot count the values Y takes, therefore Y is a *continuous* random variable.

Perhaps one has a better intuition when dealing with discrete random variables as they take only finite number of distinct values. They include outcomes of simple experiments, such as tossing a coin, rolling a die or drawing a card from a standard deck. Continuous random variables, on the other hand, take an infinite number of possible values. Their value is never an exact point, it is always in the form of an interval, though the interval can be very small. They are usually measurements, such as height, weight, temperature, etc.

A discrete random variable maps outcomes to values of countable sets and each value has a probability greater than or equal to zero. A continuous random variable maps outcomes to values of uncountable set (see Definition 4.3). The probability of any specific value is zero, whereas the probability of some set of values may be positive.

To get an intuition on the types of random variables, consider the following examples.

Example 2.4. *(Discrete)*
Again, consider tossing a fair coin. Now consider a bet that pays you 1 if you toss heads and you pay back 1 if you toss tails. So in case of "heads", we assign a value 1 and in the case of "tails" we assign a value -1. The random variable X is then given by

$$X = \begin{cases} 1 & \text{if heads} \\ -1 & \text{if tails} \end{cases}$$

Example 2.5. *(Continuous)*
Consider picking a random real number in $[0, 1]$ with all parts in the range being equally likely. Since there are infinitely many real numbers in that interval, any real number has probability zero of being selected and X cannot be a specific value. However, we can assign positive probability to any range of values. For example choosing the number to be between $[0, 0.5]$ has a probability $\frac{1}{2}$. So in this

case X = selected interval, unlike the previous example, where X was always some specific value.

2.1.2 Probability Distribution

Random variables are always associated with some probability. As we observed from the previous examples, in the discrete case each value of X has some probability of occurring, and in the continuous case the values of X have some probability of being contained in an interval. In probability theory this is known as the *probability distribution*.

The probability distribution identifies either the probability of each value of a random variable (discrete) or the probability that the value falls within a particular interval (continuous). We do distinguish these different cases, but ultimately our aim is to define a function that gives us the aforementioned probabilities. For discrete random variables it is called the *probability mass function* and for continuous random variables, it is called the *probability density function*. Some authors use the latter term to denote both functions, since they describe relatively similar concepts. In this text, we distinguish between the two terms. We denote the probability mass function by $p(x)$ and the probability density function by $f(x)$.

For a discrete random variable, the *probability mass function* gives the probability that the variable is exactly equal to some value. So in the example of tossing a coin and assigning 1 to "heads" and -1 to "tails", the probability mass function should tell us that $\mathbb{P}(X = 1) = \frac{1}{2}$ and $\mathbb{P}(X = -1) = \frac{1}{2}$. Essentially, if other outcomes existed, their probability should also be given by the same function. Precise definition is given below.

Definition 2.3. *Let X be a discrete random variable. The probability mass function $p_X : \mathbb{R} \to [0, 1]$ is defined as*

$$p_X(x) = \mathbb{P}(X = x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) = x\})$$

In the case where X is a continuous random variable, we interpret the *probability density function* as the relative chance of X taking values within the infinitesimal interval $[x, x + dx]$. The probability for a random variable to fall within a given interval is given by the integral of its density over the set.

Definition 2.4. *Let X be a continuous random variable. The random variable has density function f if*

$$\mathbb{P}[a \leq X \leq b] = \int_a^b f(x)dx$$

Some important probability distributions

- Binomial distribution: Number of successes of an experiment with n trials and success probability $p \in [0,1]$

$$\mathbb{P}(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

- Geometric distribution: Waiting time

$$\mathbb{P}(N = n) = p(1-p)^{n-1} \text{ where } n \in \mathbb{N}.$$

- Poisson distribution:

$$\mathbb{P}(X = k) = e^{-\lambda} \frac{\lambda^k}{k!} \text{ where } \lambda \geq 0.$$

- Uniform distribution:

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

- Exponential distribution:

$$f(x) = \lambda e^{-\lambda x} \mathbf{1}_{\{[0,\infty)\}}(x).$$

- Normal distribution $N(\mu, \sigma)$:

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

- Standard Normal distribution, $N(0,1)$ (a special case of the Standard Normal Distribution with $\mu = 0$ and $\sigma = 1$:

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

Remark
The normal distribution is often used to describe real-valued random variables that tend to cluster around a single mean value. The graph of the associated probability density function is "bell"-shaped, and is known as the Gaussian function or bell curve (see Figure 1). The normal distribution is one of the most fundamental continuous probability distribution due to its role in the central limit theorem, which we discuss in section 2.1.5. The standard normal distribution is a special case of the normal distribution with $\mu = 0$ and $\sigma^2 = 1$. It is often used to compare two or more sets of data and to estimate probabilities of events involving normal distributions.

In probability theory the *cumulative distribution function* , denoted by $F_X(x)$, gives the probability that a random variable X takes on a value less than or equal to a number x, that is $F_X(x) := \mathbb{P}(X \leq x)$. Intuitively it gives the area below the density functions, up to point x and thus describes completely the probability distribution of the random variable.

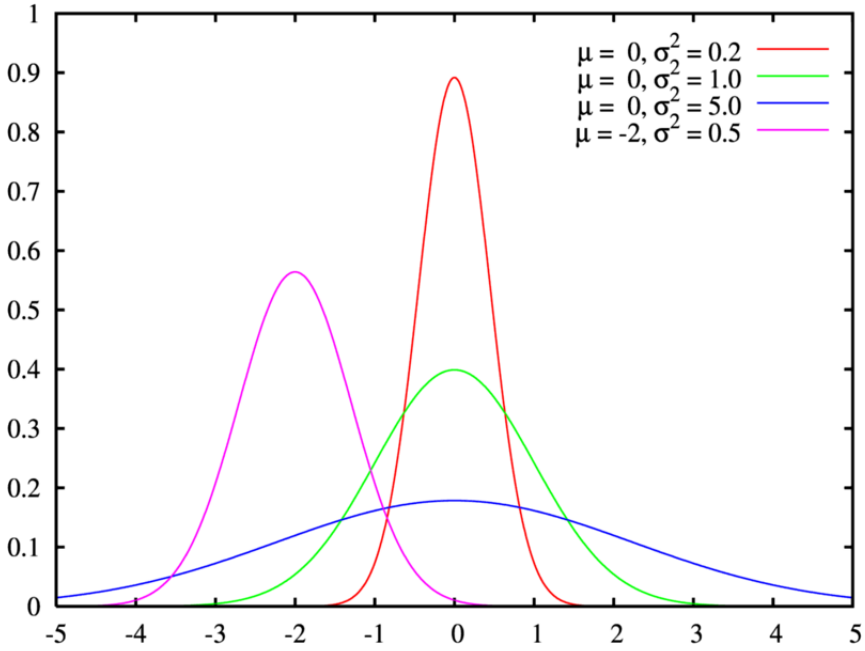


Figure 1: Probability Density Function of Normal Distribution

Definition 2.5. *For every real number x, the Cumulative Density Function of a real-valued random variable X is given by*

$$x \mapsto F_X(x) := \mathbb{P}(X \leq x)$$

Example 2.6. *Let X be discrete random variable, taking values 0 and 1 with equal probability. If $x < 0$, the probability $\mathbb{P}(X \leq x)$ is 0, since X can be either 0 or 1 and therefore it is never less than x. On the other hand, if $x > 1$, than $\mathbb{P}(X \leq x) = 1$, because independent of whether X is 0 or 1, it would always be less than x. Finally, if $x \in [0, 1]$, X would be less x only when $X = 0$; that occurs only half of the times, since X takes 0 and 1 with equal probability. Therefore $\mathbb{P}(X \leq x) = \frac{1}{2}$. The cumulative density function is given by*

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{1}{2} & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

Example 2.7. *Let X be a continuous random variable with uniform distribution on the interval $[0, 1]$ (i.e. all intervals of the same length are equally probable). Let us observe the cumulative density function. For any point $x < 0$, the probability $\mathbb{P}(X \leq x)$ is 0, since X only takes value in $[0, 1]$. If $x \in [0, 1]$ than the probability of X being less or equal to x is equal to the length of the interval. For example, if $x = \frac{1}{2}$, X takes values between 0 and $\frac{1}{2}$ with probability*

$\frac{1}{2}$ *and between* $(-\infty, 0)$ *with probability* 0. *Finally, if* $x > 1$, *than* X *is smaller, thus yielding probability of 1. This can be written as*

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x > 1 \end{cases}$$

Note that regardless of whether X is discrete or continuous, $F_X(x)$ is defined in the same way. Also, note that using definition 2.5 and definition, we can express $F_X(x)$ in terms of $f(x)$.

If X is a discrete random variable that attains values $x_1, x_2, ...$ with probability $\mathbb{P}(x_i)$, then

$$F(x) = \mathbb{P}(X \leq x) = \Sigma_{x_i \leq x} \mathbb{P}(X = x_i) = \Sigma_{x_i \leq x} p(x_i)$$

where $p(x_i)$ is the probability mass function.

If X is a continuous random variable then

$$F(x) = \mathbb{P}(-\infty \leq X \leq x) = \int_{-\infty}^{x} f(x)dx$$

where f is the probability density function.

2.1.3 Expected Value

The term expected value may be confusing in the sense that it is not used to describe the most probable value. In fact it describes the long-term average outcome of a given experiment. For example, if we get a 1 every time we toss heads and pay 1 every time we toss tails with a fair coin, than after having made 1000 tosses we expect to have about as much heads as tails, i.e. we have an expected result of around 0. However, having some expected value does not mean that we will not end up in a situation where we have a result of say 10 or -23. Intuitively, the more experiments we make, i.e. the larger our data set is, the closer we should be getting to the expected value.

If we consider the mean value of a large set of realizations, than intuitively the expected value is the limit of this mean, as the size of the data set increases to infinity. This comes from an important theorem (Law of Large Numbers) which we will state later.

The expected value is not an outcome of the experiment we would expect. For example, a standard six-sided die has an expected value of 3.5, however we cannot expect to get 3.5 when rolling! But if we roll long enough, on average our result would amount to 3.5. Note that we cannot define exactly how long "*long enough*" is, as it approaches infinity.

As another example, casino games usually have a negative expected value. The expected winnings from 1$ bet on a roulette is about -0.05 cents. That does not mean we would always lose; occasionally we can get lucky and win. What the expected value tells us is that the longer we play, the closer our average result would be to the negative 5%.

Now that we have the intuition of what expected value is, we move on to give a proper definition.

A measure determines an integral. A probability measure $\mathbb{P}$, being a special kind of measure determines a special kind of integral, called an expectation or expected value.

Definition 2.6. *The expectation $\mathbb{E}X$ of a random variable X on $(\Omega, \mathcal{F}, \mathbb{P})$ is defined by*

$$\mathbb{E}X := \int_\Omega X(\omega) d\mathbb{P}(\omega), \text{ or } \int_\Omega X d\mathbb{P}.$$

If the integral does not exist (does not converge absolutely), than the random variable does not have an expected value.

If X is a discrete random variable, taking values $x_n (n = 1, 2, \ldots)$ with probability mass function $p(x_n)$, than $\mathbb{E}X$ is given by

$$\mathbb{E}X = \sum_{i=1}^{n} x_i p(x_i).$$

If X is continuous random variable with a density function f, than $\mathbb{E}X$ is given by

$$\mathbb{E}X = \int x f(x) dx$$

Example 2.8. *Let the country Fantasyland assume the following family planning strategy: each family can have children until they either have a girl or two boys. Let X denote the number of boys. We would like to compute the expected number of boys per family .*
There are three possible outcomes: the first child is a girl (happens with probability 0.5, so $\mathbb{P}(X = 0) = 0.5$); the first child is a boy, the second is a girl (probability is $0.5 \times 0.5 = 0.25$ and $\mathbb{P}(X = 1) = 0.25$); both children are boys (probability is again $0.5 \times 0.5 = 0.25$ and $\mathbb{P}(X = 2) = 0.25$). Using the distribution of X, we can compute the expected value as:

$$\mathbb{E}X = 0 \times 0.5 + 1 \times 0.25 + 2 \times 0.25 = 0.75$$

That means out of 100 children, 75 will be boys.

Example 2.9. *Let X have a probability density function*

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{else} \end{cases}$$

i.e. X is uniformly distributed on the interval $[a, b]$. Than the expected value of X is given by

$$\mathbb{E}X = \int_a^b xf(x)dx = \int_a^b \frac{x}{b-a}dx = \frac{x^2}{2(b-a)}\Big|_a^b = \frac{a+b}{2}$$

The expectation - sometimes called the mean - gives the centre of the distribution of the variable. It describes the location of a distribution (and so is called a location parameter). Information about how far the values lie from the mean (the scale of a distribution) is obtained by considering the variance.

$$Var(X) := \mathbb{E}X(X - \mathbb{E}(X))^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2 \tag{1}$$

Two variables with the same probability distribution will have the same expected value and variance.

Some Properties

- Monotonicity: If X and Y are random variables and $X \leq Y$ almost surely, than also $\mathbb{E}X \leq \mathbb{E}Y$
- Linearity: For a constant $a \in \mathbb{R}$, $\mathbb{E}(aX) = a\mathbb{E}X$ and $\mathbb{E}(X+Y) = \mathbb{E}X + \mathbb{E}Y$

To empirically estimate the expected value of a random variable, one repeatedly measures observations of the variable and computes the arithmetic mean of the results. If the expected value exists, this procedure estimates the true expected value. The law of large numbers demonstrates that, as the size of the sample gets larger, the average of the results should be close to the expected value and will tend to become closer the more trials are performed. We will give a precise statement of the law of large numbers later, after we introduce the concept of independence.

2.1.4 Independence

Intuitively, two random events are independent if occurrence of one event does not make it neither more, nor less probable, that the other event occurs. For example, getting a 3 on the first roll of a six-sided die and getting 6 on the second roll are independent events. On the other hand, the event of getting a 3 on the first roll and the event of getting a total sum of 8 from two rolls are dependent. The standard definitions says two events A and B are independent if and only if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$

Considering random variables, let X be a real-valued random variable and A be a set of outcomes of some sort (for example the events $\{X \leq a\}$ for $a \in A$). Since

this set has some probability, it makes sense to refer to events of this sort being independent of other events of this sort. That means, two random variables X and Y with outcome sets A_1 and A_2 are independent if and only if the events $X \in A_1$ and $Y \in A_2$ are independent as described above. Mathematically this is defined as follows

Definition 2.7. *The random variables $X_1, \ldots, X_n$ are independent if whenever $A_i \in \mathcal{B}$ (the Borel σ-algebra) for $i = 1, \ldots n$ we have*

$$\mathbb{P}\left(\bigcap_{i=1}^{n}\{X_i \in A_i\}\right) = \prod_{i=1}^{n}\mathbb{P}(\{X_i \in A_i\}).$$

Example 2.10. *Consider the joint event from two experiments: tossing heads with a fair coin and rolling 3 with a standard die. Let X and Y be random variables giving the outcomes of the coin toss and the die roll respectively. Regardless of the order of the experiments, the probability of this event is $\frac{1}{12}$, since there are 12 possible outcomes in total, and all of them have equal probability, since both the coin and the die are fair. Separately, each outcome has probability $\mathbb{P}(X = heads) = \frac{1}{2}$ and $\mathbb{P}(Y = 3) = \frac{1}{6}$ respectively, whose product gives exactly the probability of the joint event. From the definition, the random variables X and Y are independent*

Theorem 2.1 (Multiplication Theorem)**.** *If $X_1, \ldots, X_n$ are independent and $\mathbb{E}|X_i| < \infty, i = 1, \ldots, n$, then*

$$\mathbb{E}\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n}\mathbb{E}(X_i).$$

Theorem 2.1 is especially useful in practice, since it allows us to compute the total expectation as a product of smaller expectations, which should be easier the compute.

A collection of random variables is said to be independent identically distributed (often abbreviated i.i.d.) if each random variable has the same probability distribution as the others and all are mutually independent. A precise definition is given below.

Definition 2.8. *Two random variables X and Y are said to be identically distributed if they are defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and the distribution function F of X is the same as the distribution function F of Y.*

Definition 2.9. *A collection of random variables X_i is said to be independent identically distributed if the X_i's are identically distributed and mutually independent (every finite subfamily of X is independent)*

Example 2.11. *The outcomes of spins of a roulette are i.i.d. If the roulette ball lands on "red", for example, 20 times in a row, the next spin is no more or less likely to be "black" than on any other spin.*

Example 2.12. *The outcomes of coin flips and die rolls are i.i.d.*

2.1.5 The Central Limit Theorem

In this section we introduce the law of larger numbers and the central limit theorem. We use advanced concepts such as almost surely convergence and distribution convergence, which are beyond the scope of this text, therefore we would not define them.

We already have some intuition on the law of large numbers from the discussion in the end of section 2.1.3. It is important as it guarantees stable long-term results for random events. It helps predict the estimation of some random variable for a long period of trials via the expected value. As the name suggests, the law of large numbers applies only when a large number of observations are considered. There is no principle that a small number of observations will converge to the expected value.

Theorem 2.2 (Strong Law of Large Numbers)**.** *If $X_1, X_2, \ldots$ are independent and identically distributed with mean μ, then the sample average converges almost surely to the expected value, that is*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mu \ (n \to \infty) \ \textit{almost surely.}$$

An alternative way to state that is

$$\mathbb{P}(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mu) = 1$$

This reads as: there is a 100% probability that the average of the sample approaches the theoretical mean as the number of observations increases to infinity.

The main result of this section is the same argument carried one stage further. The central limit theorem is one of the most important results of the theory of probability. In its simplest form, it states a remarkable result: given a distribution with a mean μ and a variance σ^2, the distribution of the mean approaches a normal distribution with a mean μ and a variance σ^2/N as the sample size N, increases. This results holds even when the distribution from which the average is computed is decidedly non-normal.

For example, suppose an ordinary coin is tossed 100 times and the number of heads is counted. This is equivalent to scoring 1 for a head and 0 for a tail and computing the total score. Thus, the total number of heads is the sum of 100 independent, identically distributed random variables. By the central limit theorem, the distribution of the total number of heads will approximately be normal. This can be illustrated graphically if one repeats the same experiment many times (the closer to infinity, the better) and plotting a histogram of the

number of heads per trial. After a large number of repetitions, the histogram curve starts to look like the bell-shaped curve of the normal distribution.

We are now ready to properly state the Central Limit Theorem.

Theorem 2.3 (Central Limit Theorem)**.** *If $X_1, X_2, \ldots$ are independent and identically distributed with mean μ and finite variance σ^2, then*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu)/\sigma \to N(0,1) \ (n \to \infty) \text{ in distribution.}$$

where $N(0,1)$ denotes the standard normal distribution. That is, for all $x \in \mathbb{R}$,

$$\mathbb{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu)/\sigma \leq x \right) \to \Phi(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}y^2} dy \ (n \to \infty).$$

Essentially theorem 2.3 states that for a sequence of n i.i.d. random variables, as the sample size increases, the distribution of the sample average of these random variables approaches to the normal distribution with mean μ and variance σ^2/n irrespective of the shape of the common distribution of the individual terms X_i.

3 Statistics Background

3.1 Basics

A major reason for using statistics is to describe and summarize a set of data. A mass of numbers is usually not very informative and the larger the data set, the harder it is to interpret. Therefore we need to find ways that allow us to present the data in clear and comprehensible form. We start with a simple example.

One hundred students sit an examination. After the examination, the papers are marked and given a score between one and a hundred. We are presented with the following results:

22	65	49	56	59	34	9	56	48	62
55	52	78	61	50	62	45	51	61	60
54	58	59	47	50	62	44	55	52	80
51	49	58	46	32	59	57	57	45	56
90	53	56	53	55	55	41	64	33	0
38	57	62	15	48	54	60	50	54	59
67	58	60	43	37	54	59	63	68	60
46	52	56	32	75	57	58	47	45	52
55	51	50	50	69	63	64	49	56	52
37	60	71	26	30	57	56	55	58	61

We want to interpret the results of the exam. The sort of questions we want to answer are:

- How can we describe the results
- Is there a single mark that best describes the results?
- How representative of the overall performance is such a mark?
- If we have the results from last year, how does this year's performance compare?

The answers are not immediately obvious from the raw data we are given. We need to calculate some statistics in order to make it clearer. One thing we can easily do is arrange the numbers in ascending order.

0	9	15	22	26	30	32	32	33	34
37	37	38	41	43	44	45	45	45	46
46	47	47	48	48	49	49	49	50	50
50	50	50	51	51	51	52	52	52	52
52	53	53	54	54	54	54	55	55	55
55	55	55	56	56	56	56	56	56	56
57	57	57	57	57	58	58	58	58	58
59	59	59	59	59	60	60	60	60	60
61	61	61	62	62	62	62	63	63	64
64	65	67	68	69	71	75	78	80	90

It is now easier to see that the scores are between 0 and 90. What's more, we can observe the frequency of each mark. For example 3 people scored 49 and only one scored 71. When we work this out, we see that 56 is the most "popular" mark with frequency of 7. Also, there are some marks that never appear, such as 28, so each of these marks has frequency 0.

We can present this information into a graphical form using a histogram, where the frequency of each mark is represented as a vertical bar.

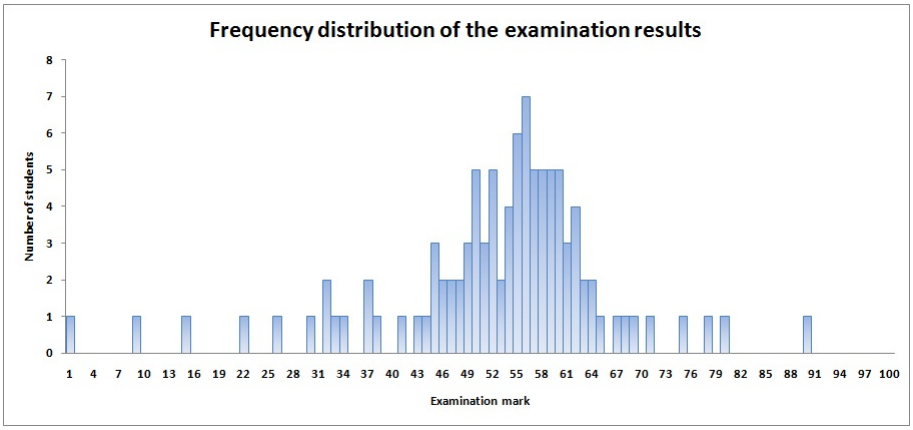


Figure 2: Histogram of marks distribution

In the histogram shown in Figure 2 we list all the possible marks a student could get and draw a bar above each mark with length corresponding to the frequency of the mark. So for mark 56 we draw a bar of length 7 (as 7 students obtained a mark of 56) and for mark 49 we draw a bar of length 3. This gives us a better visual representation of the results. Such frequency distributions are very important in statistical analysis, as they provide the basic representation of the given information.

3.1.1 Measures of central tendency

Is there a single mark that best describes the results? A reasonable approach is to find a "central" mark, that is, to find the center of our data. What we are looking for here is a measure for central tendency.

One possible answer is to select the most frequent mark, which corresponds to the longest bar in the histogram. This measure is called the *mode*. In our case, the mode of the distribution is 56 and it appears to be a reasonable estimate of the central mark. However, mode is rarely used for a number of reasons. It might be the case that we have two marks appearing with the same frequency - than we have no way of choosing between them. Also, it is possible that the mode clearly does not represent the central mark. We go back to the mark distribution from out example, but this time imagine that the ten weakest students all scored zero on the exam. The marks would be clustering around 50, still the mode would be 0. In this case the mode is a poor measure of central tendency.

Another measure of central tendency is called the *median*. It is the score that comes up in the middle of the list when we have ordered it from lowest to highest. For example, if we had only 9 students, the median would have been the fifth mark in our ordered list. If we had 10 students, we would not have a central mark; in this case the median lies halfway between the fifth and the sixth mark. Lying halfway means we are looking for a number that is equally spaced from both marks; to find such a number we add up the marks and divide them by two. In our example, we have 100 students, so the median would be halfway between the fiftieth and fifty-first mark. Since they are both 55, the median is just 55; if they were 55 and 56, the median would have been 55.5.

The median is a good measure of central tendency as it picks the score in the middle position of the distribution. Its weakness is that it takes into account only the central mark (marks), with no regard to the rest of the data. That is, the median does not include all the information given by the marks. If the mark of a student who scored 22, is corrected to 29, the median does not change. It would still not change if more than one marks are changed, as long as marks below the median were not changed to marks higher than the median or vice versa. The median is simply the score where where we cut the list in two halves.

The median can be regarded as a better choice of central value than the mode, there is a third measure of central tendency that takes into account all the information in the list and is used more often than the either of the above measures. This is the *mean* and we denote it by μ. In order to compute the mean, we use the formula

$$\mu = \frac{\sum X}{N} \tag{2}$$

where X indicates a score (in our example an examination mark) and N is the

number of scores (in our example the number of marks or students). So in order to find the mean of the distribution in our example, we add up all the marks and divide them by the number of students. This gives us a mean of 52.62.

When we talk of an "average", we are usually referring to the mean. One way to regard the mean is to see it as the value which balances the scores on both sides. Imagine the horizontal axis of the distribution given in Figure 2 as a long wooden plank with length 100. The students are sitting on the plank at a position specified by their mark; so there is one student sitting on 90, two students sitting on 32, etc. In order to balance the plank perfectly, we have to put a balancing rod exactly at the mean position. If we change any mark (equivalent to moving a student along the plank), than the mean also changed (the position of the rod shifts in order to maintain the balance). Thus the mean is determined by all the scores, unlike the median.

From the plank analogy, we can see that the mean is sensitive to extreme values. A very large or a very small score would have a greater effect on where the supporting rod stays, than a mark in the middle of the distribution would. That is, a balanced plank would tip much easier if a new person sits near either end rather than near the middle.

3.1.2 Comparing measures of central tendency

So far we have three measures of central tendency - mode, median and mean. Which is the one we should choose? The answer is - the one that is most appropriate. We want the one that best represents the central value of our distribution. Usually that results in choosing the mean, but there are occasions when the mode or the median present a better measure.

The mode is quick and easy to determine once we have the distribution, so it might be used as a rough measure without need of any other computations. Also, with some types of data we cannot calculated the mean and the median. For example, when choosing an exam date, a teacher would suggest some range of possible dates and would probably select the date chosen by the largest number of students. Note that in this case it does not make sense to calculate median and mean; a value of, for example, 18th of July is not useful if it is not in the suggested range of possible dates.

A median is often used if there are abnormally large or small values in the frequency distribution, which would result in the mean giving a distorted idea of the central tendency. As an example, five Volkswagen cars have the following maximal speeds: 190 km/h, 210 km/h, 220 km/h, 250 km/h and 400 km/h (the last one is a sports model). We see most have maximum speed of around 220 km/h, but with the inclusion of the fast sports model gives a mean of 254 km/h. This number might not be appropriate for a central value, as four out of the five cars are slower. Taking the median here is more representative for the central

value.

However, in most cases of data collection, the mean is the chosen value as it takes into account all the scores.

3.1.3 Measures of spread

A useful statistic for summarizing data is the measure of spread. It is important to know how spread the scores are. Two groups of students taking the same exam can have different distributions but the same means. In order to express the difference between the two distributions, we need to use the spreads - it is very likely that the marks for one group are more spread out than that of the other. A small spread of the results is normally seen as a good thing - it means that people are behaving similarly and hence the mean value represents the scores well. A large spread indicates that there are large differences between individual scores and the mean is therefore not representative.

The simplest measure of spread is the *range*. It is the difference between the largest and the smallest score. In the example the highest score is the mark 90 and the lowest is 0, thus the range is 90. This measure is crude, it only sets the boundaries of the scores, but tells us nothing about their general spread. Therefore marks spread evenly between 0 and 90 would have the same range as the ones from our example.

Another way to describe the spread is to calculate *quartiles*. We saw earlier that the median divides the ordered data into halves; the quartiles simply divide it into quarters. The first quartile indicates the score that is one quarter on the way up the list, starting from the lowest score. The second quartile is the score that is two quarters up the list, which is in fact halfway up and is therefore equivalent to the median. The third quartile is the score three quarters up the list and the forth quartiles is the score all the way up the list, namely the highest score. We use Q_i to denote the i^{th} quartile.

In our example, one quarter up the list of a hundred scores lies between the twenty-fifth and the twenty-sixth mark, that is between 48 and 49. Therefore, the first quartile is $Q_1 = 48.5$. Similarly, the third quartile is $Q_3 = 59.5$, that is the mark lying between the seventy-fifth and the seventy-sixth mark. We know the second quartiles is $Q_2 = 55$ as we have already computed the median; the forth quartile is $Q_4 = 90$, the highest score.

A slightly more sophisticated measure of spread than the range is the *interquartile range* - the difference between the first and the third quartile ($Q_3 - Q_1$). This is the range of half the scores, those that are in the middle of the distribution. The reason why it is interesting is that unlike the range, it is not affected by a very high or low score and would therefore represent the spread of the distribution more appropriately.

Calculating quartiles does not use all the information available from the scores in the data. As with the median, some of the scores could be different and yet we would arrive at the same interquartile range. The aim is then to devise a measure that takes into account all the available scores. There are a number of measures of spread that have been developed in this direction. Their common feature is that all begin with the mean. Their idea is as follows: if we take the mean as the central position of our distribution we can compare each score to it and find out how far it deviates from it. If we add up the deviation of each of the scores from the mean, we will have a measure of total variability in the data. If we went, we can then divide by the total number of scores, thus finding the average deviation of each score from the mean.

We can calculate the deviation of a score from the mean simply by computing $X - \mu$, where X is a score μ is the mean. The problem here is that when we add them up, the deviations tend to cancel each other out. In our example, a mark of 55 gives a deviation of $55 - 52.62 = 2.38$ and a mark of 50 gives a deviation of $50 - 52.62 = -2.62$. If we add them up, we get a deviation of -0.24. Both scores are over two marks away from the mean, but they give a total variation less than one. We do not want this - it is not a statistic that represents variability as it really is.

The problem here comes from the minus sign. In fact what it tells us is that the mark is lower than the mean. We are not actually interested in that; what we want to know is how far the score is from the mean. There are two ways to add up deviations so they do not cancel each other out and we end up with a reasonable estimate of the real variability of the scores.

One way is to ignore the negative signs altogether, i.e. to take the *absolute deviation* from the mean, that is $|X - \mu|$. Thus each deviation contributes a positive number to the total deviation. To find the average deviation, we add up all the deviations for all scores and divide by the number of scores N. We call this mean absolute deviation and compute it by the formula $\frac{\sum |X - \mu|}{N}$. In our example, the mean absolute error is 9.15.

An alternative way to taking absolute values is to square the deviations since the square of a number is always positive. We add up the square of each of the deviations. Then we can divide by the number of scores to find the average of the squared deviations. This value is called *variance*:

$$\text{Variance} = \frac{\sum (X - \mu)^2}{N}$$

In our example the variance is 176.52. The variance does what we want - it gives a large figure when the scores are spread out and a smaller one when they are closer together. Since we are dealing with squared deviations, values that are away from the mean would have more weight than values that are closer to

it. For example, if a score deviates by 2 from the mean, in contributes 4 to the variance; however, a score deviating by 4 would contribute 16, so even though the second score is only twice as far as from the mean as the first, it contributes four times as much to the variance.

The variance provides a good measure of variability. But note that in our example, the variance we calculated cannot be placed on the frequency distribution as a distance from the mean. This is because the variance is the average of the *squared* deviations. We need to undo the squaring in order to go back to the terms we started with. That is we take a square root of the variance and call this statistic the *standard deviation* σ.

$$\sigma = \sqrt{\frac{\sum(X-\mu)^2}{N}} \tag{3}$$

In our example, the standard deviation of the one hundred marks is 13.29. The standard deviation gives us a measure of the spread about the mean. In many cases, most of the scores would lie within one standard deviation form the mean, that is in the range $X - \sigma$ to $X + \sigma$

3.1.4 Comparing measures of spread

As with the measures of central tendency, the measure of spread that is most useful depends on the reasons for calculating it. The range and the interquartile range are both easy to calculate, giving limited but potentially adequate measures of the spread. Their weakness is that they do not take into account all the scores and may not represent the true variability of the scores.

The variance is a good measure of variation of the data - it takes into account all the scores and gives a small number if the scores are clustered together and a large number if they are spread out. However, when describing a set of data, the variance might not be particularly useful, as it produces a number that is of different order than the scores.

The mean absolute deviation and the standard deviation are both good descriptive statistics of the spread of a set of scores. They both use all the available information and produce a value that expresses the average deviation from the mean in terms that we want (in our example - marks). As they are expressed in the same terms as the scores, both are easy to handle.

3.1.5 Comparing two sets of data

So far we have managed to briefly describe the given data, but in most cases we would like to use that information to make a certain point. In our example we

might be concerned that students perform worse, which might be due to failing standards of education, stricter grading or poor selection criteria. The statistics we have so far can help us make a decision about such question. This, of course, requires a comparison with the previous year's results. The calculation of statistics is often used not only for description but to allow us to answer specific research question and thus involves comparing two sets of results.

To extend our example, assume we are given the results or last year's exam, where again 100 students sat the exam (table of results is not given here). We already know we can order the data in increasing order and create a frequency distribution. If we compare the results from the two years just by looking at Figure 2 and Figure 3 we note that the distribution looks similar over the two years. This can be important observation, indicating a consistency in the performance between consecutive years. However, just by looking at the distributions, we cannot tell how similar they are and we might miss many subtle differences between them.

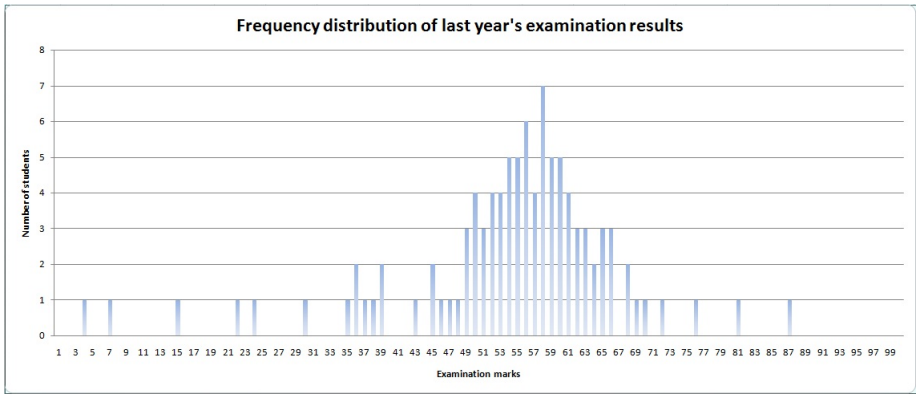


Figure 3: Histogram of last year's marks distribution

We can use the measures of central tendency to compare the two years directly.

	Last year	This year
Mode	58	56
Median	56.5	55.0
Mean	54.25	52.62

We can see that all three measures have dropped a little since the last year. The change in the mode could easily have been caused by just a few students, so in this case it is not the most useful measure. The median indicates that the central point was higher last year. Most importantly, the mean value shows a drop of 1.63 marks. It is important to recall that the mean takes into account all the students, so that means there is a drop of 1.63 marks *per student*. This could be due to a number of reasons that are worth investigating - perhaps the

exam was easier last year; or students were better last year; or perhaps there were a few particularly good students last year or a few poor students this year, which would change the mean, but would not indicate that the standards are failing. To check the last hypothesis, we need to compare the measures of spread

	Last year	This year
Range	83	90
Interquartile range	10.5	11.0
Mean absolute deviation	8.82	9.15
Variance	169.93	176.52
Standard deviation	13.04	13.29

There was a narrower range last year with no one scoring as low as 0 and as high as 90. However, there is not much difference in the interquartile range and, more importantly, in the standard deviation. It might be worth researching further to see why there was reduction in the mean performance. Note that the results alone tell us a difference has occurred, they do not tell us what the reasons for that difference are. Interpretation of the results is up to one's own judgement.

3.1.6 Comparing scores from different distributions

If a student took an exam and scored 59, how good had he done, relative to his classmates? Was he among the best or the worst? If we know the mean and the standard deviation, we can begin answering that question. For example, if the mean is 50 and the standard deviation is 5, than the score was one of the best. However, if the mean is 60 with standard deviation 3, the score is slightly lower than the mean; however, the results are clustered around 60, so probably there are a lot of other students with similar grades.

If a student took two exams, and receive 58 in Mathematics and 49 in Finance, which mark would the student be happier with? 58 is numerically higher so that might be a good first guess. However, if the student finds out most of the students who took Mathematics scored more than 60 and most of those who took Finance scored no more than 45, than things look different. The distributions of the two sets are different, making 49 in Finance a very high mark, compared with the rest of the class and 58 in Mathematics a very low mark.

Assume the student finds out that for the Mathematics exam the mean is 61 and the standard deviation is 6, and for Finance exam the mean is 45 and the standard deviation is 4. To compare the two scores, we need to standardise them. To do so, we compute the *standard score* or *z-score*. This expresses the score relative to the mean in terms of the standard deviation. So for example, a score of 58 is 3 marks away from the mean of 61. The standard deviation is 6 marks, so the score is 3/6th, or half a standard deviations away from the

mean. Essentially, the standard score tells us how many standard deviations the score is from the mean of the distribution. We use the following formula for the standard score:

$$z = \frac{X - \mu}{\sigma}$$

where X is the score we want standardised, μ is the mean and $sigma$ is the standard deviation of the distribution.

We can compare standard scores, because no matter what distribution we start with, converting the scores to z-scores results in distribution with mean of 0 and standard deviation of 1. We compare the standard scores and see which result is higher.

In Mathematics $X = 58$, $\mu = 61$, $\sigma = 6$

$$z = \frac{X - \mu}{\sigma} = \frac{58 - 61}{6} = -0.5$$

In Finance, $X = 49$, $\mu = 45$, $\sigma = 4$

$$z = \frac{X - \mu}{\sigma} = \frac{49 - 45}{4} = 1$$

In Mathematics the student is half a standard deviation below the mean (due to the negative sign) and in Finance the student is a standard deviation above the mean. The higher z-score in Finance means that the student is higher in the class results for Finance that for Mathematics.

In the previous example we compared two sets of examination results, from this year and from last year. For this year a score of 59 yields the following z score:

$$z = \frac{59 - 52.62}{13.29} = 0.48$$

For last year's distribution, a score of 59 produces

$$z = \frac{59 - 54.25}{13.04} = 0.36$$

From the two z-scores we can see a score of 59 is higher up the distribution this year, than it was last year, so 59 is a better score this year.

3.2 Hypothesis Testing

So far we have seen that frequency distributions can be described by choosing appropriate statistics, usually the mean and the standard deviation. Furthermore, we can compare scores from different distributions using standard scores. Now we need to see how to use this information to help us answer the question we wish our research to answer. In this chapter we move from simply describing the data to how we can use it to test hypotheses.

A hypothesis is a supposition - we state something we suppose to be true and then collect evidence towards proving it. Imagine we are given a coin and asked to determine whether it is fair or biased. If we flipped the coin 100 times and it came heads 55 times, it is likely to say the coin is fair. If the coin landed heads only 4 times, we would be inclined to think it is biased towards tails. Both statements are hypotheses we want to test. To do that we need a procedure called *hypothesis testing*. Hypothesis testing is a way of systematically quantifying how certain we are of the result of a statistical experiment. It follows a logical sequence of stages from proposing to hypothesis to deciding whether to reject it.

The hypothesis we want to test is also called the *null hypothesis* and is denoted by H_0. It is the statement that is believed to be correct throughout the analysis. The main goal of hypothesis testing is to tell us whether we have enough evidence to reject the null hypothesis. Rejecting the null hypothesis tells us the alternative should be true. In our example, we want to test whether the coin is biased or not, so our null hypothesis should be "the coin is fair". If we manage to reject it, we can assume the alternative - the coin is *not* fair.

It is important to note, that if we cannot reject the null hypothesis, it does not mean we can automatically accept it and claim with certainty that the coin is fair. Not being able to reject the null means we have not found enough evidence in our data, which certainly is different from demonstrating that the null hypothesis is true. What we can do is test the alternative, that is test whether the coin is fair or not. Then the null hypothesis would become "the coin is not fair". If we manage to reject it, we can safely assume the coin is in fact fair.

Let us go back to the example. Our null hypothesis is "the coin is fair". We flip the coin 100 times and it comes up heads 51 times. Our intuition tells us the coin is probably fair, but nothing more. The expected number of heads is 50 and 51 is close enough. But what if we had flipped the coin 100 000 times and it came up heads 51 000 times? In both cases we have 51% heads, but in the second case the coin seems more likely to be biased.

Let us try to quantify our intuition. Let X be a random variable, describing the outcome of a coin toss. It lands heads with probability p and tails with probability $1-p$. X takes a value of 1 if the coin lands heads, and 0 if the coin

lands tails. This is equivalent to writing

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

Now assume we have made 100 coin flips. If X_i is the outcome of the i^{th} coin flip, the random variable

$$Y = \sum_{i=1}^{100} X_i$$

represents the number of heads in the 100 flips.

Let us denote the set of outcomes from the 100 coin flips as O. What we want to calculate is the probability that we observed O, given the null hypothesis is true, or $\mathbb{P}(O|H_0)$. If this probability is sufficiently small, it would mean the outcomes are very unlikely to occur if the null hypothesis is true; therefore we can conclude that the null hypothesis is false. So for example, the probability of observing only 1 head out of 100 coin tosses, given the coin is fair, is very low (it is 0.5^{100}), much lower than 1%. Therefore, with certainty higher than 99% we can reject the null hypothesis and claim the coin is not fair.

The certainty with which we reject the null hypothesis is called *level of confidence*. We can use whatever level of confidence we want before rejecting the null hypothesis, but most often we use 90%, 95%, or 99%. If we choose a 95% confidence level, we reject the null hypothesis if

$$\mathbb{P}(O|H_0) \leq 1 - 0.95 = 0.05$$

The Central Limit theorem plays a main part here. Let us briefly recall the idea of Theorem (2.3): the sum of any number of independent and identically distributed random variables approximates the standard normal distribution. Let $p = \frac{Y}{N}$ be the proportion of heads in the sample of 100 coin flips. In our case $p = 0.51$ or 51%. But by the central limit theorem, p approximates the standard normal distribution, as p is in fact sum of the i.i.d. X_i's. This means we can estimate the standard deviation of p as

$$\sigma = \sqrt{\frac{p - p^2}{N}}$$

Our null hypothesis is that the coin is fair, or in other words $p_0 = 50\%$. Let us look at the normal curve:

A 95% level of confidence means we reject the null hypothesis if $p < 0.05$ or equivalently if p falls outside 95% of the area of the normal curve. Looking at that chart we see that this area corresponds to approximately 1.96 standard

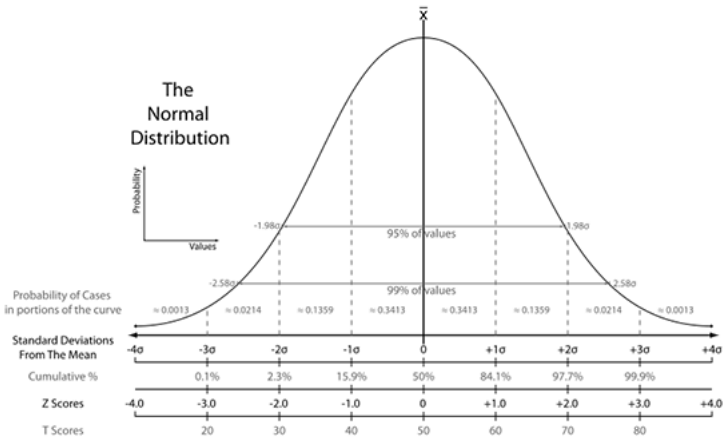


Figure 4: Normal Distribution

deviations. Therefore, if p is more than 1.96 standard deviations away from p_0, we can reject the null hypothesis with 95% confidence. If we want, we can use 99.9% level of confidence. In this case we reject the null hypothesis if p falls outside 99.9% of the area of the normal curve, which is approximately 3 standard deviations.

To check how many standard deviations away from p_0 our p is, we need to compute the z-score. Recall that $p_0 = 0.5$ and the standard deviation of p_0 is computed as $\sqrt{\frac{p_0(1-p_0)}{N}}$. We compute the z-score the same way we did in section 3.1.6:

$$z = \frac{p - 0.5}{\sqrt{\frac{0.5(1-0.5)}{N}}}$$

We can now compute that for hundred coin flips and 51 heads or $N = 100$ and $p = 0.51$, we get a z-score of 0.2, meaning p is 0.2 standard deviations away from p_0. This is not greater that 1.96, therefore we cannot reject the null hypothesis. If we had two more coins from which we have obtained 60 and 70 heads in 100 coin flips, we would get z-scores of 2 and 4 respectively. Than with 95% confidence, we could reject the null hypothesis that the coin is fair and conclude that they are biased. Note that in the case of getting 70 heads, we could reject the null hypothesis with even higher confidence - 99.9%, since p is more than 3

Coin	Flips	Heads	Z-score	95% confidence	99.9% confidence
Coin 1	100	51	0.2	not rejected	not rejected
Coin 2	100	60	2	rejected	not rejected
Coin 3	100	70	4	rejected	rejected

Table 1: Results for 100 flips of a coin

standard deviations away from p_0. These results are summarized in Table 1

This concludes our hypothesis testing.

3.3 Sampling

3.3.1 Population and samples

So far we have only considered what is know as *populations* , that is the complete set of things we are interested in. The frequency distribution have included all the scores we are interested in, such as the scores of all students who took the exam in the example of section 3.1.5. A population is not necessarily a collection of people; it can be a complete set of anything, such as the IQ of fifteen-year olds living in Ulm, the number of goals scored in each football league on a particular Sunday or the number of books in each library in Europe. The population is simply every member of a certain category that we wish to study.

Often due to the vast size of the population, we cannot study it all. In this case we select a *sample* . A sample is a subset of the population. Usually we want to know about populations rather than about samples, but in most cases it is only possible to test samples. This is a fundamental problem in statistical analysis - how can we generalise the information a sample gives us to the entire population? We illustrate the difficulty with an example.

A doctor wishes to know the risk of developing respiratory problems in German man over the age of 50 years. This is a very large population and it is quite difficult to test them all. Therefore, a sample must be tested instead. But what the doctor is not interested in the sample itself, but in what it tells him about the population. If it is not possible to estimated details of the population from the sample, than it is not worth studying it. What the doctor needs to find is sample information that is useful for estimating details about the population.

One of the difficulties of using samples to represent population is selecting the sample members. In most cases we want our sample to truly represent the population, so we can generalise our findings to the population and claim the population will behave like the sample. When conducting a survey on a sample of the voting population, we should make sure that we have, for example, the same proportion of men and women in the sample as they are in the population.

Consider the example of respiratory problems. Would any group of men over 50 be an acceptable sample? If we took only man from hill top villages, where the air is clear, or from mining towns polluted with coal dust, we are likely to have a biased sample, as not all members of the population live in such places. We would need to take the sample from a range of locations. We need to consider age as well. If our sample contained only men between 50 and 60, would it be representative for men over 60 in the population?

Any difference between the sample and the population can lead to a problem in generalisation: location, age, whether they smoke or not, occupation and so on. It is almost impossible to obtain a truly representative sample, where every characteristic of the sample matches the population characteristics. Here researchers should do the best with what they have and try to be aware of any difference between the sample and population. Here the judgement is not entirely statistical but also depends on the researcher's expertise in the subject. A doctor would know that certain factors are important with respect to respiratory problems and will try to select a sample representative of the population on these key factors.

An alternative way of selecting the sample to represent the population is through random selection. With a *random sample* , the sample members are selected at random from the entire population with each member having an equal chance of being selected. For example, when doing a survey, we might select names at random from the phone book. We have no idea who the people are, we leave it all to chance. By random selection, we make sure the sample is not deliberately biased, so any differences between the sample and the population are random and therefore not systematically influencing the data.

However, even the random sampling might not be so random. If we perform a survey in the streets on random passers by, we exclude all those people not passing by. If we perform the survey at 10 a.m., we exclude all the people whose occupation keep them at work at this time. Selecting randomly people listed in the telephone book excludes all that are not listed. Often it is hard to collect a truly random sample, but again it we must do the best we can by deciding on the key factors and selecting randomly within those factors.

3.3.2 Sample statistics and population parameters

Of the various measures of spread, the mean absolute deviation and the standard deviation both use information from all the scores. However, it has been found that the sample mean absolute deviation is an unstable estimator of the population figure, that is, there is no consistent relationship between the measure for the mean and for the population. On the other hand the standard deviation is a much more reliable estimator of the population value. Therefore when we do not know the population standard deviation, we use the sample

standard deviation to estimate it.

The formula for the standard deviation is given by equation (3). However, when applied to the sample scores, this formula underestimates the population value. To improve the estimate we change the formula for the sample standard deviation:

$$\text{Sample standard deviation } (s) = \sqrt{\frac{\sum (X - \bar{X})^2}{n-1}}, \tag{4}$$

where n is the sample size, $\bar{X}$ is the sample mean and s is the sample standard deviation (to distinguish from the population mean μ and standard deviation σ).

We also want to know the central figure in a population, but when we only have sample, rather than details on the population, we have to estimate it. Of the various measures of central tendency, the sample mean is the best estimate of the population value. But how good an estimate of μ is the sample mean $\bar{X}$ depends on the size of the sample. The larger it is, the better the sample mean is as an estimate of the population mean. We can see from a simple example.

The population of IQ scores is normally distributed with mean 100 and standard deviation of 15. If we took 20 people's IQ scores, would their mean be 100? The answer is probably not. We might have taken a sample of very smart people only, then the sample mean would be higher than 100. So sample mean will have different values, depending on the sample we have selected.

Now imagine we are able to select every possible sample of 20 IQ scores and compute their sample means. If we plot those means as a frequency distribution, we get the distribution of the sample means. Note that we are not interested in the individual scores, but in the mean of every sample of size 20. It turns out that the distribution of the sample means has some very useful characteristics.

First, we find that as we obtain more samples, the mean of the sample means get closer to the population mean. When we have collected all the possible samples, we find that their mean is the same as the population mean. So if we collect all samples of 20 IQ scores, then the mean of the samples is 100. We denote this mean by $\mu_{\bar{X}}$.

Second, the distribution of sample means will tend to be a normal distribution. If the population of scores is normally distributed, than the distribution of sample means would definitely be normally distributed. Even if the distribution of scores is not normally distributed the distribution of sample means will still look rather like a normal distribution. The larger the samples we select, the closer the distribution approaches normal distribution. This is a consequence of the Central Limit theorem (Theorem 2.3). This is an extremely useful piece of

information in our statistical analysis.

Third, as the distribution of the sample means is approximately normally distributed, we can compute the probability of finding a sample with a particular mean by calculating a z-score for our sample.

Finally, we can easily compute the standard deviation of the distribution of sample means using the standard deviation of the individual scores. We call this new standard deviation the *standard error of the mean* and denote it by $\sigma_{\bar{X}}$. The standard error provides us with the standard deviation of a sample mean from the population mean.

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

where σ is the standard deviation of the population and n is the sample size.

The standard error of the mean is precisely the distance that the sample mean is from the population mean. It tells us how good an estimate the sample mean is of the population mean. Notice that as the sample size (n) gets larger, the standard error gets smaller

The distribution of the sample means is now something we know a lot about without having to calculate the means for all the samples. The distribution of the sample means is normal distribution with mean $\mu_{\bar{X}}$, the same as the population mean μ and a standard deviation $\sigma_{\bar{X}}$, equal to the population standard deviation divided by the square root of the sample size.

In the IQ example, the distribution for samples of 20 scores will be normal with mean 100 and a standard error of $\frac{15}{\sqrt{20}} = 3.35$. As we have a normal distribution with known mean and standard deviation, we can calculate the z-scores and work out probability values. Suppose we have obtained a sample mean of 95 in our sample of 20 IQ scores. We compute the z-score:

$$z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{95 - 100}{3.35} = -1.49$$

We can look up the probability of the z-score in standard normal distribution tables as our sampling distribution is normally distributed and we get a probability of 6.80%. This tells us that the probability of obtaining a sample mean as low as 95 from our sample of 20 IQ scores is only 6.80%. If we used a larger sample size, say 30 instead of 20, this probability decreases to 3.39%. This is what we need for hypothesis testing.

4 Basic Concepts and Notation

Definition 4.1. *A set Ω is a collection of distinct objects, called "elements" of Ω.*

Interpretation

- The elements of a set can be anything - numbers, colours, letters and so on. For example $A = \{1, 2, 3, 4, 5\}$ is the set of the first 5 positive integers and $B = \{A, B, C, D, ..., X, Y, Z\}$ is the set of the alphabet.
- Two sets are said to be equal if and only if they have precisely the same elements; for example, $C = \{2, 1, 4, 3, 5\}$ is equal to the set A described above, but the set $C' = \{6, 2, 1, 4, 3, 5\}$ is not. Note that the order of the elements does not matter.
- A set may contain no elements at all, then it is called the empty set and is denoted by $\emptyset$.
- A set can have infinitely many elements, for example the set of all odd positive numbers $D = \{1, 3, 5, 7, ...\}$.

Definition 4.2. *Given two sets X and Y, we say that X is a subset of Y, if every element of X is also an element of Y. It is denoted as $X \subseteq Y$ and it implies that*

$$x \in X \Rightarrow x \in Y$$

Definition 4.3. *A set Ω is called countable if there exists an injective function $f : \Omega \to \mathbb{N}$ from Ω to the natural numbers $\mathbb{N}$*

Remark

- If a set is countable, then the number of elements of this set is some subset of the natural numbers. This means the elements can be counted one at a time and each element is associated with a natural number. However that does not mean that a countable set is finite - just like the set of natural number is not.
- If a set is not countable, it is call uncountable.

Example 4.1. *Recall the set of all positive odd numbers D as defined above. Now consider a set E consisting of all odd positive numbers, divisible by 3. Than the set $E = \{3, 9, 15, 21, 27, ...\}$ is a subset of D since all the elements in E are also elements in D.*

Example 4.2. *Consider the set $E = \{0, 1, 2\}$ and the set A as defined above. Than E is not a subset of A or $E \not\subseteq A$ since $0 \in E$, but $0 \notin A$.*

Example 4.3. *Every set is a subset of itself and the empty set is a subset of every other set. That means* $\emptyset \subseteq F$ *and* $F \subseteq F$ *for every set* F*.*

Example 4.4. *If* $X \subseteq Y$ *and* $Y \subseteq X$*, than it must be the case that* $X = Y$*.*

Definition 4.4. *Given a set* Ω*, the power set of* Ω *is the set of all subsets of* Ω*. It is denoted by* $\mathcal{P}(\Omega)$*.*

Interpretation

- If we have a set $\Omega = \{x, y, z\}$, then $\{x\}$, $\{y\}$ and $\{z\}$ each form a subset of Ω. Also, so do $\{x, y\}$, $\{x, z\}$ and $\{y, z\}$. Finally, the empty set $\{\emptyset\}$ and the whole set Ω are also considered subsets of Ω. Therefore the power set of $\{x, y, z\}$ is given by

$$\mathcal{P}(\Omega) = \{\{\emptyset\}, \{x\}, \{y\}, \{z\}, \{x, y\}, \{x, z\}, \{y, z\}, \{x, y, z\}\}$$

- The power set is therefore a set, whose elements are again sets, that consists of a different selection of items in Ω. Note that due to the set properties, the order of the items does not matter, so $\{x, y\}$ is the same as $\{y, x\}$. Also, repetitions do not matter, i.e. $\{x, x, y\} = \{x, y\}$.
- Number of subsets: if the set Ω has n elements, than the power set of Ω has 2^n elements. In the example above, Ω has 3 elements, therefore the power set should have 8 elements in total, which we can see to be true.

Definition 4.5. *A set* O *is called open if for every* $x \in O$ *there exists a real number* $\epsilon > 0$ *such that, any point* y *in* $\mathbb{R}^n$ *with distance from* x *is smaller than* ϵ*, is also in* O *(*$y \in O$*).*

Interpretation

- Consider the interval $(0, 1)$. The endpoints 0 and 1 are not in the interval; if we take a point x that is arbitrarily close to either of the endpoints, we can still find a small enough number $\epsilon \in \mathbb{R}$, such that all points with a distance from x smaller than ϵ are in $(0, 1)$. For example, consider $x = 0.9999$, so the distance from 1 is 0.0001. Still, we can take, say, $\epsilon = 0.0001/2$ and it is small enough to ensure the above requirements. Therefore $(0, 1)$ is an open interval.
- Now consider the interval $[0, 1]$. Unlike the previous example, the endpoints 0 and 1 belong to the interval. Now if we take $x = 1$, than we cannot find a positive ϵ such that any y with distance from x less than ϵ is in $[0, 1]$. Indeed, consider a very small epsilon, say $\epsilon = 0.00001$. Than for $y = 1.000001$, the distance between x and y is less than ϵ, yet clearly $y \notin [0, 1]$

Definition 4.6. *A Borel set on $\mathbb{R}$ is any set that can be formed from open sets through the operations of countable union and countable intersection. The collection of Borel sets on X forms a σ-algebra, known as the Borel algebra. This is the smallest σ-algebra containing all open sets (or equivalently, containing all closed sets).*

Interpretation

- Every "reasonable" subset of $\mathbb{R}$, in particular each interval, open set, closed set, finite set, countable set, is a Borel set. For example, $(0, 1)$ is a Borel set, so is $(1, 2)$ and so is their union.

Definition 4.7. *A simple function is*

Definition 4.8. *The indicator function of a subset $\mathcal{A}$ of a set X is a function $\mathbf{1}_{\mathcal{A}} : X \to \{0, 1\}$ defined as*

$$\mathbf{1}_{\mathcal{A}} = \begin{cases} 1, & \text{if } x \in A, \\ 0, & \text{if } x \notin A \end{cases}$$

Index