Probability Theory Summary

1. Basic Concepts

1.1 Set Theory

Before we venture into the depths of probability, we start our journey with a field trip on set theory.

1.1.1 What is a set?

Sets are collections of elements or samples. Sets are denoted by capital letters (A, B). Elements are denoted by ω_i. The set containing all elements is the **sample space** Ω. The set with no elements is called the **empty set** $\emptyset$.

Now let's discuss some notation. If we say that $A = \{\omega_1, \omega_2, \ldots, \omega_n\}$, then we say that A consists of the elements $\omega_1, \omega_2, \ldots, \omega_n$. $\omega_i \in A$ means that ω_i is in A, whereas $\omega_i \notin A$ means that ω_i is not in A.

1.1.2 Comparing sets

We can compare and manipulate sets in many ways. Let's take a look at it. We say that A is a **subset** of B (denoted by $A \subset B$) if every element in A is also in B. When (at least) one element in A is not in B, then $A \not\subset B$. If $A \subset B$ and $B \subset A$ (they consist of the same elements), then A and B are equal: $A = B$. Sets are said to be **disjoint** if they have no common elements.

We can't just add up sets. But we can take the **intersection** and the **union** of two sets. The **intersection** of A and B (denoted by $A \cap B$) consists of all elements ω_i that are in both A and B. (So $\omega_i \in A \cap B$ if $\omega_i \in A$ **and** $\omega_i \in B$.) On the other hand, the **union** of A and B (denoted by $A \cup B$) consists of all elements ω_i that are in either A or B, or both. (So $\omega_i \in A \cup B$ if $\omega_i \in A$ **or** $\omega_i \in B$.)

The **set difference** $A \setminus B$ consists of all elements that are in A, but not in B. There is also the **complement** of a set A, denoted by A^c. This consists of all elements that are not in A. Note that $A^c = \Omega \setminus A$ and $A \setminus B = A \cap B^c$.

There's one last thing we need to define. A **partition** of Ω is a collection of subsets A_i, such that

- the subsets A_i are disjoint: $A_i \cap A_j = \emptyset$ for $i \neq j$.
- the union of the subsets equals Ω: $\bigcup_{i=1}^{n} A_i = A_1 \cup A_2 \cup \ldots \cup A_n = \Omega$.

1.2 Introduction to Probability

It's time to look at probability now. Probability is all about experiments and their outcomes. What can we say about those outcomes?

1.2.1 Definitions

Some experiments always have the same outcome. These experiments are called **deterministic**. Other experiments, like throwing a dice, can have different outcomes. There's no way of predicting the outcome. We do, however, know that when throwing the dice many times, there is a certain **regularity** in the outcomes. That regularity is what probability is all about.

A **trial** is a single execution of an experiment. The possible **outcomes** of such an experiment are denoted by ω_i. Together, they form the **probability space** Ω. (Note the similarity with set theory!) An **event** A is a set of outcomes. So $A \subset \Omega$. We say that Ω is the **sure event** and $\emptyset$ is the **impossible event**.

Now what is the **probability** $P(A)$ of an event A? There are many definitions of probability, out of which the **axiomatic definition** is mostly used. It consists of three axioms. These axioms are rules which $P(A)$ must satisfy.

1. $P(A)$ is a nonnegative number: $P(A) \geq 0$.
2. The probability of the sure event is 1: $P(\Omega) = 1$.
3. If A and B have no outcomes in common (so if $A \cap B = \emptyset$), then the probability of $A \cup B$ equals the sum of the probabilities of A and B: $P(A \cup B) = P(A) + P(B)$.

1.2.2 Properties of probability

From the three axioms, many properties of the probability can be derived. Most of them are, in fact, quite logical.

- The probability of the impossible event is zero: $P(\emptyset) = 0$.
- The probability of the complement of A satisfies: $P(A^c) = 1 - P(A)$.
- If $A \subset B$, then B is equally or more likely than A: $P(A) \leq P(B)$.
- The probability of an event A is always between 0 and 1: $0 \leq P(A) \leq 1$.
- For any events A and B, there is the relation: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Using the probability, we can also say something about events. We say two events A and B are **mutually independent** if

$$P(A \cap B) = P(A)P(B). \tag{1.2.1}$$

Identically, a series of n events $A_1, \ldots, A_n$ are called **mutually independent** if any combination of events $A_i, A_j, \ldots, A_k$ (with $i, j, \ldots, k$ being numbers between 1 and n) satisfies

$$P(A_i \cap A_j \cap \ldots \cap A_k) = P(A_i)P(A_j)P(A_k). \tag{1.2.2}$$

1.2.3 Conditional Probability

Sometimes we already know some event B happened, and we want to know what the chances are that event A also happened. This is the **conditional probability** of A, given B, and is denoted by $P(A|B)$. It is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{1.2.3}$$

The conditional probability satisfies the three axioms of probability, and thus also all the other rules. However, using this conditional probability, we can derive some more rules. First, there is the **product rule**, stating that

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_2 \cap A_1) \ldots P(A_n|A_{n-1} \cap \ldots \cap A_2 \cap A_1). \tag{1.2.4}$$

Another rule, which is actually quite important, is the **total probability rule**. Let's suppose we have a partition $B_1, \ldots, B_n$ of Ω. The total probability rule states that

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(B_i)P(A|B_i). \tag{1.2.5}$$

By combining this rule with the definition of conditional probability, we find another rule. This rule is called **Bayes' rule**. It says that

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}. \tag{1.2.6}$$

2. Random Variables and Distributions

2.1 Random Variable Definitions

Suppose we know all the possible outcomes of an experiment, and their probabilities. What can we do with them? Not much, yet. What we need, are some tools. We will now introduce these tools.

2.1.1 Random variables

It is often convenient to attach a number to each event ω_i. This number is called a **random variable** and is denoted by $\underline{x}(\omega_i)$ or simply $\underline{x}$. You can see the random variable as a number, which can take different values. For example, when throwing a dice we can say that $\underline{x}(\text{head}) = 0$ and $\underline{x}(\text{tail}) = 1$. So $\underline{x}$ is now a number that can be either 0 or 1.

Random variables can generally be split up in two categories: discrete and continuous random variables. A random variable $\underline{x}$ is **discrete** if it takes a finite or countable infinite set of possible values. (With countable finite we mean the degree of infinity. The sets of natural numbers $\mathbb{N}$ and rational numbers $\mathbb{Q}$ are countable finite, while the set of real numbers $\mathbb{R}$ is not.)

Both types of random variables have fundamental differences, so in the coming chapters we will often explicitly mention whether a rule/definition applies to discrete or continuous random variables.

2.1.2 Probability mass function

Let's look at the probability that $\underline{x} = x$ for some number x. This probability depends on the random variable "function" $\underline{x}(\omega_i)$ and the number x. It is denoted by

$$P_{\underline{x}}(x) = P(\underline{x} = x). \tag{2.1.1}$$

The function $P_{\underline{x}}(k)$ is called the **probability mass function** (PMF). It, however, only exists for discrete random variables. For continuous random variables $P_{\underline{x}}(k) = 0$ (per definition).

2.1.3 Cumulative distribution function

Now let's take a look at the probability that $\underline{x} \leq x$ for some x. This is denoted by

$$F_{\underline{x}}(x) = P(\underline{x} \leq x). \tag{2.1.2}$$

The function $F_{\underline{x}}(x)$ is called the **cumulative distribution function** (CDF) of the random variable $\underline{x}$. The CDF has several properties. Let's name a few.

- The limits of $F_{\underline{x}}(x)$ are given by

$$\lim_{x \to -\infty} F_{\underline{x}}(x) = 0 \qquad \text{and} \qquad \lim_{x \to \infty} F_{\underline{x}}(x) = 1. \tag{2.1.3}$$

- $F_{\underline{x}}(x)$ is increasing. If $x_1 \leq x_2$, then $F_{\underline{x}}(x_1) \leq F_{\underline{x}}(x_2)$.
- $P(\underline{x} > x) = 1 - F_{\underline{x}}(x)$.
- $P(x_1 < \bar{x} \leq x_2) = F_{\underline{x}}(x_2) - F_{\underline{x}}(x_1)$.

The CDF exists for both discrete and continuous random variables. For discrete random variables, the function $F_{\underline{x}}(x)$ takes the form of a staircase function: its graph consists of a series of horizontal lines. For continuous random variables the function $F_{\underline{x}}(x)$ is continuous.

2.1.4 Probability density function

For continuous random variables there is a continuous CDF. From it, we can derive the **probability density function** (PDF), which is defined as

$$f_{\underline{x}}(x) = \frac{dF_{\underline{x}}(x)}{dx} \qquad \Leftrightarrow \qquad F_{\underline{x}}(x) = \int_{-\infty}^{x} f_{\underline{x}}(t)dt. \tag{2.1.4}$$

Since the CDF $F_{\underline{x}}(x)$ is always increasing, we know that $f_{\underline{x}}(x) \geq 0$. The PDF does not exist for discrete random variables.

2.2 Discrete Distribution types

There are many distribution types. We'll be looking at discrete distributions in this part, while continuous distributions will be examined in the next part. But before we even start examining any distributions, we have to increase our knowledge on combinations. We use the following paragraph for that.

2.2.1 Permutations and combinations

Suppose we have n elements and want to order them. In how many ways can we do that? The answer to that is

$$n! = n \cdot (n-1) \cdot \ldots \cdot 2 \cdot 1. \tag{2.2.1}$$

Here $n!$ means n **factorial**. But what if we only want to order k items out of a set of n items? The amount of ways is called the amount of **permutations** and is

$$\frac{n!}{(n-k)!} = n \cdot (n-1) \cdot \ldots \cdot (n-k+1). \tag{2.2.2}$$

Sometimes the ordering doesn't matter. What if we just want to select k items out of a set of n items? In how many ways can we do that? This result is the amount of **combinations** and is

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+1)}{k \cdot (k-1) \cdot \ldots \cdot 2 \cdot 1}. \tag{2.2.3}$$

2.2.2 The binomial distribution and related distributions

Now we will examine some types of discrete distributions. The most important parameter for discrete distributions is the probability mass function (PMF) $P_{\underline{x}}(k)$. So we will find it for several distribution types.

Suppose we have an experiment with two outcomes: success and failure. The chance for success is always just p. We do the experiment n times. The random variable $\underline{x}$ denotes the amount of successes. We now have

$$P_{\underline{x}}(k) = P(\underline{x} = k) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{2.2.4}$$

This distribution is called the **binomial distribution**.

Sometimes we want to know the probability that we need exactly k trials to obtain r successes. In other words, the r^{th} success should occur in the k^{th} trial. The random variable $\underline{x}$ now denotes the amount of trials needed. In this case we have

$$P_{\underline{x}}(k) = P(\underline{x} = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}. \tag{2.2.5}$$

This distribution is called the **negative binomial distribution**.

We can also ask ourselves: how many trials do we need if we only want one success? This is simply the negative binomial distribution with $r = 1$. We thus have

$$P_{\underline{x}}(k) = P(\underline{x} = k) = p(1-p)^{k-1}. \tag{2.2.6}$$

This distribution is called the **geometric distribution**.

2.2.3 Other discrete distributions

Let's discuss some other discrete distributions. A random variable $\underline{x}$ follows a **Poisson distribution** with parameter $\lambda > 0$ if

$$P_{\underline{x}}(k) = e^{-\lambda}\frac{\lambda^k}{k!}. \tag{2.2.7}$$

This distribution is an approximation of the binomial distribution if $np = \lambda$, $p \to 0$ and $n \to \infty$.

A random variable $\underline{x}$ has a **uniform distribution** if

$$P_{\underline{x}}(k) = \frac{1}{n}, \tag{2.2.8}$$

where n is the amount of possible outcomes of the experiment. In this case every outcome is **equally likely**.

A random variable has a **Bernoulli distribution** (with parameter p) if

$$P_{\underline{x}}(k) = \begin{cases} p & \text{for } k = 1, \\ 1-p & \text{for } k = 0. \end{cases} \tag{2.2.9}$$

Finally there is the **hypergeometric distribution**, for which

$$P_{\underline{x}}(k) = \frac{\binom{r}{k}\binom{m-r}{n-k}}{\binom{m}{n}}. \tag{2.2.10}$$

2.3 Continuous Distribution Types

It's time we switch to continuous distributions. The most important function for continuous distributions is the probability density function (PDF) $f_{\underline{x}}(k)$. We will find it for several distribution types.

2.3.1 The normal distribution

We start with the most important distribution type there is: the **normal distribution** (also called **Gaussian distribution**). A random variable $\underline{x}$ is a **normal random variable** (denoted by $\underline{x} \sim N(\overline{x}, \sigma_x^2)$) if

$$f_{\underline{x}}(x) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{1}{2}\left(\frac{x-\overline{x}}{\sigma_x}\right)^2}. \tag{2.3.1}$$

Here $\overline{x}$ and σ_x are, respectively, the mean and the standard deviation. (We will discuss them in the next part.) It follows that the cumulative distribution function (CDF) is

$$F_{\underline{x}}(x) = \frac{1}{\sqrt{2\pi}\sigma_x} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{t-\overline{x}}{\sigma_x}\right)^2} dt. \tag{2.3.2}$$

The above integral doesn't have an analytical solution. To get a solution anyway, use is made of the **standard normal distribution**. This is simply the normal distribution with parameters $\overline{x} = 0$ and $\sigma_x = 1$. So,

$$\Phi(z) = P(\underline{z} < z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{1}{2}t^2}\, dt. \tag{2.3.3}$$

There are a lot of tables in which you can simply insert z and retrieve $\Phi(z)$. To get back to the variable x, you make use of the transformation

$$z = \frac{x - \overline{x}}{\sigma_x} \qquad \Leftrightarrow \qquad x = \sigma_x z + \overline{x}. \tag{2.3.4}$$

2.3.2 Other continuous distributions

There is also a continuous **uniform distribution**. A random variable $\underline{x}$ has a uniform distribution (denoted by $\underline{x} \sim U(a, b)$) on the interval (a, b) if

$$f_{\underline{x}}(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b, \\ 0 & \text{otherwise.} \end{cases} \tag{2.3.5}$$

A random variable has an **exponential distribution** if

$$f_{\underline{x}}(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0. \end{cases} \tag{2.3.6}$$

Finally, a random variable has a **gamma distribution** if

$$f_{\underline{x}}(x) = \begin{cases} \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0, \end{cases} \tag{2.3.7}$$

where Γ is the **gamma function**, given by

$$\Gamma(a) = \int_0^{\infty} x^{a-1} e^{-x}\, dx. \tag{2.3.8}$$

2.4 Important parameters

Certain parameters apply to all distribution types. They say something about the distribution. Let's take a look at what parameters there are.

2.4.1 The mean

The **mean** is the expected (average) value of a random variable $\underline{x}$. It is denoted by $E(\underline{x}) = \overline{x}$. For discrete distributions we have

$$E(\underline{x}) = \overline{x} = \sum_{i=1}^{n} x_i P_{\underline{x}}(x_i), \tag{2.4.1}$$

with $x_1, \ldots x_n$ the possible outcomes. For continuous distributions we have

$$E(\underline{x}) = \overline{x} = \int_{-\infty}^{\infty} x f_{\underline{x}}(x)\, dx. \tag{2.4.2}$$

By the way, $E(\ldots)$ is the mathematical **expectation operator**. It is subject to the rules of linearity, so

$$E(a\underline{x} + b) = aE(\underline{x}) + b, \tag{2.4.3}$$

$$E(g_1(\underline{x}) + \ldots + g_n(\underline{x})) = E(g_1(\underline{x})) + \ldots + E(g_n(\underline{x})). \tag{2.4.4}$$

2.4.2 The variance

The **variance** or **dispersion** of a random variable is denoted by σ_x^2. Here σ_x is the **standard deviation**. If $\underline{x}$ is discrete, then the variance is given by

$$\sigma_x^2 = D(\underline{x}) = E\left((\underline{x} - \overline{x})^2\right) = \sum_{i=1}^{n} (x_i - \overline{x})^2 P_{\underline{x}}(x_i) \tag{2.4.5}$$

If $\underline{x}$ is continuous, then it is given by

$$\sigma_x^2 = D(\underline{x}) = E\left((\underline{x} - \overline{x})^2\right) = \int_{-\infty}^{\infty} (x - \overline{x})^2 f_{\underline{x}}(x)\, dx. \tag{2.4.6}$$

Here $D(\ldots)$ is the mathematical **dispersion operator**. It can be shown that σ_x^2 can also be found (for both discrete and continuous random variables) using

$$\sigma_x^2 = E(\underline{x}^2) - \overline{x}^2. \tag{2.4.7}$$

Note that in general $E(\underline{x}^2) \neq \overline{x}^2$. The value $E(\underline{x}) = \overline{x}$ is called the **first moment**, while $E(\underline{x}^2)$ is called the **second moment**. The variance σ_x^2 is called the **second central moment**.

This is all very nice to know, but what is it good for? Let's take a look at that. The variance σ_x^2 tells something about how far values are away from the mean $\overline{x}$. In fact, **Chebyshev's inequality** states that for every $\epsilon > 0$ we have

$$P(|\underline{x} - \overline{x}| \geq \epsilon) \leq \frac{\sigma_x^2}{\epsilon^2}. \tag{2.4.8}$$

2.4.3 Other moments

After the first and the second moment, there is of course also the **third moment**, being

$$E\left((\underline{x} - \overline{x})^3\right). \tag{2.4.9}$$

The third moment is a measure of the symmetry around the center (the **skewness**). For symmetrical distributions this third moment is 0.

The **fourth moment** $E\left((\underline{x} - \overline{x})^4\right)$ is a measure of how peaked a distribution is (the **kurtosis**). The kurtosis of the normal distribution is 3. If the kurtosis of a distribution is less than 3 (so the distribution is less peaked than the normal distribution), then the distribution is **platykurtic**. Otherwise it is **leptokurtic**.

2.4.4 Median and mode

Finally there are the median and the mode. The **median** is the value x for which $F_{\underline{x}}(x) = 1/2$. So half of the possible outcomes has a value lower than x and the other half has values higher than x.

The **mode** is the value x for which (for discrete distributions) $P_{\underline{x}}(x)$ or (for continuous distributions) $f_{\underline{x}}(x)$ is at a maximum. So you can see the mode as the value x which is most likely to occur.

3. Multiple Random Variables

3.1 Random Vectors

Previously we have only dealt with one random variable. Now suppose we have more random variables. What distribution functions can we then define?

3.1.1 Joint and marginal distribution functions

Let's suppose we have n random variables $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$. We can put them in a so-called **random vector** $\underline{\mathbf{x}} = [\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n]^T$. The **joint distribution function** (also called the **simultaneous distribution function**) $F_{\underline{\mathbf{x}}}(\mathbf{x})$ is then defined as

$$F_{\underline{\mathbf{x}}}(x_1, x_2, \ldots, x_n) = F_{\underline{\mathbf{x}}}(\mathbf{x}) = P(\underline{x}_1 \leq x_1, \underline{x}_2 \leq x_2, \ldots, \underline{x}_n \leq x_n). \tag{3.1.1}$$

(You should read the commas "," in the above equation as "and" or, equivalently, as the intersection operator $\cap$.) From this joint distribution function, we can derive the **marginal distribution function** $F_{\underline{x}_i}(x_i)$ for the random variable $\underline{x}_i$. It can be found by inserting ∞ in the joint distribution function for every x_j other than x_i. In an equation this becomes

$$F_{\underline{x}_i}(x_i) = F_{\underline{\mathbf{x}}}(\infty, \infty, \ldots, \infty, x_i, \infty, \ldots, \infty). \tag{3.1.2}$$

The marginal distribution function can always be derived from the joint distribution function using the above method. The opposite is, however, not always true. It often isn't possible to derive the joint distribution function from the marginal distribution functions.

3.1.2 Density functions

Just like for random variables, we can also distinguish discrete and continuous random vectors. A random vector is **discrete** if its random variables $\underline{x}_i$ are discrete. Similarly, it is continuous if its random variables are continuous.

For discrete random vectors the **joint (mass) distribution function** $P_{\underline{\mathbf{x}}}(\mathbf{x})$ is given by

$$P_{\underline{\mathbf{x}}}(\mathbf{x}) = P(\underline{x}_1 = x_1, \underline{x}_2 = x_2, \ldots, \underline{x}_n = x_n). \tag{3.1.3}$$

For continuous random vectors, there is the **joint density function** $f_{\underline{\mathbf{x}}}$. It can be derived from the joint distribution function $F_{\underline{\mathbf{x}}}(\mathbf{x})$ according to

$$f_{\underline{\mathbf{x}}}(x_1, x_2, \ldots, x_n) = f_{\underline{\mathbf{x}}}(\mathbf{x}) = \frac{\partial^n F_{\underline{\mathbf{x}}}(x_1, x_2, \ldots, x_n)}{\partial x_1 \, \partial x_2 \, \ldots \, \partial x_n}. \tag{3.1.4}$$

3.1.3 Independent random variables

In the first chapter of this summary, we learned how to check whether a series of events $A_1, \ldots, A_n$ are independent. We can also check whether a series of random variables are independent. This is the case if

$$P(\underline{x}_1 \leq x_1, \underline{x}_2 \leq x_2, \ldots, \underline{x}_n \leq x_n) = P(\underline{x}_1 \leq 1)P(\underline{x}_2 \leq 2) \ldots P(\underline{x}_n \leq n). \tag{3.1.5}$$

If this is, indeed the case, then we can derive the joint distribution function $F_{\underline{\mathbf{x}}}(\mathbf{x})$ from the marginal distribution functions $F_{\underline{x}_i}(x_i)$. This goes according to

$$F_{\underline{\mathbf{x}}}(\mathbf{x}) = F_{\underline{x}_1}(x_1)F_{\underline{x}_2}(x_2) \ldots F_{\underline{x}_n}(x_n) = \prod_{i=1}^{n} F_{\underline{x}_i}(x_i). \tag{3.1.6}$$

3.2 Covariance and Correlation

Sometimes it may look like there is a relation between two random variables. If this is the case, you might want to take a look at the covariance and the correlation of these random variables. We will now take a look at what they are.

3.2.1 Covariance

Let's suppose we have two random variables $\underline{x}_1$ and $\underline{x}_2$. We also know their joint distribution function $f_{\underline{x}_1,\underline{x}_2}(x_1, x_2)$. The **covariance** of $\underline{x}_1$ and $\underline{x}_2$ is defined as

$$C(\underline{x}_1, \underline{x}_2) = E\left((\underline{x}_1 - \overline{x}_1)(\underline{x}_2 - \overline{x}_2)\right) = \int_{-\infty}^{\infty} (\underline{x}_1 - \overline{x}_1)(\underline{x}_2 - \overline{x}_2) f_{\underline{x}_1,\underline{x}_2}(x_1, x_2) dx_1\, dx_2 = E(\underline{x}_1\underline{x}_2) - \overline{x}_1\overline{x}_2. \tag{3.2.1}$$

The operator $C(\ldots,\ldots)$ is called the **covariance operator**. Note that $C(\underline{x}_1, \underline{x}_2) = C(\underline{x}_2, \underline{x}_1)$. We also have $C(\underline{x}_1, \underline{x}_1) = D(\underline{x}_1) = \sigma^2_{x_1}$.

If the random variables $\underline{x}_1$ and $\underline{x}_2$ are independent, then it can be shown that $E(\underline{x}_1, \underline{x}_2) = E(\underline{x}_1)E(\underline{x}_2) = \overline{x}_1\overline{x}_2$. It directly follows that $C(\underline{x}_1, \underline{x}_2) = 0$. The opposite, however, isn't always true.

But the covariance operator has more uses. Suppose we have random variables $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$. Let's define a new random variable $\underline{z}$ as $\underline{z} = \underline{x}_1 + \underline{x}_2 + \ldots + \underline{x}_n$. How can we find the variance of $\underline{z}$? Perhaps we can add up all the variances of $\underline{x}_i$? Well, not exactly, but we are close. We can find σ^2_z using

$$\sigma^2_z = \sum_{i=1}^{n}\sum_{j=1}^{n} C(\underline{x}_i, \underline{x}_j) = \sum_{i=1}^{n} \sigma^2_{x_i} + 2\sum_{1\le i<j\le n}\sum C(\underline{x}_i, \underline{x}_j). \tag{3.2.2}$$

We can distinguish a special case now. If the random variables $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$ are all independent, then $C(\underline{x}_i, \underline{x}_j) = 0$ for every i, j $(i \neq j)$. So then we actually are able to get the variance of $\underline{z}$ by adding up the variances of $\underline{x}_i$.

3.2.2 The correlation coefficient

Now let's make another definition. The **correlation coefficient** is defined as

$$\rho(\underline{x}_1, \underline{x}_2) = \frac{C(\underline{x}_1, \underline{x}_2)}{\sigma_{x_1}\sigma_{x_2}}. \tag{3.2.3}$$

This function has some special properties. Its value is always between -1 and 1. If $\rho(\underline{x}_1, \underline{x}_2) \approx \pm 1$, then $\underline{x}_2$ is (approximately) a linear function of $\underline{x}_1$. If, on the other hand, $\rho(\underline{x}_1, \underline{x}_2) = 0$, then we say that $\underline{x}_1$ and $\underline{x}_2$ are **uncorrelated**. This doesn't necessarily mean that they are independent. Two variables can be uncorrelated, but not independent. If two variables are, however, independent, then $C(\underline{x}_1, \underline{x}_2) = 0$, and they are therefore also uncorrelated.

3.3 Conditional Random Variables

In chapter 1 of this summary, we have seen conditional probability. We can combine this with functions like the cumulative distribution function, the probability density function, and so on. That is the subject of this part.

3.3.1 Conditional relations

Given an event B, let's define the conditional CDF as

$$F_{\underline{x}}(x|B) = P(\underline{x} \leq x|B) = \frac{P(\underline{x} \leq x, B)}{P(B)}. \tag{3.3.1}$$

Here the event B can be any event. Also, the comma once more indicates an intersection. The conditional PDF now follows as

$$f_{\underline{x}}(x|B) = \frac{dF_{\underline{x}}(x|B)}{dx}. \tag{3.3.2}$$

The nice thing is that conditional probability has all the properties of normal probability. So any rule that you've previously seen about probability can also be used now.

Let's see if we can derive some rules for these conditional functions. We can rewrite the **total probability rule** for the conditional CDF and the conditional PDF. Let $B_1, B_2, \ldots, B_n$ be a partition of Ω. We then have

$$F_{\underline{x}}(x) = \sum_{i=1}^{n} F_{\underline{x}}(x|B_i)P(B_i) \qquad \Rightarrow \qquad f_{\underline{x}}(x) = \sum_{i=1}^{n} f_{\underline{x}}(x|B_i)P(B_i). \tag{3.3.3}$$

From this we can derive an equivalent for **Bayes' rule**, being

$$f_{\underline{x}}(x|A) = \frac{P(A|x)f_{\underline{x}}(x)}{\int_{-\infty}^{\infty} P(A|x)f_{\underline{x}}(x)dx}. \tag{3.3.4}$$

Here the event A can be any event. The probability $P(A|x)$ in the above equation is short for $P(A|\underline{x} = x)$.

3.3.2 The conditional probability density function

In the previous paragraph, there always was some event A or B. It would be nice if we can replace that by a random variable as well. We can use the random variable $\underline{y}$ for that. By doing so, we can derive that

$$f_{\underline{x}}(x|y) = \frac{f_{\underline{x},\underline{y}}(x, y)}{f_{\underline{y}}(y)}, \tag{3.3.5}$$

where $f_{\underline{x},\underline{y}}(x, y)$ is the joint density function of $\underline{x}$ and $\underline{y}$. Note that if $\underline{x}$ and $\underline{y}$ are independent, then $f_{\underline{x},\underline{y}}(x, y) = f_{\underline{x}}(x)f_{\underline{y}}(y)$ and thus $f_{\underline{x}}(x|y) = f_{\underline{x}}(x)$.

We can also rewrite the **total probability rule**. We then get

$$f_{\underline{y}}(y) = \int_{-\infty}^{\infty} f_{\underline{y}}(y|x)f_{\underline{x}}(x)dx. \tag{3.3.6}$$

Similarly, we can rewrite **Bayes' rule** to

$$f_{\underline{x}}(x|y) = \frac{f_{\underline{y}}(y|x)f_{\underline{x}}(x)}{f_{\underline{y}}(y)} = \frac{f_{\underline{y}}(y|x)f_{\underline{x}}(x)}{\int_{-\infty}^{\infty} f_{\underline{y}}(y|x)f_{\underline{x}}(x)dx}. \tag{3.3.7}$$

3.3.3 The conditional mean

The **conditional mean** of $\underline{y}$, given $\underline{x} = x$, can be found using

$$E(\underline{y}|x) = \int_{-\infty}^{\infty} y\, f_{\underline{y}}(y|x)\, dy. \tag{3.3.8}$$

Note that this mean depends on x, and is therefore a function of x. Now let's look at $E(\underline{y}|\underline{x})$. We know that $\underline{x}$ is a random variable, and $E(\underline{y}|x)$ is a function of x. This implies that $E(\underline{y}|\underline{x})$ is a random variable. We may ask ourselves, what is the mean of this new random variable? In fact, it turns out that

$$E(E(\underline{y}|\underline{x})) = E(\underline{y}) = \overline{y}. \tag{3.3.9}$$

3.3.4 n random variables

Suppose we have n random variables $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$. We can then also have a conditional PDF, being

$$f(x_n, \ldots, x_{k+1}|x_k, \ldots, x_1) = \frac{f(x_1, \ldots, x_k, x_{k+1}, \ldots, x_n)}{f(x_1, \ldots, x_k)}. \tag{3.3.10}$$

From this, the so-called **chain rule** can be derived, being

$$f(x_1, \ldots, x_n) = f(x_n|x_{n-1}, \ldots, x_1)f(x_{n-1}|x_{n-2}, \ldots, x_1)\ldots f(x_2|x_1)f(x_1). \tag{3.3.11}$$

4. Transformations

4.1 Transformations of Random Variables

Let's suppose that we have got some random variables. We can make new random variables out of those, using functions. How do those new random variables behave? We will take a look at that now. This chapter of the summary is rather difficult, while it is not incredibly important. So do not stare yourself blind on this part.

4.1.1 Finding the CDF

Suppose we have a random variable $\underline{x}$. We can define a new variable $\underline{y}$ to be a function of $\underline{x}$, so $\underline{y} = g(\underline{x})$. Now we would like to know how we can find the CDF $F_{\underline{y}}(y)$. It can be found using

$$F_{\underline{y}}(y) = P(\underline{y} \leq y) = P(g(\underline{x}) \leq y) = P(\underline{x} \in I_y), \tag{4.1.1}$$

where the set I_y consists of all x such that $g(x) \leq y$. So, to find the CDF $F_{\underline{y}}(y)$, we first need to find I_y: We need to know for what x we have $g(x) \leq y$. The intervals that are found can then be used to express $F_{\underline{y}}$ in $F_{\underline{x}}$.

Let's look at a special case. When $g(x)$ is strictly increasing, or strictly decreasing, we have

$$F_{\underline{y}}(y) = F_{\underline{x}}(g^{-1}(y)) \quad \text{for increasing } g(x) \text{ and} \qquad F_{\underline{y}}(y) = 1 - F_{\underline{x}}(g^{-1}(y)) \quad \text{for decreasing } g(x). \tag{4.1.2}$$

Here the function $g^{-1}(y)$ is the inverse of $g(x)$. It is defined such that if $y = g(x)$, then $x = g^{-1}(y)$.

4.1.2 Finding the PDF

Now that we've got the CDF $F_{\underline{y}}(y)$, it's time to find the PDF $f_{\underline{y}}(y)$. You probably remember that the PDF is simply the derivative of the CDF. That rule can be used to find the PDF.

Let's consider the special case that $g(x)$ is either strictly increasing or strictly decreasing. Now we have

$$f_{\underline{y}}(y) = \frac{dF_{\underline{y}}(y)}{dy} = \begin{cases} f_{\underline{x}}(g^{-1}(y))\frac{dg^{-1}(y)}{dy} & \text{for increasing } g(x) \\ -f_{\underline{x}}(g^{-1}(y))\frac{dg^{-1}(y)}{dy} & \text{for decreasing } g(x) \end{cases} = f_{\underline{x}}(g^{-1}(y))\left|\frac{dg^{-1}(y)}{dy}\right|. \tag{4.1.3}$$

Note that we have simply taken the derivative of equation (4.1.2), using the chain rule. Also note that if $g(x)$ is decreasing, also $g^{-1}(y)$ is decreasing, and thus $dg^{-1}(y)/dy$ is negative. This explains the last step in the above equation, where the absolute stripes $|\ldots|$ suddenly appear.

Now what should we do if $g(x)$ is not increasing or decreasing? In this case no inverse function $g^{-1}(y)$ exists. Let's suppose that for a given y the equation $y = g(x)$ has n solutions $x_1, x_2, \ldots, x_n$. Now we can say that

$$f_{\underline{y}}(y) = \sum_{i=1}^{n} \frac{f_{\underline{x}}(x_i)}{\left|\frac{dg(x_i)}{dx}\right|}. \tag{4.1.4}$$

If only one solution x_i is present, then this equation reduces back to equation (4.1.3).

4.1.3 Functions of two random variables

Let's suppose we now have two random variables $\underline{x}_1$ and $\underline{x}_2$. Also, let's define $\underline{y} = g(\underline{x}_1, \underline{x}_2)$. In this case, we can find the CDF $F_{\underline{y}}(y)$ using

$$F_{\underline{y}}(y) = P(\underline{y} \leq y) = P(g(\underline{x}_1, \underline{x}_2) \leq y) = P((\underline{x}_1, \underline{x}_2) \in D_y), \tag{4.1.5}$$

where the set D_y consists of all the pairs (x_1, x_2) such that $g(x_1, x_2) \leq y$. To find the PDF, we can use

$$f_{\underline{y}}(y) = \int_{-\infty}^{\infty} f_{\underline{x}_1,\underline{x}_2}(x_1, g^{-1}(x_1, y)) \left| \frac{dg^{-1}(x_1, y)}{dy} \right| dx_1 = \int_{-\infty}^{\infty} f_{\underline{x}_1,\underline{x}_2}(g^{-1}(y, x_2), x_2) \left| \frac{dg^{-1}(y, x_2)}{dy} \right| dx_2. \tag{4.1.6}$$

4.1.4 Transformations of two random variables

Now let's not only define $\underline{y}_1 = g_1(\underline{x}_1, \underline{x}_2)$, but also $\underline{y}_2 = g_2(\underline{x}_1, \underline{x}_2)$. Now we can find the joint CDF using

$$F_{\underline{y}_1,\underline{y}_2}(y_1, y_2) = P(\underline{y}_1 \leq y_1, \underline{y}_2 \leq y_2) = P(g_1(\underline{x}_1, \underline{x}_2) \leq y_1, g_2(\underline{x}_1, \underline{x}_2) \leq y_2) = P((\underline{x}_1, \underline{x}_2) \in D_{y_1,y_2}), \tag{4.1.7}$$

where the region D_{y_1,y_2} is the intersection of the regions D_{y_1} and D_{y_2}. We can now find the joint PDF by differentiating the CDF. We then get

$$f_{y_1,y_2}(y_1, y_2) = \frac{\partial^2 F_{\underline{y}_1,\underline{y}_2}(y_1, y_2)}{\partial y_1 \, \partial y_2} = \frac{\partial^2}{\partial y_1 \, \partial y_2} \int \int_{D_{y_1,y_2}} f_{\underline{x}_1,\underline{x}_2}(x_1, x_2) \, dx_1 \, dx_2. \tag{4.1.8}$$

There is, however, another way to find the joint PDF. For that, let's examine the matrix

$$\mathbf{g}(x_1, x_2)\partial_{\mathbf{x}}{}^T = \begin{bmatrix} g_1(x_1, x_2) \\ g_2(x_1, x_2) \end{bmatrix} \begin{bmatrix} \frac{\partial}{\partial x_1} & \frac{\partial}{\partial x_2} \end{bmatrix} = \begin{bmatrix} \frac{\partial g_1(x_1,x_2)}{\partial x_1} & \frac{\partial g_1(x_1,x_2)}{\partial x_2} \\ \frac{\partial g_2(x_1,x_2)}{\partial x_1} & \frac{\partial g_2(x_1,x_2)}{\partial x_2} \end{bmatrix}. \tag{4.1.9}$$

The determinant of this matrix is called the **Jacobian** of $\mathbf{g}$. The joint PDF can now be found using

$$f_{\underline{y}_1,\underline{y}_2}(y_1, y_2) = \frac{f_{\underline{x}_1,\underline{x}_2}(x_1, x_2)}{\left| \det\left(\mathbf{g}(x_1, x_2)\partial_{\mathbf{x}}{}^T \right) \right|}. \tag{4.1.10}$$

The above equation also works for dimension higher than 2. In fact, it works for any pair of n-dimensional vectors $\mathbf{y}$ and $\mathbf{x}$ for which $\mathbf{y} = \mathbf{g}(\mathbf{x})$.

4.1.5 The multi-dimensional mean

Let's suppose we have an n-dimensional random vector $\underline{\mathbf{x}}$, an m-dimensional random vector $\underline{\mathbf{y}}$ and a function $G(\mathbf{x})$ such that $\underline{\mathbf{y}} = G(\underline{\mathbf{x}})$. It would be interesting to know the **expectation vector** $E(\underline{\mathbf{y}})$. It can be found using

$$E(\underline{\mathbf{y}}) = \int_{\mathbb{R}^m} \mathbf{y} f_{\underline{\mathbf{y}}}(\mathbf{y}) \, dy \qquad \Leftrightarrow \qquad E(\underline{y}_i) = \int_{\mathbb{R}^m} y_i f_{\underline{\mathbf{y}}}(\mathbf{y}) \, dy. \tag{4.1.11}$$

Using the right part of the above equation, you can find one component of $E(\underline{\mathbf{y}})$. The left part is the general (vector) equation. Note that in both cases you need to integrate m times. Once for every component of $\underline{\mathbf{y}}$.

Generally, we don't know $f_{\underline{\mathbf{y}}}(\mathbf{y})$ though. But we do know $f_{\underline{\mathbf{x}}}(\mathbf{x})$. So to find $E(\underline{\mathbf{y}})$, we can first find $f_{\underline{\mathbf{y}}}(\mathbf{y})$. This is, however, not always necessary. There is a way to find $E(\underline{\mathbf{y}})$ without finding $f_{\underline{\mathbf{y}}}(\mathbf{y})$. You then have to use

$$E(\underline{\mathbf{y}}) = E(G(\underline{\mathbf{x}})) = \int_{\mathbb{R}^n} G(\mathbf{x}) f_{\underline{\mathbf{x}}}(\mathbf{x}) \, dx. \tag{4.1.12}$$

The above equation is called the **expectation law**. If the function $G(\mathbf{x})$ is linear (so you can write it as $G(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$ for a constant matrix A), then the above equation simplifies greatly. In that case we have $\overline{\mathbf{y}} = A\overline{\mathbf{x}} + \mathbf{b}$, where $\overline{\mathbf{y}} = E(\underline{\mathbf{y}})$ and $\overline{\mathbf{x}} = E(\underline{\mathbf{x}})$.

4.1.6 Multi-dimensional variance and covariance

Calculating the variance $D(\underline{\mathbf{y}})$ of $\underline{\mathbf{y}}$ goes more or less similar to calculating the mean. There is a slight difference though. While $E(\underline{\mathbf{y}})$ was an $m \times 1$ vector, $D(\underline{\mathbf{y}})$ is an $m \times m$ matrix. To find this matrix, we can use either of the following two equations

$$D(\underline{\mathbf{y}}) = E\left((\mathbf{y} - \overline{\mathbf{y}})(\mathbf{y} - \overline{\mathbf{y}})^T\right) = \int_{\mathbb{R}^m} (\mathbf{y} - \overline{\mathbf{y}})(\mathbf{y} - \overline{\mathbf{y}})^T f_{\underline{\mathbf{y}}}(\mathbf{y})\, dy, \tag{4.1.13}$$

$$D(\underline{\mathbf{y}}) = D(G(\underline{\mathbf{x}})) = E\left((G(\mathbf{x}) - \overline{\mathbf{y}})(G(\mathbf{x}) - \overline{\mathbf{y}})^T\right) = \int_{\mathbb{R}^n} (G(\mathbf{x}) - \overline{\mathbf{y}})(G(\mathbf{x}) - \overline{\mathbf{y}})^T f_{\underline{\mathbf{x}}}(x)\, dx. \tag{4.1.14}$$

If $G(\mathbf{x})$ is, once more, linear, we can simplify the above equation. In this case we have

$$D(\underline{\mathbf{y}}) = AD(\underline{\mathbf{x}})A^T \qquad \Leftrightarrow \qquad Q_{yy} = AQ_{xx}A^T, \tag{4.1.15}$$

where $Q_{yy} = D(\underline{\mathbf{y}})$ and $Q_{xx} = D(\underline{\mathbf{x}})$. From these two matrices, we can also find the **covariance matrices** Q_{yx} and Q_{xy}, according to

$$C(\underline{\mathbf{y}}, \underline{\mathbf{x}}) = Q_{yx} = AQ_{xx} \qquad \text{and} \qquad C(\underline{\mathbf{x}}, \underline{\mathbf{y}}) = Q_{xy} = Q_{xx}A^T. \tag{4.1.16}$$

Here Q_{yx} is an $m \times n$ matrix, while Q_{xy} is an $n \times m$ matrix. So in the multi-dimensional situation we do not have $C(\underline{\mathbf{y}}, \underline{\mathbf{x}}) = C(\underline{\mathbf{x}}, \underline{\mathbf{y}})$. However, since Q_{xx} is symmetric, we do have $Q_{yx} = Q_{xy}^T$.

4.2 The Central Limit Theorem

If we put together multiple random variables, interesting things start happening. And it has something to do with the normal distribution. If you want to know more about it, then quickly read the chapter below.

4.2.1 The central limit theorem

Let's suppose we have a number of (possibly different) independent random variables $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$. Now let's define a new random variable $\underline{y}$ as the sum of all these variables, so $\underline{y} = \underline{x}_1 + \underline{x}_2 + \ldots + \underline{x}_n$. Let's suppose we know the mean $\overline{y}$ and the standard deviation and σ_y. The **central limit theorem** states that as n increases, we have

$$F_{\underline{y}}(y) \approx \Phi\left(\frac{y - \overline{y}}{\sigma_y}\right). \tag{4.2.1}$$

In words, we see that as n increases, $\underline{y}$ behaves like a normal distribution with average $\overline{y}$ and standard deviation σ_y. The corresponding PDF then is

$$f_{\underline{y}}(y) \approx \frac{1}{\sigma_y\sqrt{2\pi}} e^{-\frac{(y-\overline{y})^2}{2\sigma_y^2}}. \tag{4.2.2}$$

Let's now look at a special case. Suppose $\underline{x}_1 = \underline{x}_2 = \ldots = \underline{x}_n = \underline{x}$. Also, all these distributions have mean $\overline{x}$ and standard deviation σ_x. In this case we can find $\overline{y}$ and σ_y. We have $\underline{y} = n\underline{x}$. The average of $\underline{y}$ evidently becomes $\overline{y} = n\overline{x}$. To find the standard deviation of $\underline{y}$, we first look at the variance of $\underline{y}$. Since the random variables $\underline{x}$ are independent, we find that $\sigma_y^2 = n\sigma_x^2$. From this follows that $\sigma_y = \sqrt{n}\sigma_x$. The random variable $\underline{y}$ thus behaves like a normal distribution with the just found mean $\overline{y}$ and standard deviation σ_y.

4.2.2 The De Moivre-Laplace theorem

Let's suppose the random variable $\underline{x}$ is binomially distributed. So it is a discrete variable with as mean $\overline{x} = np$ and as variance $\sigma_x^2 = np(1-p)$. The **De Moivre-Laplace theorem** now states that for certain conditions the (discrete) binomial distribution of $\underline{x}$ also starts behaving like a (continuous) normal distribution. So,

$$P_{\underline{x}}(k) = \binom{n}{k} p^k (1-p)^{n-k} \approx \frac{1}{\sigma_x\sqrt{2\pi}} e^{-\frac{(k-\overline{x})^2}{2\sigma_x^2}} = \frac{1}{\sqrt{2\pi np(1-p)}} e^{-\frac{(k-np)^2}{2np(1-p)}}. \tag{4.2.3}$$

The condition for which the above equation is accurate is that k must be in the interval $(\overline{x}-3\sigma_x, \overline{x}+3\sigma_x)$. Of course k can't be smaller than 0 or bigger than n.

Suppose we do n experiments, with $\underline{x}$ denoting the amount of successes. We would like to know the chance that we have exactly k successes. We now know how to calculate that (approximately). We simply insert $\frac{k-\overline{x}}{\sigma_x}$ in the PDF of the standard normal distribution. But what should we do if we want to know the chance that we have at least k_1 successes, and at most k_2 successes? In this case we have

$$P(k_1 \leq \underline{x} \leq k_2) = \Phi\left(\frac{k_2 + 1/2 - \overline{x}}{\sigma_x}\right) - \Phi\left(\frac{k_1 - 1/2 - \overline{x}}{\sigma_x}\right). \tag{4.2.4}$$

Note the halves in the above equation. They are present because the binomial distribution is discrete, while the normal distribution is continuous. If we, for example, want to have at least 42 successes, and at most 54, then for the normal distribution we should take as boundaries 41.5 and 54.5.

4.3 Composed Distributions

There are some distributions we haven't treated yet. That was because they were a bit too difficult to start with right away. Often this was because they are composed of multiple other distributions. But now the time has come to take a look at them.

4.3.1 The multivariate normal distribution

Suppose we have an n-dimensional random vector $\underline{\mathbf{x}}$ with mean $\overline{\mathbf{x}}$ and variance matrix Q_{xx}. We say that $\underline{\mathbf{x}}$ has a **multivariate normal distribution** ($\underline{\mathbf{x}} \sim N_n(\overline{\mathbf{x}}, Q_{xx})$) if its PDF has the form

$$f_{\underline{\mathbf{x}}}(\mathbf{x}) = \frac{1}{\sqrt{\det(2\pi Q_{xx})}} e^{\left(-\frac{1}{2}(\mathbf{x}-\overline{\mathbf{x}})^T Q_{xx}^{-1}(\mathbf{x}-\overline{\mathbf{x}})\right)}, \tag{4.3.1}$$

where the variance matrix Q_{xx} has only positive entries.

Let's suppose $\underline{\mathbf{x}}$ is 2-dimensional. If we plot $f_{\underline{\mathbf{x}}}(\mathbf{x})$, we get a 3-dimensional graph. For this graph, we can draw **contour lines** (lines for which $f_{\underline{\mathbf{x}}}(\mathbf{x})$ is constant). This implies that

$$(\mathbf{x}-\overline{\mathbf{x}})^T Q_{xx}^{-1}(\mathbf{x}-\overline{\mathbf{x}}) = r^2, \tag{4.3.2}$$

for some constant r. The shapes we then get are ellipses. We can do the same if $\underline{\mathbf{x}}$ is 3-dimensional. However, we then draw contour areas, which take the shape of ellipsoids. In situations with even more dimensions, we get hyper-ellipsoids. All these shapes are called the **ellipsoids of concentration**.

The shape of these ellipsoids depends on the variance matrix Q_{xx}. If Q_{xx} is the identity matrix I_n, or a multiple of it, then the ellipsoids will be circles/spheres/hyperspheres. If Q_{xx} is just a diagonal matrix, then the principal axes of the ellipsoids will be the axes $x_1, x_2, \ldots, x_n$ itself. In other cases, the axes of the ellipsoid will have shifted.

Many things can be derived from the PDF, for which we just gave the equation. Examples are the marginal distributions and the conditional distributions. An interesting thing is that those distributions are, in turn, also normal distributions. And if that wasn't interesting enough, also all linear transformations of a multivariate normal distribution are (multivariate) normal distributions.

4.3.2 The χ^2 distribution

Let's suppose $\underline{x}_1, \underline{x}_2, \ldots \underline{x}_n$ are all normally distributed random variables with mean $\overline{x}_i$ and variance 1, so $\underline{x}_i \sim N(\overline{x}_i, 1)$. The **Chi-square** distribution with n degrees of freedom, denoted as $\chi^2(n, \lambda)$, is now defined as

$$\underline{\chi}^2 = \sum_{i=1}^{n} \underline{x}_i^2. \tag{4.3.3}$$

The **non-centrality parameter** λ is defined as

$$\lambda = \sum_{i=1}^{n} \overline{x}_i^2. \tag{4.3.4}$$

If $\lambda = 0$, we are dealing with the **central Chi-square distribution** $\chi^2(n, 0)$.

The Chi-square distribution has mean $E(\underline{\chi}^2) = n + \lambda$ and variance $D(\underline{\chi}^2) = 2n + 4\lambda$. If two (independent) Chi-square distributions $\underline{\chi}_1^2$ and $\underline{\chi}_2^2$ are added up, we once more get a Chi-square distribution, but now with $(n_1 + n_2)$ degrees of freedom and non-centrality parameter $(\lambda_1 + \lambda_2)$.

4.3.3 The t distribution

Suppose that $\underline{x} \sim N(\nabla, 1)$ and $\underline{\chi}^2 \sim \chi^2(n, 0)$ are independent random variables. The **(Student's)** t **distribution** with n degrees of freedom, denoted as $t(n, \nabla)$, is now defined as

$$\underline{t} = \frac{\underline{x}}{\sqrt{\underline{\chi}^2 / n}}. \tag{4.3.5}$$

Here ∇ is the non-centrality parameter. If $\nabla = 0$, we are dealing with the **central** t **distribution**.

4.3.4 The F distribution

Suppose that $\underline{\chi}_1^2 \sim \chi^2(n_1, \lambda)$ and $\underline{\chi}_2^2 \sim \chi^2(n_2, 0)$ are two independent Chi-square distributions. The F **distribution**, denoted as $F(n_1, n_2, \lambda)$, is then defined as

$$\underline{F} = \frac{\underline{\chi}_1^2 / n_1}{\underline{\chi}_2^2 / n_2}. \tag{4.3.6}$$

It is said to have n_1 and n_2 degrees of freedom. Also, λ is the non-centrality parameter. When $\lambda = 0$, we are dealing with a **central** F **distribution**.

5. Estimation

5.1 Introduction to Estimation

One of the powers of probability is that you can estimate the behavior of phenomena. How can we do that? That's something we will look at in this chapter.

5.1.1 Definitions

It often occurs that there is some phenomenon of which we want to know the behavior. Such a phenomenon can be modeled as a random vector $\underline{\mathbf{y}}$, with a certain size m. However, we usually don't know the distribution of such a random variable. The PDF just isn't known to us. But, given certain parameters, we can find it. In this case we can put the n unknown parameters into a vector $\mathbf{x}$. Then, once $\mathbf{x}$ is known, we can find the PDF of $\underline{\mathbf{y}}$. This PDF is written as $f_{\underline{\mathbf{y}}}(\mathbf{y}|\mathbf{x})$. (An example could be where we know $\underline{\mathbf{y}}$ is normally distributed, but we don't know the mean $\overline{\mathbf{y}}$ and the standard deviation σ_y.)

Now how can we find $\mathbf{x}$? The truth is, we can't. However, by observing the phenomenon described by $\underline{\mathbf{y}}$, we can guess it. Let's say our guess $\hat{\mathbf{x}}$ (called the **estimate** of x) is given by some function $\hat{\mathbf{x}} = G(\mathbf{y})$. Our task is to set this function $G(\mathbf{y})$. Once we have done so, we can also define a new random variable, called the **estimator**, as $\underline{\hat{\mathbf{x}}} = G(\underline{\mathbf{y}})$.

5.1.2 Finding a good estimator

So we now know we have to choose an estimator $\underline{\hat{\mathbf{x}}} = G(\underline{\mathbf{y}})$. How would we know what would be a good one? There are three criteria for that. First there is the **estimation error** $\underline{\hat{\epsilon}} = \underline{\hat{\mathbf{x}}} - \mathbf{x}$, which is also a random variable. The estimator $\hat{x}$ is said to be an **unbiased estimator** of x if, and only if $E(\underline{\hat{\epsilon}}) = E(\underline{\hat{\mathbf{x}}} - \mathbf{x}) = \mathbf{0}$, or, equivalently, $E(\underline{\hat{\mathbf{x}}}) = \mathbf{x}$. If the estimator $\underline{\hat{\mathbf{x}}}$ is not unbiased, then we say that its **bias** is $E(\underline{\hat{\epsilon}})$.

Another measure of quality is the **mean squared error** (MSE), defined as $E\left(\left|\underline{\hat{\mathbf{x}}} - \mathbf{x}\right|^2\right)$, or, equivalently, $E\left(\underline{\hat{\epsilon}}^2\right)$. Of course the mean squared error should be as small as possible.

Finally, let's look at the third measure of quality. It is defined as $P(\left|\underline{\hat{\mathbf{x}}} - \mathbf{x}\right|^2 \leq r^2)$, for some radius r. In words, this is the probability that the vector $\underline{\hat{\epsilon}}$ is in the (hyper-)sphere with radius r. This chance should be as big as possible.

There are three common ways of finding an estimator. Which one to use depends on the data that you have and the accuracy that you want. We will take a look at them in the rest of this chapter.

5.2 Least-Squares Estimation

One of the most well-known methods of determining an estimation is the least-squares estimation method. Let's take a look at how it works.

5.2.1 The consistent case

Let's suppose we have a set of measurements $\mathbf{y}$, having size m, and a set of unknown parameters $\mathbf{x}$, having size n. To apply the **least-squares method**, we assume that

$$\mathbf{y} = A\mathbf{x}, \tag{5.2.1}$$

where the $m \times n$ matrix A is known. In words, this means that the measured parameters (in $\mathbf{y}$) are linear functions of the unknown variables (in $\mathbf{x}$). However, now the question arises whether, given a (measured) $\mathbf{y}$, there is an $\mathbf{x}$ which satisfies the above equation. If there is, then the system is **consistent**. Otherwise it is **inconsistent**. The inconsistent case will be treated in the next paragraph. Now we'll take a closer look at the consistent case.

So suppose the system $\mathbf{y} = A\mathbf{x}$ is consistent. In this case we know that there is at least one $\mathbf{x}$ satisfying $\mathbf{y} = A\mathbf{x}$. If there is exactly one solution, then this solution is our estimate $\hat{\mathbf{x}}$. It can be found using

$$\hat{\mathbf{x}} = A^{-1}\mathbf{y}. \tag{5.2.2}$$

The corresponding **least-squares solution** $\hat{\mathbf{y}}$ can be found using $\hat{\mathbf{y}} = A\hat{\mathbf{x}} = \mathbf{y}$.

It is, however, also possible that there are infinitely many solutions $\mathbf{x}$. In this case we can't be sure which $\mathbf{x}$ to choose. This often means we need more measurement data. This is the case if the columns of A aren't all linearly independent.

5.2.2 The inconsistent case

Now let's suppose the system $\mathbf{y} = A\mathbf{x}$ is inconsistent. In this case there is no solution $\mathbf{x}$. We now refer to the system as an **overdetermined system**, denoted as $\mathbf{y} \approx A\mathbf{x}$. Assuming that there are no linearly dependent columns in A, we define the **redundancy** of the system as $m - n$.

To make sure there are solutions, we add a measurement error vector $\mathbf{e}$, such that $\mathbf{y} = A\mathbf{x} + \mathbf{e}$. We now want to choose $\mathbf{e}$ such that $\mathbf{e}^2 = \mathbf{e}^T\mathbf{e}$ is minimal. This is the **least-squares principle**. The minimal $\mathbf{e}$ is denoted by $\hat{\mathbf{e}}$. With $\hat{\mathbf{e}}$ chosen correctly, there is a solution $\mathbf{x}$, called the estimate $\hat{\mathbf{x}}$. It can be found using

$$\hat{\mathbf{x}} = \left(A^T A\right)^{-1} A^T \mathbf{y}. \tag{5.2.3}$$

The corresponding least-squares solution can once more be found using $\hat{\mathbf{y}} = A\hat{\mathbf{x}}$. The difference $\hat{\mathbf{e}} = \mathbf{y} - A\hat{\mathbf{x}} = \mathbf{y} - \hat{\mathbf{y}}$ is the **least-squares residual vector**. The value of $\hat{\mathbf{e}}^2$ is a measure of the inconsistency of the system.

The vector $\mathbf{y}$ generally consists of measurement data. Sometimes we know that some measurement data is more accurate than other. That data should thus be taken into account more. For this, there is the **weighted least-squares method**. In this case we don't want to minimize $\mathbf{e}^2 = \mathbf{e}^T\mathbf{e}$. This time we should minimize $\mathbf{e}^T W\mathbf{e}$, where W is the **weight matrix**. The minimum value $\hat{\mathbf{e}}^T W\hat{\mathbf{e}}$ is now the measure of inconsistency of the system. The corresponding estimate can then be found using

$$\hat{\mathbf{x}} = \left(A^T W A\right)^{-1} A^T W\mathbf{y}. \tag{5.2.4}$$

Generally W is a positive diagonal matrix. This is, however, not always the case.

5.2.3 Orthogonal projectors

Let's take a closer look at what variables we got now. We have a measurement vector $\mathbf{y}$. In the inconsistent case $\mathbf{y}$ can't be written as $A\mathbf{x}$. So we search for a $\hat{\mathbf{y}}$ close to $\mathbf{y}$ that can be written as $A\hat{\mathbf{x}}$. The least-squares residual vector $\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}}$ is as small as possible.

It can now be shown that $\hat{\mathbf{y}}$ and $\hat{\mathbf{e}}$ are orthogonal vectors. While $\hat{\mathbf{y}}$ lies in the column space of A, $\hat{\mathbf{e}}$ is orthogonal to the column space of A. (So $\hat{\mathbf{e}}$ lies in the column space of $A^\perp$.) Now let's define the two matrices P_A and $P_A^\perp$ as

$$P_A = A\left(A^T W A\right)^{-1} A^T W \qquad \text{and} \qquad P_A^\perp = I_m - P_A. \tag{5.2.5}$$

These two matrices are **orthogonal projectors**. What they do is, they project vectors on the column spaces of A and $A^\perp$. We therefore have

$$\hat{\mathbf{y}} = P_A\mathbf{y} \qquad \text{and} \qquad \hat{\mathbf{e}} = P_A^\perp\mathbf{y}. \tag{5.2.6}$$

5.2.4 Implementing random vectors

Previously we saw the system of equations $\mathbf{y} = A\mathbf{x}+\mathbf{e}$. Here, the vector $\mathbf{y}$ represented a series of measurements. However, we can take more measurements from a phenomenon. Every time, these measurements can take different values. So it would be wise to represent $\mathbf{y}$ by a random vector $\underline{\mathbf{y}}$. Equivalently, the vector $\mathbf{e}$ also has a different value every time. So it should also be a random vector $\underline{\mathbf{e}}$. We thus get

$$\underline{\mathbf{y}} = A\mathbf{x} + \underline{\mathbf{e}}. \tag{5.2.7}$$

We assume we have chosen our estimate $\mathbf{x}$ (which does stay constant during different measurements) such that $E(\underline{\mathbf{e}}) = \mathbf{0}$ or, equivalently, $E(\underline{\mathbf{y}}) = A\mathbf{x}$. If this is indeed the case, then we say that the above linear system is called a **linear model** of $E(\underline{\mathbf{y}})$.

From $\underline{\mathbf{y}}$, we can derive the random variables $\hat{\underline{\mathbf{x}}}$, $\hat{\underline{\mathbf{y}}}$ and $\hat{\underline{\mathbf{e}}}$. For that, we can use relations we actually already know. They are

$$\hat{\underline{\mathbf{x}}} = \left(A^T W A\right)^{-1} A^T W \underline{\mathbf{y}} = \mathbf{x} + \left(A^T W A\right)^{-1} A^T W \underline{\mathbf{e}}, \tag{5.2.8}$$

$$\hat{\underline{\mathbf{y}}} = P_A \underline{\mathbf{y}} = A\mathbf{x} + P_A \underline{\mathbf{e}} \qquad \text{and} \qquad \hat{\underline{\mathbf{e}}} = P_A^{\perp} \underline{\mathbf{y}} = \mathbf{0} + P_A^{\perp} \underline{\mathbf{e}}. \tag{5.2.9}$$

And that's all we need to know if we want to estimate what the outcome of the experiment will be next.

5.3 Best linear unbiased estimation

The second method of finding $\hat{\underline{\mathbf{x}}}$ we will look at is the BLUE method. To use it, you also need the variance matrix Q_{yy} of $\underline{\mathbf{y}}$. Because the BLUE method considers Q_{yy}, it can be a rather accurate estimation method. Let's find out how it works.

5.3.1 Definitions and conditions

Let's consider the linear system of equations

$$E(\underline{\mathbf{y}}) = A\mathbf{x}. \tag{5.3.1}$$

Just like in the previously discussed method, we want to find an estimate $\mathbf{x}$ (or more precise, an estimator $\underline{\mathbf{x}}$) that corresponds to certain conditions. Before we look at those conditions, we first make some definitions.

Let's define the vector $\mathbf{z}$ (having size k) as a linear combination of $\mathbf{x}$. So, $\mathbf{z} = F^T\mathbf{x} + \mathbf{f_0}$, for some known $n \times k$ matrix F and k-vector $\mathbf{f_0}$. Just like we want to find an estimator $\hat{\underline{\mathbf{x}}}$ for $\mathbf{x}$, we can also be looking for an estimator $\underline{\mathbf{z}}$ for $\mathbf{z}$. This gives us also the relation $\underline{\mathbf{z}} = F^T\underline{\mathbf{x}} + \mathbf{f_0}$. So to find $\underline{\mathbf{x}}$ we might as well try to find $\underline{\mathbf{z}}$. The estimator $\underline{\mathbf{z}}$ depends on $\underline{\mathbf{y}}$. So let's define $G(\underline{\mathbf{y}})$ such that $\underline{\mathbf{z}} = G(\underline{\mathbf{y}})$.

Enough definitions. Let's look at what conditions we want the estimator $\underline{\mathbf{z}}$ to have. First of all we want it to be **unbiased**. This means that $E(\underline{\mathbf{z}}) = \mathbf{z}$ or, equivalently, $E(G(\underline{\mathbf{y}})) = F^T\mathbf{x} + \mathbf{f_0}$. For reasons of simplicity, we also want it to be **linear**. This is the case if $G(\underline{\mathbf{y}})$ is a linear function, and can thus be written as $G(\underline{\mathbf{y}}) = L^T\underline{\mathbf{y}} + \mathbf{l_0}$ for some $m \times k$ matrix L and a k-vector $\mathbf{l_0}$. This linearity condition can be rewritten to two conditions, being

$$A^T L = F \qquad \text{and} \qquad \mathbf{l_0} = \mathbf{f_0}. \tag{5.3.2}$$

Any estimator $\underline{\mathbf{z}}$ of $\mathbf{z}$ that is both linear and unbiased is called a **linear unbiased estimator** (LUE).

5.3.2 Finding the best linear unbiased estimator

There is one slight problem. There are many LUEs. We want only the so-called **best linear unbiased estimator** (BLUE), denoted by $\hat{\underline{\mathbf{z}}}$. But what is the best LUE? We now define the best LUE (the BLUE) to be the LUE with the smallest mean squared error (MSE). So we have

$$E\left(|\hat{\underline{\mathbf{z}}} - \mathbf{z}|^2\right) \leq E\left(|\underline{\mathbf{z}} - \mathbf{z}|^2\right) \tag{5.3.3}$$

for every LUE $\underline{\mathbf{z}}$. Now the question arises how we can find this BLUE $\hat{\underline{\mathbf{z}}}$ and the corresponding best estimator $\hat{\underline{\mathbf{x}}}$ for $\mathbf{x}$. For that, we use the **Gauss-Markov** theorem, which states that

$$\hat{\underline{\mathbf{x}}} = \left(A^T Q_{yy}^{-1} A\right)^{-1} A^T Q_{yy}^{-1} \underline{\mathbf{y}} \qquad \text{and} \qquad \hat{\underline{\mathbf{z}}} = F^T \hat{\underline{\mathbf{x}}} + \mathbf{f_0}, \tag{5.3.4}$$

where Q_{yy} is the variance matrix of $\underline{\mathbf{y}}$ (so $Q_{yy} = D(\underline{\mathbf{y}})$). It is interesting to note that the final value of $\hat{\underline{\mathbf{x}}}$ does not depend on the matrix F or the vector $\mathbf{f_0}$ at all. It just depends on A and $\underline{\mathbf{y}}$.

Another interesting fact to note is the link with the weighted least-squares estimation method. If the weight matrix W in the WLSE method is equal to the inverse of the variance matrix Q_{yy} (so $W = Q_{yy}^{-1}$) in the BLUE method, you would find exactly the same estimator $\underline{\mathbf{x}}$.

5.4 Maximum Likelihood Estimation and Confidence Regions

The third method of finding $\hat{\underline{\mathbf{x}}}$ uses the PDF of $\underline{\mathbf{y}}$. It can therefore not always be applied. But its advantage is that it can be applied in the case where $\underline{\mathbf{y}}$ can't be written as $A\mathbf{x}$.

5.4.1 The condition

To make a **maximum likelihood estimation** (MLE) we need to know the PDF of $\underline{\mathbf{y}}$, given a certain unknown vector $\mathbf{x}$. We write this as $f_{\underline{\mathbf{y}}}(\mathbf{y}|\mathbf{x})$.

Now suppose we have some measurement $\mathbf{y}$. The idea behind the MLE is to choose the value of $\mathbf{x}$ for which it is most likely that $\mathbf{y}$ occurred. The **likelihood of y** then is $f_{\mathbf{y}}(\mathbf{y}|\mathbf{x})$. Note that this likelihood is a function of the unknown vector $\mathbf{x}$. This likelihood should be maximized. The corresponding $\mathbf{x}$, now denoted as $\hat{\mathbf{x}}$, is the maximum likelihood estimation.

5.4.2 Finding the maximum likelihood estimation

There is no general method of finding the MLE. For relatively easy PDFs of $\underline{\mathbf{y}}$, simple logic can often lead to the MLE. For more difficult PDFs, finding the MLE might even require complicated (numerical) computation.

If, however, $f_{\underline{\mathbf{y}}}(\mathbf{y}|\mathbf{x})$ is sufficiently smooth, then $\hat{\mathbf{x}}$ can be found using the conditions

$$\partial_x f_{\underline{\mathbf{y}}}(\mathbf{y}|\hat{\mathbf{x}}) = \mathbf{0} \qquad \text{and} \qquad \partial^2_{xx^T} f_{\underline{\mathbf{y}}}(\mathbf{y}|\hat{\mathbf{x}}) < 0. \tag{5.4.1}$$

If the PDF simply gives a scalar result, then the above states that the first derivative must be zero, indicating that there is a local minimum/maximum. The second derivative must be smaller than zero, indicating it is in fact a maximum.

If, however, the PDF returns a vector, then things are a bit more difficult. Then the first condition requires that the gradient has to be zero. The second condition states that the so-called **Hessian matrix** (the matrix of derivatives) needs to be negative definite. In other words, all its eigenvectors need to be negative.

Finally, when multiple values satisfy the above conditions, just insert their values $\hat{\mathbf{x}}$ into $f_{\underline{\mathbf{y}}}(\mathbf{y}|\hat{\mathbf{x}})$ and see which value gives the highest likelihood of $\mathbf{y}$.

5.4.3 Confidence Regions

Suppose we have finally acquired a good estimate $\hat{\mathbf{y}}$ of $\mathbf{y}$ (using any of the three discussed methods). How can we indicate how good this estimate actually is? We can do this, using **confidence regions**.

Suppose we have a region S. For example, S can be defined as the interval $[\hat{\mathbf{y}} - \epsilon, \hat{\mathbf{y}} + \epsilon]$. We now take a new measurement $\mathbf{y}$. Let's examine the chance that $\mathbf{y}$ is in the region S. This chance then is

$$P(\mathbf{y} \in S) = 1 - \alpha. \tag{5.4.2}$$

We now say that S is a $100(1-\alpha)$ **percent confidence region**. (For example, if $\alpha = 0.01$, then it is a 99% confidence region.) Also $1 - \alpha$ is called the **confidence coefficient**.

6. Hypothesis Tests

6.1 Basic Concepts of Hypothesis Tests

We will now examine hypothesis tests. To become familiar with them, we first look at some basic concepts. After that, we consider the simple case where there are only two hypotheses.

6.1.1 Definitions

Let's suppose we have a random vector $\underline{\mathbf{y}}$. Its PDF is $f_{\underline{\mathbf{y}}}(\mathbf{y}|\mathbf{x})$, where the vector $\mathbf{x}$ is not known. We can now assume a certain value for $\mathbf{x}$. Afterwards, we can use a measurement $\mathbf{y}$ to check whether our original assumption of $\mathbf{x}$ made sense. We now have a **statistical hypothesis**, denoted by $H : \underline{\mathbf{y}} \sim f_{\underline{\mathbf{y}}}(\mathbf{y}|\mathbf{x})$.

The **(process) state space** contains all the possible values for $\mathbf{x}$. Often an hypothesis H states that $\mathbf{x}$ has a certain value. It now completely specifies the distribution of $\underline{\mathbf{y}}$. It is therefore called a **simple hypothesis**. H can also state that $\mathbf{x}$ is among a certain group of values. In this case $\underline{\mathbf{y}}$ is not completely specified. H is then called a **composite hypothesis**. In this course we only deal with simple hypotheses.

6.1.2 The binary decision problem

Usually, when you examine hypotheses, you have two hypothesis. It is possible to have multiple hypothesis $H_1, H_2, \ldots, H_n$, but we will treat that later. For now we assume we have just two hypotheses. First there is the **null hypothesis** $H_0 : \underline{\mathbf{y}} \sim f_{\underline{\mathbf{y}}}(\mathbf{y}|\mathbf{x_0})$, representing the **nominal state**. Second, there is the **alternative hypothesis** $H_a : \underline{\mathbf{y}} \sim f_{\underline{\mathbf{y}}}(\mathbf{y}|\mathbf{x_a})$. Both distributions state that the random variable $\underline{\mathbf{y}}$ has a certain PDF $f_{\underline{\mathbf{y}}}$.

Let's examine the **binary decision problem**. We have a single observation $\mathbf{y}$. Based on this observation, we have to choose whether we accept H_0 (assume it to be correct) or reject it. The procedure used to decide whether to accept H_0 or not is called a **test**.

How do we decide whether to accept H_0? For that, we define the **critical region** K. If $\mathbf{y} \in K$, then we reject H_0. On the other hand, if $\mathbf{y} \notin K$ (or equivalently, $\mathbf{y} \in K^c$), then we accept H_0. We can also define the **test statistic** $T = h(\mathbf{y})$, where T is a scalar and $h(\mathbf{y})$ some function. Corresponding to the (multi-dimensional) region K is also a scalar region K. We now reject H_0 if $T \in K$ and accept H_0 is $T \notin K$.

6.1.3 Four situations

In the binary decision problem, we have two options: accept or reject. In this choice, we can be either right or wrong. There are now four possible situations:

- We reject H_0, when in reality H_0 is true. So we made an error. This is called the **type 1 error**. Its probability, called the **probability of false alarm** α, can be found using
$$\alpha = P(\underline{\mathbf{y}} \in K|H_0) = \int_K f_{\underline{\mathbf{y}}}(\mathbf{y}|H_0)\,dy = \int_K f_{\underline{\mathbf{y}}}(\mathbf{y}|x_0)\,dy, \tag{6.1.1}$$
where the latter part is just a different way of writing things. α is also called the **size** of the test, or the **level of significance**.
- We accept H_0, when in reality H_0 is false. This time we made a **type 2 error**. The so-called **probability of missed detection** β is given by
$$\beta = P(\underline{\mathbf{y}} \notin K|H_a) = \int_{K^c} f_{\underline{\mathbf{y}}}(\mathbf{y}|H_a)\,dy = 1 - \int_K f_{\underline{\mathbf{y}}}(\mathbf{y}|H_a). \tag{6.1.2}$$

- We accept H_0, and were right in doing so. The so-called **probability of correct dismissal** (which doesn't have its own sign) is now given by

$$P(\underline{\mathbf{y}} \notin K|H_0) = \int_{K^c} f_{\underline{\mathbf{y}}}(\mathbf{y}|H_0)\,dy = 1 - \int_K f_{\underline{\mathbf{y}}}(\mathbf{y}|H_0)\,dy = 1 - \alpha \tag{6.1.3}$$

- We reject H_0, and were right in doing so. The **probability of detection**, also called the **power** of the test γ, now is

$$\gamma = P(\underline{\mathbf{y}} \in K|H_a) = \int_K f_{\underline{\mathbf{y}}}(\mathbf{y}|H_a)\,dy = 1 - \beta. \tag{6.1.4}$$

6.1.4 Defining the critical region

The size of α, β and γ depends on the critical region K. It is our task to define K. According to what criteria should we do that? Often, in real life, we want to minimize costs. A false alarm has a certain (positive) cost c_0, while a missed detection has a certain (also positive) cost c_a. The **average cost** in an experiment, also called the **Bayes risk**, is then $c_0\alpha + c_a\beta$. (We assume a correct choice has no costs, nor any special benefits.) We want to find the K for which the costs are at a minimum. (In fact, the **Bayes criterion** states that the Bayes risk should be minimized.) So we want to minimize

$$c_0\alpha + c_a\beta = c_0 \int_K f_{\underline{\mathbf{y}}}(\mathbf{y}|H_0)\,dy + c_a\left(1 - \int_K f_{\underline{\mathbf{y}}}(\mathbf{y}|H_a)\,dy\right) = c_a + \int_K \Big(c_0 f_{\underline{\mathbf{y}}}(\mathbf{y}|H_0) - c_a f_{\underline{\mathbf{y}}}(\mathbf{y}|H_a)\Big)\,dy. \tag{6.1.5}$$

We know that c_a is constant. So we should minimize the integral on the right. Note that an integral is something that adds up infinitely many numbers. By choosing K, we choose what numbers this integral adds up. We want to minimize the value of the integral. So we should make sure it only adds up negative numbers. (Any positive number would make its value only bigger.) So, we only have $\mathbf{y} \in K$ if

$$c_0 f_{\underline{\mathbf{y}}}(\mathbf{y}|H_0) - c_a f_{\underline{\mathbf{y}}}(\mathbf{y}|H_a) < 0. \tag{6.1.6}$$

The critical region K thus consists of all $\mathbf{y}$ for which

$$\frac{f_{\underline{\mathbf{y}}}(\mathbf{y}|H_0)}{f_{\underline{\mathbf{y}}}(\mathbf{y}|H_a)} < \frac{c_a}{c_0}. \tag{6.1.7}$$

In other words, if the above equation holds for the measurement $\mathbf{y}$, then we reject H_0.

6.1.5 A-priori probabilities

Let's complicate the situation a bit more. Previously we have made an assumption. We assumed that we didn't have a clue whether H_0 or H_a would be true in reality. Let's suppose we do have a clue now. The probability that H_0 is correct (and thus that $\mathbf{x} = \mathbf{x_0}$) is $P(\mathbf{x} = \mathbf{x_0})$. (Abbreviated this is $P(\mathbf{x_0})$.) Similarly, we know that the chance for H_a to be correct is $P(\mathbf{x_a})$. The probabilities $P(\mathbf{x_0})$ and $P(\mathbf{x_a})$ are called **a-priori probabilities** — probabilities we already know before the experiment. We know that either H_0 or H_a is true, so we have $P(\mathbf{x_0}) + P(\mathbf{x_a}) = 1$.

If H_0 is true, then we have a chance α to lose c_0. Similarly, if H_a is true, then we have a chance β to lose c_a. Therefore our average costs now become $P(\mathbf{x_0})c_0\alpha + P(\mathbf{x_a})c_a\beta$. From this we can find that $\mathbf{y} \in K$ (we should reject H_0) if

$$\frac{f_{\underline{\mathbf{y}}}(\mathbf{y}|H_0)}{f_{\underline{\mathbf{y}}}(\mathbf{y}|H_a)} < \frac{P(\mathbf{x_a})c_a}{P(\mathbf{x_0})c_0}. \tag{6.1.8}$$

If we don't have a clue which hypothesis will be correct, then $P(\mathbf{x_0}) = P(\mathbf{x_a}) = 1/2$. Note that, in this case, the above equation reduces to the result of the previous paragraph.

6.2 Multiple Hypothesis

Previously we have only considered two hypotheses, being H_0 and H_a. But what should we do if we have p hypotheses $H_1, H_2, \ldots, H_p$? How do we know which one to choose then? Let's take a look at that.

6.2.1 Deciding the hypothesis

First we will make a few definitions. Let's define the **discrete decision** δ as our choice of vector $\mathbf{x_i}$. Also, there is the **cost function** $C(\mathbf{x}, \delta)$ (not be confused with the covariance operator). Suppose we accept hypothesis H_j (and thus $\delta = \mathbf{x_j}$), but in reality we have $\mathbf{x} = \mathbf{x_i}$. In this case our costs are $C(\mathbf{x} = \mathbf{x_i}, \delta = \mathbf{x_j})$. This can also be written as $C(\mathbf{x_i}, \mathbf{x_j})$, or even as C_{ij}. We also assume that $C(\mathbf{x_i}, \mathbf{x_i}) = 0$. In words, this says that if we accept the right hypothesis, we don't have any costs.

Suppose we have a measurement $\mathbf{y}$. It is now rather difficult to decide which hypothesis we accept. We therefore make an assumption. We assume that the costs for all errors are equal. (So C_{ij} = constant for all i, j with $i \neq j$.) This is part of the so-called **Maximum A Posteriori probability criterion** (MAP). Now we can decide which hypothesis to accept. We should accept H_i if for all $j \neq i$ we have

$$f_{\underline{\mathbf{y}}}(\mathbf{y}|\mathbf{x_j})P(\mathbf{x_j}) < f_{\underline{\mathbf{y}}}(\mathbf{y}|\mathbf{x_i})P(\mathbf{x_i}). \tag{6.2.1}$$

So we should accept H_i if the number i gives the maximum value for $f_{\underline{\mathbf{y}}}(\mathbf{y}|\mathbf{x_i})P(\mathbf{x_i})$. This is, in fact, quite logical. If costs don't matter, we should simply choose the hypothesis which gives us the biggest chance that we're right. This also causes the chance that we're wrong to be the smallest.

6.2.2 Acceptance regions

Given a measurement $\mathbf{y}$, we now know which hypothesis H_i to choose. Let's look at all $\mathbf{y}$ for which we will accept H_i. These $\mathbf{y}$ form the **acceptance region** A_i. We can also look at this definition the other way around: If $\mathbf{y} \in A_i$, then we accept H_i.

Let's ask ourselves something. Suppose that in reality H_i is true. What is then the chance that we accept H_j? Let's call this chance β_{ij}. Its value depends on the acceptance region A_j and can be found using

$$\beta_{ij} = \int_{A_j} f_{\underline{\mathbf{y}}}(\mathbf{y}|\mathbf{x_i})\, dy. \tag{6.2.2}$$

The chance that we make any wrong decision (given that H_i is true) is denoted as β_i. It can be found by simply adding up all the β_{ij}s with $i \neq j$. So,

$$\beta_i = \sum_{j=1, j\neq i}^{p} \beta_{ij} \tag{6.2.3}$$

On the other hand, the chance that we make the right decision (given that H_i is true) is written as γ_i. Note that we have $\gamma_i = 1 - \beta_i$. (You might see that we also have $\gamma_i = \beta_{ii}$. This is correct. However, the sign β is normally used to indicate errors. So that's why we use the sign γ_i now, and not β_{ii}.)

You probably already expect the next question we will ask to ourselves. What would be the chance that we are wrong in general? This chance, called the **average probability of incorrect decision**, can be found using

$$\sum_{i=1}^{p} P(\mathbf{x_i})\beta_i = \sum_{i=1}^{p} \left(P(\mathbf{x_i}) \sum_{j=1, j\neq i}^{p} \beta_{ij} \right) = \sum_{i=1}^{p} \left(P(\mathbf{x_i}) \sum_{j=1, j\neq i}^{p} \int_{A_j} f_{\underline{\mathbf{y}}}(\mathbf{y}|\mathbf{x_i})\, dy \right). \tag{6.2.4}$$

If the acceptance regions are well defined, then this chance is minimal.

6.3 Other Methods of Testing

We just saw one way in which we can choose a hypothesis. Naturally, there are more ways. In this part we will examine another way to choose from two hypotheses H_0 and H_a.

6.3.1 The simple likelihood ratio test

You probably remember the maximum likelihood estimation (MLE) method, from the previous chapter. In that method, we looked for the $\mathbf{x}$ for which $f_{\underline{\mathbf{y}}}(\mathbf{y}|\mathbf{x})$ was maximal. We can do the same now. However, now we only have two possible values of $\mathbf{x}$, being $\mathbf{x_0}$ and $\mathbf{x_a}$. We thus accept H_0 if $f_{\underline{\mathbf{y}}}(\mathbf{y}|\mathbf{x_0}) > f_{\underline{\mathbf{y}}}(\mathbf{y}|\mathbf{x_a})$. We reject H_0 if

$$\frac{f_{\underline{\mathbf{y}}}(\mathbf{y}|\mathbf{x_0})}{f_{\underline{\mathbf{y}}}(\mathbf{y}|\mathbf{x_a})} < 1. \tag{6.3.1}$$

The critical region K can be derived from the above criterion. This testing method is called the **maximum likelihood test**.

Let's adjust the above method slightly. Instead of taking 1 as a boundary, we now take some (positive) constant c. We thus reject H_0 if

$$\frac{f_{\underline{\mathbf{y}}}(\mathbf{y}|\mathbf{x_0})}{f_{\underline{\mathbf{y}}}(\mathbf{y}|\mathbf{x_a})} < c. \tag{6.3.2}$$

We now have arrived at the **simple likelihood ratio** (SLR) test.

6.3.2 The most powerful test

We find another way of testing when we apply the **Neyman-Pearson** testing principle. To apply this principle, we should first give the probability of false alarm α (the size of the test) a certain value. We then examine all tests (or equivalently, all critical regions K) with size α. We select the one for which the probability of missed detection β is minimal. (Or equivalently, the one for which the power of the test γ is maximal.) The selected test is called the **most powerful** (MP) test of size α.

Let's take another look at the conditions. The value of α should be set, and the value of γ should be maximal. Now let's look at the simple likelihood ratio test. We can choose our ratio c such that the test has size α. This makes sure the first condition is satisfied. Now comes a surprising fact. The SLR test also always satisfies the second condition. In other words, the SLR test is always the test with maximal γ — it is always the most powerful test.

So, although we may have believed we had two new testing methods, we only have one. But we do always know which test is the most powerful one: the simple likelihood ratio test.