Notes for Introduction to Business Statistics

Sarah Thandi Dippenaar
University of Vienna

version: 20 May 2013

These notes are work in progress. I would be very grateful for comments and identification of errors. Please send these to me at: sarah.dippenaar@univie.ac.at

Contents

1 Sets (*Mengen*)

For an excellent (more detailed) introduction to sets in German see Schichl, Hermann and Roland Steinbauer (2009) 'Einführung in das mathematische Arbeiten', Springer-Verlag. Chapter 4.

1.1 Introduction

[1]

The idea of sets is very important for what we will do in this course. The main concepts that will be required are outlined in this section.

Definition 1 ***(Set)****. We define a set (Menge) as a well-defined (wohldefiniert) collection of elements (Elemente).*

The basic idea can be expressed by comparing a set to a bag filled with objects. This bag can contain other bags; i.e. a set can contain other sets. The bag can also be empty. We call this the empty set. Important to the idea of a set is the concept 'well-defined'. This simply means that an object in a set cannot occur more than once. A set is thus a collection of *different* objects or elements (this will become clearer later).

Examples of Sets.

- The set of all students in a room.
- The set of natural numbers $\mathbb{N}$
- The set of all positive real numbers $\mathbb{R}_+$

1.2 Definitions and Notation

SPECIFYING SETS

There are two ways in which we can specify sets.

1. list the elements

   $A = \{1, 2, 5, 10\}$

   This set contains the integers 1, 2, 5 and 10.

   $B = \{1, 2, 3, 4, \ldots\}$

   This set includes the natural numbers $\mathbb{N}$. We, of course, cannot write down all the natural numbers, but this notation assumes that we know how this set continues due to '$\ldots$'.

[1] In this course we restrict ourselves to 'naive set theory' (*naive Mengenlehre*).

2. property method

   With this method we describe the set in 'mathematical language'. We can define the set B from above as

   $B = \{x : x \in \mathbb{N}\}$

   This means that the set B is made up of all x's that are contained in the natural numbers. We read ':' B is the set of all x's such that x is contained in the natural numbers, where ':' means 'such that'. The information following : defines the properties of that which precedes it.

Definition 2 ***(element in (element aus))****. If an element a belongs to the set A, we write*

$a \in A$

and say a is in A, a is an element of A or a is contained in a.

Definition 3 ***(Subset, Superset)****. The set B is called a subset (Teilmenge) of the set A if the set B only contains elements also contained in A. A is then called the superset (Obermenge) of B.*

We denote this as $B \subset A$. We will use this notation for a proper subset (*echte Teilmenge*). This means that A cannot be equal to B; i.e A contains all elements of B, but B does not contain all elements of A. An alternative representation would be $B \subseteq A$, which allows A and B to be equal.

Examples.

- Let $A = \{2, 5, 7, 8\}$ and $B = \{2, 5\}$. Then $B \subseteq A$. It is also correct to write $B \subset A$ as B is a real subset ob A. Further, we can write $A \subseteq A$; i.e. A is a subset of itself. However, it it incorrect to say $A \subset A$ as a set cannot be a real subset of itself.
- $\mathbb{N} \subseteq \mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R} \subseteq \mathbb{C}$, where $\mathbb{N}$ denotes the natural numbers, $\mathbb{Z}$ integers, $\mathbb{Q}$ the rational numbers, $\mathbb{R}$ the real numbers and $\mathbb{C}$ the complex numbers.

Definition 4 ***(Equality of sets)****. Two sets A and B are equal (gleich) if and only if they contain the same elements.*

We write $A = B$.

Examples.

- Let $A = \{apple, orange, pear\}$ and $B = \{pear, apple, orange\}$. Then obviously $A = B$. Note that the order of the elments in the set does not matter.

- Let $C = \{1, 3, 17, 300\}$ and $= D\{1, 3, 3, 17, 3, 17, 300\}$. Although in the set D the element 3 is listed three times and the element 17 twice, C and D remain equal. This follows from the definition of sets, where we required the elements of a set to be well defined, in other words different from one another.

Definition 5 ***(Empty set).*** *The empty set (leere Menge) is the set that contains no elements.*

The empty set is commonly denoted as $\emptyset$ or $\{\}$. An important property of the empty set is that it is a subset of every set.

Definition 6 ***(union (Vereinigung)).*** *The union of two sets A and B is the set of elements contained in A or B. Formally*
$A \cup B = \{a : x \in A \;\; or \;\; x \in B\}$

Examples.

- Again, let $A = \{2, 5, 7, 8\}$ and $B = \{2, 5\}$. Then the union of the two sets is $A \cup B = \{2, 5, 7, 8\} = A$.
- Let $C = \{x : x \in (0, 1)\}$ and $D = \{x : x \in [1, 3)\}$, then $C \cup D = \{x : x \in (0, 3)\}$.

Definition 7 ***(intersection (Durchschnitt)).*** *The intersection of two sets A and B is the set of elements that are contained in both A and B. Formally,*
$A \cap B = \{a : x \in A \;\; and \;\; x \in B\}$

Examples.

- Again, let $A = \{2, 5, 7, 8\}$ and $B = \{2, 5\}$. Then the intersection of the two sets is $A \cap B = \{2, 5\} = A$.
- Let $C = \{x : x \in (0, 2)\}$ and $D = \{x : x \in [1, 3)\}$, then $C \cap D = \{x : x \in [1, 2)\}$.

Definition 8 ***(disjoint sets(disjunkte Mengen)).*** *If two sets A and B are disjoint (disjunkt) they contain no elements in common; i.e.*
$A \cap B = \emptyset$

Examples.

- Let $A = \{7, 8\}$ and $B = \{2, 5\}$. Then the intersection of the two sets is $A \cap B = \emptyset$. So, A and B are disjoint sets
- Let $C = \{x : x \in (0, 1)\}$ and $D = \{x : x \in (1, 3)\}$, then $C \cap D = \emptyset$ and the two sets are disjoint.

Definition 9 ***(complement (Komplement))****. The complement A^c is the set of elements not contained in A.*

Examples.

- Let $A = \{1, 23, 5\}$ and $B = \{1, 5\}$, which is a subset of A. Then B^c with regards to A is $B^c = \{23\}$.
- Let U be the universal set (the set that contains all elemts). The compelent of the universal set $U^c = \emptyset$.
- Again, let U be the universal set. The complement of the empty set $\emptyset^c = U$.

Definition 10 ***(difference (Differenz))****. The difference of the sets A and B, written as $A \setminus B$, is the set of elements contained in A but not in B. Formally,*
$A \setminus B = \{x : x \in A, x \notin B\}$

Examples.

- Let $A = \{1, 23, 5\}$ and $B = \{1, 5\}$. Then $A \setminus B = \{23\}$.
- Let $C = \{x : x \in (-1, 2)\}$ and $D = \{x : x \in (0, 2)\}$. Then $C \backslash D = \{x : x \in (-1, 0]\}$.

1.3 Algebra of Sets

Let A, B and C be sets and U the universal set.

1. $A \cup A = A$
2. $A \cap A = A$
3. $(A \cup B)C = A \cup (B \cup C)$
4. $(A \cap B)C = A \cap (B \cap C)$
5. $A \cup B = B \cup A$
6. $A \cap B = B \cap A$
7. $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
8. $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
9. $A \cup \emptyset = A$
10. $A \cap \emptyset = \emptyset$
11. $A \cap U = A$

12. $A \cup U = U$

13. $(A^c)^c = A$

14. $A \cap A^c = \emptyset$

15. $A \cup A^c = U$

16. $U^c = \emptyset$

17. $\emptyset^c = U$

There are two more important rules that carry the name **de Morgan's laws** (without proof):

$(A \cup B)^c = A^c \cap B^c$

$(A \cap B)^c = A^c \cup B^c$

2 Basics of Probability

2.1 Sample Space and Events

Probability theory helps us to look at things that are connected to chance. We will call such things experiments (*Zufallsexperiment*). We do not know what the outcome of an expereiment will be in advance, but often we konw what the possible outcomes are. We call this set of possible events the **sample space (*Ereignisraum*)** and will denote it as Ω. Each element of the sample space is called an **event (*Ereignis*)**.

Example. If we throw a coin it can land on either heads (H) or tails (T). Then

$\Omega = \{H, T\}$

$\{H\} \subset \Omega$ and $\{T\} \subset \Omega$ are events

Example. If we throw two coins

$\Omega = \{(H, H), (H, T), (T, H), (T, T)\}$

$\{(H, H), (H, T), (T, H), (T, T)\}$, $\{(H, H)\}$, $\{(H, T)\}$, $\{(T, H)\}$, $\{(T, T)\}$, $\{(H, H), (H, T)\}$, $\{(H, H), (H, T), (T, T)\}$ and $\{(H, H), (T, H), (T, T)\}$ etc. are events

Example. The lifetime of a car
$\Omega = \{x : 0 \leq x < \infty\}$

Let us take a closer look at the second example from above where we threw two coins. We can define the event that at least one of the coins lands on H and let us call this event A.

$A = \{(H, H), (T, H), (H, T)\}$

The complement of A will then be $\Omega \setminus A$. Let us call this B.

$A^c = B = \{(T, T)\}$

Note that A and B are disjoint sets whose union is equal to the sample space; i.e.

$A \cup B = \Omega$ and $A \cap B = \emptyset$

Further we define two sets C and D as

$C = \{(H, H)\}$
$D = \{(H, H), (H, T)\}$

Then the union is

$C \cup D = \{(H, H), (H, T)\}$

and the intersection

$C \cap D = \{(H, H)\}$

2.2 Axioms of Probability

We are interested in the probability of events. To be able to talk more formally about this we will denote the probability that an event E will occur as $P(E)$. $P(E) = 1$ will mean that the event is sure to happen and $P(E) = 0$ that the event cannot occur.

Most of probability theory is based on an axiomatic approach. An axiom is a statement that is taken to hold from which one deduces other things that we call theorems and propositions. What is remarkable is that for that which we will look at in this course we require but three axioms.

Axiom 1

Let E be an event.

$0 \leq P(E) \leq 1$

This means that the probability of an event lies between (including) 0 and 1.

Axiom 2

Let Ω be the sample space.

$P(\Omega) = 1$

This axiom states that the probility of an event in the sample space occuring is 1. In other words, it is sure that at least one of the events the sample space will occur.

Axiom 3

Let A and B be two disjoint subsets (events) of the sample space Ω.

For two mutually exclusive events (*einander ausschlißende Ereignisse*) A and B

$P(A \cup B) = P(A) + P(B)$

This means that the probability of the union of two mutually exclusive events (two disjoint sets) is equal to the sum of the probability of the individual events.

This axiom can be written in the general form for the mutually exclusive events E_i

$P(\cup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} E_i$

Example. If we throw a fair coin

$$P(\{H\}) = P(\{T\}) = \frac{1}{2}$$

Example. In the case of two fair coins

$$P(\{(H,H)\}) = P(\{(H,T)\}) = P(\{(T,H)\}) = P(\{(T,T)\}) = \frac{1}{4}$$

We will learn later how to derive such probabilities.

2.3 Some Propositions

We state some important and useful propositions. They can be proven using the three axioms from above. However, we will simply state them and not give any proofs.

1. $P(\emptyset) = 0$
2. $P(E^c) = 1 - P(E)$
3. if $E \subseteq F$, then $P(E) \leq P(F)$
4. $P(E \cup F) = P(E) + P(F) - P(E \cap F)$
5. $P(E \setminus F) = P(E) - P(E \cap F)$

2.4 Sample spaces with equally likely outcomes

After having discussed the very basics of probability, we conclude this chapter with one last definition that will accompany us through much of this course. We introduce the idea of sample space Ω in which all outcomes have equally likely probability.

Definition 11 ***(Laplace'sche Wahrscheinlichkeit)***. *Let all elements of the sample space Ω have the same probability of occuring. Let A be an event in Ω, then*

$$P(A) = \frac{|A|}{|\Omega|}, \text{ where}$$

$|A|$ is defined as the number of elements in A, which is called the cardinality (Mächtigkeit) of A.

Example. Previously we concluded that the sample space of throwing a die is

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

Then $|\Omega| = 6$. Let us define the events A and B as the events that the number two and an even number is cast respectevily. Then $|A| = 1$ and $|B| = 3$. Thus,

$$P(A) = \frac{|A|}{|\Omega|} = \frac{1}{6} \text{ and}$$

$$P(B) = \frac{|B|}{|\Omega|} = \frac{3}{6} = \frac{1}{2}.$$

3 Conditional Probablitly

3.1 The basic idea

Sometimes it may occur in an experiment that we come by more information than we had in the beginning and that we can use this to calculate the probability of particular events given this information.

To motivate this idea, let us go back to the eternal example of rolling a die. Let us further assume that each of the six numbers of the die have equal probability of coming out on top. We already konw that

$$\Omega = \{1, 2, 3, 4, 5, 6\} \text{ and } P(i) = \frac{1}{6} \text{ for all } i \in \{1, 2, 3, 4, 5, 6\}$$

Imagine now, that we know that the outcome is an even number; we define $E = \{2, 4, 6\}$. Given this information, we are interested in finding out the probability that the number that fell is four; i.e. $P(4)$. To calculate this, we introduce the following:

Definition 12 ***(Conditional Probability (Bedingte Wahrscheinlichkeit)).*** *Let A and B be events in the sample space Ω. The probability that the event A occurs given that B has occured is called the conditional probability of A given B (bedingte Wahrscheinlichkeit von A gegeben B). We write $P(A|B)$. If $P(B) > 0$, then*

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{1}$$

We can see this as the situation where we have compressed the original sample space Ω to the reduced sample space B.

We return to our example from above. We know that $P(E) = P(2) + P(4) + P(6) = \frac{1}{2}$. We define the event that the number four occurs as $F = \{2\}$. Then $F \cap E = \{2\}$. Further, we know from what we have already learned, that $P(F \cap E) = P(F) = \frac{1}{6}$. We are now ready to apply definition 12:

$$P(A|E) = \frac{P(A \cap E)}{P(E)} = \frac{P(A)}{P(E)} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$$

So, given the information that the number that fell is even, we were able to deduce that the probability that the number four fell is two thirds. We look at further cases of this example. We now let $B = \{2, 5\}$. $B \cap E = \{2\}$. Then,

$$P(B|E) = \frac{P(B \cap E)}{P(E)} = \frac{P(\{2\})}{P(E)} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$$

Now, let $C = \{3, 5\}$. Then, $C \cap E = \emptyset$.

$$P(C|E) = \frac{P(C \cap E)}{P(E)} = \frac{P(\emptyset)}{P(E)} = \frac{0}{\frac{1}{2}} = 0$$

3.2 Multiplication theorem for conditional probabilities

If we multiply both sides of 12 with $P(E)$, we get

$$P(A \cap E) = P(E)P(A|E)$$

This is the multiplication theorem for conditional probabilities (*Multiplikationssatz für bedingte Wahrscheinlichkeiten*) and comes quite in handy sometimes.

Naturally, we can generalise the multiplication theorem to cover more than just two sets.

Let $E - 1, E_2 \ldots, E_n$ be n different events. More succinctly we can write these events as E_i, where $i = 1, 2, \ldots, n$. Then the multiplication theorem can be rewritten as

$$P(E_1 \cap E_2 \cap \ldots \cap E_n) = P(E_1)P(E_2|E_1)P(E_3|E_2 \cap E_1) \ldots P(E_n|E_1 1 \cap E_2 \cap \ldots E_{n-1})$$

To verify this equation, apply the definition of conditional probability to the right hand side of the equation and simplify.

Example. Three are 12 batteries in a box. 4 of these are broken. Without replacement, 3 batteries will be randomly drawn from this box. We want to calculate the probability that all 3 of these batteries are functions. To go about this we define the following events:

A... the first battery we pull out is working
B... the second battery we pull out is working
C... the third battery we pull out is working

We now want to calculate the probability that A, B and C will all occur. In other words, we want to know the probability described by $P(A \cap B \cap C)$.
From what we have learnt we know that we can write this using the multiplication theorem as follows:

$$P(A \cap B \cap C) = P(A)P(A|B)P(A|B \cap C)$$

We have to now find the probabilities on the right hand side of the equation.

$P(A) = \frac{8}{12}$
$P(A|B) = \frac{7}{11}$
$P(A|B \cap C) = \frac{6}{10}$
Then $P(A \cap B \cap C) = P(A)P(A|B)P(A|B \cap C) = \frac{8}{12}\frac{7}{11}\frac{6}{10} = \frac{14}{55} \approx 0.25$

3.3 Total Probability

Let A_i, with $i = 1, 2, \ldots, n$ be mutually exclusive events (i.e. $A_i \cap A_j = \emptyset$ for all $i \neq j$). Further, let the event E be a subset of the union of the $A_i s$ ($E \subset \bigcup_{i=1}^{n} A_i$).

Then it holds that
$P(E) = P(E|A_1)P(A_1) + P(E|A_2)P(A_2) + \ldots P(E|A_n)P(A_)$
Why does this hold? Recall that we can write
$E = (E \cap A_1) \cup (E \cap A_2) \cup \ldots (E \cap A_n)$

and from the Axiom 3, we know that the probability of the union of events is the same as the sum of the probablities of their individual events. We call this total probability (*Satz der totalen Wahrscheinlichkeit*).

Example. (taken from Futschick et al, (2010)) A factory has three machines X, Y and Z that each produce different objects.

- X produces 50% of the objects of which 3% don't funtion
- Y produces 30% of the objects of which 4% don't funtion
- Z produces 20% of the objects of which 5% don't funtion

We define D as the event that arbitrary object produced in the factory does not function. Now we want to calculate the probability of D with the information that we have:

$$\begin{aligned} P(D) &= P(X)P(D|X) + P(Y)P(D|Y) + P(Z)P(D|Z) \\ &= (0.5)(0.03) + (0.3)(0.04) + (0.2)(0.03) \\ &= 0.037 \end{aligned}$$

3.4 Bayes

In this section we look at Bayes' Formula (*Bayes-Formel*) and the Theorem of Bayes (*Theorem von Bayes*). But before we get to these important results, we will look at a couple of other things to help us along the way.

Let E and F be events. And as we have seen previously, we can write

$E = (E \cap F) \cup (E \cap F^c)$

From Axiom 3, we further know that

$$\begin{aligned} P(E) &= P(E \cap F) + P(E \cap F^c) \\ &= P(E|F)P(F) + P(E|F^c)P(F^c) \\ &= P(E|F)P(F) + P(E|F^c)[1 - P(F)] \end{aligned}$$

Before we procede to define Bayes' Formula we need to take another look at the definition of conditional probablitiy. Recall that for two events E and F, we said that

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

Further we noted that we can thus rewrite the above as

$$P(E \cap F) = P(E|F)P(F)$$

We are now ready to define Bayes' Formula

Definition 13 ***(Bayes' Formula (Bayes-Formel)).*** *Let E and F be two events. Then*

$$P(E|F) = \frac{P(F|E)P(E)}{P(F)} = \frac{P(F|E)P(E)}{P(E|F)P(E) + P(F|E^c)P(E1c)}$$

Take a close look at this formula to make sure that you know where we used which result that we derived at some point earlier.

The Theorem of Bayes, which we will define next, follows the same ideas and principles as those we used for definition 13. But it also allows us to work with more than two events.

Definition 14 ***(Bayes' Theorem (Theorem von Bayes)).*** *Let E and A_i, with $I = 1, 2, \ldots, n$ be events, such that $E \subset \bigcup_{i=1}^{n}$. Then it holds that*

$$P(A_i|E) = \frac{P(E|A_i)P(A_i)}{\sum_{i=1}^{n} P(E|A_i)P(A_i)}$$

3.5 Independent Events

Independence of events will prove to be very useful later on and we will define it here.

Definition 15 . *Let A and B be two events. A and B are said to be independent (unabhängig) if the following holds*

$$P(A \cap B) = P(A)P(B)$$

and dependent otherwise.

Three events E, F and G are said to be independent if the following hold:

$P(E \cap F \cap G) = P(E)P(F)P(G)$
$P(E \cap F) = P(E)P(F)$
$P(E \cap G) = P(E)P(G)$
$P(G \cap F) = P(G)P(F)$

3.5.1 A note of caution

Independent and mutually exclusive events are often confused. We will show here that the two are absolutely **not** the same.

Let A and B be two independent, nonempty events. Then

$$P(A \cap B) = P(A)P(B)$$

Can A and B be mutually exclusive events (i.e. disjoint sets)? No! Why?

If the two sets are mutually exclusive, it follows that $A \cap B = \emptyset$. We learnt at the beginning that $P(\emptyset) = 0$. From this it would have to follow that

$$0 = P(A)P(B)$$

But this cannot be as $P(A) = 0$ or $P(B) = 0$ (this implies that the sets A or B would be empty) would have to hold and this cannot be as we required both these sets to be nonempty in the beginning. Therefore, two nonempty, independent sets cannot be mutually exclusive!

4 Inferential Statistics Basics

Up until this point we have been looking at elements of probability theory. With this chapter we begin to deal with inferential statistics (*Inferenzstatistik*). Inferential statistics is that which most people understand by the term 'statistics'.

So then, what is the difference between the two and why didn't we start with inferential statistics in the first place? The latter question is easier to answer: one cannot understand inferential statistics without a certain understanding of probability. The first question is not difficult to answer either, but requires a little more space.

We began the previous chapters by defining the sample space, learned how to present its elements as random variables and then found that one can describe their 'behaviour' with different distributions. The central point is that in these cases we knew exactly what is contained in the sample space and then we were able to calculate the probability of certain events of interest. We often looked at examples of throwing fair coins. We were able to describe and in this case even list all the possible outcomes. Recall the example of throwing a fair coin twice. We wrote down the sample space and then calculated the probability of the event that both times the coins would show heads.

In inferential statistics, we do pretty much the opposite. We now no longer know what the sample space looks like. Instead, we have certain information about that which is of interest to us that we were able to observe. This information is called data. Using these data we want to make conclusions about the entire sample space.

Before we can really get started, we still need to define a few concepts and introduce some new ideas.

4.1 Definitions and Concepts

In this section we introduce some terms and concepts that are important when talking about inferential statistics.

Population (*Grundgesamtheit*). In inferential statistics, the population refers to the set about which we are trying to understand certain things. A population can be many things: all people living in Austria, all students in the world that are studying business economics or drosophila.

Sample (*Stichprobe*). A sample is a subset of the population. In inferential statistics samples provide the data we use to make conclusions about certain characteristics of the population.

Random sample (*Zufallsstichprobe*). A random sample is a sample of the population of interest that was randomly produced. In many cases it is desirably - as we will discuss later - that a sample be random.

Estimator (*Schätzer*). In the sections on probability we introduced parameters for random variables such as the expected value ($E(X) = \mu$) and the variance ($Var(X) = \sigma^2$). We were able to calculate these with the information we had. In inferential statistics we very often do not know what these parameters for a population are. So, we try to estimate them from the samples we have from the population. An estimator is a 'formula' that we apply to the data from the sample. A common estimator for μ is the mean $\overline{x}$ (see chapter 5).

Estimate (*Schätzung*). An estimate for a population parameter is simply the numerical outcome of applying data to an estimator. We may calculate that the mean hight of people in a sample is 1.7m; ie. $\overline{x} = 1.7$. The formula for $\overline{x}$ is the estimator and 1.7 is the estimate.

In essence, the story about inferential statistics as we will treat it here is that we have a population about which we would like to know something that we don't know. Because we cannot look at every element of the population, we look at a sample taken from this population and estimate the things we want to know from the data generated by the sample.

4.2 An illustrative example

Assume a population is given by a set $S = \{4, 7, 10\}$ and each element of the set has equal probability of occuring. We can then actually calculate the population mean μ and variance σ^2.

$\mu = 7$ and $\sigma^2 = 6$

Now imagine we take samples of size $n = 2$ with replacement. The following table lists all possible samples from the population together with the sample average $\overline{x}$.

(a,b)	$\overline{x}$
(4,4)	4
(4,7)	5.5
(4,10)	7
(7,4)	5.5
(7,7)	7
(7,10)	8.5
(10,4)	7
(10,7)	8.5
(10,10)	10

And now we derive the sampling distribution for $\overline{x}$.

$\overline{x}$	$P(\overline{x})$
4	$\frac{1}{9}$
5.5	$\frac{2}{9}$
7	$\frac{3}{9}$
8.5	$\frac{2}{9}$
10	$\frac{1}{9}$

With this we can calculate the expected value and variance of $\overline{x}$.

$E(\overline{x}) = 7$ and $\sigma^2_{\overline{x}} = 3$

There are a number of things that are exemplified in this simple example. First, the sample mean of 7 (i.e. $\overline{x} = 7$) has a higher probability of occuring than any of the other sample means. This is no coincidence; 7 is the population mean. Second, the expected value of $\overline{x}$ is also equal to the true population mean. The sample variance $\sigma^2_{\overline{x}}$ is equal to 3. This is not the same as the population variace, but as is discussed in chapter 6 on sample distributions, the variance of $\overline{x}$ is $\frac{\sigma^2}{n}$. In this example this would be $\frac{6}{2}$, which is 3.

4.3 There's more to it...

What is described as inferential statistics above is just a snippet of the whole story. Yes, there's more to it than that. A lot more, actually. But we will not venture that far in this class and hence not in this text[2]. Just to give an idea: we have not at all talked about causal relationships between variables and much of statistics is dedicated to such questions.

[2] Of course, the same holds for the treatment of probability theory in this text.

5 Some Descriptive Statistics

When we collect data to measure something or gather information, the first thing we would want to do is to describe it. This section briefly presents some of the most common types of descriptive measures that we encounter in textbooks, the media, but also in academic research. We can also be confronted with many different types of data and we first discuss these. After this, we look at different types of ways to numerically describe data.

We define the following notation:
n . . . number of observations
x_i . . . observation i

5.1 Levels of Measurement

Nominal level (Nominalskaliert) Nominal data are items that are differentiated by a naming system. The names refer to different characteristics (*Ausprägungen*) that something can take on. Examples are things such as eye colour, countries or names of people. Data at the nominal level are qualitative. It does not make sense to calculate something like the mean or standard deviation (see below) of nominal data.

Ordinal level (Ordinalskaliert) Ordinal data are data that have an order (nominal data do not). For example, placement in a race - first, second, et. - has an order, but no meaning can be given to the difference in the placements; i.e. the placements cannot be used to answer the question of 'how much more?'. Ordinal data should also not be used to calculate some of the measures we will discuss below.

Interval level (Intervallskaliert) In the case of interval scaled data, the idea of difference does have meaning, but it does not have starting point. The most commonly cited example is temperature. $30°C$ is $20°C$ warmer than $10°C$, but $0°C$ does not mean that there is no temperature.

Ratio level (Verhältnisskaliert) For data at the ratio level the difference between two values makes sense and there is a starting point. For example between 10 and $30km$ there is a difference of $20km$ and the idea of $0km$ has meaning in the sense that it is the absence of distance.

Data are metric (*metrisch*) if they are at interval or ratio level. These can be discrete (*diskret*) or continuous (*kontinuierlich*).

5.2 Measures of Central Tendency

Here we look at different measures for the centre or middle of data.

Definition 16 ***(arithmetic mean(arithmetisches Mittel))***

$$\overline{x} = \frac{x_1 + x_2 \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2}$$

Definition 17 ***median((Median))***

$$\begin{array}{ll} n\ odd: & \widetilde{x} = x_{(n+1)/2} \\ n\ even: & \widetilde{x} = (x_{n/2} + x_{(n+2)/2})/2 \end{array} \tag{3}$$

Definition 18 ***(geometric mean(geometrisches Mittel))***

$$\overline{x_G} = (x_1 x_2 \ldots x_n)^{\frac{1}{n}} \tag{4}$$

5.3 Measures of Dispersion

Definition 19 ***(variance(Varianz))***

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x}) \tag{5}$$

Definition 20 ***(standard deviation(Standardabweichung))***

$$s = \sqrt{s^2} \tag{6}$$

6 Sample Distributions

In the previous chapter, we said that we are interested in certain characteristics that we cannot directly observe of a given population. In the sections on probability theory we saw that certain parameters were central for describing the behaviour of random variables.

In the case of binomial random variables, it was the parameter p describing the probability of success. For both the Poisson and the exponential distributions it was the parameter λ that was essential for determining the probabilities of outcomes. And for the normal distribution there were two parameters of interest; the expected value μ and the variance σ^2. (In this section we will start to get a sense of why the normal distribution is so important.)

But now we are faced with the situation where we do not know these parameters. All we have is a (hopefully good) sample from the population. From this sample, we can produce estimates for the unknown population parameters. For this, we use estimators.

6.1 Estimators for Parameters

We will concentrate on the parameters p, μ and σ^2 and present the estimators for these here[3]. (We have already seen estimators for the last two in chapter 5.)

Sample mean

The sample mean $\overline{x}$ serves as an estimator for μ. It is given by

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{7}$$

Sample Variance

The sample variance, denoted by s^2 is given by

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x}) \tag{8}$$

The question as to why we divide by $n-1$ and not simply n will probably arise. Suffice it to say, dividing by $n-1$ instead of simply n provides us with a better estimator for the variance[4].

[3]We will not discuss how these estimators are derived or what their characteristics are, nor how good they are. Nevertheless, they are by and large intuitive

[4]Dividing by $n-1$ makes s^2 unbiased (erwartungstreu).

Sample proportion

When we looked at binomially distributed random variables, the parameter p gave us the probabiliy for the success of an experiment. If we have a sample of size n, we can count how many successes and failures occured. If we there were m successes in the sample, a plausible estimator, which we will call $\widehat{p}$, for p would be

$$\widehat{p} = \frac{m}{n} \tag{9}$$

From this, it is clear why we talk about sample proportions (*Anteilswerte*) in such a situation.

It is important to note that these estimators are also random variables. Intuitively, we can make sense of this if we consider that we can take many different samples from a population and that p, μ and σ^2 will vary accross these samples. It can then happen that one sample provides an estimate for a parameter that is much closer to the true value of the population parameter than another sample. Therefore, methods were developed to test 'how good' an estimate from a sample is. We will discuss some of these later.

For now, it is important to understand that these estimators are random variables and that these random variables also have distributions. We now discuss the distributions of our three estimators.

6.2 Sample Mean

The sample mean $\overline{x}$ is an estimator for the true population mean μ. A mean usually comes with a variance. For the distribution of the sample mean, we have to distinguish between two possible cases:

1. the true population variance σ^2 is known
2. the true population variance σ^2 is unknown

Note that in both cases we do not know μ, which is why we are looking at the 'behaviour' of its estimator $\overline{x}$.

6.2.1 variance known

Let $X_1, \ldots, X_n$ denote random variables with mean μ and variance σ^2. The sample mean $\overline{x}$ then has mean μ, and variance $\frac{\sigma^2}{n}$; that is

$$E(\overline{x}) = \mu \text{ and } Var(\overline{x}) = \frac{\sigma^2}{n}$$

What are we to understand by $E(\overline{x}) = \mu$? It simply means that if we repeatedly took a sample from the population, the average value of the sample mean would equal the actual population mean. (Section 4.2 illustrates this point.)[5]

If in addition the random variables $X_1, \ldots, X_n$ are all normally distributed mean μ and variance σ^2, then $\overline{x}$ is also normally distributed with mean μ and variance $\frac{\sigma^2}{n}$.

This is very convenient, as we know how to work with the normal distribution. But what if the random variables $X_1, \ldots, X_n$ are not normally distributed? In this case the Central Limit Theorem (*Zentraler Grenzwertsatz*) comes to the rescue:

Central Limit Theorem
Suppose that a random variables $X_1, \ldots, X_n$ follow a distribution other than the normal distribution with mean μ and variance σ^2 (all X_i must have the same distribution). If n is large enough, then the sample mean $\overline{x}$ is normally distributed with ean μ and variance $\frac{\sigma^2}{n}$, i.e.

$\overline{x} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$

Of course, this scenario, where we know the true population variance, is less realistic than the one we discuss next.

6.2.2 variance unknown - the t-distribution

If the population variance σ^2 is unknown - which is in most cases very likely - we have to use the sample variance s^2 as an estimator. In this case, it was shown, that a distribution other than the normal distribution fares better than the normal in describing the 'behaviour' of $\overline{x}$. This distribution is the t-distribution (also known as the student-distribution). In other words, when we standardise $\overline{x}$ with

$$T = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

this random variable follows the t-distribution with $n - 1$ degrees of freedom (*Freiheitsgrade*). We shall denote the degrees of freedom as q.

Characteristics of the t-distribution
The t-distribution resembles the normal distribution in shape. Like the normal distribution it is a bell shaped curve and is symmetric about its mean. The mean is 0 and the variance 1.

[5] One may also ask why we divide the variance by n. This is not all too difficult to derive, but shall not be done here.

Its cumulative distribution function[6] (*Verteilungsfunktion*) is given by

$$F_q^{(t)}(t) = P(T \leq t),$$

where q represents the degrees of freedom. To find the values we are looking for, there are again - like for the standard normal distribution - tables where we can look them up. However, the tables for the t-distribution vary with the degrees of freedom q. So we need to look at the table with the 'right' degrees of freedom. From the symmetry of the distribution it follows that

$$F_q^{(t)}(-t) = 1 - F_q^{(t)}(t).$$

For a random variable T that follows a stundent distribution with q degrees of freedom, it holds that

$$\begin{aligned} P(|T| \leq x) &= P(-x \leq T \leq x) \\ &= F_q^{(t)}(x) - F_q^{(t)}(-x) \\ &= F_q^{(t)}(x) - (1 - F_q^{(t)}(x)) \\ &= 2F_q^{(t)}(x) - 1 \end{aligned}$$

Now, we know how to find the probability that a radom variable T (recall that T is the standardised sample mean) lies below a certain value or between two given values. All this is familiar to us from our discussion of the standard normal distribution.

But, instead, we may want to find the value for which T lies below with a certain probability or the interval in which T will be contained with a certain probability. This we can do much in the same way as we did for the standard normal distribution.

For example, we may want to know in which interval of the t-distribution the random variable T may lie with a particular probability $\alpha \in [0, 1]$. Let's look at a concrete example. We want to know for a random variable T with 24 degrees of freedom for which value x

$$P(-x \leq T \leq x) = 0.95$$

will hold. How do we go about this? Much in the same way as we did for the standard normal distribution. We already know that this is the same as solving the following for x

$$\begin{aligned} &0.95 = 2F_{24}^{(t)}(x) - 1 \\ &\Rightarrow 0.95 + 1 = 2F_{24}^{(t)}(x) \\ &\Rightarrow \frac{0.95 + 1}{2} = F_{24}^{(t)}(x) \end{aligned}$$

[6] We do not look at the density function of this distribution. For our purposes, it suffies to look only at the distribution function.

To complete this, we introduce the inverse function (*Umkehrfunktion*) for $F_q^{(t)}(x)$ which is denoted by $\mathcal{Q}_q^{(t)}$. Applying the inverse, we get

$$x = \mathcal{Q}_{24}^{(t)}\left(\frac{0.95+1}{2}\right) = \mathcal{Q}_{24}^{(t)}(0.975)$$

$\mathcal{Q}_{24}^{(t)}(0.975)$ can be found on the table for the t-distribution with 24 degrees of freedom and therefore

$$x = 2.064.$$

6.3 Sample Proportion

Suppose we want to look at the proportion in a sample that exhibits a certain characteristic. We learned previously that the binomial distribution [7] can be used in some such situations.

In 1812, Pierre-Simon Laplace proved that if a random variable has a binomial distribution with parameters p and n, its standardised form (subtracting its mean np and dividing by its standard deviation $\sqrt{np(1-p)}$) converges towards the standard normal distribution. Due to this we can state the[8]

DeMoivre-Laplace Theorem

When n independent trials, where each has a probability p for success, are performed and S_n denotes the number of successes, then for any $a < b$ it holds that

$$P\left(a \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq b\right) \overset{n\to\infty}{\to} \Phi(b) - \Phi(a)$$

This looks very helpful as we know how to deal with the standard normal distribution. However, there are two things to keep in mind. First, we don't know p. That is not such a problem; we can just use our estimator $\widehat{p}$. (Note: the m in the formula we gave earlier is simply the S_n in the statement of theorem above.) Second, binomial variables are discrete, whereas the normal distribution is continuous.

On the one hand, from the just stated theorem, we know that binomial random variables can be approximated by the normal distribution for n and p that are large enough. In this case, we have a rule of thumb. We consider n and p to be large enough if $n\widehat{p}(1-\widehat{p}) > 9$. Further, the results are improved if we apply a continuouity correction (*Stetigkeitskorrektur*). Putting all these things together, for any $a < b$ we can state the following for a binomial variable X

[7] Another suspect for such a situation would be the hypergeometric distribution. However, we omit discussing it here.

[8] Why DeMoivre? Abraham DeMoivre published on this in 1733 as well.

$$P(a \leq X \leq) \approx \Phi\left(\frac{b + 0.5 - n\widehat{p}}{\sqrt{n\widehat{p}(1-\widehat{p})}}\right) - \Phi\left(\frac{a - 0.5 - n\widehat{p}}{\sqrt{n\widehat{p}(1-\widehat{p})}}\right)$$

In the above, 0.5 was added to b and subtracted from a. This is the continuouity correction.

7 Confidence Intervals

If we have a random sample from a population infront of us, we can use the data from this to calculate estimates for unknown population parameters. This estimators do not give any indication of their precision. We can make statements about the precision of these estimates with **confidence intervals** (*Konfidenzintervalle*). A confidence interval is an interval derived from the data in which the true parameter lies with the probability chosen for the interval. For example, a 95% confidence interval for the mean calculated from the sample contains the true population mean with a probability of 0.95.

7.1 Confidence Intervals for Means

7.1.1 known variance

$$\left[\overline{x} - \mathcal{Q}^{(N)}(1-\tfrac{\alpha}{2})\tfrac{\sigma}{\sqrt{n}}, \overline{x} + \mathcal{Q}^{(N)}(1-\tfrac{\alpha}{2})\tfrac{\sigma}{\sqrt{n}}\right]$$

$$L = 2\left[\mathcal{Q}^{(N)}(1-\tfrac{\alpha}{2})\tfrac{\sigma}{\sqrt{n}}\right]$$

$$n \geq \frac{4[\mathcal{Q}^{(N)}(1-\frac{\alpha}{2})]^2\sigma^2}{L^2}$$

7.1.2 unknown variance

$$\left[\overline{x} - \mathcal{Q}^{(t)}_{n-1}(1-\tfrac{\alpha}{2})\tfrac{s}{\sqrt{n}}, \overline{x} + \mathcal{Q}^{(t)}_{n-1}(1-\tfrac{\alpha}{2})\tfrac{s}{\sqrt{n}}\right]$$

7.2 Confidence Intervals for Proportions

$$\left[\hat{p} - \mathcal{Q}^{(N)}(1-\tfrac{\alpha}{2})\sqrt{\tfrac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + \mathcal{Q}^{(N)}(1-\tfrac{\alpha}{2})\sqrt{\tfrac{\hat{p}(1-\hat{p})}{n}}\right]$$

$$n \geq \left(\frac{\mathcal{Q}^{(N)}(1-\frac{\alpha}{2})}{L}\right)^2$$

7.3 Confidence Intervals for the Difference of two Means

7.3.1 before-after comparisons

Generate sample mean of before and after differences and then apply relevant formulae from section 7.1.

7.3.2 independent samples

3 cases

known variance

$$\left[(\overline{x}_1 - \overline{x}_1) - \mathcal{Q}^{(N)}(1 - \tfrac{\alpha}{2})\sigma_D, (\overline{x}_1 - \overline{x}_1) + \mathcal{Q}^{(N)}(1 - \tfrac{\alpha}{2})\sigma_D\right],$$

where

$$\sigma_D = \sqrt{\frac{\sigma_1^2}{n_1} + d\frac{\sigma_2^2}{n_2}}$$

unknown variance: case 1

Assumption: $\sigma_1^2 = \sigma_2^2$

$$\left[(\overline{x}_1 - \overline{x}_2) - \mathcal{Q}^{(t)}_{n_1+n_2-2}(1 - \tfrac{\alpha}{2})s_P, (\overline{x}_1 - \overline{x}_2) + \mathcal{Q}^{(t)}_{n_1+n_2-2}(1 - \tfrac{\alpha}{2})s_P\right],$$

where

$$s_P = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 +}{n_1 + n_2 - 2}}$$

unknown variance: case 2

Assumption: $\sigma_1^2 \neq \sigma_2^2$

$$\left[(\overline{x}_1 - \overline{x}_2) - \mathcal{Q}^{(t)}_{\nu}(1 - \tfrac{\alpha}{2})\sigma_D, (\overline{x}_1 - \overline{x}_2) + \mathcal{Q}^{(t)}_{\nu}(1 - \tfrac{\alpha}{2})\sigma_D\right],$$

where

$$\nu = \frac{s_D^4}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}$$

and

$$s_D = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

7.4 Confidence Intervals for Proportion Differences

$$\left[(\hat{p}_1 - \hat{p}_2) - \mathcal{Q}^{(N)}(1 - \tfrac{\alpha}{2})\sigma_D, (\hat{p}_1 - \hat{p}_2) + \mathcal{Q}^{(N)}(1 - \tfrac{\alpha}{2})\sigma_D\right],$$

where

$$s_D = \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

8 Hypothesis Testing

In previous chapters we presented some estimators for unknown population means, variances and proportions. In chapter 7, we learnt to construct confidence intervals for estimates in order to be able to say more about their precision. In this chapter another method method for making inferences about population parameters is introduced; namely that of hypothesis testing.

A hypothesis test (*Hypothesentest*) is a statistical rule for deciding between two rival hypotheses. The two hypothesis between which we want to decide are the **null hypothesis** (*Nullhypothese*) and the **alternative hypothesis** (*Alternativhypothese*). The null hypothesis is commonly labeled H_0 and the alternative as H_A[9].

The null hypothesis is "accepted" (so to say, by default) unless there is sufficient evidence in the data to make us decide otherwise.

The procedure that is followed for a hypothesis test can be stipulated as follows:

1. Formulate the null and alternative hypotheses.
2. Decide on a significance level α.
3. Choose an appropriate test.
4. Construct a test statistic based on the data.
5. Based on the results, decide whether to reject or accept the nullhypothesis.

This is all good and well, but to actually apply this information we need to know more about these steps and the concepts contained in them. We discudd them below.

8.1 The Hypotheses

Let us call the true parameter that is unknown to us Θ and Θ_0 the value that this parameter takes on under the null hypothesis. Hypotheses can be constructed for a **one sided** (*einseitig*) or a **two sided** (*zweiseitig*) test. We use the following example to help us understand the differences:

Example (bread). `A factory bakes bread every morning which it then delivers to different customers. This bread factory is confronted with a number of issues that it has to deal with.`

Hypotheses for one sided tests.

[9] Some texts write H_1 for the alternative hypothesis.

The owner of the factory thinks that the machine for a particular loaf of bread makes the loaves too large (that is, too heavy) and this would reduce his profits. This would be a case for hypotheses of the following form (Θ in this case would be the true average weight of the loaves and Θ_0 the weight that the average is supposed to be.)

H_0: $\Theta = \Theta_0$
H_A: $\Theta > \Theta_0$

An important customer comes along and complains that the bread he has been buying weighs less than it should. To test this claim, this time the hypotheses would look like this:

H_0: $\Theta = \Theta_0$
H_A: $\Theta < \Theta_0$

Hypotheses for a two sided test.

Now, the owner of the bakery wants to check if, on average, the baguettes have the weight they are supposed to have. This would mean that on average, their weight should be neither above or below the intended weight. To test this, we need to construct a two sided test. The hypotheses have the following form:

H_0: $\Theta = \Theta_0$
H_A: $\Theta \neq \Theta_0$

8.2 The Significance Level

The significance level $\alpha \in (0, 1)$ is the probability of rejecting H_0 when the null hypothesis is true that we find acceptable (rejecting the null hypothesis when it is true is called the type 1 error and α is the probability of this error). The most common values for α are 0.05 and 0.01. The significance level α should be decided upon before calculating any test statistics with the data.

8.3 Tests, Test Statistics and Decisions

Here we will look at hypothesis test for the mean. Let us assume that we have a random sample of size n from a population for which the true mean or expected value μ is unknown. We use $\overline{x}$ as an estimator for population mean. From chapter 6 we know about the distribution of this random variable. Given that all the required conditions discussed there hold, we can again differentiate between two cases. One, where the population variance is known and two, where it is unknown and estimated from the data.

8.3.1 Test for the mean: population variance known

The **test statistic** (*Teststatistik*), which we will call T is then given by

$$T = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

where μ_0 is the mean under the null hypothesis. From chapter 6 we know that the test statistic follows a standard normal distribution. So, how do we decide once we have calculated the test statistic? That depends on the α we have chosen and the set up of the two rival hypotheses. For a given significance level α we would decide as follows in the three cases below:

- $H_0 : \mu = \mu_0$ vs. $H_A : \mu \neq \mu_0$
  Reject H_0 in favour of H_A when: $|T| > \mathcal{Q}^{(\mathcal{N})}(1 - \frac{\alpha}{2})$
- $H_0 : \mu \leq \mu_0$ vs. $H_A : \mu > \mu_0$
  Reject H_0 in favour of H_A when: $T > \mathcal{Q}^{(\mathcal{N})}(1 - \alpha)$
- $H_0 : \mu \geq \mu_0$ vs. $H_A : \mu < \mu_0$
  Reject H_0 in favour of H_A when: $T < -\mathcal{Q}^{(\mathcal{N})}(1 - \alpha)$

8.3.2 Test for the mean: population variance unknown

The test statistic for this case is given by

$$T = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

where s is the sample standard deviation (recall $s^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})$). Again from chapter 6, we know that in this case we do not use the normal distribution, but instead the student or t-distribution with $n-1$ degrees of freedom (*Freiheitsgrade*). Our decision criteria then look as follows:

- $H_0 : \mu = \mu_0$ vs. $H_A : \mu \neq \mu_0$
  Reject H_0 in favour of H_A when: $|T| > \mathcal{Q}^{(t)}_{n-1}(1 - \frac{\alpha}{2})$
- $H_0 : \mu \leq \mu_0$ vs. $H_A : \mu > \mu_0$
  Reject H_0 in favour of H_A when: $T > \mathcal{Q}^{(t)}_{n-1}(1 - \alpha)$
- $H_0 : \mu \geq \mu_0$ vs. $H_A : \mu < \mu_0$
  Reject H_0 in favour of H_A when: $T < -\mathcal{Q}^{(t)}_{n-1}(1 - \alpha)$