Foundations and Methods 2:

Basics of Probability Theory

Christian Hilbe, christian.hilbe@it-u.at

0 Motivation: Some recap and motivation

Remark 0.1 (**Recap - interactive)**
The last two weeks, Bernd and Mona introduced you to descriptive statistics. In particular you learned:

- There are different types of data variables (e.g., *nominal*, *ordinal*, *metric* scale)
- There are different ways to summarize aspects of this data. For example, if you want to describe the (average) location of your data, there is the *mode*, the *median*, or the *(arithmetic) mean*; if you want to describe the variation in your data, there is the *range*, the *variance*, or the *standard deviation*; and, if you want to look at how two variables co-vary, there are different *correlation coefficients* (Importantly, you also learned that correlations can be spurious, and that they do not imply causation).
- You learned how to take that data and to visualize it appropriately

The general idea was that you took a given data set, and you wanted to summarize some aspects of that data as effectively as possible.

Remark 0.2 (**Motivation for today's class)**
In many applications of statistics (e.g., scientific research, market research, quality control), the data set describes a random sample taken from a larger population. In that case, you often like to draw inferences beyond the specific participants that happened to be in your sample. For example, you might want to know people's voting preferences from asking a sample of $n = 500$ Austrians. Or you may want to know the efficacy of some new medical treatment, after having treated $n = 500$ patients. In each case, you would like to make statements that go beyond those 500 people: statements that apply to the entire Austrian population, or to all future patients that may get this treatment.

In the next two classes, we thus address the following question: *How is it possible that by asking just 500 people, you can draw inferences about 6.3 million people?* (all Austrians eligible to vote). Or more generally, *to which extent can the results of a random sample represent the whole population?* To this end, it is useful to ask the converse question: *How likely is it that your random sample is misleading, by producing very extreme or unrepresentative results?* To get some basic idea of how to even address this question, today we talk about probability theory. In the next class (done by Spiros), you will learn some key concepts and key terminology of inferential statistics in more detail.

Example 0.3 (Inferring people's happiness)
To give you an intuition for the underlying problem, here is a (hypothetical) example. Suppose it is known that 57% of the Germans agree or very much agree to the question 'Are you satisfied with your everyday life.' For brevity, let's refer to this result by saying *'57% of Germans are happy'*. Now you would like to know whether Austrians are more or less happy than Germans. So you ask a random sample of $n = 100$ Austrians, and you learn that 68 of them would describe themselves as happy. This is how the situation represents itself to you:

Austrian population	Your random sample
Proportion p_A of happy people in Austria	68/100
(This is what you would like to know)	(This is what you know based on your data and descriptive statistics)

To interpret this situation, keep in mind there are two possible ways how the data of your sample could have come about:

1. In reality, Austrians are indeed more happy than Germans (i.e., $p_A > 0.57$), and your sample correctly captures this trend.
2. On average, Austrians are just about as happy, or even less happy than Germans (i.e., $p_A \leq 0.57$). It just so happened that when taking our sample, we randomly happened to pick more of the happy ones.

We would like to estimate how likely we might have ended up in the second case, i.e., to have a sample mean that suggests Austrians are more happy, even though that's not the ground truth.

Remark 0.4 (Prerequisites – interactive)
A standard course on probability theory would require quite a bit of math (e.g., calculus, measure theory). *Who in the audience knows math, or even probability theory?*
In the following, I focus on the main ideas, rather than on the technical aspects. Moreover, I focus on those aspects that are most relevant to inferential statistics. If you would like to learn more about probability theory, feel free to ask me, or to read a book. For an informal book that you could read in the bus, start with Haigh (2012); if you are looking for more mathematics, one source is Feller (1968).

1 The vocabulary and basic rules of probability theory

Remark 1.1 (Some basic concepts)
In probability theory you are interested in the outcome of *chance experiments* (for example, the number of happy subjects when you draw a random sample of 100 Austrians). A possible outcome of such an experiment is called an *event.* Such events are often described by a capital letter (e.g., $A = 68/100$). The *probability* $\mathbb{P}(A)$ of an event A can then be interpreted as how frequently you would observe A if you repeated the chance experiment indefinitely ('frequentist interpretation'; there are other interpretations).

For the probability function to make sense, we require $0 \leq \mathbb{P}(A) \leq 1$ for all events A. Moreover, if the event A is impossible to occur, then $\mathbb{P}(A) = 0$; if it occurs with certainty, then $\mathbb{P}(A) = 1$.

Example 1.2 (**Rolling a dice – interactive)**
Just to get some intuition for these concepts, suppose the chance experiment is rolling a dice. Then here are some possible events and their probabilities:

- $A =$ *'Rolling a 3'*. Then $\mathbb{P}(A) = 1/6$.
- $A =$ *'Rolling an even number'*. Then $\mathbb{P}(A) = 1/2$.
- $A =$ *'Rolling some number from 1,...,6'*. Then $\mathbb{P}(A) = 1$.

If an event cannot be further decomposed into smaller constituent events (e.g., *'Rolling a 3'*), we call it an elementary event.
What could be other interesting chance events and there associated probabilities?

Remark 1.3 (**Some simple properties of probabilities – interactive)**
1. For some event A, let $\bar{A}$ denote the event that happens when A does not happen ($\bar{A}$ is the *complement* of A). Then $\mathbb{P}(\bar{A}) = 1 - \mathbb{P}(A)$.

2. We say two events A and B are mutually exclusive if they cannot both happen simultaneously (e.g., the event of rolling a 3 and the event of rolling an even number). Let $A \cup B$ denote the event that one of them occurs. Then for mutually exclusive events, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

3. Two events A, B are called independent if the fact that A happened does not affect the likelihood of B. Let $A \cap B$ denote the probability that both happen. Then for independent events, $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$.

Example 1.4 (**Independent events – interactive)**
1. Consider the two events
   $A \ldots$ There are above-30 degree temperatures in Linz on June 10.
   $B \ldots$ There are above-30 degree temperatures in Linz on June 11.
   These events are not independent.

2. Suppose you are throwing a (perfect) dice twice in a row, and you are interested in the following events,
   $A \ldots$ Getting a '6' in your first throw.
   $B \ldots$ Getting a '6' in your second throw.
   These events are independent, and $\mathbb{P}(A \cap B) = \frac{1}{6}\frac{1}{6} = \frac{1}{36}$.

Remark 1.5 (**Random variables)**
If the elementary events of a random experiment correspond to (real) numbers, one refers to the outcome as a *random variable*. For example, the maximum temperature (in Celsius) in Linz on June 10 is (ex ante) a random variable. The winner of the next football world cup is not a random variable (even though you can think of the next world cup as a chance experiment). It is common to use the capital letter X to denote random variables (as opposed to A, B which denoted events more generally). To describe particular events, we can now use equations or inequalities, e.g., $X = 3$ or $X \leq 3$.

2 Probability distributions

Remark 2.1 (**Probability distribution**)
For a random variable that can only take finitely many values, we can look at a table that assigns to each possible outcome its associated probability. That table is called the *probability distribution* of the chance experiment.

Example 2.2 (**Throwing dice – interactive**)
For the example of throwing a dice, the table looks like this

Number thrown k	1	2	3	4	5	6
$f(k) = \mathbb{P}(X = k)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

We refer to such a probability distribution $f(k)$ where every outcome is equally likely the *uniform distribution*. Sometimes one is also interested in how likely it is that the outcome is at most a given number:

Number thrown k	1	2	3	4	5	6
$F(k) = \mathbb{P}(X \leq k)$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{6}$	$\frac{4}{6}$	$\frac{5}{6}$	1

This function $F(k)$ is called the *cumulative distribution* of $f(k)$. By its definition, $F(k)$ is monotonically increasing, and its final value is 1.

Example 2.3 (**Recruiting a few subjects for a sample – interactive**)
Suppose the true fraction of happy Austrians is $p_A = 70\%$. Now you draw a random sample of three Austrians. Let X be the random variable that refers to the number of happy people in your sample. *What is the probability distribution of X?* By using the rules for computing probabilities of independent events (see **Remark 1.3**), we get:

Possible outcomes k	0	1	2	3
$f(k) = \mathbb{P}(X = k)$	$0.3^3 \approx 0.027$	$3 \cdot 0.3^2 \cdot 0.7 \approx 0.189$	$3 \cdot 0.3 \cdot 0.7^2 \approx 0.441$	$0.7^3 \approx 0.343$

[The factor of 3 in the column for $k = 1$ and $k = 2$ captures the three different ways how one (two) of the subjects could be happy while two (one) of them are not.]

Remark 2.4 (**Binomial distribution – interactive**)
The previous example is an instance of a more general situation: Suppose you repeatedly draw a sample of size n and each outcome independently has a given property with probability p. Let X be the random variable referring to the number k of draws that have the property. Can we generalize the previous calculations and give a formula for the probability $\mathbb{P}(X = k)$?

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

A random variable with this probability distribution is called a *binomially distributed.*

Exercise 2.5 (**Inferring happiness – interactive)**
Initially, we looked at an example where we knew:

1. The true proportion of happy Germans is $p_D = 0.57$, and
2. In our sample of $n = 100$ Austrians, we got 68 happy ones (but we don't know the true proportion p_A of happy Austrians).

We want to know: How likely is it that Austrians are generally more happy (even though we only asked 100 out of 8 millions). One approach is to assume to the contrary that Austrians are just as happy as Germans are, and $p_A = 0.57$. Under that assumption, we can actually compute how likely it is that among 100 randomly drawn Austrians, 68 or even more would be happy. Specifically, let X be the number of happy subjects in the sample. Then use Excel, or Python, or the language of your choice to verify (see also **Figure 1**):

$$\mathbb{P}\big(X \geq 68 \mid p_A = 0.57\big) = \sum_{k=68}^{100} \binom{100}{k} p_A^k (1-p_A)^{100-k} \approx 0.0160.$$

Thus, under the assumption that Austrians are just as happy as Germans, it would be very unlikely to draw a random sample with 68/100 happy Austrians. On these grounds, we might reject the assumption that Austrians are just as happy as Germans. That is, we may conclude that indeed Austrians are generally more happy than Germans. The probability to be wrong is less than 2%.
Bonus question: Suppose instead of a sample of hundred subjects, we only took a sample of $100/4 = 25$ *subjects, and we observed* $68/4 = 17$ *happy subjects. How would this change our previous conclusions?* [For an answer, see **Figure 2**]

Remark 2.6 (**The basic idea of hypothesis testing)**
The previous example illustrates the basic idea of hypothesis testing (testing hypotheses based on a restricted sample). Suppose you want to know whether some hypothesis H_1 is true, given a random sample with n observations (*Are Austrians different from Germans in terms of their happiness?*). Then you may instead consider the opposite hypothesis H_0, the negation of H_1 (*Austrians are just as happy as Germans*). Based on the assumption that H_0 were true, you can often calculate (or simulate) how likely it is to observe an outcome that is at least as extreme as in your sample. If that likelihood is very small, you may take that as evidence in favor of H_1.

Remark 2.7 (**On the effect of sample size)**
In general, larger samples often make it easier to draw general conclusions. That observation is related to the fact that as you draw more subjects, it becomes increasingly unlikely that your sample is very unrepresentative of the whole population. We formalize this idea in the next section (*if there is time left*).

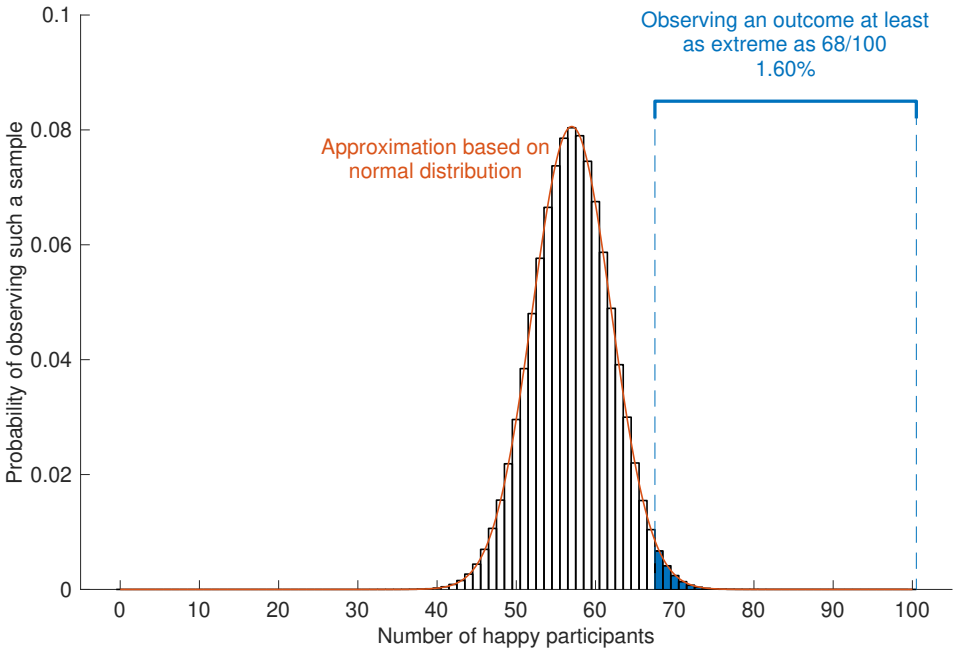

Figure 1: Drawing inferences from a sample. This figure illustrates the basic setup in **Exercise 2.5**. Under the assumption that the proportion of happy Austrians is $p_A = 57\%$, we compute the exact probability distribution when drawing a sample of $n = 100$ subjects. Then we ask how likely it is to observe an outcome that is at least as extreme as observing 68/100 happy subjects in the sample. The respective event is comparably unlikely (note that here we look at one-sided deviations, as opposed to the more common two-sided comparisons). The red curve shows that for the given sample size, the distribution of this chance experiment can be well approximated by an appropriate normal distribution. For that normal distribution, we assume a mean of 57, and a standard deviation of $\sqrt{100 \cdot 0.57 \cdot 0.43}$, as described in more detail later on, see **Remark 3.8**.

3 Bonus: Some key results of probability theory

Remark 3.1 (Expected value and variance of a distribution)
Before we can start, we first need to define two key measures to quantify the location and the spread of probability distributions (analogous to the arithmetic mean and the variance that we already encountered in descriptive statistics). To this end, suppose the random variable X can only take finitely many values within some set I and let $f(k) = \mathbb{P}(X = k)$ be its probability distribution. Then the *expected value* of X is

$$\mathbb{E}(X) = \sum_{k \in I} k f(k). \tag{1}$$

Similarly we define its *variance* as

$$\text{VAR}(X) = \mathbb{E}\big[\big(X - \mathbb{E}(X)\big)^2\big] = \sum_{k \in I} \big(k - \mathbb{E}(X)\big)^2 f(k). \tag{2}$$

So the variance is the mean of the squared deviations from X around its expected value. The square root of $\text{VAR}(X)$ is again referred to as the *standard deviation*. It is quite common to use the greek letter μ for

the expected value of a random variable, and σ for its standard deviation.

Exercise 3.2 (**Expected value and variance of throwing a dice – interactive**)
Based on these definitions, can you compute the expected value and the variance as you throw a dice?

Remark 3.3 (**Law of large numbers**)
Based on this terminology, we can now formulate a first key result of probability theory. Suppose you repeatedly draw one element of an (infinitely) large population. Let X_i denote the random variable that corresponds to the outcome of the i-th draw (e.g., whether or not the i-th person in your sample is happy or not, $X_i = 1$ or $X_i = 0$). Suppose the expected value of all these X_i is μ. Suppose we repeatedly calculate the mean of our sample, denoted by $S_n = \frac{1}{n}\sum_{i=1}^{n} X_n$. This mean is itself a random variable (it is a number that depends on the outcome of a chance experiment). We would like to know: how likely is it that this mean is very different from the true expected value μ? That is, how likely is it in the long run, that we get a value of S_n that is more than ε away from μ, where ε should be thought of as a very small number. Then the (weak) law of large numbers says:

$$\lim_{n\to\infty} \mathbb{P}\big(|S_n - \mu| > \varepsilon\big) = 0 \qquad \text{for all } \varepsilon > 0. \tag{3}$$

Somewhat informally, this means: As your sample becomes sufficiently large, it is increasingly unlikely that your estimate S_n of the population mean differs substantially from the actual population mean μ.

Remark 3.4 (**On the law of large numbers**)
From the viewpoint of inferential statistics, the law of large numbers has perhaps one weakness. It tells you that as your draw larger samples, your sample mean S_n becomes increasingly reliable; but it does not tell you *how* reliable this estimate is (i.e., by how much S_n might vary around μ for a given n). For that we would need to know the distribution of S_n (remember: S_n is itself a random variable!) To quantify that distribution, we'll need another result. However, first we need to introduce some further terminology.

Remark 3.5 (**Continuously distributed random variables**)
Suppose you have a random variable X that can assume arbitrary real values – for example, all values in some interval, or all values in $\mathbb{R}$. Moreover, suppose there is a function $f(x)$ such that the probability that X is in any given interval (a, b) is

$$\mathbb{P}(a < X < b) = \int_a^b f(x)dx. \tag{4}$$

Then $f(x)$ is called the random variable's probability density (it is the analogue of probability distributions in the previous case of finitely many outcomes). Accordingly, we define the expected value of X as

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx \tag{5}$$

and its variance as

$$VAR(X) = \int_{-\infty}^{\infty} \big(x - \mathbb{E}(x)\big)^2 f(x), \tag{6}$$

provided these integrals exist.

Exercise 3.6 (**Uniform distribution – interactive**)
Suppose you have a random variable that is uniformly distributed in the interval [0,1] and that cannot take any values outside of that interval. What's the density function of that random variable? What is its expected value?

Remark 3.7 (**The normal distribution**)
Suppose you have a random variable X and numbers μ and σ such that X has the probability density

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{7}$$

Then we say X follows a normal distribution with expected value μ and standard deviation σ. One reason why normal distributions are important: Many other distributions can be approximated by normal distributions, because of the following result (and more general variants thereof), see also **Figure 1**.

Remark 3.8 (**Central limit theorem**)
Suppose you have a similar setup as in **Remark 3.3**. That is, we repeatedly draw one element of an (infinitely) large population. Let X_i denote the random variable that corresponds to the outcome of the i-th draw. Suppose all X_i follow the same probability distribution (density function), and their means and standard deviations exist and are given by μ and $\sigma > 0$. Suppose we repeatedly calculate the mean of our sample, denoted by $S_n = \frac{1}{n}\sum_{i=1}^{n} X_n$. Then S_n approaches a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$. Or more precisely: If you consider the random variable $Z_n = (S_n - \mu)/(\sigma/\sqrt{n})$, then that random variable converges (in distribution) towards a standard normal distribution (one with mean 0 and standard deviation 1).

Remark 3.9 (**On the importance of the central limit theorem**)

1. The central limit theorem is perhaps the crown jewel of probability theory. It is the reason why normal distributions are important, and why they often arise in nature (e.g., distribution of human height). If you would like to see a more in-depth explanation of this theorem, I recommend the video by *3Blue1Brown*, `https://www.youtube.com/watch?v=zeJD6dqJ5lo`
2. What is remarkable about the theorem: We did not assume that the random variables X_i themselves follow a normal distribution. Rather they could follow *any* distribution, as long as this distribution has a well-defined mean and standard deviation. No matter what that distribution of the X_i is, the sample mean S_n approaches a normal distribution.
3. In the context of inferential statistics, the population's mean value μ and its standard deviation σ are usually unknown. Hence, they first need to be estimated from the sample. To estimate μ one takes the mean value of the sample; however, to compute σ, one needs to divide by $n-1$ instead of n, in order to not introduce certain biases in your estimate.

Remark 3.10 (Summary)

The aim of this lecture was to give a gentle introduction into probability theory. However, the broader aim was two-fold:

1. Give you an idea *why* it is even possible to infer properties of large populations, by just considering a comparably small sample. The idea is: if your sample is sufficiently large, it is unlikely that your sample is very unrepresentative of the underlying population. You can even quantify that likelihood.
2. Give an idea for the intuitive observation that larger samples diminish uncertainty in your estimates.

In the next class, Spiros will tell you more how to do inferential statistics in practice, including the respective terminology (e.g., *p-values*, *one-sided* and *two-sided tests*, *confidence intervals*, etc).

References

Feller, W. (1968). *An Introduction to Probability Theory and its Applications*. Wiley, NY, 3rd edition.

Haigh, J. (2012). *Probability: A very short introduction*. Oxford University Press, Oxford, UK.

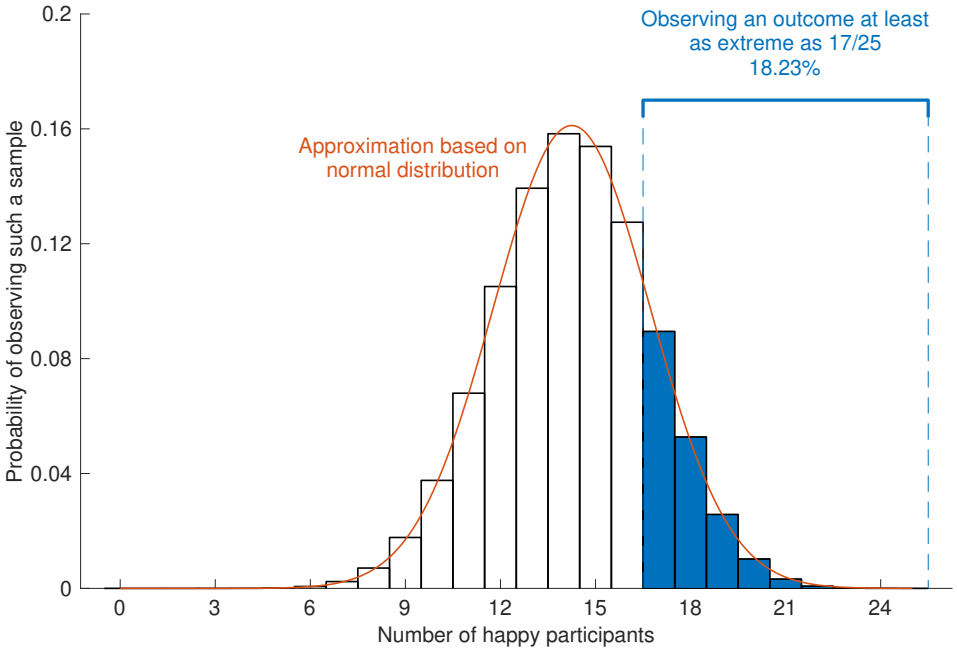

Figure 2: Illustration of inference from a sample. The figure is analogous to **Figure 1**, but this time we assume a sample size of 25, and $68/4 = 17$ positive observations in our sample. Observing such a sample is still somewhat unlikely, but now such a sample might occur in roughly one out of five cases.