Chapter 5

Statistics: what it is and why it works

5.1 Motivation

By now, we have learned a whole lot about probability theory. In the previous chapter we have seen how to compute the distribution of a RV given another RV. Moreover, we know how to factor joint distributions and simplify them by making independence assumptions. In principle, this puts us in a good position to start formulating our own probabilistic models. However, our models will be pretty useless if we do not know their parameters. And as it so happens, we virtually never know them in real life. So what we are going to talk about next is how to **estimate** these parameters from data that we observe. The tools we are going to use for estimation come from statistics.

Statistics is a relatively broad term. There are many ways of doing it and chances are that different people from different fields mean different things when they use the term *statistics*. This is mostly so because the goals that people want to achieve using statistics are different. The underlying mechanics do in fact not differ that much. In this course, we are going to focus on the basics of statistics that you need to know no matter what your goals are. However, allow us to give you a quick birds-eye view on statistics.

There are two main goals you can have using statistics (this is grossly oversimplified, but hey, we said it was the birds-eye view). On the one hand you can do **descriptive statistics** which means that you are gathering information about a phenomenon that you are interested in and report that information. For example, you might be interested in how many faculty members at your university are alcoholics. What you do is you go to each faculty member, check whether they are an alcoholic and report the total number of alcoholics at your institute. Crucially, you are not going to draw any conclusions (such as that people working in the humanities are more

likely to be alcoholics than those working at the science faculty).

Another, somewhat milder example of descriptive statistics are housing advertisements. If you are looking for a flat, you will usually find descriptions of the offered flats in terms of square metres, storey, the presence of a balcony, etc. All of these descriptions can be seen as descriptive statistics. Again, when reviewing these ads, your goal will not be to make a statement like "flats with a balcony are more habitable than flats without one". This may be your own preconception, but it is nothing that you would be trying to get out of your data.

The second big part of statistics is **inferential statistics**. Here you are actually interested in drawing conclusions (inferences). So if you do an alcoholism survey amongst faculty at your university, you would like to find some way of determining the relationship between area of research, say, and the chance that someone is an alcoholic. The rest of the course will mostly be about inferential statistics. In particular, our main question will be the following: given that we observe some data and we know (or assume) that the data is distributed according to some distribution (e.g. a multinomial) what are the parameters of that distribution? Hence, we will try to infer the parameters.

Within inferential statistics, there is a further distinction one can make. In **statistical (data) analysis** people are interested in analysing the properties of a given data set. Say you have obtained questionnaires from 500 faculty members. Then your goal is to make statements about these 500 questionnaires. The scope of your study does not extend beyond those 500 data points and all statements that you make are in principle limited to this data set. Obviously, this is not what people do in research papers. Researchers often try to generalize the results they obtain on their data set to a bigger population, like all university employees or even all of humankind. In the following sections we are going to give some indication for why such generalisations may be justified and at the same time warn you that they are often not.

Finally, there is the field of **prediction**. In prediction you again analyse your data, but what you actually want to do is to predict future data of the same kind. Your current data set is of no actual interest to you except that it allows you to gather information that may turn out to be valuable for making your predictions. Again, if you have compiled 500 questionnaires from faculty, you would like to predict what the rate of alcoholics for the following 100 questionnaires is. After you have extracted the information you need from your original data set, you could even discard it in this setting. In practice, of course, you should NEVER discard your data. Instead, you should make it publicly available, so that other people can reproduce your study.

This latter field of prediction is nowadays most commonly known as **machine learning**. However, statistical analysis and prediction are closely

intertwined and share a lot of their methodology. It is therefore not always easy to make the distinction.

5.2 Statistics and Sample Means

In the previous section we have introduced the word **statistic** and also alluded to the fact that we often assume that our observed data is distributed according to some distribution. The way we usually conceptualize data is that each data point is an instantiation of a random variable. This means that when you are observing 1000 data points, we conceptualize this as observing the outcomes of 1000 random variables. Importantly, each data point could potentially have taken on a different value and it just so happens that in our specific **data sample** it took on the value that it did.

There is one further assumption that we usually make about our data, namely that it is **i.i.d.** (identical and independently distributed). This just means that we assume that all the random variables that generated our data points follow the same distribution and that they are independent of each other. When we say they follow the same distribution, we do not just mean the same class of distributions (e.g. multinomial), but really the same distribution with identical parameters. We often call that distribution the **data-generating distribution**, but you also find the terms *underlying distribution* or *true distribution* in the literature. We have in fact already used the i.i.d. assumption before. When we do repeated Bernoulli trials (as in Section **??**), the total probability of the resulting sequence is computed as a product of independent RVs. We can encode the i.i.d. assumption for n Bernoulli trials as follows:

$$\forall i \text{ s.t. } 1 \leq i \leq n : X_i \sim Bernoulli(\theta)\,. \tag{5.1}$$

All X_i follow the same distribution since the parameter θ does not depend on i but is constant throughout. By the same token, we get independence as the distribution does also not depend on other RVs. Thus, repeated Bernoulli trials, such as repeated coin flips, do actually invoke the i.i.d. assumption. When working with real data this assumption will often be violated but we are going to make it nonetheless for mathematical convenience or if we can motivate it based on our knowledge of the data set.

After we have described our conception of data, let us move on to defining what a statistic is.

Definition 5.1 *A* statistic *is the value of any function of a data sample. If we have sampled n data points that we assume are instantiations of RVs X_1^n, a statistic is the value of a function g on those RVs, i.e. $g(X_1^n)$.*

Arguably the most important statistic in all of statistics is the **sample mean**. The sample mean is just the average of the values of the RVs X_1^n, i.e. of the data points. It is usually denoted by $\overline{\mu}$. The sample mean can be seen as guess of the expectation of X_i. The expectation is sometimes also called mean and $\overline{\mu}$ estimates it from a data sample; hence the name sample mean. Some distributions even have their mean μ as a parameter. To indicate that we are just making a guess at μ we put a horizontal bar on top. This same indicator (or a similar one, like a caret, in which case we would write $\hat{\mu}$) can be used for other quantities, as well.

Definition 5.2 *The sample mean of i.i.d. random variables $X_1, \ldots, X_n$ is defined as*

$$\overline{\mu} := \frac{1}{n}\sum_{i=1}^{n} X_i \ .$$

Notice that since $\overline{\mu}$ is the average of a collection of random variables, it is itself a random variable. Thus, we can compute its expectation.

$$\mathbb{E}[\overline{\mu}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \tag{5.2}$$

$$= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] \tag{5.3}$$

$$= \frac{1}{n} \times n\mathbb{E}[X] \tag{5.4}$$

$$= \mathbb{E}[X] = \mu \tag{5.5}$$

This result is huge! Before we interpret it, let us be clear about how we computed it: lines (5.3) and (5.4) follow from the linearity of expectation and the fact that the RVs are i.i.d.

So why is this result so important? It says that the expectation of the sample mean is equal to the true mean. This means that sampling data gives us a way to estimate the true mean. Since the sample means are random variables and therefore distributed according to some distribution, each sample mean should show up in proportion to its probability. Thus, the mean of sample means will approximate the true mean. Conceptually this may be quite a bit to chew on, but the practical implications are compelling. If you are running one experiment (i.e. if you take one data sample) you basically have no clue how probable that sample mean is according to the distribution of sample means. It could be very improbable and thus not be representative at all of the population you are investigating. So what should you do? The above result tells us that you should just repeat your experiment *enough* times, so that you get *enough* sample means. The mean of those sample means will in turn be pretty close to the true mean of the distribution that

underlies your population of interest. This is the mathematical reason why in science we want our experimental results to be replicable. If I get a result and several other people get the same or reasonably close results, we can be fairly sure that we obtained them from a high-probability region in the distribution of sample means, i.e. the results are indeed representative for the population under scrutiny.

Exercise 5.3 *Under this link you find a file that contains 1000 random samples from a binomial distribution with parameter $n = 100$. The file contains 1000 numbers and the i^{th} number is number of successes in the i^{th} sample. Write a Python script that computes an approximation to the parameter θ of the binomial.*

We have just said that repetition of results is the gold standard in science because if several sample means are close to each other, each one of them (and their average) is likely to be a reasonable approximation to the true mean of the underlying distribution. Unfortunately, more often than not, we only draw one data sample and thus only have one sample mean. The natural question to ask is whether we can somehow ensure that this one sample mean is informative about the true mean. The strict answer is no. We can always be unlucky and obtain a very improbable and thus non-representative sample mean. On the bright side, we can take measures to reduce or chance of being unlucky (and thereby make the one sample mean more trustworthy). What these measures are is explained in Section 5.4. Let us first review the concept of limits in preparation for the proofs that we are going to see in that Section.

5.3 Limits

To help our understanding of the theorem in Section 5.4 we have to recall how mathematical limits are defined. For an infinite sequence of numbers we can ask ourselves whether the sequence will eventually come close to a single point or whether it will just keep moving through the space of real numbers. This question can be formalized with the concept of limits.

Definition 5.4 (Finite Limit of a sequence) *Take any sequence of real numbers (a_n) where $a_n := a(n)$ for some function $a : \mathbb{N} \to \mathbb{R}$. We say that $L \in \mathbb{R}$ is the* limit *of that sequence as n goes to infinity if for any $\varepsilon > 0$ we can find an $n_0 \in \mathbb{N}$ such that for any $n \geq n_0$*

$$|L - a_n| \leq \varepsilon \ .$$

We write $\lim_{n\to\infty} a_n = L$ to express this fact.

Definition 5.5 (Infinite Limit of a sequence) *Take any sequence of real numbers (a_n). We say that the sequence diverges (to $\pm\infty$) if for every $K \in \mathbb{R}$ there is an $n_0 \in \mathbb{N}$ such that for all $n > n_0$ it holds that*

$$|a_n| \geq K \ .$$

We write $\lim_{n\to\infty} a_n = \pm\infty$ to express this fact.

Definition 5.4 tells us that if a sequence converges to a limit L, then for all but finitely many elements (those at the beginning of the sequence) the difference between each element and L will be $\leq \varepsilon$. More informally, we can say that the difference between the elements of the sequence and L can be made arbitrarily small if we are willing to walk far enough down the sequence. Definition 5.5 has been included for completeness' sake but will not be of much relevance in the remainder of the course.

Notice that it is possible that a sequence has no limit at all (neither finite nor infinite). We will not deal with this case here, though.

Example of a limit calculation To give you some more feeling for limits, here is an example. Consider the sequence $a_n = \frac{1}{n}$. What is its limit? Intuitively, a_n becomes smaller as n becomes larger. Moreover, all a_n are non-negative. A good guess for the limit thus seems to be $L = 0$. Let us show that it is indeed the limit of this sequence. Choose any real $\varepsilon > 0$. Then for $n \geq n_0$ we want that

$$|L - a_n| = |a_n| = \frac{1}{n} \leq \varepsilon \tag{5.6}$$

We solve this inequality to get $n \geq 1/\varepsilon$. Thus we set $n_0 = \lceil 1/\varepsilon \rceil$ which is the smallest integer $n_0 \in \mathbb{N}$ such that $1/\varepsilon \leq n_0$. Since ε was chosen arbitrarily we conclude that indeed $\lim_{n\to\infty} \frac{1}{n} = 0$. $\square$

Instead of limits of sequences, we will actually need limits of functions. However, notice that limits of functions are simply the limits of sequences of function outputs.

Definition 5.6 (Limit of a function) *Consider a function f that is defined on the reals. We say that the limit of $f(x)$ as x approaches x_0 is L if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that $0 < |x - x_0| \leq \delta$ implies*

$$|L - f(x)| \leq \varepsilon \ .$$

We write $\lim_{x\to x_0} f(x) = L$ to express this fact.

5.4 The Weak Law of Large Numbers

The weak law of large number states that as we increase our sample size, the probability tends to 0 that our estimated mean $\overline{\mu}$ will be further than a small amount ε away from the true expectation $\mathbb{E}[X]$ of the data-generating distribution P_X. In other words, the more sample points we take, the smaller is the chance that we commit a large error when estimating the mean from our sample. To become clear about what we need to prove, let us first state the weak law of large numbers.

Theorem 5.7 (Weak law of large numbers) *For $n \in \mathbb{N}$, let X_1^n, be i.i.d. distributed random variables with distribution P_X, expectation $\mathbb{E}[X] \in \mathbb{R}$ and variance $var(X) = \sigma^2 \in \mathbb{R}$. Further let X_1^n have sample mean $\overline{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then for any real $\varepsilon > 0$, it holds that*

$$\lim_{n\to\infty} P(|\mathbb{E}[X] - \overline{\mu}| \geq \varepsilon) = 0 \ .$$

At this point it may be good to just pause for a moment, stare at the theorem and try to connect it to the verbal explanation from above. The significance of the theorem derives from the fact that it basically provides us with the theoretical underpinning that allows us to draw inferences from data.

In order to prove the weak law of large numbers, we use two auxiliary lemmas. Once we have proven those, Theorem 5.7 will follow easily.

Lemma 5.8 (Markov's inequality) *For any random variable X and any $a > 0$ it holds that*

$$P(|X| \geq a) \leq \frac{\mathbb{E}[|X|]}{a} \ .$$

Exercise 5.9 *Prove Markov's inequality.*

Besides having established Lemma 5.8, Andrey Markov has made many significant contributions to probability theory. For example, if you go on to study information theory and/or any computational linguistics courses, you are guaranteed to encounter Markov chains. For now, let us move on to our second auxiliary lemma.

Lemma 5.10 (Chebyshev's inequality) *Let X be a RV with expectation $\mathbb{E}[X]$ and variance $var(X) = \sigma^2 \in \mathbb{R}$. Furthermore, let $\varepsilon > 0$.*

Then

$$P(|\mathbb{E}[X] - X| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2} \ .$$

Proof of Lemma 5.10 By Markov's inequality we have that

$$P((\mathbb{E}[X] - X)^2 \geq \varepsilon^2) \leq \frac{\mathbb{E}[(\mathbb{E}[X] - X)^2]}{\varepsilon^2} = \frac{var[X]}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2} \ . \tag{5.7}$$

□

The final step in proving the weak law of large numbers is to apply Chebyshev's Inequality in the case where the random variable of interest is the sample mean.

Proof of Theorem 5.7 We assume i.i.d. RVs X_1^n with sample mean $\overline{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$. This means that $\mathbb{E}[X] - \overline{\mu}$ is a RV depending on X_1^n whose variance is

$$var(\mathbb{E}[X] - \overline{\mu}) = var(\overline{\mu}) \tag{5.8}$$

$$= var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \tag{5.9}$$

$$= \frac{1}{n^2} var\left(\sum_{i=1}^{n} X_i\right) \tag{5.10}$$

$$= \frac{1}{n^2}\ n\ var(X) = \frac{var(X)}{n} \ . \tag{5.11}$$

Chebyshev's Lemma 5.10 then implies that

$$\lim_{n \to 0} P(|\mathbb{E}[X] - \overline{\mu}| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} \ . \tag{5.12}$$

If we fix $\varepsilon > 0$ and increase n, the number of i.i.d. samples, this probability will go to 0. More formally, choose $\delta > 0$. One way to show $P(|\mathbb{E}[X] - \overline{\mu}| \geq \varepsilon)$ to be smaller than δ, is to show $\frac{\sigma^2}{n\varepsilon^2} < \delta$ which happens whenever $\frac{\sigma^2}{\delta\varepsilon^2} < n$. Thus, if we sample more than $\frac{\sigma^2}{\delta\varepsilon^2}$ data points, we can ensure that $P(|\mathbb{E}[X] - \overline{\mu}| \geq \varepsilon) < \delta$. Since δ is arbitrary, this shows that $\lim_{n \to \infty} P(|\mathbb{E}[X] - \overline{\mu}| \geq \varepsilon) = 0$, which is exactly what the weak law of large numbers states. □

The proof of the weak law of large numbers also sheds some new light on the importance of the variance of the underlying distribution. We want that variance to be as small as possible since we then need fewer samples in order to have a close estimate of the true mean of the data-generating distribution.

Thus, it is helpful if the variance of the data-generating distribution is small. Otherwise we need a lot of samples.

Notice also the relationship between the variance σ^2 and the number of samples n in the above proof. For a fixed variance, the probability $P(|\mathbb{E}[X] - \overline{\mu}| \geq \varepsilon)$ reduces by a factor that is proportional to n. Hence, we need to take more and more samples as we want to decrease the probability that the sample mean deviates from the actual mean by more than ε.

A more common and more practical interpretation of the relationship between σ^2 and n is the following: suppose we have collected n data points where n is fixed, meaning we have no quick and cheap way to obtain more data points. Then the lower the variance of the data-generating distribution, the more confident we can be that $\overline{\mu}$ is a reasonably good estimate of the true mean of the data-generating distribution. This insight also establishes a relationship to the expectation of sample means (see Equations (5.2)-(5.5)). If the distribution of sample means has low variance, the interval of sample means that are likely to occur is relatively tight. Thus we will only need few sample means in order to compute a good approximation to the true mean.

To visualise the effect of the sample size for a fixed variance, let us take a look at Figure 5.1. The plot on the top shows the same function applied to different data samples from the same distribution, all of which are of relatively small size. The plots therefore show large spread as witnessed by the width of the curves. The plot at the bottom shows the same function applied to a larger data set. One clearly sees that the spread is much smaller.

5.5 Sufficient Statistics

A popular interpretation of statistics is that they are data summaries. Obviously, some statistics summarize the data better than others. For example, the constant function $\mathbf{6}(\cdot)$ which returns the value 6 on all inputs delivers an extremely poor summary of most data (in fact it is not sensitive to the data at all). The mean is a more useful summary as it captures some overall tendency in the data.

Is there any statistic that captures all the necessary information in the data? This questions is hard to answer in general but when it comes to capturing the information about the parameters of the underlying distribution, the answer is yes[1]. If a statistics conveys all the information about the parameters that the data contain, we call it a **sufficient statistic**. The formal definition is somewhat less obvious.

Definition 5.11 (Sufficient Statistics (discrete)) *Given some discrete RV X over data and a statistical model with parameters θ, a statis-*

[1] At least for the distributions that we are concerned with in this script, which are all in the exponential family.

tic $t(x)$ is sufficient if $P(X = x|t(X) = t(x), \Theta = \theta)$ does not depend on θ, i.e. if

$$P(X = x|t(X) = t(x), \Theta = \theta) = P(X = x|t(X) = t(x)) \ .$$

The above definition captures exactly what it means for a statistic to contain all the information about the parameters. Once we know the sufficient statistic, we can simply ignore the parameters. As an example, consider the Bernoulli distribution. For set of i.i.d Bernoulli trials $x = x_1^n$, the statistic $t(x) = \sum_{i=1}^n x_i$ is sufficient. To see this, consider the distribution $P(X|\sum_{i=1}^n x_i, \Theta = \theta)$. It takes the following form

$$P\left(X = y|\sum_{i=1}^n x_i, \Theta = \theta\right) = \begin{cases} 0 & \text{if } \sum_{i=1}^n y_i \neq \sum_{i=1}^n x_i \\ \frac{1}{\binom{k}{n}} & \text{otherwise, where } k = \sum_{i=1}^n x_i \end{cases}$$

Clearly, neither of the two left-hand-side terms depends on θ. Thus, $t(x) = \sum_{i=1}^n x_i$ is a sufficient statistic according to our definition. Notice that the sum does not capture the order in which the events occurred. This shows that the sufficient statistic is not a perfect summary in general (after all, we might care about the order of events). It only captures all the information about the parameters that the data contain.

A statistic that is non-sufficient for the Bernoulli distribution is the value of the first outcome, i.e. $t(x) = x_1$. There are several binary sequences that have the same starting value but different numbers of ones and zeros. The corresponding conditional distribution is

$$P(X = y|x_1, \Theta = \theta) \begin{cases} = 0 & \text{if } y_1 \neq x_1 \\ \propto \binom{k}{n}\theta^k(1-\theta)^{n-k} & \text{otherwise, where } k = \sum_{i=1}^n x_i \end{cases}$$

where the proportionality follows because we need to renormalise the probability to all those sequences whose starting values is equal to x_1. The crucial point, however, is that the parameter shows up in the right hand side and thus x_1 is not a sufficient statistic for the binomial distribution.

Exercise 5.12 *Assume a RV $X = X_1^n$ whose observations are i.i.d. according to a Poisson distribution with parameter λ. Recall that the p.m.f. of the Poisson distribution for one data point is*

$$P(X = x|\Lambda = \lambda) = e^{-\lambda}\frac{\lambda^x}{x!} \ .$$

Show that $t(x) = \sum_{i=1}^n x_i$ is a sufficient statistic for the Poisson distribution as well.

Definition 5.11 is constrained to discrete distributions. It is also not very useful when one does not yet know the sufficient statistics and has to find them. We will remedy both these problems with the following theorem.

Theorem 5.13 (Factorisation Theorem) *Assume a discrete RV over data X and a statistical model $P(X = x|\Theta = \theta)$ with parameters θ. A statistic $t(X)$ is sufficient if and only if the model can be written as*

$$P(X = x|\Theta = \theta) = g(\theta, t(x)) \cdot h(x, t(x)) \ .$$

Proof Assume $t(x)$ is sufficient. Then

$$\begin{aligned}
P(X = x|\Theta = \theta) &= P(X = x, t(X) = t(x)|\Theta = \theta) && (5.13)\\
&= P(X = x|t(X) = t(x), \Theta = \theta) \cdot P(t(X) = t(x)|\Theta = \theta) && (5.14)\\
&= P(X = x|t(X) = t(x)) \cdot P(t(X) = t(x)|\Theta = \theta) && (5.15)\\
&= h(x, t(x)) \cdot g(\theta, t(x)) \ ,
\end{aligned}$$

where in (5.13), we used the fact that $t(\cdot)$ is a deterministic function, the second equality is the chain rule and in (5.15), we use that $t(x)$ is sufficient. Now assume that $P(X = x|\Theta = \theta) = g(\theta, v(x)) \cdot h(x, v(x))$ for some statistic $v(x)$. We need to show that $v(x)$ is sufficient in the sense of Definition 5.11. Observe that $P(X = x, V(X) = v(x)|\Theta = \theta) = P(X = x|\Theta = \theta)$ since $v(\cdot)$ is deterministic. Thus

$$\begin{aligned}
P(v(X) = v|\Theta = \theta) &= \sum_{x:v(x)=v} P(X = x, v(X) = v|\Theta = \theta) && (5.16)\\
&= \sum_{x:v(x)=v} P(X = x|\Theta = \theta) = \sum_{x:v(x)=v} g(v, \theta)h(x, v)
\end{aligned}$$

where the last equality follows from our assumption that the p.m.f. of X can be rewritten in the desired way. We now use the p.m.f. for $v(x)$ in the next step.

$$\begin{aligned}
P(X = x|v(X) = v, \Theta = \theta) &= \frac{P(X = x, v(X) = v|\Theta = \theta)}{P(v(X) = v|\Theta = \theta)} && (5.17)\\
&= \frac{P(X = x|\Theta = \theta)}{P(v(X) = v|\Theta = \theta)} && (5.18)\\
&= \frac{g(\theta, v)h(x, v)}{\sum_{x:v(x)=v} g(v, \theta)h(x, v)} = \frac{h(x, v)}{\sum_{x:v(x)=v} h(x, v)} && (5.19)
\end{aligned}$$

Clearly, the conditional does not depend on θ and thus $v(X)$ is sufficient. □

Now that we have shown the factorisation theorem for discrete RVs we will use it as a general definition for sufficiency, thereby also including continuous RVs.

Definition 5.14 (Sufficient Statistic) *Given some RV X over data and a statistical model with parameters θ, a statistic $t(x)$ is sufficient if $P(X = x|t(X) = t(x), \Theta = \theta)$ can be written as*

$$P(X = x|\Theta = \theta) = g(\theta, t(x)) \cdot h(x, t(x)) \ .$$

5.6 Parameter Estimation

It is time for us to meet Sir Ronald Fisher, one of the founding fathers of statistics. Many of the methods that Fisher introduced for statistical testing and **parameter estimation** are still in wide-spread use today. One of his biggest achievements was proposing the **Maximum Likelihood Principle**. To understand this principle, we first have to introduce likelihood functions.

Recall that we can informally write Bayes' Rule as

$$\text{posterior} \propto \text{likelihood} \times \text{prior} \ .$$

Recall further that every distribution P_X that we have seen so far depends on a number of parameters. Once these parameters are set, we can compute the probability of any event that is captured by a value of the RV X. But what can we do if the parameters are not known? It turns out that we can estimate them. In order to estimate our parameters, we will make the dependence of P_X on its parameters explicit by letting

$$P(X = x) = P(X = x|\Theta = \theta) \ . \tag{5.20}$$

This means we regard Θ itself as a random variable (over parameters) and use the distribution $P_{X|\Theta=\theta}$ instead of P_X. Notice that for all x we have $P(X = x) = P(X = x|\Theta = \theta)$ as long as the parameters of P_X are set to θ. Again, the purpose of this substitution is to make the dependence of the distribution on its parameters explicit.

Definition 5.15 (Likelihood Function) *For $n \in \mathbb{N}$ and a* fixed *set of n data points or observations $x = x_n^1 = x_1, x_2, \ldots, x_n$, we define the* likelihood function *of a family of distributions $P_{x|\Theta}$ as*

$$L_x(\theta) := P(X = x|\Theta = \theta) \ .$$

There are two crucial things to note about the likelihood function. First, the data set x is assumed to be fixed. Thus, the only random variable that can take on different values is the parameter RV Θ. This convention also tells us that the likelihood function is a function of the parameters *and not of the data*! This is the reason that we index it with the specific data set x and not with a random variable X.

Moreover, the likelihood function is based on conditional probability distributions. If we were to sum over all $x \in \text{supp}(X)$ the result would be one since we would be summing over the support of a distribution with parameter vector θ. Instead, however, the likelihood forces us to leave x fixed and only allows us to sum over all values of Θ. This sum is by no means guaranteed to yield 1 as a result! The important lesson here is that the likelihood function is generally *not a probability distribution*! This is a tough pill to swallow in the beginning and you should maybe take a moment to let this sink in and convince yourself that this is indeed so.

With these (important!) remarks in mind, let us quickly elaborate on notation. Since the data set x is fixed anyway, many authors do not even bother to include it as a subscript of the likelihood function and just write $L(\theta)$. Other authors have adopted the unfortunate convention to write $L(\theta; x)$. While this notation is ok when you know what they are talking about, it may also give you the wrong impression that x is an argument of the likelihood function.

Furthermore, we have made the choice to represent the dependence of the distribution on its parameters as $P(X = x|\Theta = \theta)$. This is the Bayesian way of writing the dependence. A frequentist statistician would rather write $P(X = x; \theta)$ which reads as "the probability of x parametrised by θ". The crucial difference is that the frequentist would feel uncomfortable to regard the parameters θ as a realisation of a random variable because he would claim not to know how to find "the correct distribution" P_Θ for that RV.

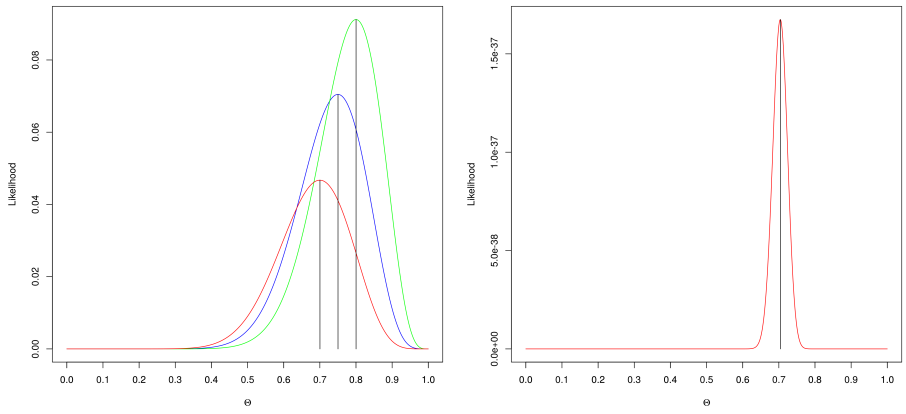
The Bayesian statistician, on the other hand, wants to do exactly that: he wants to impose a distribution P_Θ over the parameters. If we look back at Bayes' rule, we see that this distribution P_Θ would play the part of the prior. If, as in the present case, the prior is a distribution over parameters, we also call it a *parameter prior* or *prior over parameters*. We are siding with the Bayesian view here as it is much easier to interpret and do mathematics with.

After choosing a parameter prior we can compute the posterior distribution over parameters.

$$P(\Theta = \theta|X = x) \propto P(X = x|\Theta = \theta) \times P(\Theta = \theta) \tag{5.21}$$

Notice that we do not compute the distribution $P(\Theta = \theta|X = x)$ itself but rather a quantity that is proportional to it. It turns out that this proportional quantity will be all we need in the remainder of this chapter. Let us emphasize however what makes Bayes' rule so important: it gives us a principled way to compute a distribution over parameters from data!

With the posterior over parameters at hand, we can formulate the parameter inference problem: it is the problem of picking a *good* parameter. Notice that exactly this learning problem is referred to when people talk about machine learning. What the machine is trying learn from data are



(a) (b)

Figure 5.1: Plots of likelihood functions for different data samples. All data samples were randomly drawn from a binomial distribution with parameters $n = 10$ and $\theta = 0.7$. Figure 5.1a depicts the likelihood functions for data samples consisting of 2 draws each. The vertical lines indicate at which value of θ the likelihoods reach their maximums. Two draws contain relatively little information about the underlying distribution and thus the likelihood function is fairly spread, indicating that there are several parameter values that are about equally good. Figure 5.1b shows the likelihood function for a data sample consisting of 50 draws. 50 draws convey much more information about the underlying distribution and therefore the likelihood function is peaked around the MLE (which is also close to the true parameter 0.7. Deviating slightly from the MLE in his scenario leads to a huge drop in likelihood, effectively ruling out a huge part of the parameter space as candidates for the true parameter.

good parameters. Where the statistician would talk about **parameter estimation**, the computer scientist talks about parameter learning. Both expressions refer to the same thing, but machine learning just sounds a lot sexier, doesn't it?

We are left with the question what *good* parameters are. This question does not have a single definitive answer and is actually a constant matter of debate. We are going to present one classic (and still very relevant) answer which is given by the maximum likelihood principle.

5.7 The Maximum Likelihood Principle

The maximum likelihood principle simply states that we should try to maximize the likelihood function of our data, that is, we should pick the parame-

ter value that achieves this maximisation. This parameter value is known as the **maximum likelihood estimate** (or as maximum likelihood estimates if there are several).

Definition 5.16 (Maximum Likelihood Estimate) *A maximum likelihood estimate (MLE) for a parameter set* $\text{supp}(\Theta)$ *on a data sample* x *is a value* θ^* *such that*

$$\theta^* = \arg\max_{\theta} L_x(\theta)$$

In Figure 5.1 we have plotted some likelihood functions. They are based on data samples that were randomly generated from a binomial distribution with parameters $n = 10$ and $\theta = 0.7$. The likelihood plots in Figure 5.1a are based on data sets of only two samples, that is 20 i.i.d. Bernoulli trials. The plot in Figure 5.1b is based on a data set of 50 samples, that is 500 i.i.d. Bernoulli trials. The vertical lines connect the maximum likelihood estimate for each data set with its likelihood value. Notice the different scales of the two plots. The likelihood value depends on the data size. Therefore, it does not make sense to compare likelihood values across data sets. After all, each data set comes with its own likelihood function and the likelihood functions of different data sets are indeed different functions. This difference is also the reason why the maximum likelihood principle tells us to only find the parameter value θ at which the maximum of the likelihood function is achieved. There is no point in even looking at the numerical likelihood value of that maximum.

Notice further that in Figure 5.1, there is only one MLE for each of these functions. We will shortly see why it is desirable for the likelihood function to only have one maximum.

Let us explain how to compute a MLE. By definition, the likelihood function maps the MLE to one of its maximums. From calculus we know that the derivative of any differentiable function is 0 at the function's maximums[2]. Hence, all we need to do in order to find the MLE is to differentiate the likelihood function with respect to θ and check where the derivative is 0. First, however, we need to write down the likelihood function. Looking back at Definition 5.15, we recall that the likelihood is defined only with respect to a probabilistic model. In order to write down a likelihood function, we need to first define such a model. In other words, we have to define concrete likelihood functions on a case-by-case basis. We present one such case as an example below. Before we do so, let us note that for mathematical convenience, one often uses the logarithm of the likelihood function instead of the likelihood function itself. Taking the logarithm has two advantages:

[2]Technically we also require that the function be defined on a closed (as opposed to open) interval. At this stage it is ok to assume that our parameter sets are always closed intervals.

1) logarithms turn products into sums and sums are often easier to handle and 2) when using a computer, very small values may be rounded down to 0. Logarithms mitigate this problem as they turn very small numbers into negative numbers that have rather large absolute values.

Definition 5.17 (Log-Likelihood) *For $n \in \mathbb{N}$ and a* fixed *set of n data points or observations $x = x_n^1 = x_1, x_2, \ldots, x_n$, we define the log-likelihood function over a family of distributions $P_{x|\Theta}$ as*

$$\mathcal{L}_x(\theta) := \log\left(L_x(\theta)\right) \ .$$

Notice that finding the maximum of the logarithm of any function f that only takes on positive values is the same as finding the maximum of the original function f. This is so because the logarithm function is strictly increasing, meaning that for any two possible arguments $x > y > 0$, we have $\log(x) > \log(y)$ and hence, if $f(w) > f(z)$, we have $\log(f(w)) > \log(f(z))$.

In order to make it easier to find MLEs of your own, we give you a procedure that you can apply in most cases and show you an example of how to use it.

1. Define a probabilistic model.
2. Write down the functional form of $L_x(\theta)$ for the parameters of that model.
3. Write down $\mathcal{L}_x(\theta)$.
4. Compute $\frac{d}{d\theta}\mathcal{L}_x(\theta)$. This is also called the score function.
5. Solve $0 = \frac{d}{d\theta}\mathcal{L}_x(\theta)$ for θ.

Example of finding the MLE Assume we have to do market research for a clothing store in Amsterdam. The store wants to expand its size and the manager wonders whether he should use the extra space to exhibit more men's or more women's clothing. To help him with this decision, we are going to count how many male and female customers are coming in on a given day. We treat the gender of each of the n customers as a random variable X_i and interpret $X_i = 0$ as male and $X_i = 1$ as female. We assume that the underlying distribution that determines the gender of customers is the same for all customers. We also assume that a customer's gender is independent of the gender of any other customer. Hence, we stipulate that $X_i \perp X_j$ whenever $i \neq j$.

As an aside, notice that these assumptions are not without problems in real applications. The first assumption, that the gender distribution is the same for all customers, may not always be justifiable. Depending on the time of the day, it is possible that more men or women will come in. The second

assumption, that the gender of one customer does not depend on the gender of other customers, may also not always be true. If couples come to shop at the store, then the gender of one partner will determine the gender of the other partner (in which way the partners in a couple determine each other's genders depends on whether the couple is homo- or heterosexual). For the sake of the example we will nevertheless assume that our assumptions hold true.

Step 1: As the day is over, we have observed k women and $n - k$ men. Above, we postulated a model according to which the occurrences are distributed according a binomial distribution whose parameter n we already know: it is simply the total number of our observations (a.k.a. the total number of customers who entered the shop that day). What we want in order to facilitate the manager's decision is to estimate θ, the probability that a random customer is female.

Step 2: Our likelihood function looks as follows:

$$L_x(\theta) = \binom{n}{k}\theta^k \times (1-\theta)^{n-k} \ . \tag{5.22}$$

As a mnemonic that θ is unknown we can informally write

$$L_x(\theta) = \binom{n}{k}?^k \times (1-?)^{n-k} \ .$$

Step 3: We can now take the logarithm of our likelihood function.

$$\log(L_x(\theta)) = \log\left(\binom{n}{k}\theta^k \times \theta^{n-k}\right) \tag{5.23}$$

$$= \log\left(\binom{n}{k}\right) + k\log(\theta) + (n-k)\log(1-\theta) \tag{5.24}$$

Step 4: To find the MLE, we first differentiate the log-likelihood with respect to θ.

$$\frac{d}{d\theta}\mathcal{L}_x(\theta) = \frac{d}{d\theta}\log\left(\binom{n}{k}\right) + k\frac{d}{d\theta}\log(\theta) + (n-k)\frac{d}{d\theta}\log(1-\theta) \tag{5.25}$$

$$= \frac{k}{\theta} - \frac{n-k}{1-\theta} \tag{5.26}$$

Step 5: Finally, we want to find a point where this derivative vanishes

in order to find the maximum of $\mathcal{L}_x$.

$$0 = \frac{k}{\theta} - \frac{n-k}{1-\theta} \qquad \Leftrightarrow \tag{5.27}$$

$$\frac{n-k}{1-\theta} = \frac{k}{\theta} \qquad \Leftrightarrow \tag{5.28}$$

$$(n-k)\theta = k(1-\theta) \qquad \Leftrightarrow \tag{5.29}$$

$$n\theta - k\theta = k - k\theta \qquad \Leftrightarrow \tag{5.30}$$

$$n\theta = k \qquad \Leftrightarrow \tag{5.31}$$

$$\theta = \frac{k}{n} \tag{5.32}$$

And we are done! You know once and for all that the MLE for the parameter θ of *any* binomial distribution with parameter n (and having observed k occurrences) is $\frac{k}{n}$.

Exercise 5.18 *A coin is getting flipped 1000 times and comes up heads 600 times. According to the MLE, would you say that this coin is fair?*

Those who are already familiar with calculus may have felt a bit uncomfortable, because we simply state that setting the derivative of the log-likelihood to 0 will give us a maximum of that function. In general, this technique will only give us an extremum which might as well be a minimum. How can we be so sure that we really got a maximum? We will just do a proof by picture here and let you do the math.

Take another look at Figure 5.1. Clearly, we only see one maximum per function and that one is unique in all cases. What you cannot see in the plots is that the likelihood is never 0 for any $\theta \in (0, 1)$. It is just really, really small, that is why it looks as if it was 0 in many places in the plot. What actually happens is that the likelihood is constantly decreasing as θ approaches 0 and 1. This constant decrease implies that the derivative is not equal to 0 anywhere other than at the maximum.

Finally, notice that the likelihood is only equal to 0 when $\theta = 0$ or $\theta = 1$. Thus, the likelihood function does indeed have two minima. However, these occur at the boundary points of the interval $[0, 1]$ and since we just stated that $L_x(\theta)$ is constantly decreasing as θ approaches those points, we can safely conclude that the maximum does not lie on the boundary points. Hence, the only point at which the derivative of the likelihood function of the binomial distribution is 0 is at the sole maximum.

We have mentioned above that having only one maximum is a desirable property. If we want to find the MLE, we have but one choice in this case. If there were several maxima, we would have to compute all of them and pick amongst them. If there are two maxima that have equal likelihood values, we have no way of choosing between them and thus no way of determining

the single best parameter estimate. Moreover, we will encounter situations in the next chapter where it is very hard (if not impossible) to find the global MLE (i.e. the MLE at the highest maximum).

You may rightfully wonder what distributions have the desirable property of only having one maximum in their likelihood functions. It turns out that those are exactly the distributions in the exponential family. The link includes a list of those distributions and you will be happy to see that most commonly used distributions are members of the exponential family and thus their likelihoods only have one maximum. Another feature of exponential family distributions is that they all have sufficient statistics. Hence, whenever we are doing maximum likelihood estimation for an exponential family distribution, all we need to know about the data are the sufficient statistics. This point will become important in the following chapter.

Exercise 5.19 *You are given some likelihood function L_x for the binomial distribution along with its MLE θ^*. Show rigorously that θ^* is indeed a maximum. That is, show that $\mathcal{L}''_x(\theta^*) < 0$, where $\mathcal{L}''_x$ is the second derivative of the log-likelihood function.*

5.8 Maximum a Posteriori Estimation

Recall that in Section 5.6 we set ourselves the goal of finding a good parameter estimate from the posterior distribution over parameters. Let us say that the best parameter estimate is the one with the highest posterior probability. Then the question is: do we actually accomplish our goal with the MLE? Does the MLE give us the parameter with the highest posterior probability? Unfortunately, the general answer is no, which becomes evident from Bayes' rule.

$$\arg\max_\theta P(\Theta = \theta | X = x) = \arg\max_\theta P(X = x | \Theta = \theta) \times P(\Theta = \theta) \tag{5.33}$$

The MLE only maximises over the likelihood term but not over the prior! Hence it will in general be different from the **maximum a posteriori** estimate.

Definition 5.20 (Maximum a posteriori estimate) *A maximum a posteriori (MAP) estimate for a parameter set* $\text{supp}(\Theta)$ *on a data sample x is a value θ^* such that*

$$\theta^* = \arg\max_\theta P(\Theta = \theta | X = x) \ .$$

The MAP estimate can be derived with exactly the same steps as the MLE. However, parameters are usually real values which means that distributions over them are continuous distributions. Since we have not dealt with

continuous distributions in this course, we will stop just short of actually imposing (non-uniform) prior distributions over parameters and computing MAP estimates.

Notice that up to now, we justified the maximum likelihood principle merely from intuition by postulating that high likelihood should somehow be an indicator for good parameter values. We will now justify the maximum likelihood principle more formally, by showing that the MLE is just a special kind of MAP estimate. Since this realisation implies that under certain condition the MLE will also have the highest posterior probability, we can safely argue for the MLE based on its posterior probability.

Let us assume that our prior distribution P_Θ over parameters is uniform. Then we can write the MAP estimator[3] as

$$\arg\max_\theta \frac{\partial}{\partial\theta} \log(P(\Theta = \theta|X = x)) \tag{5.34}$$

$$= \arg\max_\theta \frac{\partial}{\partial\theta} \log(P(X = x|\Theta = \theta) \times P(\Theta = \theta)) \tag{5.35}$$

$$= \arg\max_\theta \frac{\partial}{\partial\theta} \log(P(X = x|\Theta = \theta)) + \log(P(\Theta = \theta)) \tag{5.36}$$

$$= \arg\max_\theta \frac{\partial}{\partial\theta} \log(P(X = x|\Theta = \theta))\,, \tag{5.37}$$

where we used the uniformity of Θ to get from (5.36) to (5.37). We can neglect $\log(P(\Theta = \theta)$ when finding the maximum, because if we add a constant to all likelihood values, the MLE will not change.

This derivation tells us that if we choose a uniform parameter prior, the MLE will be the MAP estimate. Hence, what we are actually assuming whenever we choose the MLE as our parameter estimate, is that our prior over parameters is uniform. This is a very strong assumption to make and one may rightfully criticise the maximum likelihood principle because of this assumption (actually one should).

Let us conclude this section by saying that for the rest of this course we will make the assumption that we are using uniform parameter priors with justification. This means that we can safely use the MLE in order to find the parameter value with the highest posterior probability.

Further reading

If your calculus is a bit rusty or you have never taken a calculus class before, you should consult OSU's excellent (and very fun) online course which allows you to learn at your own pace. The course also comes with a lecture script

[3] If you know some continuous probability theory: the MAP estimate is taken on the posterior density function in this case since the distribution over parameters is continuous.

that contains many insightful examples and exercises. Let us emphasise that doing statistics (including machine learning) without a solid understanding of calculus is close to impossible.

If you are looking for a good introduction to parameter estimation that covers everything we have discussed here and also goes a considerable stretch further, we refer you to Gregor Heinrich's widely cited tutorial. Heinrich's examples come from text modelling but the techniques he describes can be applied anywhere. One of the model he discusses, Latent Dirichlet Allocation, originates from text modelling but has wide-spread applications in biology, as well.