Chapter 5

Statistics: what it is and why it works

5.1 Motivation

We have by now learned a whole lot about probability theory. In the last chapter we even learned how to predict the value of a random variable given some observations. 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 component distributions. 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 distributions 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. This means that you are gathering information about a phenomenon that you are interested in and report that information. So for example, you might be interested in how many faculty members at your university are alcoholics. What you do then is you go to each faculty member, check whether they are an alcoholic and report that. Crucially, you are not going to draw any conclusions (such as that people working in the humanities are more likely to be alcoholics than science

faculty).

Another, somewhat milder examples 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, does or does not have a balcony, etc. All of these 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 a preconception that you have, but it is nothing that you would be trying to get out of your data.

This brings us to the second big part of statistics, which 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 this: 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?

Within inferential statistics, there is a further distinction one can make. In statistical analysis people are interested in analysing the properties of a given data set. Say you have obtained surveys from 500 faculty members. Then your goal would be to make statements about these 500 surveys. The scope of your study would not extend beyond those 500 and all statements that you make will in principle be limited to this data set. Obviously, if you read a research paper this is not how people do it. 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 this time around 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 surveys from faculty, you would like to predict what the rate of alcoholics for the next incoming 100 surveys 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 1000 values of 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 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 have in fact already used the i.i.d. assumption before. When we do repeated Bernoulli trials, 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:

$$X_i \sim Bernoulli(\theta); 1 \leq i \leq n \tag{5.1}$$

Now 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 meet 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.

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. $g(X_1^n)$.*

Arguable 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 mean μ is a parameter of some distributions. For others, it is identified with the expectation. In both cases, μ is a theoretical quantity assumed to be known (or at least computable). In statistics, we can never be quite sure which distribution underlies our data and therefore we estimate the mean from a sample; hence

the name sample mean. To indicate that we are just making a guess at μ we put a stroke on top. This same indicator (or a similar one) can obviously 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*

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

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.

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

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

$$= \frac{1}{n} \times n E_{P_X}[X] \tag{5.4}$$

$$= E_{P_X}[X] = \mu \tag{5.5}$$

This result is huge! Before we interpret it, let us be clear about how we computed it. We take the expectation with respect to P_X. We can do this because all RVs are identically distributed, namely according to P_X. 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? Well, it basically says that if you are getting *enough* data samples and take their sample means, then the expectation of the sample means will be the true mean underlying your distribution. Since the sample means are distributed according to some distribution, if you take many data samples, the samples means should show up in proportion to their 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 now 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 will in turn be pretty close to the true mean of the distribution that underlies your population of interest. This is basically 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 5.1 *Under* ***some url from course website*** *you find a file that contains 1000 sequences of 100 numbers each. Those sequences are i.i.d. samples from a Bernoulli distribution. Write a Python program that computes (an approximation to) the parameter of of that Bernoulli.*

The open question at this point is of course how many repetitions are *enough*? We are going to attack this question in the next section.

5.3 Limits

To help our understanding of the theorem in the next section we are first going to remind ourselves what limits are. For a sequence of numbers we can ask ourselves whether the sequence will eventually oscillate a single point or whether it will just keep growing and growing. This question can be formalized with the concept of limits.

Definition 5.3 (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 $\epsilon > 0$ we can find an $n_0 \in \mathbb{N}$ such that for any $n > n_0$*

$$|L - a_n| < \epsilon \ .$$

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

Definition 5.4 (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 $n_0, m \in \mathbb{N}$ such that $|a_{n_0}| > m$ there is some $n_1 > n_0$ such that*

$$|a_{n_1}| > m \ .$$

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

Definition 5.3 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 $\leq \epsilon$. 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 5.4 has been included for completeness' sake but will not be of much relevance in the remainder of the course. Notice, however, that it is possible that a sequence does not have a limit at all. 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 $\epsilon > 0$. Then for $n > n_0$ we want that

$$|L - a_n| = |a_n| = \frac{1}{n} < \epsilon \tag{5.6}$$

We solve this inequality to get $n > 1/\epsilon$. Thus we set $n_0 = 1/\epsilon$. Since ϵ was chosen arbitrarily we conclude that indeed $\lim_{n\to\infty} \frac{1}{n} = 0$. □

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

Definition 5.5 (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 all sequences (a_n) with limit x_0 and all $\epsilon > 0$ there is an $n_0 \in \mathbb{N}$ such that for all $n > n_0$*

$$|x_0 - f(a_n)| < \epsilon \ .$$

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

***Philip:** Actually, this should be something like "A function defined on [a,b] ... all sequences in [a,b]..." but I feel that this would complicate matters unnecessarily.*

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 $E[X]$ of the data-generating distribution P_X. In other words, the more sample points we take, the smaller is the chance that we will 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 5.1 (Weak law of large numbers) *Take i.i.d. distributed random variables $X_1, \ldots, X_n, n \in \mathbb{N}$ with distribution P_X and expectation $\mathbb{E}[X]$. Further let X_1^n have estimated mean $\overline{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then for any real $\epsilon > 0$*

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

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 any inferences from observed data in the first place.

In order to prove the weak law of large numbers, we use to auxiliary lemmas. Once we have proven those, Theorem 5.1 will follow easily.

Lemma 5.1 (Markov's inequality) *For any random variable X and any $a \in \mathbb{R}$ it holds that*

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

Exercise 5.2 *Proof Markov's inequality.*

Besides having established Lemma 5.1, Andrey Markov has made many other significant contributions to probability theory. For example, if you go on to take any computational linguistics courses, you are guaranteed to encounter Markov chains. For now, let us move on to our second auxiliary lemma.

Lemma 5.2 (Chebyshev's inequality) *Let X be a RV with expectation $\mathbb{E}[X]$ and variance $var(X) = \sigma^2$. Furthermore, let $\epsilon > 0$. Then*

$$Pr(|\mathbb{E}[X] - X| > \epsilon) \leq \frac{\sigma}{\epsilon} \ .$$

Proof of Lemma 5.2 First we note that $Z = |\mathbb{E}[X] - X|$ is itself a random variable. Furthermore, we easily see $\sigma = \mathbb{E}[Z]$. So if we plug in $\mathbb{E}[Z]$ for σ, Lemma 5.2 is proven by Markov's inequality. □

Proof of Theorem 5.1 We assume i.i.d. RVs X_1^n with expectation $\mathbb{E}[X]$, and sample mean $\overline{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$, and sample variance $\overline{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (\mathbb{E}[X] - X_i)^2$.

We note that

(5.7)

$$\mathbb{E}\left[|\mathbb{E}[X] - \overline{\mu}|\right] = \mathbb{E}\left[|\mathbb{E}[X] - \frac{1}{n}\sum_{i=1}^{n} X_i|\right] = \frac{1}{n}\mathbb{E}\left[|\sum_{i=1}^{n}(\mathbb{E}[X] - X_i)|\right] = \frac{\overline{\sigma}}{\sqrt{n}} \ .$$

For the probability $P(|\mathbb{E}[X] - \overline{\mu}| < \epsilon)$ this mean by Lemma 5.2 that

$$P(|\mathbb{E}[X] - \overline{\mu}| < \epsilon) < \frac{\overline{\sigma}}{\sqrt{n}\epsilon} \ . \tag{5.8}$$

If we fix $\mathbb{E}[X]$ and $\epsilon > 0$ and increase the number of i.i.d. samples, this probability will go to 0. More formally, choose $\delta > 0$. One way to show $P(|\mathbb{E}[X] - \overline{\mu}| < \epsilon)$ to be smaller than δ, is to show $\frac{\overline{\sigma}}{\sqrt{n}\epsilon} < \delta$. This implies that $\left(\frac{\overline{\sigma}}{\delta\epsilon}\right)^2 < n$. Thus, if we sample more than $\left(\frac{\overline{\sigma}}{\delta\epsilon}\right)^2$ data points, we can ensure that $P(|\mathbb{E}[X] - \overline{\mu}| < \epsilon) < \delta$. Since δ is arbitrary, this shows that $\lim_{n\to\infty} P(|\mathbb{E}[X] - \overline{\mu}| < \epsilon) = 0$, 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 sample variance. We want our sample variance to be as small as possible since we will then need fewer samples in order to estimate a close approximation to the true mean of the data-generating distribution. However, the sample variance is a random variable and hence we can apply the weak law of large numbers to it. This means that the more samples we take, the smaller the probability that the sample variance deviates from the variance of the data-generating distribution by a large margin. Thus, it it of crucial importance that the variance of the data-generating distribution be small. Otherwise we will need a lot of samples.

Notice also the relationship between sample variance $\overline{\sigma}^2$ and the number of samples n in the above proof. For a fixed sample variance, the probability $P(|\mathbb{E}[X] - \overline{\mu}| < \epsilon)$ reduces by a factor that is proportional to the square of n. This is to say, that 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 $\overline{\sigma}^2$ 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 sample variance, the more confident we can be that $\overline{\mu}$ is a reasonably good approximation to the true mean of the data-generating distribution.

5.5 Parameter Estimation

It is now time for us to meet Sir Ronald Fisher, one of the founding fathers of statistics. Many of the methods that Fisher introduced for statistical test-

ing and **parameter estimation** are still in wides-spread use today. One of his biggest achievements was the proposal of 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 that we have seen so P_X 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

$$P(X = x) = P(X = x|\Theta = \theta) \ . \tag{5.9}$$

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 5.6 (Likelihood Function) *For a* fixed *set of data points or observations $x = x_n^1, n \in \mathbb{N}$, 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 also tells us that the likelihood function is a function of the parameters *and not of the data*! Moreover, the likelihood function is based on conditional probability distributions. If we were to sum over all $x \in supp(X)$ the result would be one since we would be summing over the support of one single 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, some authors do not even bother to include it as an argument of the likelihood function and just write $L(\theta)$. Also, 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 this. 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 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.

$$P(\Theta = \theta|X = x) \propto P(X = x|\Theta = \theta) \times P(\Theta = \theta) \tag{5.10}$$

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 will be all we need in in the remainder of this section. Let us emphasize however that this is 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 now formulate the parameter inference problem. The problem of parameter inference is the problem of picking a *good* parameter vector Notice that it is this exact problem that is the referred to as learning problem when people talk about machine learning. What the machine is trying learn from data are *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 now 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.

5.6 The Maximum Likelihood Principle