# Parameter Estimation

Our first example of the use of Probability in a statistical problem addresses the following problem:

> We have observed a sample from a distribution that we have broadly identified, up to one or more parameters. We want to use our sample to narrow down the estimation of this (or these) parameters as much as possible

Recall that a "sample" is, mathematically speaking, a collection of independent identically distributed random observations.

For example, we might have reason to believe that the distribution at work is exponential, and want to determine the value of its parameter (or, equivalently, of its mean). Similarly, if the distribution can be assumed to be normal, we might want to determine its mean, variance or both. We will concentrate on one approach to this problem: *Interval Estimation*. That is, we will look for an upper and lower bound for the value of this parameter. Given the nature of our model, these bounds cannot be 100% certain, except in trivial cases, so we will have to be content for these bounds to have a high degree of likelihood, but no certainty.

As a simple example, suppose we toss a coin a large number of times, in order to determine whether it a fair coin or not. Assume also that the coin *is in fact fair*, even if this fact is unknown to us. The count of the number of heads follows a *Binomial Distribution*, and so, even if the coin is fair, it is not impossible to, say, observe only heads over all tosses. Though, admittedly, this is a very unlikely situation, if it actually happened, it would obviously suggest to us that the coin is badly unfair, and lead us to estimate the probability of heads to be extremely close to 1. This would be wrong, since the coin was fair, and we were simply victims of a case of extreme bad luck, but we would not have any way to know, without further information. Thus, in an experiment like this, there is a slim probability of getting things very wrong, which forces us to admit that any statement we may make can only be regarded, at best, as "most likely correct, but there's always a chance it might be badly off".

In this Module, we will not cover every possible example nor even all the ones you may encounter. We will concentrate on three common cases, trying to focus on the general method, which could then be applied to many other situations

# Chapter 1
# Estimation Model

We assume we have repeated observations of a quantity. As examples, we may consider

- repeated measurements of a physical quantity (e.g., the mass of an object, or the voltage produced by an electrical generator, etc.)
- repeated sampling of a large population (e.g., polling the American public)
- repeated lifetime tests of a product (e.g., repeated observations of the time to failure of machines produced by a given assembly line)

In all these cases we construct a mathematical model of our experiment as follows.

- We posit an abstract sample space $\Omega$
- Each observation we make is considered as the observation of a random variable $X_i$, $i = 1, 2, ..., n$ (where $n$ is the number of observations), defined on this space.
- In general, we are trying to recover the **joint distribution of these random variables**
- As a general problem this is usually too ambitious, hence, we assume that our experiment was set up in a way that allows us to make several drastic simplifications, as follows.

The following additional assumptions are not always appropriate, although they are most common. In particular, some famous failures of statistical science can be ascribed to their arbitrary application to situations that did not warrant it.

- we assume that the random variables $X_i$ are **independent and identically distributed**. Independence is a delicate point, as we can all imagine, but the "identical" in the second requirement should also be carefully considered.
- We can assume that the common distribution of the variables is known, up to the determination of one or more parameters. This can be a pretty hefty assumption, that may be justified by an analysis of the features of your experiment (e.g., it is presumably reasonable to assume that the observation of a physical quantity is normally distributed, as its fluctuations are assumed to be caused by many small and independent error sources, so that the Central Limit Theorem can be safely applied).
- If we cannot be too sure of the underlying distribution, we can still try to estimate some parameter, for example its expected value, because it seems reasonable, for example, given the size of the sample, and, possibly, some qualitative assumptions on the distribution (e.g., symmetry around the mean, existence of the variance) that **it** can be assumed to have an approximate normal distribution.

With these assumptions in hand, we now proceed to construct a function of the observations (technically called a "statistic", with a not too felicitous choice of terminology) that can tell us something about the parameter(s) we are trying to determine. Typically, we will be looking at a combination of the observations that has a known distribution (at least, approximately), and is more or less centered around the

parameter in question.

# Chapter 2
# Estimating the Mean of a Normal Distribution

This is a very common case, and, thanks to the Central Limit Theorem, applies even to situations where the underlying distribution is not really Normal. In fact, if the sample is "large enough", we know from that theorem that the sample average

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

will be approximately distributed as a Normal Random Variable, with mean $\mu = E[X_k]$, and variance $\sigma^2 = \frac{\mathrm{Var}[X_k]}{n}$ (all observation have the same expected value, and the same variance, since they all have the same distribution). As we discussed in our second Probability Chapter, how large is "large enough" depends on the features of the distribution of our Random Variables. Thus, if the distribution was very skewed, the sample would have to be very large indeed for the theorem to apply. On the other hand a fairly symmetric distribution will let the theorem kick in very early.

## 2.1 Estimating the Mean when the Variance is Known

Suppose we know the value of $\mathrm{Var}[X_k]$, and want to estimate $E[X_k]$. In many cases this is an artificial example, since this is a somewhat unlikely situation, but it applies to at least two important cases:

1. Our observations are measurements obtained using an instrument with a known (as determined by the manufacturer) variance in its readings
2. We are observing a sample of Bernoulli Random Variables (and their parameter $p$ is not extremely close to 0 or 1)

In case 1, we may be measuring, for example, the voltage produced by a generator, using a voltmeter whose manufacturer has assured us that its readings are normally distributed around the "true value", with a variance of $v$. This is possible because another common procedure is to estimate the variance of a Normal Random Variable, when the mean is known (as is the case when we measure a standard source of known "true value"), or even when the mean is unknown, as we will quickly sketch at the end of this module. Looking at our sample average, we can now say that it too will be normal, with mean equal to the unknown true value, and variance $\frac{v}{n}$.

In case 2, the sum of our observations is going to have a Binomial Distribution of mean $np$, and variance $np(1-p)$. Of course, since we are looking for $p$, we don't really know the variance. However, we may note that $0 \leqslant p(1-p) \leqslant \frac{1}{4}$ (try studying the graph of the function $x(1-x) = x - x^2$, when $0 \leqslant x \leqslant 1$). Hence, the variance cannot be larger than $\frac{1}{4}$, and if we use this value, we are just making a worst-case estimate which is going to more or less pessimistic, but definitely not wrong on the optimistic side. In this case, we will work under the assumption that our sample mean is (approximately) normal, with unknown mean $p$, and variance $\frac{1}{4\sqrt{n}}$.

**Remark 2.1.** The method of assuming a "worst case scenario" of variance $\frac{1}{4}$ for a binomial distribution that we approximate with the normal is, I believe, the best: the result is "pessimistic" but systematically on the same side. In other words, you know that the confidence level you are stating is *always* a bit too

low. A popular alternative, is to use the sample mean (rebranded as $\hat{p}$, meaning your *estimate* for the true value of $p$) in the formula (that is, use $\hat{p}(1-\hat{p})$ in place of $\sigma$ in the formula below, rather than $\frac{1}{4}$). This will obviously give you a narrower interval, at the price of not knowing whether you are being pessimistic or optimistic in your assessment of the confidence level. You can argue that the whole procedure is approximate anyway (especially when the sample size is small), so a little extra fuzziness does not really change anything.

In either case, and any other case where we may assume a known variance $\sigma^2$, an unknown mean $\mu$, and observed a sample mean $\bar{X}$ of $n$ observations, we will be able to say that $\overline{X}$ has a normal distribution, with mean $\mu$, and variance $\frac{\sigma^2}{n}$, hence standard deviation $\frac{\sigma}{\sqrt{n}}$. This information allows us to assign a probability to any event of the form $\left\{a \leqslant \overline{X} - \mu \leqslant b\right\}$, and thus assign a probability to a statement like "the true value $\mu$ is within this given distance from our observed sample mean, with this degree of probability". The "degree of probability" is ours to choose, and is usually called the "confidence level" of our estimate. The usual procedure is to choose a reasonable confidence level, and adjust $a$ and $b$ consequently. Usually, these are chosen so as to determine a symmetric interval around $\overline{X}$, that is, $b>0, a=-b$, with the given confidence value. Since $Y = \overline{X} - \mu$ is now a normal variable with zero mean and variance $\frac{\sigma^2}{n}$, we can use our software to choose an interval with a desired confidence level.

We can also use a table, rather than software, and obtain the same result "by hand". From what we know about the normal distribution, the Random Variable

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \sqrt{n}\frac{\overline{X} - \mu}{\sigma}$$

has a *Standard* Normal Distribution. The tables found in every probability or statistics book, as well as everywhere on the World Wide Web, can be used to determine a symmetric interval around 0 where $Z$ will fall with an assigned probability. For example, you can see that

$$P[-1.96 < Z < 1.96] \approx 0.95$$

Thus, we may say that with approximately 95% confidence, we may state that

$$-1.96 < \sqrt{n}\frac{\overline{X} - \mu}{\sigma} < 1.96$$
$$-1.96\frac{\sigma}{\sqrt{n}} < \overline{X} - \mu < 1.96\frac{\sigma}{\sqrt{n}}$$

In common usage, confidence levels are often chosen as 0.9, 0.95, or 0.99 (the corresponding approximate numbers we read off the tables are, respectively, 1.65, 1.96, 2.58 ). These choices are due, on the one hand, to everybody's love for round numbers, and, on the other hand, to Fisher's choices, often dictated by very narrow convenience factors: you are free to choose any level. Clearly, as we lower the confidence level, we get narrower estimates (but we have a higher probability of being wrong), and, in reverse, by allowing for estimates that are not as tight, we may gain a higher confidence level.

**Remark 2.2.** When applying this method to a 2-outcome experiment, i.e., using the normal distribution in place of the theoretically correct binomial distribution, you will read about things like "continuity correction", meaning that you are worried by the fact that your random variable should be an integer, but your *approximation* takes any real value, and hence decide to approximate, say, $P[X < k+1]$, for the binomial variable $X$, by $P\left[\tilde{X} < k + \frac{1}{2}\right]$, for your normal approximation $\tilde{X}$ to $X$. See the discussion in the file on "Normal Approximation to the Binomial" in the Online Stat book - you will notice how the examples are for really small values of $n$. While this is certainly acceptable, it is one more example of splitting hairs on a side issue. If this correction makes a real difference, chances are that your approximation is not too good in the first place. If you are in real Central Limit Theorem territory, the correction should be insignificant—if it isn't, there may be much more important discrepancies between your approximation and the "exact" model. Note that, although we won't go there, it is perfectly feasible to do exact interval estimation, without recourse to the CLT, using the binomial distribution. It is more cumbersome, and less automated, and, for these reasons, is almost never done, but if we feel the need to split hairs, we might have to bite the bullet.

Of course, the discussion above applies to any discrete distribution that is being approximated through the CLT.

## 2.2 Estimating The Mean When the Variance is Unknown

This is a more common situation. Unfortunately, the usual tools are limited to the case when the underlying distribution is really normal (as opposed to the previous case, the Central Limit Theorem does not enter the picture as early). Still, the following method is the one people will almost always use, and there has been research proving that the outcome is not that off the wall, even when the underlying distribution is not normal, provided it is symmetric around the mean, and not excessively spread out.

**Remark 2.3.** The underlying fact at work in this context is that, as $n$ increases, the distribution of the *sample mean* approaches the normal distribution, and its value approaches the true expected value faster than the corresponding fact for $s^2$ (which does approach $\sigma^2$, the true variance, but at a slower rate). Hence, in general, expression involving more of the sample than $\bar{X}_n$ will not behave quite like they would if the underlying distribution was really normal.

Still, as we already mentioned, the shape of the underlying distribution makes a big difference in speed. Hence, you will notice that the conditions for reasonable applicability of the Student distribution are precisely the same that ensure that the Central Limit Theorem kick in early.

The method is based on the fact (discussed in the next chapter), that we can use the "sample standard deviation" to get a grasp on the unknown variance of a normal distribution (we discuss the curious factor of $\frac{1}{n-1}$ used in the *sample* variance in the next chapter). It turns out that, for a sample of $n$ independent **normally distributed** random variables,

$$Y_{n-1} = \sum_{k=1}^{n} \frac{\left(X_k - \bar{X}_n\right)^2}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2}$$

has a $\chi^2_{n-1}$ distribution. Now, $Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ is a standard normal variable. Hence, the quotient

$$\frac{Z}{\sqrt{\frac{Y_{n-1}}{n-1}}} = \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \cdot \frac{\sigma}{s} = \sqrt{n}\frac{\bar{X} - \mu}{s}$$

has a $t_{n-1}$ distribution. Note the formal similarity with the quantity used when the variance is known: we exchange the (unknown) variance for the sample variance, and switch to a Student distribution, but the formula is very similar.

Looking up tables for the $t$ distribution with the appropriate number of degrees of freedom, or simply using our spreadsheet, we can then work as in the previous section.

# Chapter 3

# Estimating the Variance of A Normal Variable, and a Bonus Consequence

## 3.1 Estimating the Variance When the Mean Is Known

Again, this is a bit of an artificial situation, but it applies, for example when we are calibrating an instrument by measuring a well known quantity (for example, when testing a length measuring instrument against a Bureau of Standard sanctioned length), in order to evaluate its incertitude.

The relevant observation here would be that the variables

$$Z_k = \frac{X_k - \mu}{\sigma}$$

are standard normal variables, so that $\sum_{k=1}^{n} Z_k^2$ has a $\chi_n^2$ distribution. Hence, using a table, or appropriate software, if we know that the probability of such a variable to lie between two numbers $l_{\alpha/2}$ and $h_{\alpha/2}$ is $\alpha$ (as usual, common usage is to choose $\alpha = .9, .95, .99$), we can say that

$$P\left[ l_{\alpha/2} < \frac{1}{\sigma^2} \sum \left(X_k - \mu\right)^2 < h_{\alpha/2} \right] = \alpha$$

or, defining $S^2 = \frac{1}{n} \sum_{k=1}^{n} \left(X_k - \mu\right)^2$,

$$P\left[ \frac{1}{h_{\alpha/2}} < \frac{\sigma^2}{nS^2} < \frac{1}{l_{\alpha/2}} \right] = \alpha$$

which provides us with a confidence interval for $\sigma^2$, with *confidence level* $\alpha$

## 3.2 Unknown Mean

The more common situation is when we do not know the value of $\mu$. In this case, it is natural to try to mimic the calculation above, using $\bar{X}$, instead of $\mu$. This implies a loss of information, of course, and it also calls into play the following fact (the proof is easy, but we don't really need it):

$$E\left[ \sum_{k=1}^{n} \left(X_k - \bar{X}\right)^2 \right] = (n-1)\sigma^2$$

(note that, instead, $E\left[ \sum_{k=1}^{n} \left(X_k - \mu\right)^2 \right] = n\sigma^2$). The preference of using a substitute (technically, this is called an *estimator*), whose expected value is precisely what we are interested in (such an estimator is called *unbiased*, and it is appreciated that its distribution is "centered", in a sense, around the quantity we are looking for), has led to the use of the "sample variance"

$$s^2 = \frac{1}{n-1} \sum_{k=1}^{n} \left(X_k - \bar{X}\right)^2$$

in this situation, instead of the, perhaps more natural, choice of

$$\bar{s}^{\,2} = \frac{1}{n}\sum_{k=1}^{n}\left(X_k - \bar{X}\right)^2$$

It turns out that $\frac{\sum_{k=1}^{n}\left(X_k-\bar{X}\right)^2}{\sigma^2} = \frac{n\,\bar{s}^2}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2}$ has a $\chi^2_{n-1}$ distribution. This can be used to estimate the variance, and with these results in hand, we can mimic the previous section and observe that we can use the same formula, which, after all, only involves the sum of the squares of the difference between the data points and, respectively, the "true" mean and the sample mean. If you would rather point out the standard deviation, you could simply use $\bar{s}$ in place of $S$, but the traditional usage is to refer to $s^2$ instead of $S^2$, and $n-1$ in place of $n$.

**Remark 3.1.** The choice between $s$ and $\bar{s}$ is only dictated by usage in our context—in most cases, it is only the sum $\sum_{k=1}^{n}\left(X_k - \bar{X}\right)^2$ that really enters in the formula. One can investigate the properties of these two quantities in terms of their effectiveness in providing an estimate for the true standard deviation $\sigma$. This is a more theoretical pursuit, with somewhat limited practical implications. In any case, suffice it to say that each of the two has its own theoretical justification ($s$, as indicated, is *unbiased*, and unbiased estimators are well understood as far as their optimality, while $\bar{s}$ is the *maximum likelihood estimator* for $\sigma$, a feature that carries its own advantages). Incidentally, contrary to what you may read in some textbooks, the fact that $\bar{s}$ is *biased* **does not imply at all that it will always underestimate the true variance**: both $s$ and $\bar{s}$ are random, as they depend on the particular sample you are working with, and they may over- or underestimate $\sigma$, with no possibility of knowing which way they are, since, by definition, we do not know $\sigma$. However, both are *consistent*, meaning that, ideally, if we could increase $n$ without limit, both would approach $\sigma$ better and better, as $n$ keeps growing.

## 3.3 Small Extensions

It is sometimes interesting to be able to estimate the *difference* between the means of two populations. This is an easy application of the methods above, *if the variances, are known or, if unknown, may be assumed to be equal*.

In fact, suppose the variances are known, equal to $\sigma_1^2$, and $\sigma_2^2$ and let the two sample means be $\bar{X}_1$, and $\bar{X}_2$, and the sample size, respectively, $n_1$, and $n_2$. Then we know that $\bar{X}_1 - \bar{X}_2$ is (at least approximately) distributed as a normal variable, with mean the difference of the unknown means (which is what we want to estimate), and variance the sum of the variances: $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$. Consequently,

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

can be assumed to be a standard normal variable, so that we can easily calculate a confidence interval for the difference of the means, $\mu_1 - \mu_2$.

If the variances are unknown (a much more common situation), but can be assumed to be equal (which is a little less common), we can use the same idea used in the one-mean case, since we can use the combined sample variances to construct a chi-square–distributed estimator for the common variance. Let's skip the details (available on request), but the conclusion is that if the two samples consist, respectively, of $n_1$, and $n_2$ observations, the quantity

$$\sqrt{\frac{n_1\,n_2(n_1+n_2-2)}{n_1+n_2}}\,\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{n_1\,s_1^2 + n_2\,s_2^2}}$$

will be distributed according to a $t_{n_1+n_2-2}$ Student distribution.

The most realistic situation, unknown, different, variances, is not as neat. The point is that while the two expressions in this section do have the stated distributions (under the appropriate assumptions, of course), the obvious "recipe", consisting in substituting $s_k^2$ for $\sigma_k^2$ in the expression used when variances are known,

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

does not have a simple standard distribution (its distribution actually depends on the unknown variances). However,very roughly, it happens that pretending that its distribution *was* a Student distribution, with the smallest of $n_1 - 1$ and $n_2 - 1$ as its number of degrees of freedom, does not, usually, lead to outrageous conclusions. Please, be aware of this not-so-white lie when following this common practice. Statistics is always an approximate science, by definition (as mentioned, even flipping 1000 times heads would not prove at a 100% level that the coin is not fair), so these transgressions are not as severe as they look: we don't have a real control on how good or bad our estimate will be, but, then, even in more clean situations, we would most likely still be invoking limit theorems without a clear indication of how good the approximation will be.

## 3.4 A Related Problem: Estimating the Mean of An Exponential Distribution

We mentioned that the distributions $\mathrm{EXP}\left(\frac{1}{2}\right)$, and $\chi_2^2$ are identical, as well as the fact that summing two variable with chi-squared distributions leads to a chi-square distributed variable with a number of degrees of freedom that is the sum of the degrees of the addends. Hence, the sum of $n$ $\mathrm{EXP}\left(\frac{1}{2}\right)$ variables has a $\chi_{2n}^2$ distribution.

Now, suppose we have observed $n$ copies of an exponential random variable of unknown parameter $\lambda$, $X_1$, $X_2$, ..., $X_n$. We consider the modified variables $2\lambda X_k$. From what we saw when discussing the exponential distribution, these are all distributed like $\mathrm{EXP}\left(\frac{1}{2}\right)$, and their sum (which we might write as $2n\lambda\bar{X}$, in order to keep the privileged role of the arithmetic mean) is thus distributed as $\chi_{2n}^2$. Fixing a confidence level $\alpha$, and determining the corresponding bounds for such a variable, say, $l_{\alpha/2}$, and $h_{\alpha/2}$, we will have that

$$P\left[l_{\alpha/2} < 2n\lambda\bar{X} < h_{\alpha/2}\right] = \alpha$$

or

$$P\left[\frac{l_{\alpha/2}}{2n\bar{X}} < \lambda < \frac{h_{\alpha/2}}{2n\bar{X}}\right] = \alpha$$

as a confidence interval

# Chapter 4

# A Really Short Discussion: What Does a Confidence Interval Really Mean?

When looking at confidence intervals, one often uses the following language: "the true mean $\mu$ lies between $a$ and $b$ with probability $\alpha$". For example, suppose we had a sample of 50 observations, $\bar{X} = 2.5$, and we happened to know that $\sigma^2 = 4$. Then, from our discussion,

$$\bar{X} - 1.96\frac{\sigma}{n} = 2.5 - 1.96 \cdot \frac{2}{50} \approx 2.42 < \mu < 2.58 \approx \bar{X} + 1.96\frac{\sigma}{n}$$

with 95% probability.

The statement sounds odd: in theory at least, $\mu$ is a constant, a fixed number that we just happen not to know. It is not random at all, so to say that "it lies between these two numbers with probability $\alpha$ may sound awkward. On close examination, though, what is random is our sample, and hence the value of $\bar{X}$. If we repeat the experiment, under the same conditions, we will get something different from 2.5 (hopefully not too different, but that's not for us to decide). Hence, it is not the item in the middle of the double inequality that is random, but the interval in which we are trying to constrain it.

Also, what does the "probability of 95%" mean exactly, in this context? In principle, according to the classical interpretation (you can go back to the introduction to recall how this is a fairly delicate issue) it means that, if we went and repeated the same experiment a zillion times, 95% of the time we would get that our random intervals would be such that $\mu$ seems to lie between 2.42 and 2.58. This is not a terribly useful interpretation, since we will not repeat the same experiment a zillion times: one time is enough.

A *possible* (no warranty offered) re-interpretation of the meaning of a "confidence level" could be the following:

> If we will continue measuring things with the same procedure, always under the proper conditions, our estimates will turn out to be true about 95% of the time–we have to expect to be wrong about 5% of the time

# Chapter 5
# One Last Cautionary Comment

You will have noticed that all the discussion above refers to *one random variable*: we are observing *one* quantity, and the mathematical machinery used works on probabilities related to this one quantity. In real life, we actually have to deal with several quantities simultaneously. Almost always, these quantities **will not be independent of each other**. A real-life example, from a dissertation discussed many years ago, is of measurements of joint movements of the arm of several subjects (this was part of a study into the design of prosthetic limbs). It goes without saying that the available movements of your elbow are not independent of the movements of your wrist.

The problem here is that, to deal with these observations properly, **one has to deal with all the random variables as a unit**. In other words, if you are observing five quantities, you cannot simply use five disconnected estimates (using the tools we discussed above - in particular, what are called *univariate distributions*): you have to estimate the five quantities as a whole - entering into *multivariate statistics*. Time and content limitations prevent us to get into details here, but that doesn't mean that you should not be aware of this need. For example, the interval estimates produced in that dissertation were essentially meaningless, since the researchers had completely ignored this issue (and also because the sample size was abysmally small).