| | |
|---|---|
| Course: | Mathematical Statistics |
| Term: | Fall 2017 |
| Instructor: | Gordan Žitković |

# Lecture 8
# The statistical setup and basic notions

In these notes we adopt a simple model (the "statistical setup") of how statistics is used, and then build a mathematical theory that explains some of the choices made by statisticians.

## 8.1 Populations and samples.

In our model, statisticians draw conclusions about a population by examining a sample from it. A **population** is sometimes defined as "the pool from which observations are drawn", "the set of all observations that can be made" or even "any complete group with at least one characteristic in common". While somewhat descriptive, these do not exactly *define* a concept, and we will not try to do that here either (with the hope that the examples will tell the rest of the story). A **sample**, on the other hand, is easy to define - it is simply a subset of the population. Not every subset of the population is valuable to a statistician; we are going to be interested in **representative samples**, i.e., those samples that are representative of the entire population (what that means in mathematical terms will be explained below). How to choose such a sample is a difficult question which we happily ignore in these notes by assuming that someone else has already done that for us. In reality, sampling is an extremely important and extremely hard-to-get-right element of statistical practice.

**Example 8.1.1.** A national election is held, and the voters have to decide between two candidates, let us call them $A$ and $B$. Each voter (there are about 240 million of them in the US) has a preferred candidate, and the quantity of interest is the proportion $p$ of the voters who prefer candidate $A$. Indeed, if $p < 0.5$, candidate B wins, and if $p > 0.5$, candidate $A$ wins. We cannot access $p$ directly (it would require no less than an actual election), but we can interview a subset of voters and ask them about their preferences. To make the situation a bit less abstract, suppose that we interviewed 1000 votes, randomly selected from the population, and asked them about the candidate they prefer.

In this example, we can easily identify the population and the sample:

**population** the registered voters in the US (all 240 million of them)

**sample** the 1000 voters we interviewed.

Even though it may appear to be simple, the notion of the population turns out to be quite elusive, as the following example shows:

**Example 8.1.2.** Astronomers are interested in the distance to Proxima Centauri, the second closest star to Earth. Each in its own lab, 10 different teams of scientists measure this distance using essentially the same methodology, but independently of each other. They do not communicate to each other and simply report their final results - 10 different numbers - to a single statistician.

What exactly should be called the *population* is much less clear than in Example 8.1.1 above. Here is one possibility:

**population** all possible measurements of the distance to Proxima Centauri by the methodology used by the labs. In this case, the population is infinite, and very abstract. In fact, the very notion of the population is not very useful beyond its conceptual role.

**sample** the actual measurements performed by the 10 labs.

## 8.2 Random samples and probabilistic modeling

Once we have a sample[1] in our hands, we try to use it to draw conclusions about the population. Here is where the assumption that the sample is *representative* comes in. We interpret that to mean that the sample is random, in that it is chosen randomly from the entire population, with no systematic preference for particular individuals or groups. Mathematically, we may think of the sample as an outcome of repeated independent draws from an unknown distribution which describes that whole population[2]. In the language of probability theory, we have the following definition:

[1] For simplicity, we will not make a distinction between the sample (a subset of the population) and the measurements associated with this sample (candidate preferences, e.g.).

[2] Technically speaking, this corresponds to the notion of a *simple random sample with replacement*; there are many other sampling methods and types of samples. Just to scratch the surface, think of a data set contaminated by two sources of error. One is the simple measurement error whose magnitude is random and independent across readings. The other, however, could be systematic and have something to do with the way these measurements are transmitted to the user. This, latter, error will effect all measurements in the same way and destroy their independence in the process. We do not treat such samples in these notes, even though we will have to revisit some of our assumptions when we start talking about linear models.

**Definition 8.2.1.** A **random sample** of size $n$ from distribution $D$ is a random vector

$$(Y_1, Y_2, \dots, Y_n),$$

such that

1. $Y_1, Y_2, \dots, Y_n$ are independent, and
2. each $Y_i$ has the distribution $D$.

The distribution $D$ in the definition above is typically unknown; the entire goal of statistical analysis is to make educated guesses (inferences) about $D$, based on the observed values of $Y_1, \dots, Y_n$. To make our lives simpler, we subscribe to the **parametric** paradigm in these notes, i.e., we make the assumption that the overall shape of this (unknown) distribution is known, and that the only thing to be determined are the values of a small number of numerical parameters. How to choose the overall shape of $D$ falls under the purview of **probabilistic modeling**. It requires the knowledge of the subject-matter and an experience with similar situations in the past: results of measurements are often normally distributed and light bulb lifetimes are exponential, but the values of parameters $(\mu, \sigma)$ and $\tau$ for the population of interest need to be inferred from the sample.

Let us illustrate these concepts on the two examples presented above:

**Example 8.2.2.** In the setting of the "elections" Example 8.1.1, let us encode the preference of each interviewed voter by 1, if they prefer candidate $A$, and 0 if they prefer candidate $B$. Giving each interviewee a number from 1 to 1000, we can denote their answers with random variables $Y_1, Y_2, \dots, Y_{1000}$. Assuming that the sample is random, $Y_1, \dots, Y_{1000}$ are independent and each has the same distribution $D$. In this example, we do not have much choice when it comes to probabilistic modeling - each $Y_i$ can take only two values, 0 or 1. So we are (logically) forced to use the Bernoulli distribution. The only thing left to figure out is the parameter $p \in (0, 1)$, which corresponds to the (true) proportion of the voters in the entire population who prefer candidate $A$. We usually write this as

$$(Y_1, \dots, Y_{1000}) \text{ is a random sample from } B(p).$$

In other cases modeling plays a bigger role:

**Example 8.2.3.** Continuing Example (8.1.2), we denote the ten measurements of the distance to Proxima Centauri, performed by the ten labs, by $Y_1, \dots, Y_{10}$. Assuming, as above, that the sample is random,

$Y_1, \dots, Y_n$ will be independent random variables, all drawn from the same distribution $D$. Unlike in the previous example, we have lots of choices for the distribution $D$. It could be uniform or exponential or gamma or .... To choose the appropriate one, we need to rely on experience; it tells us that measurement errors are usually normally distributed and centered around the true value. This is an empirical fact, and it is not always applicable, but we assume it does in this particular case. If we were astronomers, we would know better.

Therefore, we assume that $Y_1, Y_2, \dots, Y_{10}$ is a random sample from the normal distribution with the unknown mean $\mu$. What about the other parameter, $\sigma$? As you know, it measures the "spread" of the distribution, and, in the current example, it corresponds to the accuracy of the measurement methodology used. We are assuming that all labs use the same methodology, so we have no reason not to assume that they all have the same $\sigma$ (we would not have a random sample otherwise). We still have a choice to make. Should $\sigma$ be considered "known", just like the fact that the distribution of the error is normal is "known", or should it, too, be estimated from the data. The answer to this will depend on how well we understand the measurement technology used. If it has been around for a long time and we have plenty of data to support a particular value of $\sigma$ (say $\sigma = 0.1$ light years), we will consider $\sigma$ to be known and make it a part of our model. In other cases, where we are quite certain that the error will be normally distributed, but have less information about its magnitude, we make $\sigma$ unknown. In the first case, the unknown parameter is $\mu$, which we usually express as

$$Y_1, \dots, Y_{10} \text{ is a random sample from } N(\mu, 0.1),$$

while in the second case, it is the pair $(\mu, \sigma)$, and we write

$$Y_1, \dots, Y_{10} \text{ is a random sample from } N(\mu, \sigma),$$

## 8.3 Statistical inference

As mentioned above, the goal of statistical analysis is to make an educated guess about the value of the unknown parameter(s), based on the observations from the sample. It is intuitively clear that a guess about the distance to Proxima Centaury based on a single measurement is less accurate than a guess based on 10 (or 100) independent measurements. The second element of statistical inference is, therefore, to quantify the accuracy of our parameter estimates. In words, we need to be able to say just "how educated" our educated guess is. This information can be expressed in various ways, and we only scratch the surface in these notes. The most important concepts are those of (point and interval) estimators:

**Definition 8.3.1.** A **(point) estimator** (or **statistic**) is any function (rule, formula) of the sample $(Y_1, \dots, Y_n)$ or other known constants. It is not allowed to depend on the (unknown) parameters!

An **interval estimator** is a pair of estimators.

An estimator (point or interval) is usually an estimator *of something*. Therefore we usually talk about an estimator for (or of) the unknown parameter $\theta$, and this is usually denoted by adding a "hat" on $\theta$, as in $\hat{\theta}$. The subtle point is that - at least in theory - the same estimator could be used to estimate different parameters. We therefore, do not include the "for $\theta$" part in the definition of an estimator, but use it often.

It is useful to think of an estimator as a piece of computer code, which takes $Y_1, \dots, Y_n$ as inputs and returns a value. It is typically written before the data are gathered, and certainly before the values of the unknown parameters are discovered. Therefore, it needs to be able to run on all conceivable inputs, produced under any conceivable value of the parameters. Instead of a single value, interval estimators return a range of plausible values of the parameter so as to give not only an estimate of the parameter value, but also an idea about its accuracy.

**Example 8.3.2.** Let $(Y_1, \dots, Y_n)$ be a random sample from $N(\mu, 1)$, where $\mu$ is unknown. The following are point estimators

a) $\hat{\mu} = \frac{Y_1 + \dots + Y_n}{n}$ (the **sample mean**, also denoted by $\bar{Y}$)

b) $\hat{\mu} = Y_1$,

c) $\hat{\mu} =$"the sample median"

d) $\hat{\mu} = Y_3^n$,

e) $\hat{\mu} = \cos(Y_2 - Y_n)$

f) $\hat{\mu} = 19$.

the following are interval estimators:

a) $(\hat{\mu}_L, \hat{\mu}_R)$, where $\hat{\mu}_L = \min(Y_1, \dots, Y_n)$ and $\hat{\mu}_R = \max(Y_1, \dots, Y_n)$

b) $(\hat{\mu}_L, \hat{\mu}_R)$, where $\hat{\mu}_L = \bar{Y} - \frac{1}{\sqrt{n}}$ and $\hat{\mu}_R = \bar{Y} + \frac{1}{\sqrt{n}}$.

and the following are not estimators:

a) $\hat{\mu} = \mu$,

b) $\hat{\mu} = \frac{\bar{Y} + \mu}{2}$.

As you can see from the previous example, almost anything (as long as it does not use "illegal" quantities) qualifies for an estimator. Our intuition tells us, however, that most of these are quite useless (declaring that $\mu = 19$, without even looking at the data is clearly not sound statistics). The goal of statistical inference is to construct good estimators (and other related quantities), in the sense that they use as much of the information contained in the sample and convert it into an estimate of the parameter of interest which is “close to the true value of the parameter as often as possible”. What that means, and how to accomplish that is the subject matter of the rest of these notes.