**Inferential Statistics**

Table of Contents

[Terminology 5](#_Toc64381815)

[Population 5](#_Toc64381816)

[Sample 6](#_Toc64381817)

[Mean 7](#_Toc64381818)

[Median 8](#_Toc64381819)

[Mode 8](#_Toc64381820)

[Skewness 9](#_Toc64381821)

[Kurtosis 9](#_Toc64381822)

[Histogram 10](#_Toc64381823)

From this point forward we will be using a different book:

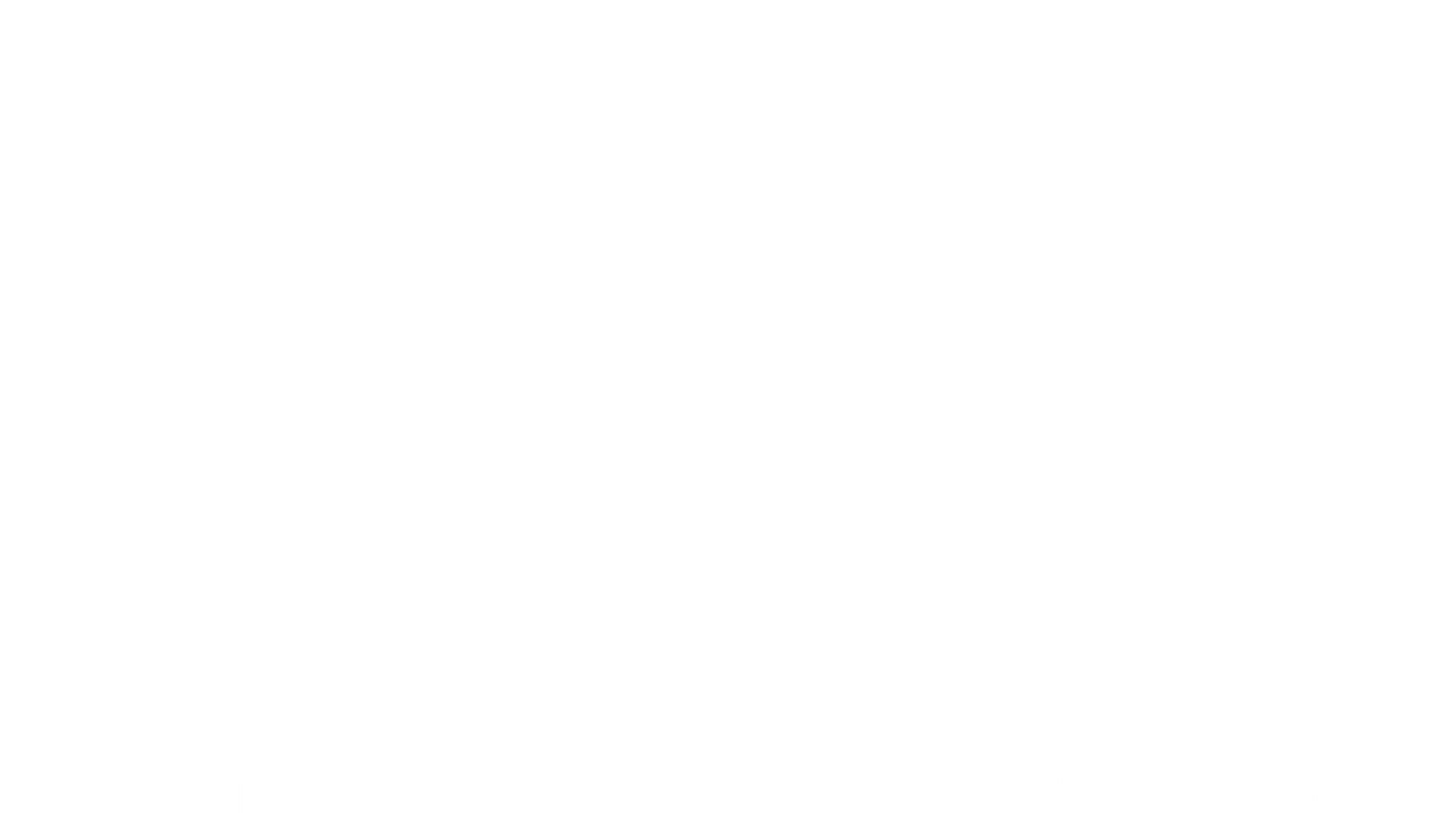
Introduction to Probability and Statistics for Engineers and Scientists - Sheldon M. Ross (3rd Edition)

Statistics is a discipline in which we study data and learn from it. To be able to do this, we need to collect data, collect the description of the data and analyse the data. Statistics is a broad field and we will only be covering a small part of it. Our goal is to use the analysis of the data to draw conclusions. This is a subbranch of statistics called Inferential Statistics. The things we study in inferential statistics or statistical inference overlaps significantly with those that are dealt with in machine learning, to an extent that many consider the two to be the same.

When we mentioned that we need to collect data and the description of the data, what we meant is that we need to find the probability model associated with data. In the first part, we covered probability models for experiments and used those models to find the probabilities of getting certain pieces of data. Now, we will need to figure out what probability model is applicable given some data and find the random experiment.

Say we toss a coin 10 times and get 7 heads. Does this mean that the coin is biased towards heads? Well, a probability does exist that we genuinely did get 7 heads from a fair coin, so we cannot just claim that the coin in biased, but out instinct still tells us it is.

We know how to make a probability model though, so let’s try that. Say is a random variable that represents the number of heads we got. If we create a graph for the PMF of , we will get something like this:



The greatest probability is for getting 5 heads. If we actually perform the calculation, we will find that is only . Even though we assumed the coin was fair, the probability of getting exactly heads is only . As we go further from this number on either direction, the probability decreases. As such, even though it feels like getting heads should make it obvious that the coin is biased, it is not necessarily so, since getting exactly heads, the ‘correct’ number, is very unlikely too.

Now consider the probability of getting somewhere between and heads. We will find that is very high, nearly . Now if someone says they got heads, we can begin to raise eyebrows since that is probabilistically unlikely.

The point we are trying to make here is that to be able to draw any conclusions about the data, we need to know some probabilistic features of the data. It is not good enough to directly make conclusions from the data itself, since the conclusion we made instinctively, that the coin was biased, was not necessarily true.

A large part of our work will be about finding average values and the probability that the average will be within a certain interval. We will be able to make claims like the average delay for some data packets being transmitted is 100ms and 90% of the time, the delay is within 10ms of this value. The more data we have, the more accurately we can make these claims.

Before continuing, we should point out a difference. We have studied probabilities so far. We ran experiments and got data. The purpose of probability was to mathematically describe the likelihood of getting some data. We knew the experiment, and we found the probability model. For example, when we toss a coin, the related probability model is what tells us that the chances of getting a head is .

In statistics, or more specifically, in statistical inference, we have data with us and we make conclusions about what experiment took place and what probability model is associated with the experiment. Again, we are dealing with probability models.

## Terminology

### Population

Whatever data we collect must come from a source. For example, if we want to collect the heights of university students in a country, the source is the students themselves. However, we do not generally collect data from the entire source. It is not possible to actually go and measure the heights of every single university student. Instead, we only measure a subset of the complete source.

The complete source is called the population. We will never encounter a piece of data that does not fall somewhere within the population. We might or might not access the entirety of the data, but we most likely will not. For the example of heights of university students, the population consists of every single university student.

When the field of statistics first came to be, data was generally collected from human beings. This is why the term ‘population’ is used here. Nowadays we can perform statistical analysis on just about anything, so the term can become a little confusing to use.

A population generally has a common family of distributions, meaning we can describe anything from the population using a PMF, CDF or PDF. Another point that crops up here is that the representations for the distributions is different in statistics. The subscripts representing the random variables are missing, so we have , and respectively. This is because multiple random variables become involved very frequently which would be confusing. If things get confusing, we might still use the older form in this course.

Each population is represented by two things, a common family of distributions and the parameters associated with the distribution. A very simple statistical inference could be to use the data available to infer the family of distributions and the values of the parameters.

### Sample

A sample is a subset of the population. We need to be careful that the sample accurately represents the population. For example, when considering the heights of university students, if we only pick students that come from countries where the average heights are greater, our data will make it seem like university students are very tall for some reason.

To ensure the sample accurately represents the population, we prefer to take a random sample. A random sample is a sample consisting of randomly selected elements from the population. A special property of a random sample is that the sample is chosen in such a way so that all possible samples of size have an equal probability of selection. This is a very vague definition, and we will study a much more specific definition later on.

### Mean

For a sample containing elements from to , the sample mean is given by:

We have previously seen a quantity called the expected value. From the name, one would imagine the expected value to be the average. We need to find a connection between the sample mean and the expected value.

Consider the term ‘variance’. We know that the formula for variance is:

The first question that crops up is why did we divide by and not . The reason lies in the difference between the population mean () and the sample mean. If we were able to calculate the mean value for the entire population, we would find that the sample mean we found is a little less than the population mean. To compensate for this, we divide by and not .

Another thing that we will learn in more detail later is that the degrees of freedom for a sample mean is . If we have pieces of data and we are told that , we can only pick pieces of data completely independently. The fifth piece of data will be determined by the 4 pieces already picked and the sample mean value. If instead we were told the population mean, we would be able to pick any data we want.

### Median

The median is the middle element if we sort the data. For an even number of pieces of data, it is the average of the two middle elements.

Consider a uniform distribution. For this distribution, the middle element will be the one which has the value . As such, the inverse of the CDF will give us the median.

This definition comes from the quantile. Here, we would need to find the smallest value of such that . Thus, the quantile is given by . This is the median.

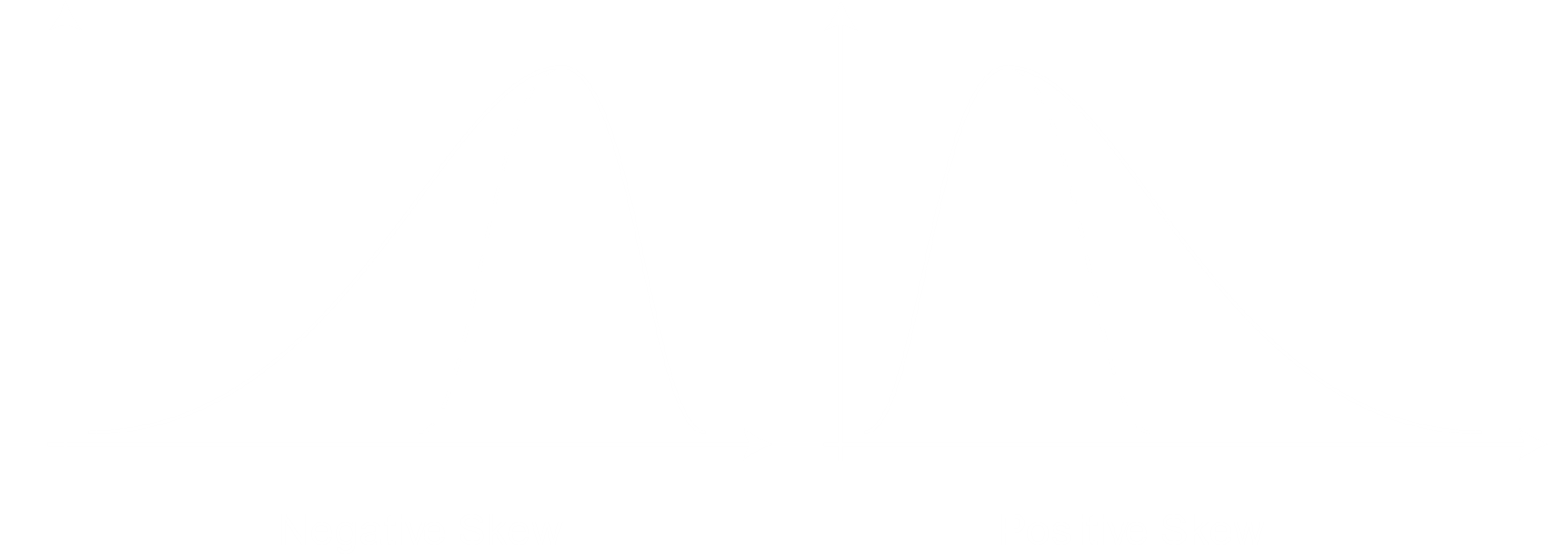
### Mode

If we multiply by when finding the quantile, what we get is the percentile, .

The mode is the maximum number of occurrences of a single piece of data. For the example we saw much earlier with 10 tosses of a coin, 5 was the mode, since it has the highest probability of occurrence.

### Skewness

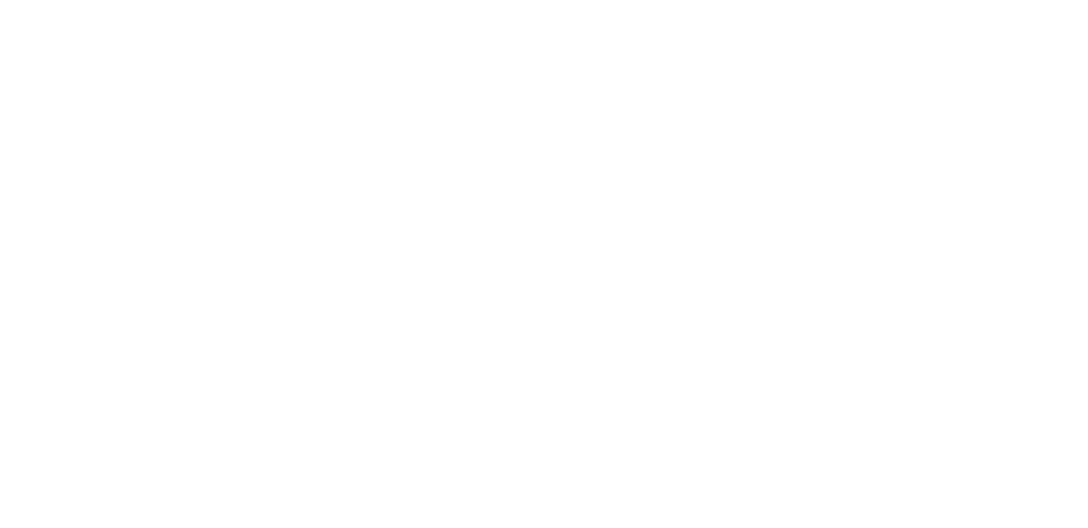
Skewness refers to the distortion or asymmetry in a symmetrical bell curve for a set of data. If the curve is shifted to the left, it is said to be positively skewed. If it is shifted to the right, it is said to be negatively skewed.



L

### Kurtosis

Kurtosis tells us how heavily the tails of a curve differ from the tails of a normal distribution.



### Histogram

A histogram is a graphical display of data using bars of different heights. Each bar groups numbers into ranges. The heights of the bars represent how much data falls into a particular range.

