

Chapter 2

Some Preliminaries

Populations, samples, parameters, and sampling distributions

In science, we are interested in collecting data that will tell us about important features of the world, such as whether two groups of measurements are different or not (e.g. the speed of light measured in one direction vs. another), or whether one variable influences another or not (e.g. the influence of axon diameter on conduction velocity). In order to truly know the exact state of some aspect of the world, we would have to make every possible measurement, and this collection of measurements is known as a *population*. Rarely, it is possible to measure a population, such as “the average height of seniors at Austin High School.” Much more often, however, is either theoretically possible but impractical (e.g. the average height of adults alive today living in Austin, Texas) or impossible (e.g. the average height of adult humans). Thus, in most experiments, we collect a *sample* of data and, if the experiment is done well, the properties of the sample will embody all of the important properties of the population that we wish to examine.

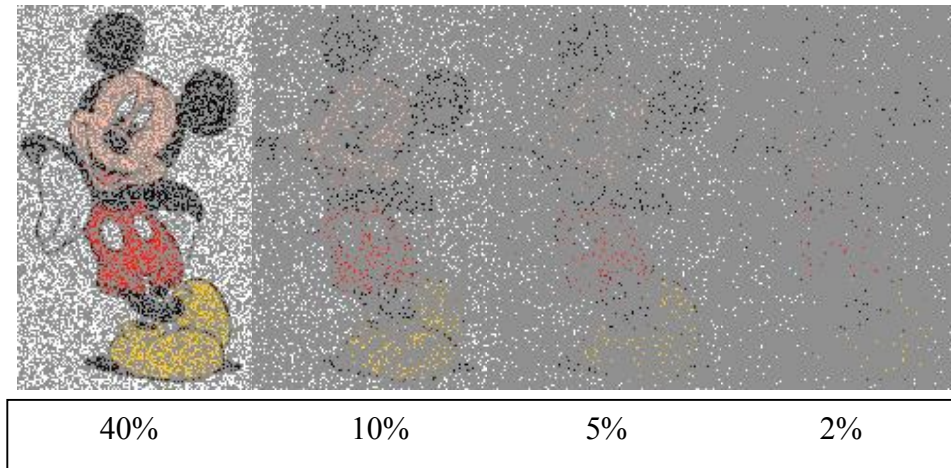
The relationship between a population and random samples is fairly easy to appreciate, and it depends most critically on the sample size. Consider the following digital image.



This full resolution image can be thought of as a “population” of pixels. By looking at the entire image – the whole population of pixels – it is easy to determine things about the image that are over and above the pixel values per se. For example, it is easy to see that Mickey is smiling or that there are two buttons on the front of his overalls.

Now consider the following four images, which consist of random samples of 40%, 10%, 5%, and 2% of the pixels (the remaining unsampled pixels have been set to gray). Notice that, as the sample size decreases, the picture becomes increasingly ragged and, importantly, it becomes progressively more difficult to make judgments about the picture. For example, our judgment about the number of buttons would probably change to “one” and “zero” for the 5% and 2% cases, respectively. Obviously, these answers would be

wrong, and they would be wrong because sample sizes were too small to allow us to correctly make the relevant decision.



The relationship between samples of experimental data and populations of interest is directly analogous to the relationship between the original Mickey image and the sampled versions. Notice that there is information even in the rightmost sampled image; the pixels are black where the ears are, yellow where the shoes are, etc. In fact, if some people might be able to guess that this was a picture of Mickey Mouse if they were told beforehand that it was a famous cartoon character. In other words, the rightmost sampled image does represent the original image to some extent; it just does not allow us to make decisions about the original image with the same accuracy as does an image containing a greater number of samples (such as the leftmost image). In exactly the same way, small samples of experimental data do not allow us to make judgments about aspects of the population with the same accuracy as do larger samples.

In science, we are generally interested in aspects of the population distribution such as the mean, the variance, the median, etc. These are called *parameters* of the distribution because, like the number of buttons Mickey has, they are not directly obtainable from the actual measurements, they must be computed or inferred from these measurements. The word parameter, in fact, means beyond (para) measurement (meter). In a scientific experiment, we collect a sample of data from a population, and then estimate some parameter of that population (the mean, for example) by computing the value of that parameter from our sample.

There is an additional step, however. Once we have estimated some parameter from our data, it would be extremely informative to know how confident we are in that estimate. Consider Mickey again. If you determine the number of buttons by looking at the 40% image, your answer will be “two.” If I do the same on the 5% image, my answer will be “one.” Clearly, your answer is better than mine is some respect, and it would be extremely valuable to quantify this in some way, that is, to compute not only the value of our parameter (the number of buttons), but also to compute how confident we are in our estimate given our sample size and other factors. This computation – the computation of how certain we are about our parameter estimates – is the key benefit that a statistical analysis yields. Simply put, we wish not only to compute the estimate of our parameter of interest, we also wish to compute the distribution of that parameter if we were to

measure it over and over again, because this distribution is what tells us how confident we can be in original estimate.

To gain an intuition about how this computation might work, consider Mickey one final time. If you were to repeatedly sample 40% of the pixels and judge the number of buttons, your answer would almost always be “two.” In other words, the distribution of your estimated parameter (the number of buttons) would be extremely narrow. If I were to repeatedly sample only 5% of the pixels, however, I would likely produce a range of answers. Sometimes, by chance, I would collect a lot of samples from the left button (as in the above image) and answer “one.” Sometimes I might get a lot of samples from the right button and give the same answer. Other times, however, I might get a fair number of samples from both buttons and answer “two,” and sometimes I might get very few samples from either button and answer “zero.” Thus, the distribution of my answers would be quite a bit wider than yours, and from that we could conclude that your original answer was better than mine in the sense that it was more stable of repeated experimentation; we could (and should) assign a much higher confidence to your answer than mine.

The distribution of a parameter estimate over repeated experimentation is called a *sampling distribution*, and it is what allows us to compute a quantitative estimate of the confidence associated with our judgments about experimental data. From the above example, it is obvious that this confidence depends critically on sample size. The core goal of statistical analysis is to figure out what the sampling distribution looks like after *having done only one experiment*. After all, we do not want to spend our time repeating an experiment 30, 50, or 100 times in order to be able make reasonable statements about, say, the mean of our data. The concept of a sampling distribution will be discussed more fully in the next chapter, and the ways in which we compute sampling distributions after having done only one experiment is the overarching topic of this book.

A little bit about data analysis and statistics

Data analysis is the process by which we determine if there are important patterns in experimental measurements.

That sentence is a little denser than you might expect, so let’s unpack it. First, consider “patterns.” We are very good at detecting patterns with our visual systems and, conveniently, it is very easy to transform patterns in numbers into patterns we can see. This is what we are doing when we graph data – we are taking information in a somewhat abstract form, and turning it into a form to which we are exquisitely sensitive. This is why the graphical analysis is – or should be – the cornerstone of data analysis.¹

Next, consider “important.” I used this word intentionally instead of “significant.” What we should always want from data is an important effect, either in a scientific or applied context. This not the same as a (statistically) significant effect, although the fact

¹ A pet peeve of mine is when a student shows me a page full of numbers – an ANOVA output, say – and says “I think there’s a problem, can you find it?” The problem is almost always obvious when the student graphs the data in a sensible way and thinks about them. More frustrating, however, is when the student says “But my advisor doesn’t like it when I bring her a bunch of graphs, she (or he) just wants the answer.”

that the two words are synonyms in the vernacular has had an insidious effect on data analysis in the behavioral sciences. This confusion has also resulted in a systemic bias such that it is primarily “significant” findings that get published and students who get “significant” results are praised, often without any regard to how important a finding is. We will have more to say on this latter but, for now, let’s define “important” and “significant.”

Consider the following somewhat-whimsical but illustrative example. We have good reason to believe that reaction time (RT) gets slower as we age, so we wished to quantify the difference between 20 year old drivers and 40 year old drivers. In a Herculean effort, we do a simple reaction time test in 20,000 people, half in each age group. The distributions of these reaction times are plotted in Figure 2.1. This figure shows that, for both groups, the vast majority of people have a reaction time right around 200 msec, and some people in both groups are either unusually fast or unusually slow. Both curves are bell-shaped (and are actually ‘normal’ or ‘Gaussian’ distributions in this artificial data set). The main point, however, is that they are virtually identical. Based on these data, I would neither want to construct a theory about reaction time and age, nor would I want to lobby for legislation to strip people over 40 of their right to drive. Thus, in this example, I would probably not consider the difference between the two age groups to have either scientific or practical importance. Put another way, if you to randomly select one person from each group and then bet on who was faster, you would basically be betting on a coin flip. In fact, I would probably go so far as to conclude that, *for all intents and purposes*, there is no difference in RT between these two age groups, and I’m sure most people would agree.

In terms of statistical significance, however, we would come to just the opposite conclusion. In fact, to many traditional statisticians, the last sentence of the previous paragraph would be considered blasphemous (and we will visit that issue later). In terms of *statistical significance*, not only is there a difference between these two groups, but it is huge. The standard threshold for statistical significance is that there is less than a 5% chance for something to have occurred by random fluctuation alone, that is, if no difference really exists. For these data, there is only a 0.0000000003 % chance of that, so the standard criterion has been exceeded by over 10 orders of magnitude! In this example, there is a clearly a big difference between what our common sense tells us and what a statistical analysis tells us, and it is not because either one is wrong.

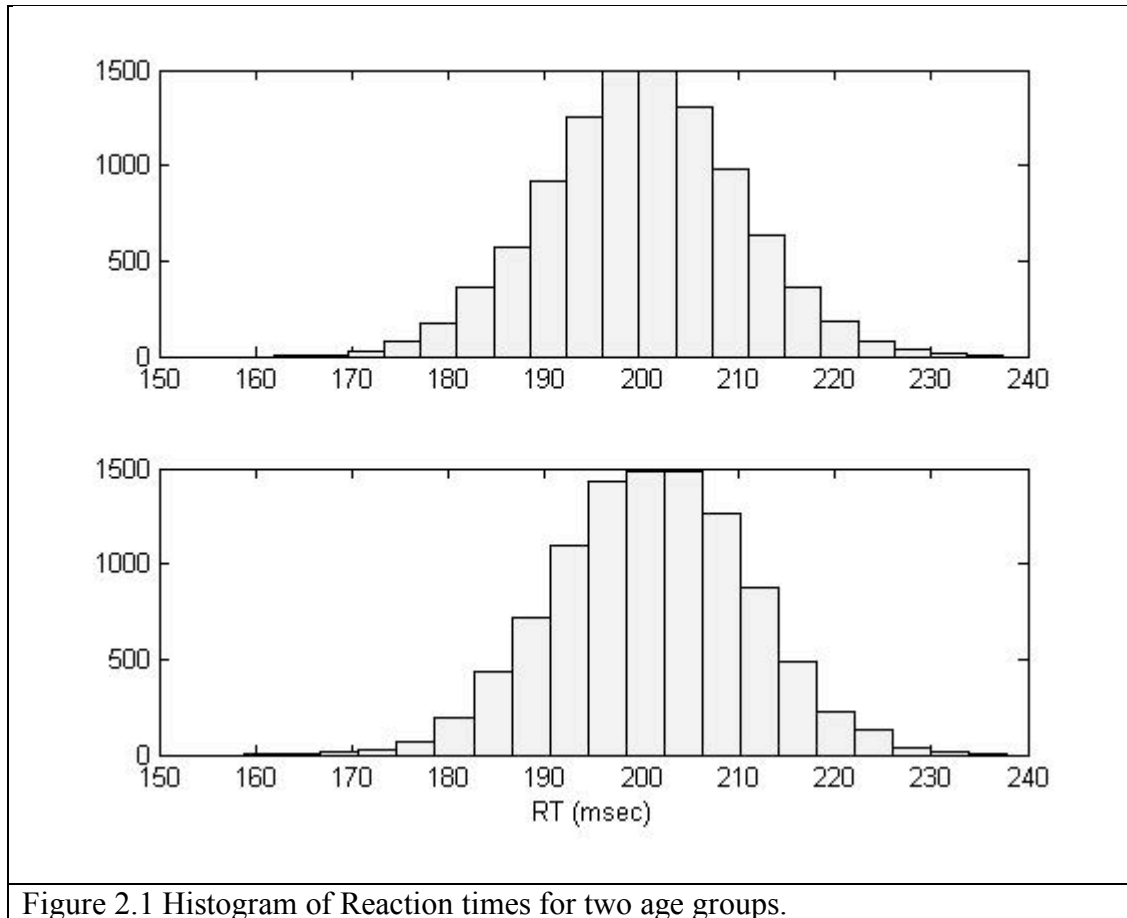


Figure 2.1 Histogram of Reaction times for two age groups.

So why are these two groups actually different in terms of statistical significance when they are obviously so similar? The answer is that tests of significance are usually concerned with quantities (parameters) like the mean and, since the sample size was so high, 10,000 measurements per group, we know the mean with a very, very high confidence or, equivalently, we have a very small margin of error.

Which brings us to what “statistical significance” and the associated probability actually mean. When you read that “... and this is significant at the 0.05 level” what the author is actually saying is “If there were actually no difference between the means of the two populations, then there is less than a 5% chance that I would observe a difference between the means of the two groups at least as large as the one I obtained.” The first part of that sentence is a statement of the null hypothesis and is, unfortunately, often substituted for a prediction-generating theory. In terms of the calculations, it is taken literally; under the null hypothesis, the two group means are identical to any decimal place. A moment’s thought should convince you that the probability of this actually being true in the real world is (usually) vanishingly small, and it is thus no surprise that, if you collect enough data, you will eventually reach statistical significance – there is almost always some difference between two groups out in the real world. The real question is whether the difference is important. So why do statistics at all? Briefly, the reason is that if we do see a pattern of data that we consider to be important, statistics can tell us how likely it is that the pattern actually arose from the vagaries of chance, or from one set of assumptions about the world vs. another. I mentioned above that we are very

good at detecting patterns. Sometimes we are too good, and statistics can help us avoid being fooled by random fluctuations.

Consider a simple situation in which we are looking for a difference between two means. Once we do a statistical test (a “null hypothesis significant test” abbreviated NHST), we will find ourselves in one four possible situations, illustrated in the following table.

Possible States of the World and an Experimental Outcome.		Is there an important effect actually present?	
		Yes	No
Was a significant effect found in an NHST analysis?	Yes	Enough data (and, hopefully) careful thought	Too much data and lack of thought, or bad luck, or both
	No	Not enough data (due to lack of thought) or bad luck (the data collected should have been sufficient for detection of the difference, but were not by chance)	Enough data and careful thought, or not enough data and lack of thought

One of the keys to being a good experimental scientist is to keep yourself in the yes/yes and no/no categories for the right reasons. The ways you do this are to carefully distinguish between a “statistically significant” effect and an important one, and to make sure that the sampling distributions that you derive from you data grant you the confidence to say whether or not an important difference between your experimental groups exists. This brings us to the next chapter, in which we will more fully explore the concept of sampling distributions.