Introduction to Statistics
Class 10, 18.05, Spring 2014
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Know the three overlapping "phases" of statistical practice.

2. Know what is meant by the term *statistic*.

2 Introduction to statistics

Statistics deals with data. Generally speaking, the goal of statistics is to make inferences based on data. We can divide this the process into three phases: collecting data, describing data and analyzing data. This fits into the paradigm of the scientific method. We make hypotheses about what's true, collect data in experiments, describe the results, and then infer from the results the *strength of the evidence* concerning our hypotheses.

2.1 Experimental design

The design of an experiment is crucial to making sure the collected data is useful. The adage 'garbage in, garbage out' applies here. A poorly designed experiment will produce poor quality data, from which it may be impossible to draw useful, valid inferences. To quote R.A. Fisher one of the founders of modern statistics,

> To consult a statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination. He can perhaps say what the experiment died of.

2.2 Descriptive statistics

Raw data often takes the form of a massive list, array, or database of labels and numbers. To make sense of the data, we can calculate *summary statistics* like the mean, median, and interquartile range. We can also visualize the data using graphical devices like histograms, scatterplots, and the empirical cdf. These methods are useful for both communicating and exploring the data to gain insight into its structure, such as whether it might follow a familiar probability distribution.

2.3 Inferential statistics

Ultimately we want to draw inferences about the world. Often this takes the form of specifying a statistical model for the random process by which the data arises. For example, suppose the data takes the form of a series of measurements whose error we believe follows a normal distribution. (Note this is always an approximation since we know the error must have some bound while a normal distribution has range $(-\infty, \infty)$.) We might then use the data to provide evidence for or against this hypothesis. Our focus in 18.05 will be on how to use data to draw inferences about model parameters. For example, assuming gestational length follows a $N(\mu, \sigma)$ distribution, we'll use the data of the gestational lengths of, say, 500 pregnancies to draw inferences about the values of the parameters μ and σ. Similarly, we may model the result of a two-candidate election by a Bernoulli(p) distribution, and use poll data to draw inferences about the value of p.

We can rarely make definitive statements about such parameters because the data itself comes from a random process (such as choosing who to poll). Rather, our statistical evidence will always involve probability statements. Unfortunately, the media and public at large are wont to misunderstand the probabilistic meaning of statistical statements. In fact, researchers themselves often commit the same errors. In this course, we will emphasize the *meaning* of statistical statements alongside the *methods* which produce them.

Example 1. To study the effectiveness of new treatment for cancer, patients are recruited and then divided into an experimental group and a control group. The experimental group is given the new treatment and the control group receives the current standard of care. Data collected from the patients might include demographic information, medical history, initial state of cancer, progression of the cancer over time, treatment cost, and the effect of the treatment on tumor size, remission rates, longevity, and quality of life. The data will be used to make inferences about the effectiveness of the new treatment compared to the current standard of care.

Notice that this study will go through all three phases described above. The experimental design must specify the size of the study, who will be eligible to join, how the experimental and control groups will be chosen, how the treatments will be administered, whether or not the subjects or doctors know who is getting which treatment, and precisely what data will be collected, among other things. Once the data is collected it must be described and analyzed to determine whether it supports the hypothesis that the new treatment is more (or less) effective than the current one(s), and by how much. These statistical conclusions will be framed as precise statements involving probabilities.

As noted above, misinterpreting the exact meaning of statistical statements is a common source of error which has led to tragedy on more than one occasion.

Example 2. In 1999 in Great Britain, Sally Clark was convicted of murdering her two children after each child died weeks after birth (the first in 1996, the second in 1998). Her conviction was largely based on a faulty use of statistics to rule out sudden infant death syndrome. Though her conviction was overturned in 2003, she developed serious psychiatric problems during and after her imprisonment and died of alcohol poisoning in 2007. See http://en.wikipedia.org/wiki/Sally_Clark

This TED talk discusses the Sally Clark case and other instances of poor statistical intuition: http://www.youtube.com/watch?v=kLmzxmRcUTo

2.4 What is *a* statistic?

We give a simple definition whose meaning is best elucidated by examples.

Definition. A *statistic* is anything that can be computed from the collected data.

Example 3. Consider the data of 1000 rolls of a die. All of the following are statistics: the average of the 1000 rolls; the number of times a 6 was rolled; the sum of the squares of the rolls minus the number of even rolls. It's hard to imagine how we would use the last example, but it is a statistic. On the other hand, the probability of rolling a 6 is *not* a statistic, whether or not the die is truly fair. Rather this probability is a property of the die (and the way we roll it) which we can *estimate* using the data. Such an estimate is given by the statistic 'proportion of the rolls that were 6'.

Example 4. Suppose we treat a group of cancer patients with a new procedure and collect data on how long they survive post-treatment. From the data we can compute the average survival time of patients in the group. We might employ this statistic as an estimate of the average survival time for future cancer patients following the new procedure. The latter is *not* a statistic.

Example 5. Suppose we ask 1000 residents whether or not they support the proposal to legalize marijuana in Massachusetts. The proportion of the 1000 who support the proposal is a statistic. The proportion of all Massachusetts residents who support the proposal is *not* a statistic since we have not queried every single one (note the word "collected" in the definition). Rather, we hope to draw a statistical conclusion about the state-wide proportion based on the data of our random sample.

The following are two general types of statistics we will use in 18.05.

1. *Point statistics*: a single value computed from data, such as the sample average $\overline{x}_n$ or the sample standard deviation s_n.
2. *Interval statistics*: an interval $[a, b]$ computed from the data. This is really just a pair of point statistics, and will often be presented in the form $\overline{x} \pm s$.

3 Review of Bayes' theorem

We cannot stress strongly enough how important Bayes' theorem is to our view of inferential statistics. Recall that Bayes' theorem allows us to 'invert' conditional probabilities. That is, if H and D are events, then Bayes' theorem says

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}.$$

In scientific experiments we start with a hypothesis and collect data to test the hypothesis. We will often let H represent the event 'our hypothesis is true' and let D be the collected data. In these words Bayes theorem says

$$P(\text{hypothesis is true} \mid \text{data}) = \frac{P(\text{data } |\text{hypothesis is true}) \cdot P(\text{hypothesis is true})}{P(\text{data})}$$

The left-hand term is the probability our hypothesis is true given the data we collected. This is precisely what we'd like to know. When all the probabilities on the right are known exactly, we can compute the probability on the left exactly. This will be our focus next week. Unfortunately, in practice we rarely know the exact values of all the terms on the right. Statisticians have developed a number of ways to cope with this lack of knowledge and still make useful inferences. We will be exploring these methods for the rest of the course.

Example 6. Screening for a disease redux
Suppose a screening test for a disease has a 1% false positive rate and a 1% false negative rate. Suppose also that the rate of the disease in the population is 0.002. Finally suppose a randomly selected person tests positive. In the language of hypothesis and data we have:
Hypothesis: H = 'the person has the disease'
Data: D = 'the test was positive.'
What we want to know: $P(H|D) = P(\text{the person has the disease} \mid \text{a positive test})$

In this example all the probabilities on the right are known so we can use Bayes theorem to compute what we want to know.

$$\begin{aligned} P(\text{hypothesis} \mid \text{data}) &= P(\text{the person has the disease} \mid \text{a positive test}) \\ &= P(H|D) \\ &= \frac{P(D|H)P(H)}{P(D)} \\ &= \frac{.99 \cdot .002}{.99 \cdot .002 + .01 \cdot .998} \\ &= 0.166 \end{aligned}$$

Before the test we would have said the probability the person had the disease was 0.002. After the test we see the probability is 0.166. That is, the positive test provides some evidence that the person has the disease.