Will Monroe
CS 109

Lecture Notes #17
August 2, 2017

Sampling and Bootstrapping

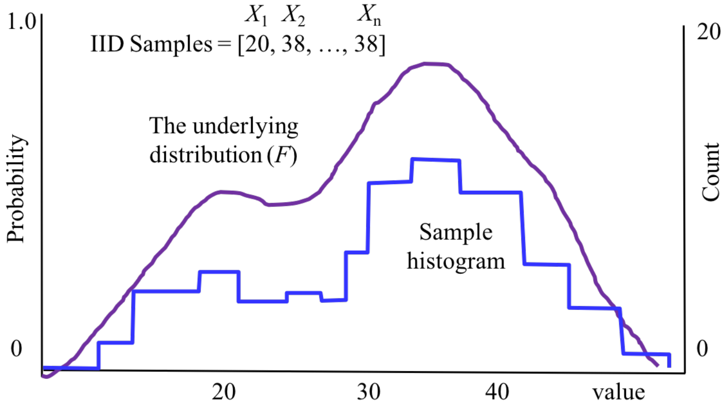
Based on a handout by Chris Piech

In this chapter we are going to talk about statistics calculated on samples from a population. We are then going to talk about probability claims that we can make with respect to the original population—a central requirement for most scientific disciplines.

Let's say you are the king of Bhutan, and you want to know the average happiness of the people in your country. You can't ask every single person, but you could ask a random subsample. In this next section we will consider principled claims that you can make based on a subsample. Assume we randomly sample 200 Bhutanese people and ask them about their happiness, on a scale of 1 to 100 (happinesses? smiles?). Our data looks like this: $72, 85, \ldots, 71$. You can also think of it as a collection of $n = 200$ I.I.D. (independent, identically distributed) random variables $X_1, X_2, \ldots, X_n$.

Understanding Samples

The idea behind sampling is simple, but the details and the mathematical notation can be complicated. Here is a picture to show you all of the ideas involved:



The theory is that there is some large population (such as the 774,000 people who live in Bhutan). We collect a sample of n people at random, where each person in the population is equally likely to be in our sample. From each person we record one number (e.g., their reported happiness). We are going to call the number from the i-th person we sampled X_i. One way to visualize your samples $X_1, X_2, \ldots, X_n$ is to make a histogram of their values.

We make the assumption that all of our X_i's are identically distributed. That means that we are assuming there is a single underlying distribution F that we drew our samples from. Recall that a distribution for discrete random variables should define a probability mass function.

Estimating Mean and Variance from Samples

We assume that the data we look at are I.I.D. from the same underlying distribution (F) with a true mean (μ) and a true variance (σ^2). Since we can't talk to everyone in Bhutan, we have to rely on our sample to estimate the mean and variance. From our sample we can calculate a sample mean ($\bar{X}$) and a sample variance (S^2). These are the best guesses that we can make about the true mean and true variance.

$$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n} \qquad\qquad S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$$

The first thing to know about these estimates is that they are *unbiased*. Having an **unbiased estimate** means that if we were to repeat this sampling process many times, the expected value each of the estimate should be equal to the true value we are trying to estimate. We will first prove that that is the case for $\bar{X}$.

$$\begin{aligned} E[\bar{X}] &= E\left[\sum_{i=1}^{n} \frac{X_i}{n}\right] = \frac{1}{n} E\left[\sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n} \sum_{i=1}^{n} E[X_i] = \frac{1}{n} \sum_{i=1}^{n} \mu = \frac{1}{n} n\mu = \mu \end{aligned}$$

The equation for the sample mean seems related to our understanding of expectation. The same could be said about sample variance except for the surprising $(n-1)$ in the denominator of the equation. Why $(n-1)$? That denominator is necessary to make sure that $E[S^2] = \sigma^2$.

The proof for S^2 is a bit more involved; you don't have to remember this, but some people may be interested in knowing it:

$$\begin{aligned} E[S^2] &= E\left[\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}\right] \\ (n-1)E[S^2] &= E\left[\sum_{i=1}^{n} (X_i - \bar{X})^2\right] = E\left[\sum_{i=1}^{n} ((X_i - \mu) + (\mu - \bar{X}))^2\right] \\ &= E\left[\sum_{i=1}^{n} (X_i - \mu)^2 + \sum_{i=1}^{n} (\mu - \bar{X})^2 + 2\sum_{i=1}^{n} (X_i - \mu)(\mu - \bar{X})\right] \\ &= E\left[\sum_{i=1}^{n} (X_i - \mu)^2 + n(\mu - \bar{X})^2 + 2(\mu - \bar{X})\sum_{i=1}^{n} (X_i - \mu)\right] \\ &= E\left[\sum_{i=1}^{n} (X_i - \mu)^2 + n(\mu - \bar{X})^2 + 2(\mu - \bar{X})(n\bar{X} - n\mu)\right] \\ &= E\left[\sum_{i=1}^{n} (X_i - \mu)^2 - n(\mu - \bar{X})^2\right] = \sum_{i=1}^{n} E\left[(X_i - \mu)^2\right] - nE\left[(\mu - \bar{X})^2\right] \\ &= n\sigma^2 - n\operatorname{Var}(\bar{X}) = n\sigma^2 - n\frac{\sigma^2}{n} = n\sigma^2 - \sigma^2 = (n-1)\sigma^2 \end{aligned}$$

So $E[S^2] = \sigma^2$.

The intuition behind the proof is that sample variance calculates the distance of each sample to the sample mean, *not* the true mean. The sample mean itself varies, and we can show that its variance is also related to the true variance.

Variance of the Sample Mean

We now have estimates for mean and variance that are not biased—that is, they are correct on average. However, the estimates change depending on the samples. How stable are they?

The sample mean is computed as an average of random variables. It takes on values probabilistically, which makes it a random variable itself. We can compute its variance:

$$\begin{aligned}
\text{Var}(\bar{X}) &= \text{Var}\left(\sum_{i=1}^{n} \frac{X_i}{n}\right) = \left(\frac{1}{n}\right)^2 \text{Var}\left(\sum_{i=1}^{n} X_i\right) \\
&= \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \text{Var}(X_i) = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \sigma^2 = \left(\frac{1}{n}\right)^2 n\sigma^2 \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

This tells us that the variance of the sample mean is proportional to the variance of the underlying distribution, but goes down with the number of samples.

Standard Error

Knowing that the variance of the sample mean is small if the number of samples is large is reassuring, but the expression for the variance of the sample mean depends on the true variance of the underlying distribution. What if we don't know that true variance? What can we say about the stability of our estimate of the mean, given only the sample we took?

We know that S^2 is an unbiased estimator for the true variance. So one reasonable thing to try is to substitute S^2 for σ^2:

$$\begin{aligned}
\text{Var}(\bar{X}) &= \frac{\sigma^2}{n} \approx \frac{S^2}{n} && \text{since } S^2 \text{ is an unbiased estimate} \\
\text{SD}(\bar{X}) &= \frac{\sigma}{\sqrt{n}} \approx \frac{S}{\sqrt{n}} && \text{since SD is the square root of Var}
\end{aligned}$$

That $\text{SD}(\bar{X})$ formula has a special name: it is called the **standard error**, and it is a common way of reporting uncertainty of estimates of means ("error bars") in scientific papers. Let's say our sample of happiness has $n = 200$ people, the sample mean is $\bar{X} = 83$, and the sample variance is $S^2 = 450$. We can calculate the standard error of our estimate of the mean to be $\frac{S}{\sqrt{n}} \approx 1.5$. When we report our results, we will say that the average happiness score in Bhutan is 83 ± 1.5, with variance 450.

If you're wondering, S^2 has a variance too; it turns out it's equal to $\frac{1}{n}\left(E[(X-\mu)^4] - \frac{n-3}{n-1}(\sigma^2)^2\right)$. We won't use that one in CS 109.

Bootstrap

The bootstrap is a statistical technique for understanding distributions of statistics. It was invented here at Stanford in 1979 when mathematicians were just starting to understand how computers, and computer simulations, could be used to better understand probabilities.

The first key insight is that if we had access to the underlying distribution (F), then answering almost any question we might have about how accurate our statistics are would become straightforward. For example, in the previous section we gave a formula for how you could calculate the sample variance from a sample of size n. We know that in expectation our sample variance is equal to the true variance. But what if we want to know the probability that the true variance is within a certain range of the number we calculated? That question might sound dry, but it is critical to evaluating scientific claims. If you knew the underlying distribution, F, you could simply repeat the experiment of drawing a sample of size n from F, calculate the sample variance from our new sample and test what portion fell within a certain range.

The next insight behind bootstrapping is that the best estimate that we can get for F is from our sample itself. The general algorithm looks like this:

```
def bootstrap(sample):
   N = number of elements in sample
   pmf = estimate the underlying pmf from the sample
   stats = []
   repeat 10,000 times:
      resample = draw N new samples from the pmf
      stat = calculate your stat on the resample
      stats.append(stat)
   stats can now be used to estimate the distribution of the stat
```

Next week we will talk in much more detail about estimating distributions from samples. For now, the simplest way to estimate F (and the one we will use in this class) is to assume that $P(X = k)$ is simply the fraction of times that k showed up in the sample. This set of probabilities defines a probability mass function for a discrete random variable, which we'll call $\hat{F}$, the "hat" indicating that $\hat{F}$ is an estimate of the probability distribution of F. This estimated distribution, formed from counts of samples, is sometimes called this **empirical distribution**.

Bootstrapping is a reasonable thing to do because the sample you have is the best and only information you have about what the underlying population distribution actually looks like. Many samples will look quite like the population they came from.

With this approach, we can compute probabilities and estimates not just for the mean, but for any statistic we want. To calculate $\text{Var}(S^2)$, for example, we could calculate ${S_i}^2$ for each resample i, and after 10,000 iterations, we could calculate the sample variance of all the ${S_i}^2$s.

You might be wondering why the resample is the same size as the original sample (n). The answer is that the variation of the variation of stat that you are calculating could depend on the size of the sample (or the resample). To accurately estimate the distribution of the stat, we must use resamples of the same size.

The bootstrap has strong theoretical guarantees, and it is accepted by the scientific community. It breaks down when the underlying distribution has a "long tail" or if the samples are not I.I.D.

Example of p-value calculation

We are trying to figure out if people are happier in Bhutan or in Nepal. We sample $n_1 = 200$ individuals in Bhutan and $n_2 = 300$ individuals in Nepal and ask them to rate their happiness on a scale from 1 to 10. We measure the sample means for the two samples and observe that people in our Nepal sample are slightly happier—the difference between the Nepal sample mean and the Bhutan sample mean is 0.5 points on the happiness scale.

Have we really shown that people in Nepal are happier? Sample means can fluctuate. How do we know that we didn't just get that difference because of the random differences among samples?

There isn't a rigorous, objective way to prove that the difference you discovered wasn't due to chance, or even to give a probability that the difference was due to chance. It is possible, however, to give a probability for the reverse statement: *if* the only difference in the samples was due to chance, what would be the probability that we get a result just as extreme?

The assumption that the difference between the samples was due to chance is an example of a **null hypothesis**. A null hypothesis says that there is no relationship between two measured phenomena or no difference between two groups. The probability we gave is known as a ***p*-value**. So a p-value is the probability that, when the null hypothesis is true, the statistic measured would be equal to, or more extreme than, than the value you are reporting.

In the case of comparing Nepal to Bhutan, the null hypothesis is that there is no difference between the distribution of happiness in Bhutan and Nepal. When you drew samples, Nepal had a mean that 0.5 points larger than Bhutan by chance.

We can use bootstrapping to calculate the p-value. First, we estimate the underlying distribution of the null hypothesis underlying distribution, by making a probability mass function from all of our samples from Nepal and all of our samples from Bhutan.

```
def pvalueBootstrap(bhutanSample, nepalSample):
   N = size of bhutanSample
   M = size of nepalSample
   universalSample = combine bhutanSamples and nepalSamples
   universalPmf = estimate the underlying pmf of universalSample
   count = 0
   repeat 10,000 times:
      bhutanResample = draw N new samples from universalPmf
      nepalResample = draw M new samples from universalPmf
      muBhutan = sample mean of bhutanResample
      muNepal = sample mean of nepalResample
      meanDifference = |muNepal - muBhutan|
      if meanDifference > observedDifference:
         count += 1
   pValue = count / 10,000
```

This is particularly nice because we never had to assume the distribution that our samples came from had a particular form (e.g., we never had to claim that happiness is normally distributed). You might have heard of a t-test. That is another way of calculating p-values, but it makes the assumptions that samples are normally distributed and have the same variance. Nowadays, when we have reasonable computer power, bootstrapping is a more versatile and accurate tool.