9. Sampling

Chris Piech and Mehran Sahami

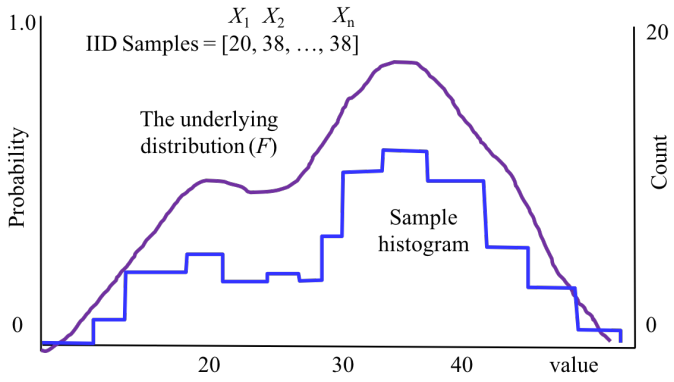
May 2017

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 and ask them about their happiness. 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 (for example 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 (for example their reported happiness). We are going to call the number from the ith 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_is 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 IID 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 question to ask is, are those unbiased estimates? Yes. Unbiased, means that if we were to repeat this sampling process many times, the expected value of our estimates should be equal to the true values we are trying to estimate. We will prove that that is the case for $\bar{X}$. The proof for S^2 is in lecture slides.

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

The equation for 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 the $E[S^2] = \sigma^2$.

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.

Standard Error

Ok, you convinced me that our estimates for mean and variance are not biased. But now I want to know how much my sample mean might vary relative to the true mean.

$$\begin{aligned} \text{Var}(\bar{X}) &= \text{Var}(\sum_{i=1}^{n} \frac{X_i}{n}) = \left(\frac{1}{n}\right)^2 \text{Var}\left(\sum_{i=1}^{n} X_i\right) \\ &= \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \text{Var}(X_i) = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \sigma^2 = \left(\frac{1}{n}\right)^2 n\sigma^2 = \frac{\sigma^2}{n} \\ &\approx \frac{S^2}{n} && \text{Since S is an unbiased estimate} \\ \text{Std}(\bar{X}) &\approx \sqrt{\frac{S^2}{n}} && \text{Since Std is the square root of Var} \end{aligned}$$

That $\text{Std}(\bar{X})$ term has a special name. It is called the standard error and its how you report uncertainty of estimates of means in scientific papers (and how you get error bars). Great! Now we can compute all these wonderful statistics for the Bhutanese people. But wait! You never told me how to calculate the $\text{Std}(S^2)$. True, that is outside the scope of CS109. You can find it on wikipedia if you want.

Let's say we calculate the our sample of happiness has n = 200 people. The sample mean is $\bar{X} = 83$ (what is the unit here? happiness score?) and the sample variance is $S^2 = 450$. We can now calculate the standard error of our estimate of the mean to be 1.5. When we report our results we will say that the average happiness score in Bhutan is 83 ± 1.5 with variance 450.

Bootstrap

The bootstrap is a newly invented statistical technique for both understanding distributions of statistics and for calculating p-values (a p-value is a the probability that a scientific claim is incorrect). 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 as to how accurate our statistics are becomes 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 simplest way to estimate F (and the one we will use in this class) is to assume that the $P(X = k)$ is simply the fraction of times that k showed up in the sample. Note that this defines the probability mass function of our estimate $\hat{F}$ of F.

```
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
```

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. Moreover most samples will, if they're randomly chosen, look quite like the population they came from.

To calculate $\text{Var}(S^2)$ 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^2s.

The bootstrap has strong theoretic grantees, and 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 Nepal are slightly happier–the difference between the Nepal sample mean and the Bhutan sample mean is 0.5 points on the happiness scale.

If you want to make this claim scientific you should calculate a p-value. 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. The null hypothesis is the hypothesis that there is no relationship between two measured phenomena or no difference between two groups.

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. The null hypothesis argument is: there is no difference in the distribution of happiness between Nepal and Bhutan. 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 the bhutanSample
    M = size of the nepalSample
    universalSample = combine bhutanSamples and nepalSamples
    universalPmf = estimate the underlying pmf of the universalSample
    count = 0
    repeat 10,000 times:
        bhutanResample = draw N new samples from the universalPmf
        nepalResample = draw M new samples from the universalPmf
        muBhutan = sample mean of the bhutanResample
        muNepal = sample mean of the nepalResample
        meanDifference = |muNepal - muBhutan|
        if meanDifference > observedDifference:
            count += 1
    pValue = count / 10,000
```

This is particularly nice because nowhere did we have to make an assumption about a parametric distribution that our samples came from (ie we never had to claim that happiness is gaussian). You might have heard of a t-test. That is another way of calculating p-values, but it makes the assumption that both samples are gaussian and that they both have the same variance. In the modern context where we have reasonable computer power, bootstrapping is a more correct and versatile tool.