10/4/2022

Go over Exam #1

Sampling Distributions
Now that we have some machinery under our belts, we can begin to approach the topic of inferential statistics. How do we take our samples and understand something about the whole population? How good are our estimates?

Estimators – biased vs. unbiased
A point estimate is a single value that estimates a population parameter. Point estimates are easy to understand, but they lack information about how accurate they are likely to be. In inferential statistics, we usually prefer a confidence interval, which is centered on a point estimate, but includes a margin of error that expresses some information about how far away from the true value of the parameter we are likely to be. You may see this interval expressed as $estimate \pm margin\ of\ error$, or as an interval $(lower\ bound, upper\ bound)$. The first form emphasizes the point estimate, the second form emphasizes the range of possible values. While these are mathematically equivalent, in this course we will prefer the interval notation.

Estimators at the center of these intervals come in many flavors. Some estimators are unbiased (the ones we use routinely in statistics are used precisely because they are unbiased). Some estimators are biased.

An example of a biased estimator is the range. The range is especially problematic because it tends to grow with the sample size. One of the reasons we discussed two formulas for the standard deviation is because when we use the population formula on a sample, it tends to underestimate the true population standard deviation. Using $n - 1$ instead of n makes the estimate unbiased.

Some population parameters may have several methods for generating an estimate. For example, we can estimate a population mean with a sample mean or a sample median. Both are unbiased, but we use the mean in part because this estimator has a small variance than the median. So, our choice of estimators generally must be both unbiased, and have the smallest variance when we have multiple methods.

When we estimate parameters from samples, we have a couple of methods we can use to obtain estimates. One method is called the Maximum Likelihood Function. A second method is called the Method of Moments.

Maximum Likelihood Functions
The maximum likelihood function is a method of estimating the most likely value of a parameter for a probability distribution given a sample of outcomes from that distribution. This handout will discuss in broad outlines the general method for constructing a maximum likelihood function and calculating the maximum likelihood estimate (MLE) from that function using calculus.

In general terms, we consider the probability distribution $f(x, \lambda)$ and collect some samples of data that obey the distribution function. For each outcome, we measure the value of x, with the parameter λ still unknown. The maximum likelihood function is the product of these outcomes, i.e. $L(f) =$

$\prod_{i=1}^{n} f(x_i, \lambda) = \prod_{i=1}^{n} f_i(\lambda)$. We will use this function to estimate the most likely value of the parameter λ. But, let's first construct the maximum likelihood function in a specific example.

Construct the maximum likelihood function for the exponential distribution modeling the time between events in a Poisson process. We take several observations and obtain the following wait-times: $x_i = \{5, 2, 1, 4, 2, 6, 3, 1, 4, 2\}$.

For the first observation, we obtained $x_1 = 5$. We substitution this into the exponential distribution $f(x, \lambda) = \lambda e^{-\lambda x}$ for x, obtaining $f_1(\lambda) = \lambda e^{-5\lambda}$. The second observation was $x_2 = 2$. So, we substitution that into the exponential distribution for x, obtaining $f_2 = \lambda e^{-2\lambda}$. And so forth.

$$f_3(\lambda) = \lambda e^{-\lambda}, f_4(\lambda) = \lambda e^{-4\lambda}, f_5(\lambda) = \lambda e^{-2\lambda}, f_6(\lambda) = \lambda e^{-6\lambda}$$
$$f_7(\lambda) = \lambda e^{-3\lambda}, f_8(\lambda) = \lambda e^{-\lambda}, f_9(\lambda) = \lambda e^{-4\lambda}, f_{10}(\lambda) = \lambda e^{-2\lambda}$$

The maximum likelihood function is the product of these expressions: $L(f) = \prod_{i=1}^{10} f_i(\lambda) =$

$$L(f) = \lambda e^{-5\lambda}\, \lambda e^{-2\lambda}\, \lambda e^{-\lambda}\, \lambda e^{-4\lambda}\, \lambda e^{-2\lambda}\, \lambda e^{-6\lambda}\, \lambda e^{-3\lambda} \lambda e^{-\lambda} \lambda e^{-4\lambda}\, \lambda e^{-2\lambda}$$
$$L(f) = \lambda^{10} e^{-30\lambda}$$

Because this probability distribution contains exponentials, we convert a product to a sum in the exponent. In this case, the exponential distribution maximum likelihood function becomes

$$L(f) = \prod_{i=1}^{n} \lambda e^{-x_i \lambda} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$$

Once we have the Maximum Likelihood function, we take the derivative and set it equal to zero to try to find the value of the parameter for which the probability is the greatest.

$$\frac{dL}{d\lambda} = n\lambda^{n-1} e^{-\lambda \sum_{i=1}^{n} x_i} - \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i} \left(\sum_{i=1}^{n} x_i \right) = 0$$

Factor out $\lambda^{n-1} e^{\lambda \sum_{i=1}^{n} x_i}$.

$$n\lambda^{n-1} e^{-\lambda \sum_{i=1}^{n} x_i} - \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i} \left(\sum_{i=1}^{n} x_i \right) = \lambda^{n-1} e^{-\lambda \sum_{i=1}^{n} x_i} \left(n - \lambda \left(\sum_{i=1}^{n} x_i \right) \right) = 0$$

It's possible that λ is 0, but that's actually a minimum. The exponential part can never be zero. So the parentheses must be zero for the maximum.

$$n - \lambda \left(\sum_{i=1}^{n} x_i \right) = 0$$

$$n = \lambda\left(\sum_{i=1}^{n} x_i\right)$$

$$\lambda = \frac{n}{\sum_{i=1}^{n} x_i}$$

This expression for λ is the reciprocal of the mean, so another way to express this is

$$\frac{1}{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

So, if we want to estimate the parameter for the distribution, we take the reciprocal of the mean of the sample. I did this derivation using the generic formula, but if we put our data back in, $\bar{x} = 3$, so the best estimate for λ is $\frac{1}{3}$.

I have a handout on my website (and linked below) that goes through this idea with other distributions. If the distribution has more than one parameter to solve for, you will have to take partial derivatives and solve both as a system of equations. Some functions, particularly some discrete distributions, will not have a smooth function we can integrate. In such cases, we can find the maximum graphically or by other means.

Method of Moments

The method of moments is another method for deriving methods of parameter estimations. For example, we saw for the gamma distribution that $\mu = \alpha\beta$, and $\sigma^2 = \alpha\beta^2$. If we have only one parameter to estimate, we can estimate it with just the first moment, but if we have two parameters, we'll need two moments and will solve it as a system of equations.

We use the discrete, sample moment for our estimate. $E(X) = \frac{1}{n}\sum_{i=1}^{n} x_i$, and $E(X^2) = \frac{1}{n}\sum_{i=1}^{n} x_i^2$. And recall the first moment is just the mean. The variance short-cut formula is expressed in terms of moments: $\sigma^2 = E(X^2) - [E(x)]^2$.

Suppose we have 10 observations from a gamma distribution:
$\{1.2, 7.0, 5.6, 3.4, 9.3, 6.2, 4.0, 3.6, 8.2, 7.9\}$. The mean is 5.64. The second moment is 37.69. So we set $\alpha\beta = 5.64$ and $\alpha\beta^2 = 37.69 - (5.64)^2 = 5.88$. We do a little algebra.

$$\alpha\beta^2 = (\alpha\beta)\beta = 5.64\beta = 5.88$$
$$\beta = \frac{5.88}{5.64} = 1.04$$

And then $\alpha\beta = \alpha(1.04) = 5.64$ making $\alpha = 5.42$.

The method of moments and the maximum likelihood function may produce slightly different estimates (or formulas) for the parameters. Sometimes they will be the same. To determine which is the best estimate to use (assuming the estimator is unbiased), we will need to look at the sampling distribution of the estimate to find the one with the smallest variance.

Central Limit Theorem

Gives us some machinery to estimate how good our estimates of parameters derived from samples are. While proving this theorem is beyond the scope of this course, we will discuss the results, and its implications for statistical inference.

In some respects, we can think of the central limit theorem as putting a numerical measurement to the law of large numbers. Recall the law of large numbers said that as we take larger and larger empirical samples, we get closer and closer to theoretical estimates of probabilities. For our unbiased estimators, we can think of the idea the same way. But the central limit theorem goes further. It allows us to say how close to the theoretical value we are likely to be for a sample of a given size. The central limit theorem forms the basis for calculating our margins of error for our confidence intervals.

A sampling distribution is the distribution of statistics collected from samples of the same size. For instance, suppose that we sample the heights of 10 women and find the mean of their heights. Then we take a sample of 10 more women and find the mean of their heights. And so forth for, let's say 100 samples of 10. We can build a histogram of those means. That is the sampling distribution. We could find the mean of the means, and measure the standard deviation of those sample means. This last value is usually referred to as the standard error. The central limit theorem establishes that the mean of the sampling distribution will tend to be centered around the true value of the mean (the parameter), and it also establishes the relationship between the population standard deviation and the sampling distribution standard deviation (the standard error).

$$\sigma_{\bar{X}} = SE = \frac{\sigma}{\sqrt{n}}$$

For proportions (whose binomial distributions can be estimated by the normal distribution), we can estimate the sample proportion the same way with the mean of the sampling distribution centered around the population parameter, and the standard error of the estimate is given by

$$\sigma_{\hat{p}} = SE = \sqrt{\frac{p(1-p)}{n}}$$

We will experiment with these sampling distributions in the lab this week.

Some statistics, especially for small samples don't start out looking much like the normal distribution. They might have thicker tails (a larger kurtosis) than the normal distribution, or they might be skewed. But as the sample sizes get larger, the central limit theorem tells us that the sampling distributions will become more normal, and more narrow.

In order to apply the central limit theorem, especially to distributions that don't start out as normal, we usually require that samples sizes be larger than a certain size. Different authors of statistics books tend to have different rules of thumb. Some books say it's safe to assume normality for $n > 30$, but some distributions need $n > 50$. For this course, we are going to assume we can apply the normal distribution safely for $n > 40$. As we will discuss when we talk about confidence intervals in the next lecture, we have another distribution, the T-distribution, that can estimate our sampling distributions with when we fail to meet that threshold.

Some sampling distributions don't have good theory to support specific measures for their variability. For that, we can also rely on simulations, or as we'll see later in the course, bootstrapping, to make estimates of those distributions.

References:

1. https://assets.openstax.org/oscms-prodcms/media/documents/IntroductoryStatistics-OP_i6tAI7e.pdf
2. https://faculty.ksu.edu.sa/sites/default/files/probability_and_statistics_for_engineering_and_the_sciences.pdf
3. http://betsymccall.net/prof/courses/spring14/cscc/2470MLEs.pdf