Statistics Review

Stock & Watson Ch 3

Andrew Bossie

NJCU

February 18, 2017

Estimating the Population Mean

So say we want to estimate μ_Y, the mean of Y. We will call this $\hat{\mu}_Y$, the hat indicates the mean of the sample distribution.

We could come up with lots of possible estimators. An obvious one is the sample average $\bar{Y}$, but its not the only one. We could just use the first observation from the same Y_1 as our estimator or the last one Y_N or any three of them and add 20, whatever. How do we chose what the best estimator is?

An **estimator**: a function of the sample of data drawn randomly from a population.

Just a reminder, that we actually have a RANDOM sample is very important.

What do we want from an Estimator?

We want the right answer and we want a precise answer!

On average you would want repeated randomly drawn samples to give you the right answer. That is $E(\hat{\mu}_y) = \mu_y$
We call this **unbiasedness**. If $E(\hat{\mu}_y) \neq \mu_y$ $\hat{\mu}_y$ is biased.

What do we want from an Estimator?

Consistency is like unbiasedness for large samples. An estimator is consistent if, as the sample get large, the probability that $\hat{\mu}_y = \mu_y$ approaches 1. It is a little different than unbiasedness, though.

Unbiasedness is about what you expect from lots of different small samples drawn out of a population. Consistency is about what you expect as one draw of a sample gets larger and larger. (Think about when the sample size=population size.)

We also worry about precision. We want to find the estimator with the smallest variance. This gives us the most "precise" estimate. The most **efficient** estimator uses the information we have most efficiently to give us the most precise estimate.

So how good is $\bar{Y}$ as an estimator?

We already went over the sample mean as an estimator or the population mean. It is both unbiased and efficient.

$$E(\bar{Y}) = \mu_Y$$

$$\bar{Y} \rightarrow \mu_Y$$

Efficiency is a new concept, though. the variance of $\hat{Y}$ is $\frac{\sigma_Y^2}{n}$ The variance of Y_1 for instance is σ_Y^2 which is "wider". There is an interesting example in the textbook (3.11) that illustrates this better.

Best Linear Unbiased Estimator (BLUE)

BLUE is going to come up a lot. It is basically a catchy phrase to describe the best estimator from all possible estimators. It is both unbiased, and also most efficient. There are often several estimators that are unbiased, so the “Best” part of BLUE is usually about efficiency.

The linear part will be more relevant when we start talking about regression estimates.

$\bar{Y}$ is the least squares estimator of μ_Y

The idea of the "least squares" estimator is going to be important. Least Squares = BLUE. We can evaluate an estimator in terms of what minimizes:

$$\sum_{i=1}^{n}(Y_i - m)^2$$

so we are looking for m (which is some estimator of E(Y)).We want the m that minimizes the gap between all the Y_i values. These "gaps" are the "mistakes" in the estimate. We want to minimize these mistakes. $m = \bar{Y}$ minimizes these mistakes.

Gun deaths example.

Hypothesis Testing

We often want to ask specific questions of the data. For instance, suppose we want to test the **null hypothesis** (the hypothesis to be tested) that the population mean (E(Y)) is equal to some value ($\mu_{Y,0}$). We will evaluate this with our sample.

$$H_0 : E(Y) = \mu_{Y,0}$$

this is contrary to the **alternative hypothesis**

$$H_1 : E(Y) \neq \mu_{Y,0}$$

This is a "two-sided alternative hypothesis".

Hypothesis Testing

A note about scientific method. We either “reject the null” hypothesis or we “fail to reject the null”. That is our “acceptance” of the null hypothesis is conditional on the assumption that it may be rejected in the future. We are never supposed to positively assert a null hypothesis is true. Though, we often get sloppy about this.

How we evaluate the null hypothesis

We never expect the sample average to be exactly equal to the hypothesis value (except by accident). So if $\bar{Y} <> \mu_{Y,0}$ it may be because our null hypothesis isn't true, but it also just may be a function of the uncertainty in random sampling.

This is where our distribution comes in. We know the sample is distributed (if its large enough) as a normal distribution. So we can calculate the probability that our $\bar{Y}$ is different than $\mu_{Y,0}$ because it really is different as opposed to just different becuase of the particular sample it is from.

How we evaluate the null hypothesis

$$p \rightarrow 0$$

We measure this with the **p-value** also called the **significance probability**. This is the probability of drawing a $\bar{Y}$ randomly that is as different from $\mu_{Y,0}$ as the estimate you have drawn.
In a distributional sense, it is the probability that $\bar{Y}$ is as far in the tails as the estimate you have drawn.

As the p-value gets smaller, the smaller the probability that you got your estimate by chance.

Saying this differently: as $p \rightarrow 0$ the probability you got the result you did randomly also goes to 0.

What do you do if you don't know the population variance?

Long story short: the sample variance is a reasonable approximation.
Sample variance

$$s_Y^2 = \frac{1}{1-n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

It goes without saying that $s_Y = \sqrt{s_Y^2}$ (sample standard deviation). You need to pay attention to the fact that this formula is slightly different than the population variance. You divded by (n-1), not n. This is a **degrees of freedom** adjustment. We need to estimate $\bar{Y}$ and this eats up some information. So we have one less degree of freedom.

What do you do if you don't know the population variance?

s_Y^2 is a consistent estimator of σ_Y^2 for essentially the same reason our estimate of the mean is consistent.

Standard Deviation: This describes the spread of values in the sample. The sample standard deviation, s, is a random quantity – it varies from sample to sample – but it stays the same on average when the sample size increases.

Standard Error is not the same thing as Standard Deviation

Standard error of the mean: This is the standard deviation of the sample mean ($\bar{Y}$) and describes its accuracy as an estimate of the population mean, μ. When the sample size increases, the estimator is based on more information and becomes more accurate, so its standard error decreases.

$$SE(\bar{Y}) = \hat{\sigma}_{\bar{Y}} = \frac{s_Y}{\sqrt{n}}$$

Notice that as $n \to \infty$ the standard error of the mean goes to zero. Again, if the sample size is the population size there is no error in the estimate.

p-values when σ_Y is unknown

We just uses the standard error of the mean instead of the standard deviation.

The t-statstic

The t-statistics is just another way of articulating you hypothesis test. It means basically the exact same thing as the p-value.

$$t = \frac{\bar{Y} - \mu_{Y,0}}{SE(\bar{Y})}$$

This a kind of **test statistic**. The t-stat is approximately distributed as N(0,1) for a large n.

You can write the p-value as a function of the t-stat. For instance t=1.96 p=5%. Or another way of thinking about this is that 95% of all possible random outcomes of $\bar{Y}$ are between t=1.96 and t=-1.96.

Prespecified Significance Levels

Usually when we do hypothesis testing we set a level at which we will decide to reject the null hypothesis or not. Scientists/economists whatever usually set it at as "95% confidence level"

That is the same as saying: $\textit{Reject } H_0 \textit{ if } p < .05$

Or saying: $\textit{Reject } H_0 \textit{ if } |t^{act}| > 1.96$

Rejecting the null

When you are testing a hypothesis you can make two errors:

type I error the null hypothesis is rejected when it is in fact true
type II error the null hypothesis is not rejected when it is false
When you pick a **significance level** you are picking the level at which you are willing to make a type I error. The p value tells you the percent chance that you got your result randomly and that you are mistaking that random result for a statement about the null hypothesis.

The **power of the test** is the probability that you are going to make a type II error. The smaller a t-stat you pick decreases the power of the test.

You are trying to balance between avoiding these two types of errors.

Common Confidence Intervals

Confidence intervals give you a “spread” around μ_Y. Confidence intervals give a range for what you would expect X% of random values to fall under:
95% confidence interval for $\mu_Y = \bar{Y} \pm 1.96SE(\bar{Y})$
90% confidence interval for $\mu_Y = \bar{Y} \pm 1.64SE(\bar{Y})$
99% confidence interval for $\mu_Y = \bar{Y} \pm 2.58SE(\bar{Y})$

We also sometimes want to do a “one-sided” hypotheses test (i.e. does education increase wages). In which case the t-stats sign matters and the t-stat needs to be adjust. For instance $t = \pm 1.64$ for a 90% confidence interval on a two tailed test, but $t = +1.64$ gives you a 95% confidence for the hypothesis $\bar{Y} > \mu_{Y,0}$. We probably wont focus on one-sided tests much.

Comparing Means from Different Populations

We often want to talk about the difference between two means:

- Do male and female college graduates earn the same amount on average?
- Do African Americans have higher rates of diabetes on Average?
- Were whites with incomes below $50k more likely to vote for Donald Trump than whites with incomes above $50k?

So we will want to test for the difference between two means. The textbook example: college educated men (m) vs women's (w) earnings

$$H_0 : \mu_m - \mu_w = d_0$$

$$H_1 : \mu_m - \mu_w \neq d_0$$

Notice d_0 does not have to be zero, but often it is.

Comparing Means from Different Populations

The information we have: $\bar{Y}_m$ and $\bar{Y}_w$. These are independent because they come from different samples. They are both distributed $N(\mu_i, \frac{\sigma_i^2}{n_i})$ where i=(m,w).

So because these are independent variables, the distribution of their linear combination $\bar{Y}_m - \bar{Y}_w$ is is the some of the two population variances. That is the distribution is:

$$N(\mu_m - \mu_w, \frac{\sigma_m^2}{n_w} + \frac{\sigma_w^2}{n_w})$$

Comparing Means from Different Populations

When you calculate the standard error for $\bar{Y}_m - \bar{Y}_w$, then you have to consider the new distribution.

$$SE(\bar{Y}_m - \bar{Y}_w) = \sqrt{\frac{s_m^2}{n_m} + \frac{s_w^2}{n_w}}$$

Where s_i^2 is the sample variance for men and women respectively.

Comparing Means from Different Populations

Calculating the t-stat and the p-value (which is just a function of the t-stat) is straightforward once you have calculated the SE.

$$t = \frac{\bar{Y}_m - \bar{Y}_w}{SE(\bar{Y}_m - \bar{Y}_w)}$$

Table(3.1)

A quick note on causality

Sometimes we are fine with just getting a “correlational” effect where we can say one variable is related to another.

However, we often want to make **causal** statements:

- Do pesticides cause cancer (not just associated with cancer)?
- Does government spending cause output to increase?
- Does ibuprofen stop headaches?

To answer causal questions the ideal are “randomized controlled experiments”. These are very difficult to pull off in economics. However, establishing causality is the primary focus of econometrics now.

A quick note on causality

When we are looking for a causal effect we are really looking for a difference in means. In the ideal “randomized controlled experiments” we are able to figure out what the **treatment effect** is. Ideally we randomly select people to participate. Then we assign treatment (X=1) randomly. The group that is not treated is the control group (X=0 means you are in the control group).

The causal effect, then is:

$$E(Y|X=1) - E(Y|X=0)$$

The estimate of the treatment effect is the estimate in the difference of the mean of Y conditional on being treated minus the mean of Y conditional on not being treated. Again, in practice it is very hard to get credible estimates of the treatment and control.

Economists often rely on “natural experiments” to approximate a random control

Small Samples and t-stats

When talking about the t-stat etc so far we have relied on the central limit theory (CLT), which makes things easier because we can assume our sample distribution is normal. The CLT means we don't have to worry about the distribution of the population.

With small samples it is more important that the underlying distribution be normal for us to say anything meaningful about our estimates

For small samples the t-stat is a **Student t distribution** with n-1 degress of freedom.

What this means in practice is that with small samples ($n < 30$?) your critical values come from the Students t distribution, not a normal distribution.

Small Samples and t-stats

You also have to make adjustments to calculate the pooled variance. You can also calculate the standard error if both groups have the same number of observations or the same variance.

Generally $n < 30$ is a low bar, so I am not going to spend much time talking about small samples. Though you need to be aware if you have $n < 30$ you need to be more cautious about your estimates. Both in calculating SEs but also whether the hypothesis testing is valid at all.

Sample Covariance and Correlation

This basically functions the same as when we were talking about sample variance. We just use the covariance of the two samples as our substitute for the population covariance. There is also a degree of freedom correction.

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

the **sample correlation** is also similarly like population covariance:

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

and again $-1 \leq r_X Y \geq 1$

The covariance of the X and Y is consistent for large samples for similar reasons everything else is consistent (i.e. what happens to the estimate of the covariance of X and Y when the samples=populations.)