Daniel Cullen & Travis Cyronek
Econ 140A
Updated: August 23, 2017

Statistics Review

1 Random Variable

When the value of a variable is the outcome of a statistical experiment, we call is a *random variable*. Random variables are usually, but not always, denoted by (capital) X. Random variables can be discrete or continuous. *Discrete* random variables are those which take on a countable number of values. For example, consider the roll of a fair dice. The values that the random variable could take on are 1, 2, 3, 4, 5, and 6. Note that discrete random variables could take on an infinite number of possible outcomes: 1, 2, 3, ..., ∞ (that is, the values that X might take can be countably infinite). *Continuous* random variables can take on infinitely many values. For example, we might consider the height of people in the class. In measuring someones height, we might find that a student is 5.763... ft. (presuming we can measure height with infinite precision).

The outcomes of a random variable are associated with some notion of frequency (to use the word loosely, "probability"). For a discrete random variable, the frequency with which a particular outcome happens is that outcome's probability. That is, let x_1, x_2, ..., x_n be the possible outcomes of the random variable X associated with the probability distribution p_1, p_2, ..., p_n, where $p_i = Pr(X = x_i)$ for $i = 1, \ldots, n$. To use our earlier example of the roll of a dice, $x_1 = 1$, $x_2 = 2$, ..., $x_6 = 6$ and $p_i = 1/6 \; \forall$ i (because it's a fair dice).

The continuous case is (slightly) different. The probability of any particular outcome of a continuous random variable is 0 (it can take on infinitely many values!). Probabilities in this environment can be thought of as calculating an area under the appropriate *probability density function (pdf)*. Recall that a pdf is a function that gives us an idea of the relative likelihood of particular values of a random variable. One of the more common pdfs that we'll encounter is that of the normal distribution (a bell curve). The formula for a random variable X that is normally distributed with mean μ and variance σ^2 is

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

2 Expected Value

One of the more important features of the distribution of a random variable is its *expected value*. This feature gives us an idea about where a distribution is centered. Some books might call the expected value the "long-run average" of an experiment. That is, it's the average of an experiment if we were to continually repeat the experiment. How we write the expected value depends on whether or not the variable is discrete or continuous (but notice the similarities).

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} x_i p_i \qquad \text{(Discrete)}$$

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x) dx \qquad \text{(Continuous)}$$

There are a few useful properties of expected values that we should be very comfortable with in this class. Let X and Y be random variables; let a and b be scalar constants.

$$\begin{aligned}
&\mathbb{E}[a] = a \\
&\mathbb{E}[aX] = a\mathbb{E}[X] \\
&\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y] \\
&\mathbb{E}[aX+bY] = a\mathbb{E}[X] + b\mathbb{E}[Y] \qquad \text{(this simply combines the previous two)}
\end{aligned}$$

This is often called the "linearity of expectations" because scalar multiplication and addition are both linear operators. As a side note, because we like to write things succinctly, we often times will define $\mu \equiv \mathbb{E}[X]$.

3 Variance, Standard Deviation, and Covariance

Another important feature of a distribution is its *variance* (and, by extension, its standard deviation). This gives us a measure of the spread, or dispersion, of the outcomes of a random variable. It is defined as follows.

$$\begin{aligned}
Var(X) &= \mathbb{E}[(X - \mathbb{E}(X))^2] \\
&= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \qquad \text{(another way to write it)}
\end{aligned}$$

Again, because we love notation, we'll often write $Var(X) = \sigma^2$. Next, the *standard deviation* of the random variable X is simply the square root of its variance. I.e.

$$St.dev(X) = \sqrt{Var(X)} \qquad \text{or} \qquad \sigma = \sqrt{\sigma^2}$$

Last, the covariance of two random variables X and Y can tell us how they move together. That is, if X is high when Y is high, then the covariance is positive. If X is high when Y is low

(or vice versa), the covariance is negative. The covariance is defined as

$$Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \equiv \sigma_{XY}$$

Just like with the expected value, there are a lot of useful properties that we should be very comfortable with.

$$\begin{aligned}
&Var(a) = 0 \\
&Var(aX) = a^2 Var(X) \\
&Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2ab\, Cov(X, Y)
\end{aligned}$$

$$\begin{aligned}
&Cov(a + X, b + Y) = Cov(X, Y) \\
&Cov(aX, bY) = ab\, Cov(X, Y) \\
&Cov(X + Y, Z) = Cov(X, Z) + Cov(Y, Z) \\
&Cov(X, X) = Var(X) \\
&Cov(X, a) = 0
\end{aligned}$$

4 Hypothesis Testing

First, a review of some terms that we will see frequently in this course.

Paramter: true characteristic / feature about something (often unknown); (e.g. $\mu = \mathbb{E}[X_i]$, $\sigma^2 = Var(X_i)$, β)

Estimator: a function of a random variable(s) used to obtain an estimate for some parameter of interest; (e.g. $\overline{X} = \frac{1}{n}\sum X_i$, $S^2 = \frac{1}{n-1}\sum (X_i - \overline{X})^2$, $\widehat{\beta}$)

Estimate: the value obtained after applying an estimator to actual data; (e.g. $\overline{x}$, s^2, b)

Hypothesis testing is a way of framing and answering "yes / no" statistical questions. To perform and implement a hypothesis test, we can follow some fairly general steps.

1. Form the null / alternative hypotheses about some parameter of interest and determine which estimator we are using to estimate this parameter.
2. Form a test statistic (this will depend on the assumptions we are making about the model and what the null / alternative hypotheses look like).
3. Find a critical value associated with your tolerance for error.
4. Determine if your calculated test-statistic falls in the rejection region and conclude the test.

Step 1.

Typically our problems will not tell you explicitly what the hypotheses we want to test are. As such, we will have to use our intuition and understanding to figure it out. Specifically what we have to look out for is what parameter we want to test. In most cases, we will also want to determine if we want a 1-tailed or a 2-tailed test. A good way to jump into this is to consider a simple example.

Suppose we want to test to see if the mean height of people in a school is 68 inches (let Y_i be person i's height). Further, suppose that the school has 1600 students in it ($n = 1600$). We will use the arithmetic average as our estimator: $\overline{Y} = \frac{1}{n}\sum_i Y_i$. We gather the data and do the above calculation, and find that $\overline{y} = 74$ with $s = 120$. In this example, our hypotheses will be

$$H_0: \ \mu = 68 \qquad \text{vs.} \qquad H_1: \ \mu \neq 68 \qquad \text{(2-sided)}$$

Step 2.

The next thing we need to do is form a test statistic. In this step what we are doing (essentially) is taking our estimator (which will have some distribution) and normalizing it so that we can quickly compare it and another, known distribution. The most common ones that we will encounter in this class are t and z-statistics. To make this a little clearer, let's revisit our example. If we recall from class, under some assumptions we have

$$\overline{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Using this we can "standardize" the estimator:

$$\overline{Y} - \mu \sim N\left(0, \frac{\sigma^2}{n}\right)$$

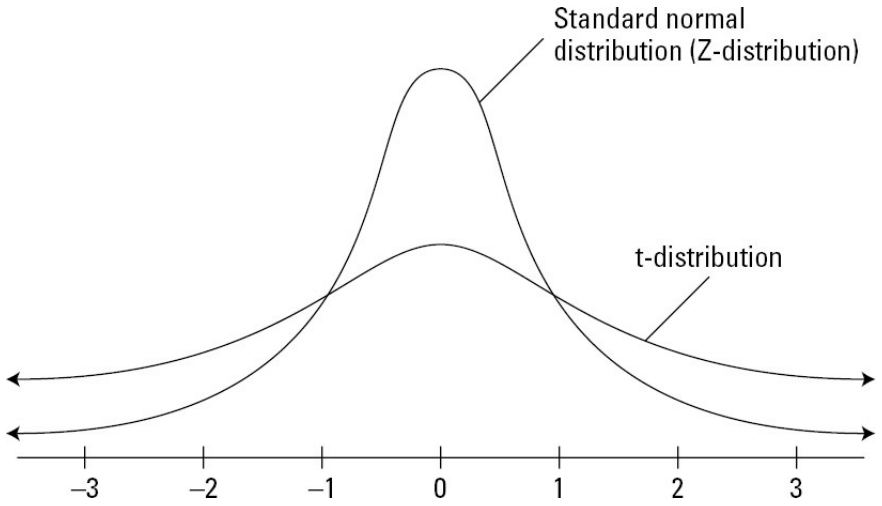
$$\underbrace{\frac{\overline{Y} - \mu}{\sigma/\sqrt{n}}}_{\equiv z} \sim N(0,1) \qquad \text{(use the rules from before to convince yourself of this)}$$

We often call the "thing" on the left hand side of the previous statement the z-statistic. It is of particular interest to us because it has a standard normal distribution. Perhaps subtle, we might not be able to use a z-statistic for a couple of reasons. First, we might not know σ, the population standard deviation. Second, our sample size might not be large enough for us to be confident that the estimator has converged to a standard normal distribution (this stems from the Central Limit Theorem (CLT))[1]. As such, we would want to use a t-statistic,

[1]This point is actually a little more nuanced than is appropriate for the discussion here. If you want to know more, come see me and I can redirect you to a good reference or two.

$$t = \frac{\overline{Y} - \mu}{s/\sqrt{n}} \sim t_{n-1},$$

which is distributed t with $n-1$ degrees of freedom. The 1 comes from number of things we are estimating from the data (we are just estimating the mean for this running example). If we were estimating more things (such as k slope parameters in a regression model), then it would be distributed t_{n-k}, for instance. What is the relationship between the distribution of z and the distribution of t? They look very similar; indeed they are both bell-shaped. It turns out that the relative likelihood of obtaining values in the tails is higher for the t-distribution.



A question we get frequently is when we should use a t or a z-statistic for a particular hypothesis test. You can use the following rules. If σ is unknown, use a t-statistic. If σ is known and $n \geq 30$, use a z-statistic, otherwise use a t-statistic. Returning back to our example, we will have to use a t-statistic because we don't know σ.

$$\widehat{t} = \frac{\overline{y} - \mu_{H_0}}{s/\sqrt{n}} = \frac{74 - 68}{120/\sqrt{1600}} = \frac{6}{3} = 2$$

Before we move on to the next step, we should note that the formula above (for the z and t-statistics) are for the sample mean specifically. We can still do all of the above with other estimators. Generally, the formula generalize to the following.

$$t = \frac{\widehat{\beta} - \beta_{H_0}}{SE(\widehat{\beta})},$$

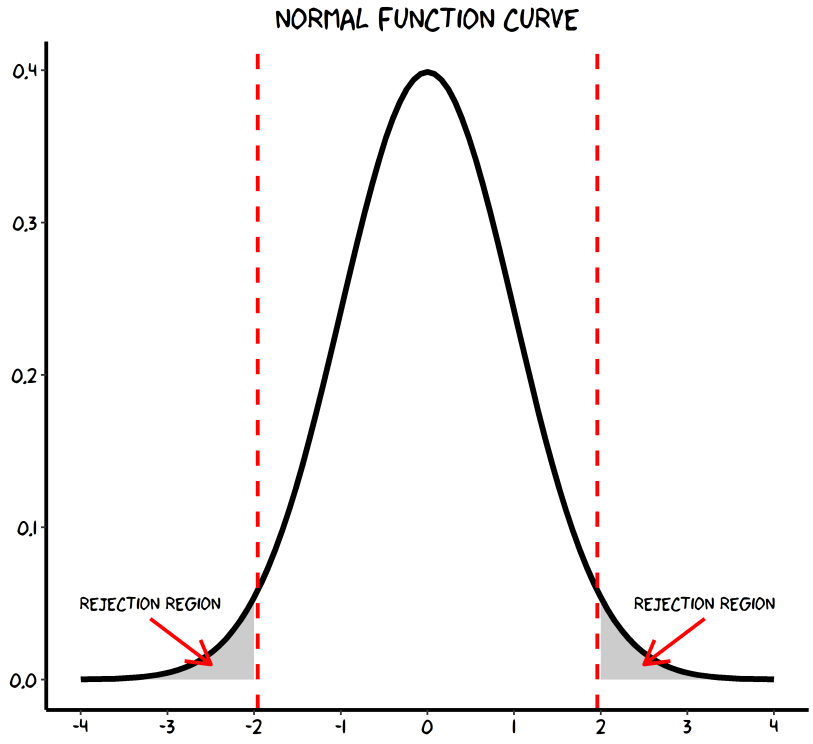
where $SE(\cdot)$ is the standard error (think of this as the standard deviation) of the estimator $\widehat{\beta}$. The standard error for the sample mean is

$$SE(\overline{Y}) = \frac{s}{\sqrt{n}}$$

Step 3.

Now we have to determine the critical value which will determine whether we *reject* or *fail to reject* the null (note that we never accept the null; think of the courtroom analogy[2]). The most common level of significance is the 5% level. That is, we will choose critical values such that the probability of rejecting the null hypothesis given that the null is correct (sometimes called type I error) is 5%. To find the critical values we would have to look it up in tables[3].

For our example (a 2-sided test with a 5% significance level and 1599 degrees of freedom) the critical values are about -1.96 and +1.96. That is, there is 0.05 in the tails beyond these values (0.025 in each tail, as the bell curve is symmetric). Those tails will make up what we call the *rejection region*. Below is a visual of this.



Step 4.

[2]That is, a jury decides whether or not a defendant is guilty or not guilty and does not directly address the issue of innocence (innocence being analogous to accepting the null).

[3]I would recommend looking at past exams and getting familiar with those tables that we give you for exam purposes.

The final step is to conclude our test. To do so, we follow a simple rule. If $\widehat{t}$ is in the rejection region formed by t_{cv}, reject the null hypothesis. For our example, because $\widehat{t} = 2$ is in the rejection region, we reject the null hypothesis at the 5% level of significance. There is evidence to suggest that the average height of students in the school is not equal to 68 inches.

Notice that I wrote more than just "reject" or "fail to reject." You might think its obvious, but we will eventually get to hypotheses that are a little more complex and less straightforward. Get into the habit of actually writing out a conclusion like I have above in order to get full credit on exams! Before moving on to the last section, let's try a similar (but slightly different) example.

Suppose, as before, that we want to test to see if the mean height of people in a class is below 68 inches. Again, we calculated that $\overline{y} = 74$ with $s = 120$ and $n = 1600$. (What is the difference between this problem and the one before?)

1. The null and alternative hypotheses will be slightly different, but we will still be using the sample average as our estimator.

$$H_0: \ \mu < 68 \qquad \text{vs.} \qquad H_1: \ \mu \geq 68 \qquad \text{(1-sided)}$$

2. Our test statistic will be the same as before:

$$\widehat{t} = \frac{\overline{y} - \mu_{H_0}}{s/\sqrt{n}} = \frac{74 - 68}{120/\sqrt{1600}} = \frac{6}{3} = 2$$

3. Let's say, as before, we want a confidence level of 5%. We now will look for a critical value where 5% is in the tail beyond it (as opposed to 2.5% from the 2-tailed test). Is the critical value for this particular 1-tailed test going to be positive or negative? Why? Indeed, the critical value will be 1.645. To determine what sign the critical value will be in a 1-tailed test, we must ask ourselves, when would we reject the null hypothesis? Recall that the null in this example is that the mean is less than 68 inches. Loosely speaking, we wouldn't believe that statement if we went out and obtained a number that was well above 68 inches. Thus (when normalizing with our test statistic) we know that the critical value here will be positive.

4. Because the calculated test statistic falls in the rejection region (i.e. $2 > 1.645$), we reject the null hypothesis at the 5% level of significance. There is evidence to suggest that the mean height of students at this school is not less than 68 inches.

5 Similar Topics

There are a few other things you guys might want to look over, though I won't go into any detail here. First, refresh yourself on what a *p-value* is. It is very closely related to a t-statistic. It tells us the probability of obtaining an estimate more extreme than the one we calculated. In the case of p-values, we will reject if the p-value is small (as opposed to relatively large).

The other thing you want to look over are *confidence intervals.* We can use these things to do the same thing as the hypothesis tests we did in the previous section. In general, the confidence interval for the estimator $\widehat{\beta}$ will be given by

$$b \pm t_{cv} \times SE(\widehat{\beta})$$

Again, I implore you to look more into these as we will use them in the class.