Chapter 2 outline, Econometrics

This is the chapter that reviews statistics. In this chapter we will cover (among other things):

- random variables
- probability distributions
- estimation
- properties of estimators
- hypothesis testing
- descriptive statistics

My suggestion is to gain a firm grasp of hypothesis testing from this chapter, as it will be the most applicable item on the list. However, you should also understand why we want our estimators to have the properties that they have.

1 Random variables

random variable – variable that takes on alternative values, each with a probability less than or equal to one (note that the variable may be greater than 1, but the PROBABILITY is less than or equal to one)

A *probability distribution* is the process that generates random variables. There are many different types of probability distributions. You may be familiar with some – the normal distribution (this is the bell curve distribution), the uniform distribution (it basically says every possible choice for the random variable has an equal chance of being chosen) and the binomial distribution (think about flipping a coin, with heads being assigned a value of 1 and tails a value of 0).

Probability distributions may be *continuous* or *discrete*. A random variable from a continuous probability distribution can take on ANY value in the real number line. A random variable from a discrete probability distribution allows only certain values to be chosen. For an example of a discrete probability distribution, think about rolling a die. If you use a 6-sided fair die, then each number (1,2,3,4,5,6) represents a random variable that can be generated from this probability distribution. What is the probability of generating any of those numbers with a fair die? It is $\frac{1}{6}$. So the random variable we obtain from rolling the die is generated from a discrete uniform distribution. As for a continuous random variable, think about a variable like temperature. Temperature can take on both positive and negative values, and depending on how fine one makes the scale, it can take on values like 92.64578 degrees.

1.1 Expected values

The probability distribution that we will use most frequently in this class as the basis for the linear regression model is the normal distribution. We can completely describe a continuous normal distribution with two pieces of information, its mean and variance. (In order to contrast this with another distribution, we can completely describe a continuous uniform distribution by its endpoints.) Typically, we will call the mean of the distribution the expected value, which is exactly what you would think it is. The expected value is the weighted average of the possible outcomes of a random process. The weights that we apply are the probabilities that correspond to each particular outcome. I will show an example of the expected value for the discrete uniform distribution of rolling the die.

$\frac{1}{6} * 1 + \frac{1}{6} * 2 + \frac{1}{6} * 3 + \frac{1}{6} * 4 + \frac{1}{6} * 5 + \frac{1}{6} * 6 = \frac{21}{6} = 3.5$

The $\frac{1}{6}$s are the expected probability of rolling each number, and the possible outcomes are 1, 2, 3, 4, 5, and 6. In generic mathematical notation, we can represent the expected value as:

$p_1X_1+p_2X_2+p_3X_3+...+p_NX_N = \sum_{i=1}^{N} p_iX_i$ – note that the p_is represent the probability of each outcome and the X_is represent the outcomes. There are two ways that we denote expected values, we use μ_X (this is the Greek letter mu, and we call the expected value "mu of X") or $E[X]$. Notice that the formula for the expected value of X is very similar to the formula that we have obtaining the arithmetic mean of X (or $\bar{X}$), which is $\frac{1}{N}\sum_{i=1}^{N} X_i$. In fact, the arithmetic mean simply assumes equal weight on each of the observations. That weight is given by $\frac{1}{N}$. We could write $\sum_{i=1}^{N} \frac{1}{N}X_i$ to make it look more like the form for the expected value. One other fact that you should recall from basic probability is that $\sum_{i=1}^{N} p_i = 1$, which just says that the sum of the probabilities of all possible outcomes is 1. Notice that $\sum_{i=1}^{N} \frac{1}{N} = 1$.

We have already defined the variance of a variable as $\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2$. We will now define the variance using the expected value operator, which is a more general method of defining the variance. We denote variance as $Var(X)$ or σ_X^2 (this is the Greek letter sigma, and we call the variance of X "sigma squared of X"). So $\sigma_X^2 = \sum_{i=1}^{N} p_i[X_i - E(X)]^2$. Alternatively, we could write that $\sigma_X^2 = E[X - E[X]]^2$. Once again, this is very similar to the formula that we already have for the variance. If we replace p_i with $\frac{1}{N}$ and $E(X)$ with $\bar{X}$, then we would have the exact same formula that I gave on the first day.

1.1.1 Why use expected values?

Why not just use the formulas for arithmetic mean and variance from the appendix to chapter 1? The reason why we use the expected value notation is because it does not assume that all the observations are equally weighted like the formulas from the appendix to chapter 1. As I already mentioned, the probability distribution that we will base most of our results on is the normal distribution. The normal distribution does not have equal probability weights for all of its outcomes, so we need to use the expected value notation because it is more general and allows for differing probability weights.

There are a few results that we have for expected values, just like we had some rules for the summation operator.

Result 1 $E[aX + b] = aE[X] + b$, where X is a random variable and a and b are constants.

Note what $E[aX + b]$ means. $E[aX + b] = \sum_{i=1}^{N} p_i[aX + b]$

$\sum_{i=1}^{N} p_i[aX + b] = \sum_{i=1}^{N} p_i aX_i + \sum_{i=1}^{N} p_i b$

We can use our summation operator rules to get:

$\sum_{i=1}^{N} p_i aX_i + \sum_{i=1}^{N} p_i b = a\sum_{i=1}^{N} p_i X_i + b\sum_{i=1}^{N} p_i$

We know $\sum_{i=1}^{N} p_i = 1$ and $\sum_{i=1}^{N} p_i X_i = E[X]$, so

$a\sum_{i=1}^{N} p_i X_i + b\sum_{i=1}^{N} p_i = aE[X] + b$

Result 2 $E[(aX)^2] = a^2E[X^2]$, where X is a random variable and a is a constant

You should practice your knowledge of expectation operators by proving result 2.

Result 3 $Var(aX + b) = a^2Var(X)$, where X is a random variable and a and b are constants.

Recall $var(X) = E[X - E[X]]^2$, so $Var(aX + b) = E[(aX + b) - E[(aX + b)]]^2$

First, work on the expected value operator inside all the brackets. We know that $E[(aX + b)] = aE[X] + b$. So now we have:

$E[(aX + b) - (aE[X] + b)]^2$

Now, distribute the negative sign through to get:

$E[(aX + b) - (aE[X] + b)]^2 = E[aX + b - aE[X] - b]^2$

$E[aX + b - aE[X] - b]^2 = E[aX - aE[X]]^2$

Now factor the a out:

$E[aX - aE[X]]^2 = E[a(X - E[X])]^2$

Distribute the square:

$E[a(X - E[X])]^2 = E[a^2(X - E[X])^2]$

Now, using our $\sum_{i=1}^{N} p_i$ form,

$E[a^2(X - E[X])^2] = \sum_{i=1}^{N} p_i[a^2(X - E[X])^2]$

We can factor out the a^2 to get:

$\sum_{i=1}^{N} p_i[a^2(X - E[X])^2] = a^2 \sum_{i=1}^{N} p_i[(X - E[X])^2]$

Returning to the $E[X]$ notation:

$a^2 \sum_{i=1}^{N} p_i[(X - E[X])^2] = a^2 E[(X - E[X])^2]$, which is just $a^2 Var(X)$

1.2 Joint distributions of random variables and 2.1.3 Independence and correlation

Suppose we have two random variables, X and Y. The joint distribution is given by the list of probabilities of all possible outcomes on both X and Y. Below is a table for the joint distribution of rolling a pair of six-sided dice.

Die 1/Die 2	1	2	3	4	5	6
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
2	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
3	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
4	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
5	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
6	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$

Note that if we sum all of the probabilities in the table the result is 1. Also note that if we sum along either a column or a row we get a probability of $\frac{1}{6}$, which is just the probability of rolling any of the possible numbers on one of the dice. Finally, note that this probability distribution is discrete. If we were working with continuous probability distributions we would not be able to write down a table like this because continuous distributions assume that *every* number along the real number line is able to be chosen (and I have yet to see anyone write down every single number).

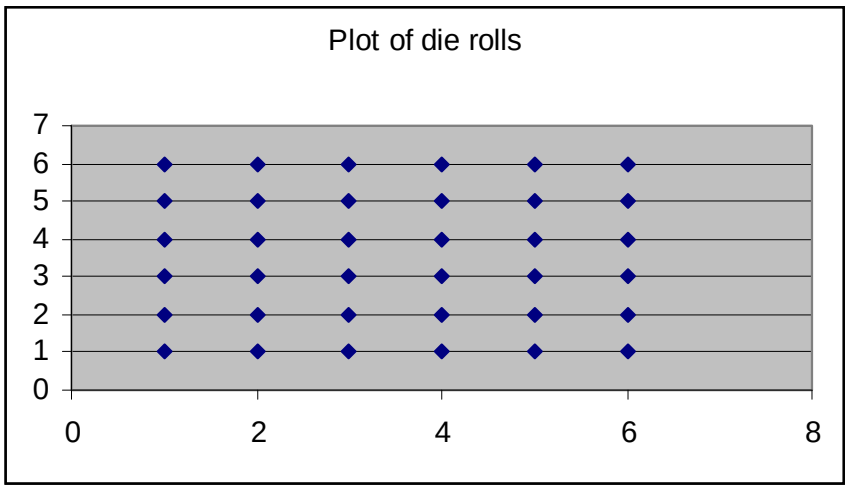
1.2.1 Covariance and the expectations operator

We can write down the covariance of two random variables using the expectations operator. Once again, we use the expectations operator instead of the simple formula for the covariance given in the appendix to chapter 1 because using the expectations operator is more general than the simple formulas.

$$Cov(X,Y) = E[(X-E[X])(Y-E[Y])] = \sum_{i=1}^{N} \sum_{j=1}^{N} (p_{ij}(X_i-E[X])(Y_j-E[Y])),$$

where p_{ij} is the joint probability of X and Y

If we calculate the covariance for the joint distribution of the dice rolling example above, we get a covariance of: zero. This is exactly what we should



expect, for two reasons. First, we know that the outcomes of the individual die rolls are *independent* of one another, which means that what occurs on one die roll does not affect the other die roll. If two random variables are independent, then they will have a covariance of zero. (You should note that a covariance of zero does NOT imply that they are independent). Second, suppose we plot the points in the joint distribution of the die rolls. The plot would look as follows:

It doesn't look like there is a very good linear relationship between X and Y, so we should expect the covariance to be near zero (if not exactly zero).

There problem we have with covariance is that the magnitude is NOT scale free; it depends on the units of measurement. Take a look at the sample data below. In the first table, the height is measured in inches and the weight is measured in pounds. In the second table, I have converted the height from inches to centimeters by multiplying the height column in the first table by 2.5 and I have converted the weight from pounds to ounces by multiplying the weight column in the first table by 16. So all I have done is rescale the variables: I have **NOT** changed the relationship between the variables.

height	weight	height	weight
64	150	160	2400
71	200	177.5	3200
67	127	167.5	2032
74	301	185	4816
61	158	152.5	2528
73	167	182.5	2672
79	241	197.5	3856
59	114	147.5	1824
62	112	155	1792
67	148	167.5	2368
75	159	187.5	2544
76	182	190	2912

If we calculate the covariance of the height and weight in the first table, we get a value of: 227.9.

If we calculate the covariance of the height and weight in the second table, we get a value of: 9076.7.

However, as I already mentioned, nothing has changed about the relationship of the variables except that they have been measured in different scales. Since we would really not like our results to depend on the scales we use to measure the variables, we have the *correlation coefficient*. The correlation coefficient is a scale free measurement of the relationship of two random variables. The correlation coefficient will lie between 1 and -1 (I have worked out a proof for you, use this link to get there); positive correlation coefficients imply a direct relationship between X and Y while negative correlation coefficients imply an inverse relationship between X and Y. A correlation coefficient of 1 implies perfect positive correlation and a correlation coefficient of -1 implies perfect negative correlation. We use the Greek letter ρ to denote the correlation coefficient.

$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{\sigma_x^2 \sigma_y^2}}$, where σ_x^2 is the variance of X and σ_y^2 is the variance of Y.

If we calculate the correlation coefficient for the height and weight in the first table, we get: 0.689951

If we calculate the correlation coefficient for the height and weight in the second table, we get: 0.689951.

Thus, regardless of the scale that we use to measure height and weight we get the same correlation coefficient, which is what we should expect since the relationship between the variables did not change, only the scale. (NOTE: If you calculate this correlation coefficient by hand or by using Excel you may get slightly different numbers depending on how you calculate the variance of X and Y. I used the formula that I gave at the beginning of the course for the variance, the one with N in the denominator. The spreadsheet packages may use $N-1$, for reasons which we will discuss shortly.)

Here are 4 more results using expectations operators.

Result 4 If X and Y are random variables, $E[X+Y] = E[X] + E[Y]$

Use this link to go to the proof.

Result 5 $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$
See the book for this proof, pg. 49.

Result 6 If X and Y are independent random variables, $E[XY] = E[X]E[Y]$
See the html file I have on the web site for the proof of this proposition. You can use this link.

Result 7 If X and Y are independent random variables, $Cov(X,Y) = 0$
See the proof in the book on pg. 49. It uses result 6 above.

2 Desirable properties of estimators

You should notice that I have put this section before section 2.2, mainly because section 2.2 uses many of the concepts in 2.3. The only information that you need from section 2.2 to understand section 2.3 is that we define a population is all the possible outcomes of a process – meaning we have every single data point. Since it is very unlikely that we ever have the true population when applying our regression analysis, we will be concerned with samples. Samples are just subsets of the population. What we would like to have are methods of estimating the population parameters (the mean, variance, and covariance, among other things) from the sample(s) that we draw. We would like those estimators to have some, if not all, of the following properties.

2.1 Lack of bias

If an estimator is unbiased, it means that, on average, the estimator provides the "true value". Suppose the true value of our parameter is β. Suppose our estimator for β is $\hat{\beta}$. An unbiased estimator is one where $E[\hat{\beta}] = \beta$, which means that on average our estimator yields the true parameter.

We define the amount of bias of an estimator as: $bias = E[\hat{\beta}] - \beta$. Note that while an estimator may be unbiased, nothing is implied about the variance of the distribution of our estimator. At this point you should note that estimators ARE random variables – they depend on the sample drawn, and as such have a sample mean (expected value) and a sample variance.

2.2 Efficiency

This is probably the 40^{th} definition of efficiency that you have seen as undergraduates in economics. The efficiency that we discuss is the efficiency of the estimator. Suppose we have 2 estimators for our "true value" of β. Call those two estimators $\hat{\beta}$ (beta hat) and $\check{\beta}$ (beta upside down hat). Also suppose that both estimators are unbiased. If $\hat{\beta}$ has a lower variance than $\check{\beta}$, then we say

that $\hat{\beta}$ is *relatively* more efficient than $\check{\beta}$, because both estimators are unbiased and $\hat{\beta}$ has the lower variance.

We say that an estimator is the most efficient estimator if it is an unbiased estimator and it has a lower variance than any other unbiased estimator. I will not go into the details of how to prove that estimators are efficient, although there are two basic methods. The first involves something called the C-R (Cramer-Rao) bound and the second involves the method of Lehman and Scheffé. They require some knowledge of matrix algebra, but if you are interested you can look up these methods in most graduate econometrics texts.

Why do we care about efficiency? Again, suppose we have $\hat{\beta}$ (beta hat) and $\check{\beta}$ (beta upside down hat). If $\hat{\beta}$ has a smaller variance than $\check{\beta}$, then I am less likely to be far away from the "true value" of β if I use $\hat{\beta}$, because $\hat{\beta}$ does not vary as much as $\check{\beta}$.

Important note: We only compare the variances of UNBIASED estimators when we look for the efficient estimator. There may be biased estimators with a lower variance than our unbiased estimator, but that biased estimator is NOT more efficient because it has some bias. For a method of comparing biased and unbiased estimators, see the next small section on minimum mean square error.

2.3 Minimum mean square error

Sometimes (although not for what we will do in this class) you may wish to use estimators that are biased but have a smaller variance than any unbiased estimators. It could be the case that the efficient estimator has a large variance and the biased estimator, while it has some bias, has a much smaller variance. The question then becomes, Which would you rather have? An estimator that is on average correct (unbiased) but may be very far away from the true parameter at times, or an estimator that is on average not correct (biased) but usually very close to the true parameter.

We define mean square error (MSE) as: $MSE = E[(\hat{\beta} - \beta)^2]$

We can "easily" rewrite this as: $MSE = [Bias(\hat{\beta})]^2 + Var(\hat{\beta})$

Note that if an estimator is unbiased, $MSE = Var(\hat{\beta})$; however, a biased estimator may have a lower MSE if its variance and bias are both small, and the unbiased estimator has a large variance.

2.4 Consistency

Consistency as a large-sample (or asymptotic) property of estimators. If our estimator is consistent, it means that as the sample size gets very large the probability that our estimator $\hat{\beta}$ is different from our "true value" β becomes very small.

Consistency is a property that we would prefer to have over all the other properties (although a consistent, efficient, minimum mean square error estimator would be better than just one that is consistent). The rationale is as

follows: suppose you have a biased but consistent estimator of a parameter. On average, you are not correct, but as you gather more and more data you get closer to obtaining the true value. If you have an unbiased but inconsistent estimator, this means that on average you are correct, but gathering more data does NOT get you closer to the true parameter.

3 Estimation

Now that we have discussed desirable properties of estimators, let's review the estimators for mean, variance, and covariance that I gave you in chapter 1.

3.1 Estimators of the mean, variance, and covariance

In this section estimators of the mean, variance, and covariance are discussed.

3.1.1 Mean

Recall that we proposed using $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ as an estimate of the mean of the variable. The question is, Is this a "good" (in terms of the criteria we defined above) estimator for the mean of a variable? In order to show that $\bar{X}$ is a good estimator of the mean, recall that the population mean of a random variable X is denoted μ_x. To show that $\bar{X}$ is an unbiased estimator of μ_x we need to show that $E[\bar{X}] = \mu_x$.

$E[\bar{X}] = E[\frac{1}{N}\sum_{i=1}^{N} X_i]$

$E[\frac{1}{N}\sum_{i=1}^{N} X_i] = \frac{1}{N}E[\sum_{i=1}^{N} X_i]$

Now, we need the expected value of the sum of $X_i's$. Think about what this means:

$E[X_1 + X_2 + ... + X_N] = E[X_1] + E[X_2] + ... + E[X_N]$

Note that the expected value of drawing any X at random is just μ_x. So we have:

$E[X_1] + E[X_2] + ... + E[X_N] = \mu_x + \mu_x + ... + \mu_x = N\mu_x$

Plugging back into our equation a few steps ago, we get:

$\frac{1}{N}E[\sum_{i=1}^{N} X_i] = \frac{1}{N}N\mu_x = \mu_x$

Thus, our estimator for the population mean, μ_x, is unbiased.

3.1.2 Variance

Recall that our estimator for the variance in chapter 1 was $Var(X) = \frac{1}{N}\sum_{i=1}^{N}[(X_i - \bar{X})^2]$

Is this a good estimator for the variance of a random variable? It looks like what we want, although we can show that it is a biased estimator of the variance. A better (in the sense that it is unbiased) estimator of the variance is given by: $Var(X) = \frac{1}{N-1}\sum_{i=1}^{N}[(X_i - \bar{X})^2]$. The book does a good job of showing this proof on pages 50-51.

3.1.3 Covariance

Recall that our estimator for the covariance in chapter 1 was $Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}[(X_i - \bar{X})(Y_i - Y)]$

Once again, this estimator is biased. An unbiased estimator of the covariance is given by:

$Cov(X,Y) = \frac{1}{N-1}\sum_{i=1}^{N}[(X_i - \bar{X})(Y_i - Y)]$

3.1.4 Sample correlation coefficient

Now that we have our unbiased estimators of the variance and covariance, we can recalculate the correlation coefficient. All we need to do is replace the biased estimators that we had previously with the new unbiased estimators.

3.2 The Central Limit Theorem

An important theorem (that we will exploit) is the central limit theorem. The central limit theorem states: If the random variable X has mean μ and variance σ^2, then the sampling distribution of $\bar{X}$ becomes approximately normal with mean μ and variance σ^2/N.

What does this mean for us? It means we can simplify our methods of statistical testing if we have a large enough sample of data by assuming that the mean of the data is normally distributed.

4 Probability Distributions

Four types of probability distributions that are useful in statistical testing are discussed below.

4.1 Normal

The normal distribution is the foundation for our statistical testing. It is commonly referred to as the "bell-shaped" distribution. We can completely define a normal by its mean and variance. Suppose we have a random variable X that is normally distributed with mean μ_x and variance σ_x^2. We would write this as:

$X \tilde{} N(\mu_x, \sigma_x^2)$

This is read just as I have written above: X is normally distributed with mean μ_x and variance σ_x^2. Now, there is a function that describes the normal distribution. It is:

$\frac{1}{\sqrt{2\pi\sigma_x^2}} \exp[\frac{-1}{2\sigma_x^2}(X_i - \mu_x)^2]$

This function tells us the what the probability is of drawing a random X_i. If we integrate over this function from $-\infty$ to ∞, we get:

$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp[\frac{-1}{2\sigma_x^2}(X - \mu_x)^2]dx = 1$, which just means that the area under the curve is equal to 1, which should make sense since evaluating this integral is similar to adding up all the probabilities from a discrete probability distribution.

What we will be primarily concerned with is how close a random draw from that distribution is to the mean. There are two basic equations you should know for hypothesis testing.

$\Pr(\mu_x - 1.96\sigma_x < X_i < \mu_x + 1.96\sigma_x) \approx 0.95$

$\Pr(\mu_x - 2.57\sigma_x < X_i < \mu_x + 2.57\sigma_x) \approx 0.99$

What these equations say is the following:

The first equation says that the probability of a random draw being within 2 standard deviations of the mean is about 0.95. The second equation says that the probability of a random draw being within 2.5 standard deviations of the mean is about 0.99. Thus, it is "very" likely that a random draw will be within 2 standard deviations and "extremely" likely that a random draw will be within 2.5 standard deviations.

Now, where do the pairs (-1.96,1.96) and (-2.57,2.57) come from? They come from the following:

$\int_{-1.96}^{1.96} \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp[\frac{-1}{2\sigma_x^2}(X - \mu_x)^2]dx \approx 0.95$

$\int_{-2.57}^{2.57} \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp[\frac{-1}{2\sigma_x^2}(X - \mu_x)^2]dx \approx 0.99$

One other note that needs to be made about the normal is that most of the other distributions we will talk about depend on the STANDARD normal distribution. We define a standard normal distribution as one where a random variable X is distributed normally with mean 0 and variance 1. So we would write this as $X \tilde{} N(0,1)$. We can transform ANY random variable that is distributed normally into a standard normal by subtracting off the mean from the random variable and dividing the result by the standard deviation of the random variable. For instance, suppose $X \tilde{} N(\mu_x, \sigma_x^2)$. Then, $\frac{X-\mu_x}{\sigma_x} \tilde{} N(0,1)$.

A nice result of the STANDARD normal is that if $Z \tilde{} N(0,1)$, then we can say that there is a 95% chance that a random draw will fall between -1.96 and 1.96. There is also a 99% chance that a random draw will fall between -2.57 and 2.57.

I will discuss more about the normal when we get to hypothesis testing. For now, we like the normal distribution because:

1. It is symmetric and bell-shaped, which is how we assume our variables are distributed.

2. The distribution is fully described by its mean and variance – it does not assume that we know very much about the distribution.

Here is a link that discusses standardizing normally distributed random variables.

4.2 Chi-square

Another useful (although not the most useful) probability distribution is the Chi-square distribution, denoted χ^2. It is useful for testing hypotheses about the variances of random variables.

Result 11 The sum of the squares of N independently distributed standard normal random variables is distributed as a chi-square with N degrees of freedom.

The chi-square distribution starts at the origin (which means that it is positive, which is always positive, which is part of the reason it is useful for testing hypotheses about variances), has a tail that goes to infinity, and is skewed to the right. The shape of the chi-square depends on the number of degrees of freedom, and as the degrees of freedom increase the chi-square begins to look more like the normal.

Suppose we have some variable Z that is distributed as a chi-square with N degrees of freedom. We would write this as: $Z \tilde{} \chi^2_N$.

4.3 t-distribution

The t-distribution is the distribution that we will use most often when we test hypotheses. We use the t-distribution to test hypotheses when the population variance of a random variable is not known.

Result 12 Assume that $X \tilde{} N(0,1)$ and $Z \tilde{} \chi^2_N$. If X and Z are independent, then $\frac{X}{\sqrt{\frac{Z}{N}}}$ has a t-distribution with N degrees of freedom.

The t-distribution is symmetric like the normal, but it has larger tails. As the degrees of freedom increase, the tails become smaller (flatter) and the t-distribution approximates the normal distribution as the degrees of freedom become very large.

Suppose we have some random variable W that is distributed as a t-distribution with N degrees of freedom. We would write that as: $W \tilde{} t_N$.

4.4 F-distribution

The F-distribution is useful in testing joint hypotheses. We will use the F-distribution in later chapters for various testing purposes.

Result 13 If X and Z are independent and distributed as chi square with N_1 and N_2 degrees of freedom, respectively, then $(\frac{X}{N_i})/(\frac{Z}{N_2})$ is distributed according to the F distribution with N_1 and N_2 degrees of freedom.

The F-distribution is similar to the chi-square in that it starts at the origin and that it is skewed to the right. If some random variable Y has the F-distribution with N_1 and N_2 degrees of freedom, we would write that as: $Y \tilde{} F_{N_1,N_2}$

5 Hypothesis testing

We do hypothesis testing to check the reliability of our estimates. What we do is use the data to make probabilistic statements about our estimates. How do we do this? We use the information on the distribution of the random variable to create confidence intervals. A confidence interval is just a range of values that is likely to contain the true value of a population parameter, given a certain level of confidence. If we have $Z \tilde{} N(0,1)$, we know that we can say that "there is a 95% chance that a random draw from this distribution will lie between -1.96 and 1.96". An alternative method of testing hypotheses involves creating a test statistic and using the tables of the distributions in the back of the book. I will go through both methods, although they are related.

5.1 Hypothesis testing with known variances

The types of hypothesis testing we do depend on the amount of information that we know. Although it will be highly unlikely that we know the population variance of a random variable from our economic data, I will go through the steps of testing hypotheses when the variance is known to set up the structure of hypothesis testing.

5.1.1 Using confidence intervals

Here are the steps to test hypotheses about the population mean when the variance of the random variable is known, using the confidence interval approach.

1. Set up your hypothesis (null and alternative)
2. Choose a significance level
3. Construct a confidence interval
4. Determine whether the hypothesized value of the falls within the bounds of the confidence interval

The first step is to set up a hypothesis. Typically, we set up the hypothesis that the true mean is equal to zero. However, we can test whatever hypothesis we want. Suppose that we wish to test the hypothesis that $\mu_x = 4$. We will call this the null hypothesis, and we will denote it as $H_o : \mu_x = 4$. We also

need to set up an alternative hypothesis (what happens if the null is not true). One alternative hypothesis that we can set up is that $\mu_x \neq 4$. We denote this as $H_A : \mu_x \neq 4$.

The second step is to determine the level of significance that you want to test the null hypothesis at. The level of significance is equal to $1-(level\ of\ confidence)$. Thus, if we want to set up a 95% confidence level, we test at a level of significance of 5%. We like to test at low levels of significance (high levels of confidence) for reasons I will discuss later.

The third step involves setting up the confidence interval. What estimate do we have for the population mean? We have $\bar{X}$, the sample mean. We know that if $X \tilde{} N(\mu_x, \sigma_x^2)$, then $\bar{X} \tilde{} N(\mu_x, \frac{\sigma_x^2}{N})$, at least for large N. This is what the Central Limit Theorem tells us. Now, we want to construct a 95% confidence interval for μ_x. What do we know? We know that if $Z \tilde{} N(0,1)$, then a 95% confidence interval is given by $-1.96 < Z < 1.96$. We know that $\bar{X}$ is NOT distributed as a standard normal, but we do know that we can transform it into a standard normal by subtracting off the mean and dividing through by the standard deviation (which is just the square root of the variance). So, if $\bar{X} \tilde{} N(\mu_x, \frac{\sigma_x^2}{N})$, then $\frac{(\bar{X}-\mu_X)\sqrt{N}}{\sigma_x} \tilde{} N(0,1)$. So a 95% confidence interval for $\frac{(\bar{X}-\mu_X)\sqrt{N}}{\sigma_x}$ would be $-1.96 < \frac{(\bar{X}-\mu_X)\sqrt{N}}{\sigma_x} < 1.96$. Now, we want to see what the range for μ_x is, so we want to isolate μ_x inside the inequalities. By doing a little algebra (use this link to see the algebra), we find:

$\bar{X} - 1.96\frac{\sigma_x}{\sqrt{N}} < \mu_x < \bar{X} + 1.96\frac{\sigma_x}{\sqrt{N}}$

Now, we know σ_x (the variance is known), we know $\bar{X}$ (we estimated it from our sample), and we know N (it is just the number of observations we used to obtain $\bar{X}$). Suppose $\sigma_x = 5$, $\bar{X} = 5.2$, and $N = 10,000$. Our 95% confidence interval would be: $5.102 < \mu_x < 5.298$.

As for step four, we had $H_o : \mu_x = 4$. Since 4 does not fall between 5.102 and 5.298, we reject the null hypothesis and conclude that there is a 95% chance that the true population mean does not equal 4.

We could follow the same steps for a 99% confidence interval. Recall that if $X \tilde{} N(0,1)$, then there is a 99% chance that a random draw from the distribution will lie between -2.57 and 2.57. Replacing the 1.96s with the 2.57s, we get our 99% confidence interval to be: $5.0715 < \mu_x < 5.3285$. Notice that is we increase our confidence level the range of values increases. However, we can still say that there is a 99% chance that the true population mean is not 4.

What if we only had $N = 100$. Our confidence intervals would be:

95%: $4.22 < \mu_x < 6.18$

99%: $3.915 < \mu_x < 6.485$

Notice that when we have a lower number of observations that the confidence intervals widen. Since we haven't gathered as much information, we are not as certain about the statements we make. Notice that we FAIL TO REJECT the null hypothesis at the 99% confidence level now, although we can still reject at the 95% confidence level.

5.1.2 Using the tables to test hypotheses

We can also use the tables in the back of the book to test hypotheses, which is what we are going to do when we begin testing the significance of regression coefficients. Here are the steps for random variables with known variances.

1. Set up your hypothesis
2. Construct your test statistic.
3. Find the associated level of significance in the table for the normal distribution
4. Accept or reject the null hypothesis

For step 1, we just set up the null and alternative hypotheses. We will again have $H_o : \mu_x = 4$ and $H_A : \mu_x \neq 4$.

For step 2, we need to construct our test statistic. The test statistic that we want to construct is $\left|\frac{(\bar{X}-\mu_X)\sqrt{N}}{\sigma_x}\right|$ (I'll talk about the absolute value sign momentarily). We know this test statistic is distributed as $N(0,1)$ and the tables in the book are for standard normals. Plugging in the values for $\bar{X}, \sigma_x$, and $\sqrt{N}$, we get $\frac{(5.2-\mu_X)*10}{5}$ (this assumes I use the example above with 100 observations). For μ_X we use the null hypothesis, and set $\mu_X = 4$. Thus, our test statistic is:

$\frac{(5.2-4)*10}{5} = 2.4$

For step 3, we find 2.4 in the normal table. The value associated with 2.4 is: 0.0082. This means that there is only 0.82% chance that a value lies in the upper tail of the normal distribution. However, this value does NOT tell us the exact level of significance of our result – we saw earlier that we could NOT reject the null hypothesis at the 99% confidence level, and this result should not change. What we have to account for is that we can have a similar value on the negative side of the distribution, because we have the alternative hypothesis set up as $H_A : \mu_x \neq 4$. So we need to allow μ_x to be lower than 4 as well as above 4. Thus, we multiply $0.0082 * 2 = 0.016\,4 = 1.64\%$. This means that 4 will fall in our confidence intervals if we set up our level of confidence at 98.36% – if our confidence interval is less than that, 4 will not fall into the range, so we can reject the null hypothesis if we set our confidence level less than 98.36%. The quick and dirty method to using the test statistic is as follows: if you get out a statistic that is greater (in absolute value) than 1.96, then we can reject the null hypothesis at the 5% significance level. If the value is greater (in absolute value) than 2.57, then we can reject the null hypothesis at the 1% level.

We use the absolute value of the test statistic because when we set up our alternative hypothesis we set it up as a two-tailed test, meaning that we want to reject the null hypothesis if the critical value is either very far above OR below the null hypothesis. We could have set up a one-tailed test, where the $H_A : \mu_x \geq 4$. We would do this if we felt that there was a good reason to believe that if the mean was not 4, then it would be greater than 4. We should only

set up one-tailed tests if we have a very, very good reason to believe that the true value will be either strictly above or strictly below the null hypothesis. In the one-tailed case, if you get a test statistic that is greater than 1.65 than you can reject the null hypothesis at the 5% level, and if it is greater than 2.33 than you can reject the null hypothesis at the 1% level. Note that if you set up the alternative hypothesis as $H_A : \mu_x \leq 4$, then if you get a test statistic that is less than -1.65 you reject the null at the 5% level and if you get a test statistic that is less than -2.33 you reject the null at the 1% level. Note that the signs on the test statistic are IMPORTANT when conducting ONE-TAILED tests.

5.2 Hypothesis testing with unknown variances

Most of the time in economics we will not know the population variances. In this case, when we construct a test statistic we replace the population variance with the estimated variance. When we do this, we CANNOT use the normal distribution to test hypotheses; instead, we use the t-distribution to test hypotheses.

5.2.1 Why the t-distribution?

Our test statistic is now: $\frac{(\bar{X}-\mu_x)\sqrt{N}}{\hat{\sigma}_x}$. We can show that $\frac{(\bar{X}-\mu_x)\sqrt{N}}{\hat{\sigma}_x} \sim t_{N-1}$. Recall that a statistic has the t-distribution if it is the ratio of a standard normal and the square root of a χ^2 divided by its degrees of freedom. So, is $\frac{(\bar{X}-\mu_x)\sqrt{N}}{\hat{\sigma}_x}$ the ratio of a standard normal and the square root of a χ^2 divided by its degrees of freedom? First, recall that:

$\frac{(\bar{X}-\mu_x)\sqrt{N}}{\sigma_x} \sim N(0,1)$ (note that this is the population standard deviation in the denominator)

Now, I will state the following without proof (the proof involves more detail than we need to go into):

$\frac{(N-1)\hat{\sigma}_x^2}{\sigma_x^2} \sim \chi^2_{N-1}$

Now, if we take the ratio of the standard normal to the square root of the χ^2 divided by its degrees of freedom, we get:

$$\frac{\frac{(\bar{X}-\mu_x)\sqrt{N}}{\sigma_x}}{\sqrt{\frac{(N-1)\hat{\sigma}_x^2}{\sigma_x^2}/(N-1)}}$$

This big mess is distributed t_{N-1}. If we simplify the denominator, we get:

$\frac{\frac{(\bar{X}-\mu_x)\sqrt{N}}{\sigma_x}}{\frac{\hat{\sigma}_x}{\sigma_x}}$ (because the $N-1$'s cancel out and the square root cancels out the squares).

But this is just:

$\frac{(\bar{X}-\mu_x)\sqrt{N}}{\sigma_x} \div \frac{\hat{\sigma}_x}{\sigma_x} = \frac{(\bar{X}-\mu_x)\sqrt{N}}{\sigma_x} * \frac{\sigma_x}{\hat{\sigma}_x} = \frac{(\bar{X}-\mu_x)\sqrt{N}}{\hat{\sigma}_x}$

Now that we have shown that $\frac{(\bar{X}-\mu_x)\sqrt{N}}{\hat{\sigma}_x} \sim t_{N-1}$, we can use the t-distribution to find critical values.

5.2.2 Hypothesis testing using the t-distribution

We could set up confidence intervals using the t-distribution, but it is much easier to just calculate the test statistic and use the table in the back of the book to test our hypothesis. When conducting a hypothesis test using the t-distribution, use the following steps.

1. Set up your hypothesis.
2. Construct your test statistic
3. Decide which level of significance you will use.
4. Check the table in the back of the book to see if the absolute value of the test statistic is greater than the critical value at your desired significance level and the degrees of freedom that you have.
5. If the absolute value of the test statistic is greater than the critical value, reject the null hypothesis. If the test statistic is less than the critical value, you fail to reject the null hypothesis at the chosen significance level.

As an example suppose that we have the 64 observations, a sample mean of 520 and a sample standard deviation of 100. We want to test the null hypothesis that the population mean of the random variable is equal to 500.

1. $H_0 : \mu = 500$, $H_A : \mu \neq 500$
2. $\frac{(\bar{X}-\mu_x)\sqrt{N}}{\hat{\sigma}_x} = \frac{(520-500)\sqrt{64}}{100} = 1.6$
3. Suppose we want to use the 5% level.
4. We look at the t-distribution table in the book, and it says that the critical value for the t-distribution at the 5% level with 63 degrees of freedom is approximately 2. Note that there is no row for 63 degrees of freedom, so I used the critical value for 60 degrees of freedom since 63 is closer to 60 than it is 120.
5. Since our test statistic is less than the critical value, we fail to reject the null hypothesis.

5.3 A few more notes on hypothesis testing

5.3.1 Type I and Type II errors

There are 4 possibilities that can occur when we perform statistical tests of hypotheses. We can fail to reject a true hypothesis, fail to reject a false hypothesis, reject a true hypothesis, and reject a false hypothesis. Two of these cases, failing to reject a true hypothesis and rejecting a false hypothesis, are correct conclusions. However, we can make errors if we fail to reject a false hypothesis or if we reject a true hypothesis. A Type I error occurs when we

reject a true hypothesis and a Type II error occurs when we fail to reject a false hypothesis. When we perform our statistical testing, the level of significance that we choose is our probability of making a Type I error. We can attempt to minimize Type I errors if we choose low levels of significance. However, as we lower the level of significance we increase the chances of making a Type II error because we are widening our confidence interval, thereby including more values in the range of our confidence interval.

5.3.2 P-values

A p-value is a probability value; it is an exact level of significance for the test statistic. If a p-value is 0.05, then this means that the estimate is significant at the 5% level. It is also significant at the 6% level, the 7% level, ..., the 12% level, ..., etc. It is NOT significant at levels below 5%. P-values measure the probability of Type I errors. When we run regressions using SAS, you may see a p-value that looks something like <.0001, which means that the result is "highly significant".

5.3.3 The power of a test

If we have a high p-value (like 0.80) this means that you should fail to reject the null hypothesis. What could be the cause of failing to reject the null hypothesis? It could be that the null hypothesis is true. It could also be that the null hypothesis is false and the data set that you are using just happens to be consistent with the null. How do we determine the likelihood that our data set is consistent with the null, causing us to fail to reject it? We use the power of a test. The power of a test is one minus the probability of making a Type II error. The power of a test depends on both the size of the effect being studied and the size of the data set. Ceteris paribus, if the size of the data set increases, then the power of the test increases. If you fail to reject a null hypothesis when you perform statistical analysis with relatively low power, you should not be concerned.

The following table summarized the relationship between Type I errors, Type II errors, p-values, and the power of a test.

Decision	H_0 True	H_0 False
Fail to reject H_0	Correct decision	Type II error (1 – power)
Reject H_0	Type I error (p-value)	Correct decision

6 Descriptive statistics

In this section I will just list a few more descriptive statistics.

6.1 Median

The median is a measure of central tendency (just like the mean) that is more robust to outliers than the mean. For data sets with odd numbers, the median is the middle observation when the observations are ranked from high to low. For even numbers, the median is typically calculated as the average of the middle two observations when the observations are ranked from high to low. As an example of how the median is more robust to outliers than the mean, consider the following. Suppose 31 households were asked what their annual income was. 30 of those households said $40,000, while 1 household said $4 million.

The mean annual income is: $((30 * 40000) + (1 * 4000000))/31 = \$167,741.94$

The median annual income is: $30,000.

Which of those numbers more accurately reflects the sample?

6.2 Skewness

Skewness provides information on the symmetry of a distribution. The formula for calculating skewness is:

$S = \frac{1}{N} \sum_{i=1}^{N} \frac{\left(X_i - \bar{X}\right)^3}{\hat{\sigma}^3}$

Skewness will be positive when the tail on the right side is thicker than the one on the left side, and negative when the tail on the left side is thicker than the one on the right side.

6.3 Kurtosis

Kurtosis provides information on the thickness of the tails of a distribution. For a normal distribution, kurtosis is equal to 3. When the tails are thicker than the normal, the statistic will be greater than 3; when they are thinner the statistic will be less than 3. The formula for kurtosis is:

$S = \frac{1}{N} \sum_{i=1}^{N} \frac{\left(X_i - \bar{X}\right)^4}{\hat{\sigma}^4}$

6.4 Jarque-Bera statistic

To "test" if a data series is normally distributed, we could check to see if the mean and median are equal, if the skewness is close to zero, and if the kurtosis is close to 3. However, a more formal test of normality is given by the Jarque-Bera test. The Jarque-Bera statistic is distributed as a chi-square with two degrees of freedom. To calculate the statistic, use:

$JB = \left[\frac{N}{6}\right] \left[\frac{S^2 + (K-3)^2}{4}\right] \tilde{} \chi_2^2$ (note that S is the skewness, NOT the standard deviation)

If the Jarque-Bera statistic is greater than the critical value of the chi-square, reject the assumption of normality.