STAT 462

Applied Regression Analysis

Lesson 1: Statistical Inference Foundations

Overview of this Lesson

This lesson provides a brief refresher of the main statistical ideas that will be a useful foundation for the main focus of this course, regression analysis, covered in subsequent lessons. To simplify matters at this stage, we consider univariate data, that is, datasets consisting of measurements of just a single variable on a sample of observations. By contrast, regression analysis concerns multivariate data where there are two or more variables measured on a sample of observations. Nevertheless, the statistical ideas for univariate data carry over readily to this more complex situation, so it helps to start as simply as possible.

Key Learning Goals for this Lesson:
▪ Review the main ways to identify and summarize data numerically and graphically. ▪ Review the process of statistical thinking, which involves drawing inferences about a population of interest by analyzing sample data. ▪ Use the normal probability distribution to make probability calculations for a population assuming known mean and standard deviation. ▪ Use the normal probability distribution to make probability calculations for a sample assuming known standard deviation. ▪ Use a t probability distribution to make probability calculations for a sample using the sample standard deviation. ▪ Calculate confidence intervals for a population mean. ▪ Conduct hypothesis tests for a population mean using the rejection region and p-value methods. ▪ Calculate prediction intervals for an individual observation.

- 1.1 - Identifying and Summarizing Data (/stat462/node/248)
- 1.2 - Population Distributions (/stat462/node/249)
- 1.3 - Selecting Individuals at Random (/stat462/node/250)
- 1.4 - Random Sampling (/stat462/node/251)
- 1.5 - Interval Estimation (/stat462/node/252)
- 1.6 - Hypothesis Testing (/stat462/node/253)
- 1.7 - Random Errors and Prediction (/stat462/node/254)

1.1 - Identifying and Summarizing Data ›
(/stat462/node/248)

STAT 462

Applied Regression Analysis

1.1 - Identifying and Summarizing Data

Statistics is a collection of methods for analyzing data to understand a problem quantitatively and to help make decisions in real-world contexts. We start by framing a problem in such a way that it will be amenable to quantitative analysis (this step lies outside the scope of this course). We assume that we have already obtained sample data relevant to the problem at hand, data that can be considered to be representative of some larger population for which we wish to make statistical inferences.

We next consider identifying and summarizing the data at hand. For example, suppose that we have moved to a new city and wish to buy a home. In deciding on a suitable home, we would probably consider a variety of factors, such as size, location, amenities, and price. For the sake of illustration we focus on price and, in particular, see if we can understand the way in which sale prices vary in a specific housing market. For this example, identifying the data is straightforward: the units of observation are a random sample of size $n = 30$ single-family homes in our particular housing market, and we have a single measurement for each observation, the sale price in thousands of dollars ($), represented using the notation $Y = Price$. Here, Y is the generic letter used for any univariate data variable, while *Price* is the specific variable name for this dataset. These data are available in the houseprice (/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/houseprice.txt) data file—they represent sale prices of 30 homes in Eugene, Oregon during 2005.

The 30 homes in this dataset have been selected randomly from the population of all single-family homes for sale in this housing market. We can simply list small datasets such as this. The values of Price in this case are:

155.5	195.0	197.0	207.0	214.9	230.0	239.5	242.0	252.5	255.0
259.9	259.9	269.9	270.0	274.9	283.0	285.0	285.0	299.0	299.9
319.0	319.9	324.5	330.0	336.0	339.0	340.0	355.0	359.9	359.9

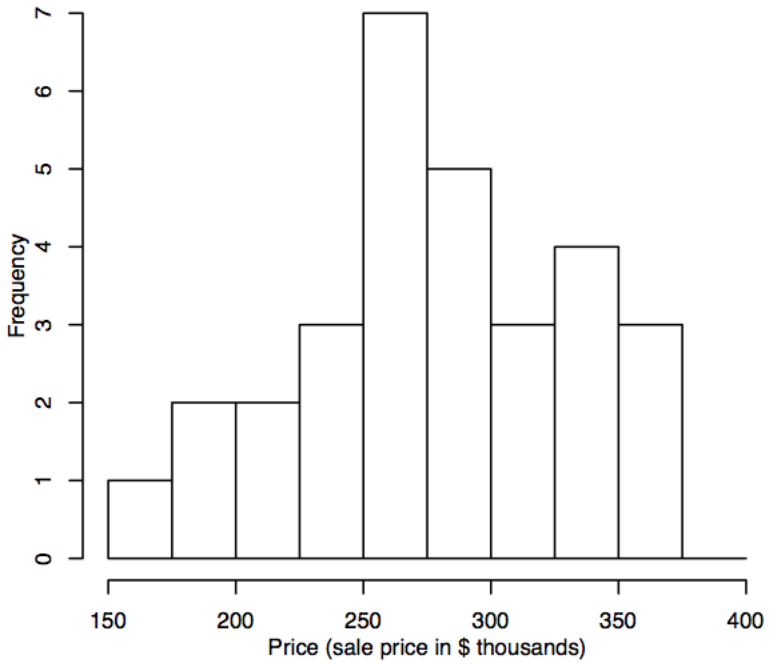
However, even for these data, it helps to summarize the numbers with sample statistics (such as the sample mean and standard deviation) or graphs. A particularly effective graph here is a stem-and-leaf plot, which places the numbers along the vertical axis of the plot, with numbers that are close together in magnitude next to one another on the plot. For example, a stem-and-leaf plot for the 30 sample prices looks like the following:

```
1 | 6
2 | 0011344
2 | 5666777899
3 | 002223444
3 | 666
```

In this plot, the decimal point is two digits to the right of the stem. So, the “1” in the stem and the "6" in the leaf represents 160, or, because of rounding, any number between 155 and 164.9. In particular, it represents the lowest price in the dataset of 155.5 (thousand dollars). The next part of the graph shows two prices between 195 and 204.9, two prices between 205 and 214.9, one price between 225 and 234.9, two prices between 235 and 244.9, and so on. A stem-and-leaf plot can easily be constructed by hand for small datasets such as this, or it can be constructed automatically using statistical software. The appearance of the plot can depend on the type of statistical software used.

The overall impression from this graph is that the sample prices range from the mid-150s to the mid-350s, with some suggestion of clustering around the high 200s. Perhaps the sample represents quite a range of moderately priced homes, but with no very cheap or very expensive homes. This type of observation often arises throughout a data analysis—the data begin to tell a story and suggest possible explanations. A good analysis is usually not the end of the story since it will frequently lead to other analyses and investigations. For example, in this case, we might surmise that we would probably be unlikely to find a home priced at much less than $150,000 in this market, but perhaps a realtor might know of a nearby market with more affordable housing.

A few modifications to a stem-and-leaf plot produce a histogram—the value axis is now horizontal rather than vertical, and the counts of observations within adjoining data intervals (called “bins”) are displayed in bars (with the counts, or frequency, shown on the vertical axis) rather than by displaying individual values with digits. The following shows a histogram for the home prices data generated by statistical software.



Histograms can convey very different impressions depending on the bin width, start point, and so on. Ideally, we want a large enough bin size to avoid excessive sampling “noise” (a histogram with many bins that looks very wiggly), but not so large that it is hard to see the underlying distribution (a histogram with few bins that looks too blocky). A reasonable pragmatic approach is to use the default settings in whichever software package we are using, and then perhaps to create a few more histograms with different settings to check that we’re not missing anything. There are more sophisticated methods, but for the purposes of the methods in this course, this should suffice.

In addition to graphical summaries such as the stem-and-leaf plot and histogram, sample statistics can summarize data numerically. For example:

- The sample mean, m_Y, is a measure of the "central tendency" of the data Y-values. [More traditional notation for the sample mean of Y uses $\bar{y}$ ("y-bar").]
- The sample standard deviation, s_Y, is a measure of the spread or variation in the data Y-values.

We won't bother here with the formulas for these sample statistics. Since almost all of the calculations necessary for learning the material covered by this course will be performed by statistical software, the course only contains formulas when they are helpful in understanding a particular concept or provide additional insight.

We can calculate sample standardized Z-values from the data Y-values:

$$Z = (Y - m_Y) / s_Y$$

Sometimes, it is useful to work with sample standardized Z-values rather than the original data Y-values since sample standardized Z-values have a sample mean of 0 and a sample standard deviation of 1.

Statistical software can also calculate additional sample statistics, such as:

- the median (another measure of central tendency, but which is less sensitive than the sample mean to very small or very large values in the data)—half the dataset values are smaller than this quantity and half are larger;
- the minimum and maximum;
- percentiles or quantiles such as the 25th percentile—this is the smallest value that is larger than 25% of the values in the dataset (i.e., 25% of the dataset values are smaller than the 25th percentile, while 75% of the dataset values are larger).

Here are the values obtained by statistical software for the home prices example:

Sample size, n	Valid	30
	Missing	0
Mean		278.6033
Median		278.9500
Std. Deviation		53.8656
Minimum		155.5000
Maximum		359.9000
Percentiles	25	241.3750
	50	278.9500
	75	325.8750

‹ Lesson 1: Statistical Inference Foundations (/stat462/node/78) up (/stat462/node/78) 1.2 - Population Distributions › (/stat462/node/249)

STAT 462

Applied Regression Analysis

1.2 - Population Distributions

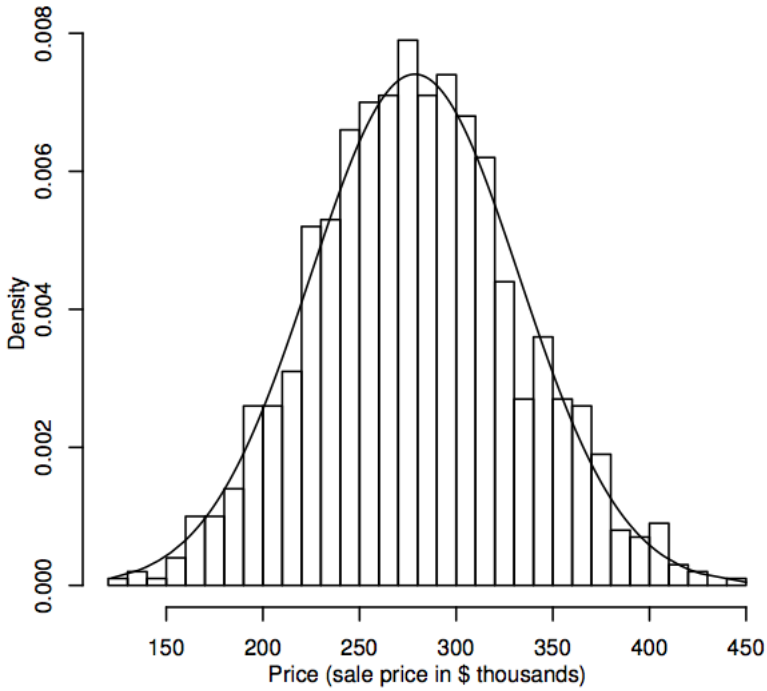
While the methods of the preceding section are useful for describing and displaying sample data, the real power of statistics is revealed when we use samples to give us information about populations. In this context, a population is the entire collection of objects of interest, for example, the sale prices for all single-family homes in the housing market represented by our dataset. We'd like to know more about this population to help us make a decision about which home to buy, but the only data we have is a random sample of 30 sale prices.

Nevertheless, we can employ "statistical thinking" to draw inferences about the population of interest by analyzing the sample data. In particular, we use the notion of a model—a mathematical abstraction of the real world—which we fit to the sample data. If this model provides a reasonable fit to the data, that is, if it can approximate the manner in which the data vary, then we assume that it can also approximate the behavior of the population. The model then provides the basis for making decisions about the population, by, for example, identifying patterns, explaining variation, and predicting future values. Of course, this process can work only if the sample data can be considered representative of the population.

Sometimes, even when we know that a sample has not been selected randomly, we can still model it. Then, we may not be able to formally infer about a population from the sample, but we can still model the underlying structure of the sample. One example would be a convenience sample—a sample selected more for reasons of convenience than for its statistical properties. When modeling such samples, any results should be reported with a caution about restricting any conclusions to objects similar to those in the sample. Another kind of example is when the sample comprises the whole population. For example, we could model data for all 50 states of the United States of America to better understand any patterns or systematic associations among the states.

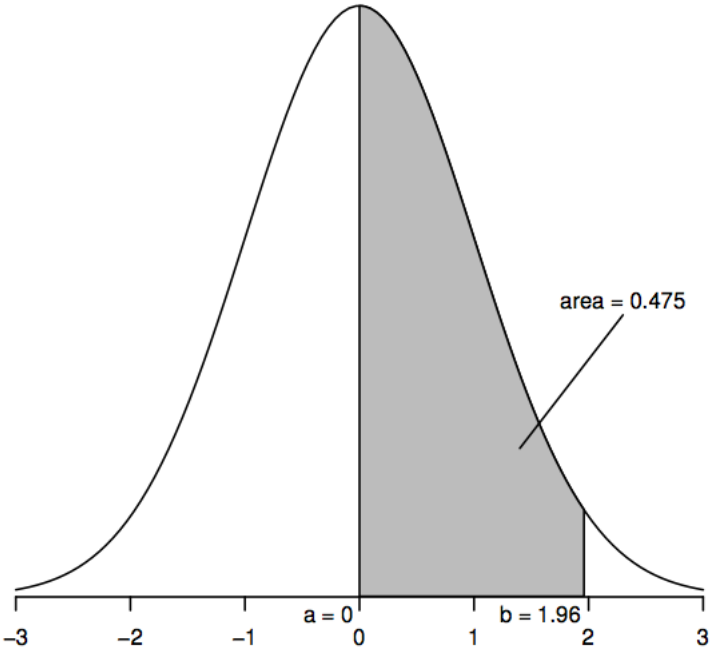
Since the real world can be extremely complicated (in the way that data values vary or interact together), models are useful because they simplify problems so that we can better understand them (and then make more effective decisions). On the one hand, we therefore need models to be simple enough that we can easily use them to make decisions, but on the other hand, we need models that are flexible enough to provide good approximations to complex situations. Fortunately, many statistical models have been developed over the years that provide an effective balance between these two criteria. One such model, which provides a good starting point for the more complicated models we consider later, is the normal distribution.

From a statistical perspective, a probability distribution is a theoretical model that describes how a random variable varies. For our purposes, a random variable represents the data values of interest in the population, for example, the sale prices of all single-family homes in our housing market. One way to represent the population distribution of data values is in a histogram, as described in Section 1.1. The difference now is that the histogram displays the whole population rather than just the sample. Since the population is so much larger than the sample, the bins of the histogram (the consecutive ranges of the data that comprise the horizontal intervals for the bars) can be much smaller, for example, the following shows a histogram for a simulated population of 1,000 sale prices.



As the population size gets larger, we can imagine the histogram bars getting thinner and more numerous, until the histogram resembles a smooth curve rather than a series of steps. This smooth curve is called a density curve and can be thought of as the theoretical version of the population histogram. Density curves also provide a way to visualize probability distributions such as the normal distribution. A normal density curve is superimposed on the histogram above. The simulated population histogram follows the curve quite closely, which suggests that this simulated population distribution is quite close to normal.

To see how a theoretical distribution can prove useful for making statistical inferences about populations such as that in our home prices example, we need to look more closely at the normal distribution. To begin, we consider a particular version of the normal distribution, the standard normal, as represented by the following density curve.



Random variables that follow a standard normal distribution have a mean of 0 (so the curve is symmetric about 0, which is under the highest point of the curve) and a standard deviation of 1 (so the curve has a point of inflection—

where the curve bends first one way and then the other—at +1 and −1). The normal density curve is sometimes called the “bell curve” since its shape resembles that of a bell.

The key feature of the normal density curve that allows us to make statistical inferences is that areas under the curve represent probabilities. The entire area under the curve is one, while the area under the curve between one point on the horizontal axis (a, say) and another point (b, say) represents the probability that a random variable that follows a standard normal distribution is between a and b. So, for example, the figure above shows that the probability is 0.475 that a standard normal random variable lies between a=0 and b=1.96, since the area under the curve between a=0 and b=1.96 is 0.475.

We can obtain values for these areas or probabilities from a variety of sources: tables of numbers, calculators, spreadsheet or statistical software, websites, and so on. Below we print only a few select values since most of the later calculations use a generalization of the normal distribution called the “t-distribution.” Also, rather than areas such as that shaded in the figure above, it will become more useful to consider “tail areas” (e.g., to the right of point b), and so for consistency with later tables of numbers, the following table allows calculation of such tail areas:

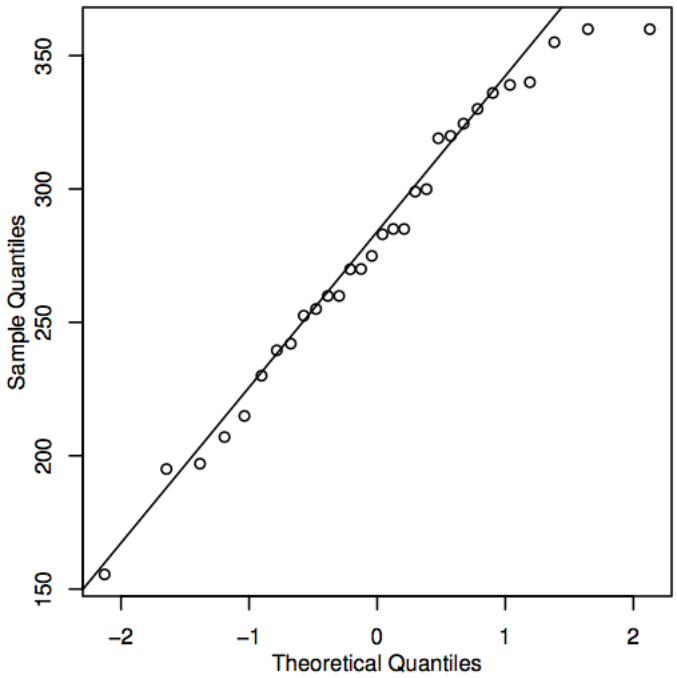
Upper-tail area	0.1	0.05	0.025	0.01	0.005	0.001
Horizontal axis value	1.282	1.645	1.960	2.326	2.576	3.090
Two-tail area	0.2	0.1	0.05	0.02	0.01	0.002

In particular, the upper-tail area to the right of 1.96 is 0.025; this is equivalent to saying that the area between 0 and 1.96 is 0.475 (since the entire area under the curve is 1 and the area to the right of 0 is 0.5). Similarly, the two-tail area, which is the sum of the areas to the right of 1.96 and to the left of −1.96, is two times 0.025, or 0.05.

How does all this help us to make statistical inferences about populations such as that in our home prices example? The essential idea is that we fit a normal distribution model to our sample data and then use this model to make inferences about the corresponding population. For example, we can use probability calculations for a normal distribution (as shown in the figure above) to make probability statements about a population modeled using that normal distribution—we’ll show exactly how to do this in Section 1.3. Before we do that, however, we pause to consider an aspect of this inferential sequence that can make or break the process. Does the model provide a close enough approximation to the pattern of sample values that we can be confident the model adequately represents the population values? The better the approximation, the more reliable our inferential statements will be.

We saw earlier how a density curve can be thought of as a histogram with a very large sample size. So one way to assess whether our population follows a normal distribution model is to construct a histogram from our sample data and visually determine whether it "looks normal," that is, approximately symmetric and bell-shaped. This is a somewhat subjective decision, but with experience you should find that it becomes easier to discern clearly nonnormal histograms from those that are reasonably normal. For example, while the histogram above clearly looks like a normal density curve, the normality of the histogram of 30 sample sale prices in Section 1.1 is less certain. A reasonable conclusion in this case would be that while this sample histogram isn’t perfectly symmetric and bell-shaped, it is close enough that the corresponding (hypothetical) population histogram could well be normal.

An alternative way to assess normality is to construct a QQ-plot (quantile–quantile plot), also known as a normal probability plot, as shown here for the home prices data:



If the points in the QQ-plot lie close to the diagonal line, then the corresponding population values could well be normal. If the points generally lie far from the line, then normality is in question. Again, this is a somewhat subjective decision that becomes easier to make with experience. In this case, given the fairly small sample size, the points are probably close enough to the line that it is reasonable to conclude that the population values could be normal.

There are also a variety of quantitative methods for asessing normality—see Section 6.3.

‹ 1.1 - Identifying and Summarizing Data (/stat462/node/248) up (/stat462/node/78) 1.3 - Selecting Individuals at Random › (/stat462/node/250)

STAT 462

Applied Regression Analysis

1.3 - Selecting Individuals at Random

Having assessed the normality of our population of sale prices by looking at the histogram and QQ-plot of sample sale prices, we now return to the task of making probability statements about the population. The crucial question at this point is whether the sample data are representative of the population for which we wish to make statistical inferences. We can then make reliable statistical inferences about the population by considering properties of a model fit to the sample data—provided the model fits reasonably well.

We saw in Section 1.2 that a normal distribution model fits the home prices example reasonably well. However, a standard normal distribution is inappropriate here, because a standard normal distribution has a mean of 0 and a standard deviation of 1, whereas our sample data have a mean of 278.6033 and a standard deviation of 53.8656. We therefore need to consider more general normal distributions with a mean that can take any value and a standard deviation that can take any positive value (standard deviations cannot be negative).

Let Y represent the population values (sale prices in our example) and suppose that Y is normally distributed with mean (or expected value), E(Y), and standard deviation, SD(Y). [More traditional notation uses Greek letters, μ and σ, for these quantities.] We can abbreviate this normal distribution as Normal(E(Y), $SD(Y)^2$), where the first number is the mean and the second number is the square of the standard deviation (also known as the variance). Then the population standardized Z-value, $Z = (Y - E(Y)) / SD(Y)$, has a standard normal distribution with mean 0 and standard deviation 1. In symbols, $Y \sim \text{Normal}(E(Y), SD(Y)^2) \iff Z = (Y - E(Y)) / SD(Y) \sim \text{Normal}(0,1^2)$.

We are now ready to make a probability statement for the home prices example. Suppose that we would consider a home as being too expensive to buy if its sale price is higher than $380,000. What is the probability of finding such an expensive home in our housing market? In other words, if we were to randomly select one home from the population of all homes, what is the probability that it has a sale price higher than $380,000? To answer this question we need to make a number of assumptions. We've already decided that it is probably safe to assume that the population of sale prices (Price) could be normal, but we don't know the mean, E(Price), or the standard deviation, SD(Price), of the population of home prices. For now, let's assume that E(Price)=280 and SD(Price)=50 (fairly close to the sample mean of 278.6033 and sample standard deviation of 53.8656). (We'll be able to relax these assumptions later in this lesson.) From the theoretical result above, $Z = (\text{Price} - 280) / 50$ has a standard normal distribution with mean 0 and standard deviation 1.

Next, to find the probability that a randomly selected Price is greater than 380, we perform some standard algebra on probability statements. In particular, if we write "the probability that a is bigger than b" as "Pr(a > b)," then we can make changes to a (such as adding, subtracting, multiplying, and dividing other quantities) as long as we do the same thing to b. It is perhaps easier to see how this works by example:

$$\Pr(\text{Price} > 380) = \Pr((\text{Price} - 280) / 50 > (380 - 280) / 50) = \Pr(Z > 2.00).$$

The second equality follows since (Price − 280) / 50 is defined to be Z, which is a standard normal random variable with mean 0 and standard deviation 1. From the table in Section 1.2, the probability that a standard normal random variable is greater than 1.96 is 0.025. Thus, Pr(Z > 2.00) is slightly less than 0.025 (draw a picture of a normal density curve with 1.96 and 2.00 marked on the horizontal axis to convince yourself of this fact). In other words, there is slightly less than a 2.5% chance of finding an expensive home (> $380,000) in our housing market, under the assumption that Price ~ Normal(280, 50^2).

We can also turn these calculations around. For example, which value of Price has a probability of 0.025 to the right of it? To answer this, consider the following calculation (based on the fact that the probability a standard normal random variable is greater than 1.96 is 0.025):

$$\Pr(Z > 1.96) = \Pr((\text{Price} - 280) / 50 > 1.96) = \Pr(\text{Price} > 1.96(50) + 280) = \Pr(\text{Price} > 378).$$

So, the value 378 has a probability of 0.025 to the right of it. Another way of expressing

this is that "the 97.5th percentile of the variable Price is $378,000."

‹ 1.2 - Population Distributions (/stat462/node/249) up (/stat462/node/78) 1.4 - Random Sampling › (/stat462/node/251)

STAT 462

Applied Regression Analysis

1.4 - Random Sampling

In the preceding section we had to make some pretty restrictive assumptions (normality, known mean, known variance) in order to make statistical inferences. We now explore the connection between samples and populations a little more closely so that we can draw conclusions using fewer assumptions.

Recall that the population is the entire collection of objects under consideration, while the sample is a (random) subset of the population. We are particularly interested in making statistical inferences not only about values in the population, denoted Y, but also about numerical summary measures such as the population mean, denoted E(Y)—these population summary measures are called *parameters*. While population parameters are unknown (in the sense that we do not have all the individual population values and so cannot calculate them), we can calculate similar quantities in the sample, such as the sample mean—these sample summary measures are called *statistics*.

Next we'll see how statistical inference essentially involves estimating population parameters (and assessing the precision of those estimates) using sample statistics. When our sample data is a subset of the population that has been selected randomly, statistics calculated from the sample can tell us a great deal about corresponding population parameters. For example, a sample mean tends to be a good estimate of the population mean, in the following sense. If we were to take random samples over and over again, each time calculating a sample mean, then the mean of all these sample means would be equal to the population mean. Such an estimate is called *unbiased* since on average it estimates the correct value. It is not actually necessary to take random samples over and over again to show this—probability theory (beyond the scope of this book) allows us to prove such theorems.

However, it is not enough to just have sample statistics (such as the sample mean) that average out (over a large number of hypothetical samples) to the correct target (i.e., the population mean). We would also like sample statistics that would have "low" variability from one hypothetical sample to another. At the very least we need to be able to quantify this variability, known as sampling uncertainty. One way to do this is to consider the sampling distribution of a statistic, that is, the distribution of values of a statistic under repeated (hypothetical) samples. Again, we can use results from probability theory to tell us what these sampling distributions are. So, all we need to do is take a single random sample, calculate a statistic, and we'll know the theoretical sampling distribution of that statistic (i.e., we'll know what the statistic should average out to over repeated samples, and how much the statistic should vary over repeated samples).

Central limit theorem—normal version

Suppose that a random sample of n data values, represented by $Y_1, Y_2, ..., Y_n$, comes from a population that has a mean of E(Y) and a standard deviation of SD(Y). The sample mean, m_Y, is a pretty good estimate of the population mean, E(Y). The sampling distribution of this statistic derives from the central limit theorem. This theorem states that under very general conditions, the sample mean has an approximate normal distribution with mean E(Y) and standard deviation SD(Y)/ $\sqrt{n}$ (under repeated sampling). In other words, if we were to take a large number of

random samples of n data values and calculate the mean for each sample, the distribution of these sample means would be a normal distribution with mean E(Y) and standard deviation SD(Y)/ √n. Since the mean of this sampling distribution is E(Y), m_Y is an unbiased estimate of E(Y).

An amazing fact about the central limit theorem is that there is no need for the population itself to be normal (remember that we had to assume this for the calculations in Section 1.3). However, the more symmetric the distribution of the population, the better is the normal approximation for the sampling distribution of the sample mean. Also, the approximation tends to be better the larger the sample size n.

The central limit theorem by itself won't help us to draw statistical inferences about the population without still having to make some restrictive assumptions. However, it is certainly a step in the right direction. Consider the home prices example again. As in Section 1.3, we'll assume that E(Price)=280 and SD(Price)=50, but now we no longer need to assume that the population is normal. Imagine taking a large number of random samples of size 30 from this population and calculating the mean sale price for each sample. To get a better handle on the sampling distribution of these mean sale prices, we'll find the 90th percentile of this sampling distribution. Let's do the calculation first, and then see why this might be a useful number to know.

First, we need to get some notation straight. In this section we're not thinking about the specific sample mean we got for our actual sample of 30 sale prices, $m_Y = 278.6033$. Rather we're imagining a list of potential sample means from a population distribution with mean 280 and standard deviation 50—we'll call a potential sample mean in this list M_Y. From the central limit theorem, the sampling distribution of M_Y is normal with mean 280 and standard deviation $50 / \sqrt{30} = 9.129$. Then the standardized Z-value from M_Y,

$$Z = (M_Y - E(Y))/ SD(Y)/\sqrt{n} = (M_Y - 280) / 9.129,$$

is standard normal with mean 0 and standard deviation 1. From the table in Section 1.1, the 90th percentile of a standard normal random variable is 1.282 (since the horizontal axis value of 1.282 corresponds to an upper-tail area of 0.1). Then

$$Pr(Z < 1.282) = Pr((M_Y - 280) / 9.129 < 1.282) = Pr(M_Y < 1.282(9.129) + 280) = Pr(M_Y < 291.703).$$

Thus, the 90th percentile of the sampling distribution of M_Y is \$292,000. In other words, under repeated sampling, M_Y has a distribution with an area of 0.90 to the left of \$292,000 (and an area of 0.10 to the right of \$292,000). This illustrates a crucial distinction between the distribution of population Y-values and the sampling distribution of M_Y—the latter is much less spread out. For example, suppose for the sake of argument that the population distribution of Y is normal (although this is not actually required for the central limit theorem to work). Then we can do a similar calculation to the one above to find the 90th percentile of this distribution (normal with mean 280 and standard deviation 50). In particular,

$$Pr(Z < 1.282) = Pr(Y - 280) / 50 < 1.282) = Pr(Y < 1.282(50) + 280) = Pr(Y < 344.100).$$

Thus, the 90th percentile of the population distribution of Y is \$344,000. This is much larger than the value we got above for the 90th percentile of the sampling distribution of M_Y (\$292,000). This is because the sampling distribution of M_Y is less spread out than the population distribution of Y—the standard deviations for our example are 9.129 for the former and 50 for the latter.

We can again turn these calculations around. For example, what is the probability that M_Y is greater than 291.703? To answer this, consider the following calculation:

$$Pr(M_Y > 291.703) = Pr((M_Y - 280) / 9.129 > (291.703 - 280) / 9.129 = Pr(Z > 1.282) = 0.10.$$

So, the probability that M_Y is greater than 291.703 is 0.10.

Central limit theorem—t-version

One major drawback to the normal version of the central limit theorem is that to use it we have to assume that we know the value of the population standard deviation, SD(Y). A generalization of the standard normal distribution called *Student's t-distribution* solves this problem. The density curve for a t-distribution looks very similar to a normal density curve, but the tails tend to be a little "thicker," so t-distributions are a little more spread out than the normal distribution. This "extra variability" is controlled by an integer number called the degrees of freedom. The smaller this number, the more spread out the t-distribution density curve (conversely, the higher the degrees of freedom, the more like a normal density curve it looks).

For example, the following table shows tail areas for a t-distribution with 29 degrees of freedom:

Upper-tail area	0.1	0.05	0.025	0.01	0.005	0.001
Critical value of t_{29}	1.311	1.699	2.045	2.462	2.756	3.396
Two-tail area	0.2	0.1	0.05	0.02	0.01	0.002

Compared with the corresponding table for the normal distribution in Section 1.2, the critical values (i.e., horizontal axis values or percentiles) are slightly larger in this table.

We will use the t-distribution from this point on because it will allow us to use an estimate of the population standard deviation (rather than having to assume this value). A reasonable estimate to use is the sample standard deviation, s_Y. Since we will be using an estimate of the population standard deviation, we will be a little less certain about our probability calculations—this is why the t-distribution needs to be a little more spread out than the normal distribution, to adjust for this extra uncertainty. This extra uncertainty will be of particular concern when we're not too sure if our sample standard deviation is a good estimate of the population standard deviation (i.e., in small samples). So, it makes sense that the degrees of freedom is lower for smaller sample sizes. In this particular application, we will use the t-distribution with $n-1$ degrees of freedom in place of a standard normal distribution in the following t-version of the central limit theorem.

Suppose that a random sample of n data values, represented by $Y_1, Y_2, \ldots, Y_n$, comes from a population that has a mean of E(Y). Imagine taking a large number of random samples of n data values and calculating the mean and standard deviation for each sample. As before, we'll let M_Y represent the imagined list of repeated sample means, and similarly, we'll let S_Y represent the imagined list of repeated sample standard deviations. Define $t = (M_Y - E(Y)) / (SY/\sqrt{n})$.

Under very general conditions, t has an approximate t-distribution with $n-1$ degrees of freedom. The two differences from the normal version of the central limit theorem that we used before are that the repeated sample standard deviations, S_Y, replace an assumed population standard deviation, SD(Y), and that the resulting sampling distribution is a t-distribution (not a normal distribution).

So far, we have focused on the sampling distribution of sample means, M_Y, but what we would really like to do is infer what the observed sample mean, m_Y, tells us about the population mean, E(Y). Thus, while the calculations in this section have been useful for building up intuition about sampling distributions and manipulating probability statements, their main purpose has been to prepare the ground for the next two sections, which cover how to make statistical inferences about the population mean, E(Y).

‹ 1.3 - Selecting Individuals at Random (/stat462/node/250) up (/stat462/node/78) 1.5 - Interval Estimation › (/stat462/node/252)

STAT 462

Applied Regression Analysis

1.5 - Interval Estimation

We have already seen that the sample mean, m_Y, is a good point estimate of the population mean, E(Y) (in the sense that it is unbiased). It is also helpful to know how reliable this estimate is, that is, how much sampling uncertainty is associated with it.

A useful way to express this uncertainty is to calculate an interval estimate or confidence interval for the population mean, E(Y). The interval should be centered at the point estimate (in this case, m_Y) since we are probably equally uncertain that the population mean could be lower or higher than this estimate (i.e., it should have the same amount of uncertainty either side of the point estimate). In other words, the confidence interval is of the form "point estimate ± uncertainty" or "(point estimate − uncertainty, point estimate + uncertainty)."

We can obtain the exact form of the confidence interval from the t-version of the central limit theorem, where $t = (M_Y - E(Y)) / (SY/\sqrt{n})$ has an approximate t-distribution with n−1 degrees of freedom. In particular, suppose that we want to calculate a 95% confidence interval for the population mean, E(Y), for the home prices example—in other words, an interval such that there will be an area of 0.95 between the two endpoints of the interval (and an area of 0.025 to the left of the interval in the lower tail, and an area of 0.025 to the right of the interval in the upper tail). Let's consider just one side of the interval first. Using the fact that 2.045 is the 97.5th percentile of the t-distribution with 29 degrees of freedom (see the table in Section 1.4), then

$$\Pr(t_{29} < 2.045) = \Pr(M_Y - E(Y)) / (SY/\sqrt{n}) < 2.045) = \Pr(M_Y - 2.045(S_Y/\sqrt{n}) < E(Y)).$$

This probability statement must be true for all potential values of M_Y and S_Y. In particular, it must be true for our observed sample statistics, $m_Y = 278.6033$ and $s_Y = 53.8656$. Thus, to find the values of E(Y) that satisfy the probability statement, we plug in our sample statistics to find

$$M_Y - 2.045(S_Y/\sqrt{n}) = 278.6033 - 2.045(53.8656/\sqrt{30}) = 258.492.$$

This shows that a population mean greater than \$258,492 would satisfy the expression $\Pr(t_{29} < 2.045) = 0.975$. In other words, we have found that the lower bound of our confidence interval is \$258,492, or approximately \$258,000.

To find the upper bound we perform a similar calculation to find that a population mean less than \$298,715 would satisfy the expression $\Pr(t_{29} < 2.045) = 0.975$. In other words, we have found that the upper bound of our confidence interval is \$298,715, or approximately \$299,000.

We can combine these two calculations as

$$\Pr(-2.045 < t_{29} < 2.045) = \Pr(-2.045 < M_Y - E(Y)) / (SY/\sqrt{n}) < 2.045)$$

$$= \Pr(M_Y - 2.045(S_Y/\sqrt{n}) < E(Y) < M_Y + 2.045(S_Y/\sqrt{n})).$$

As before, we plug in our sample statistics to find the values of E(Y) that satisfy this expression:

$$\Pr(278.6033 - 2.045(53.8656/\sqrt{30}) < E(Y) < 278.6033 + 2.045(53.8656/\sqrt{30})$$

$$= \Pr(258.492 < E(Y) < 298.715).$$

This shows that a population mean between $258,492 and $298,715 would satisfy the expression Pr $(-2.045 < t_{29} < 2.045) = 0.95$. In other words, we have found that a 95% confidence interval for E(Y) for this example is ($258,492, $298,715), or approximately ($258,000, $299,000).

More generally, using symbols, a 95% confidence interval for a univariate population mean, E(Y), results from the following:

$$\Pr(-97.5\text{th percentile} < t_{n-1} < 97.5\text{th percentile})$$

$$= \Pr(-97.5\text{th percentile} < M_Y - E(Y)) / (SY/\sqrt{n}) < 97.5\text{th percentile})$$

$$= \Pr(M_Y - 97.5\text{th percentile}(S_Y/\sqrt{n}) < E(Y) < M_Y + 97.5\text{th percentile}(S_Y/\sqrt{n}))$$

where the 97.5th percentile comes from the t-distribution with n−1 degrees of freedom. In other words, plugging in our observed sample statistics, m_Y and s_Y, we can write the 95% confidence interval as $m_Y \pm 97.5\text{th percentile } (s_Y / \sqrt{n})$.

For a lower or higher level of confidence than 95%, the percentile used in the calculation must be changed as appropriate. For example, for a 90% interval (i.e., with 5% in each tail), the 95th percentile would be needed, whereas for a 99% interval (i.e., with 0.5% in each tail), the 99.5th percentile would be needed. These percentiles are easily obtained using statistical software.

Confidence interval for a univariate mean, E(Y)

Thus, in general, we can write a confidence interval for a univariate mean, E(Y), as $m_Y \pm$ t-percentile $(s_Y /\sqrt{n})$, where the t-percentile comes from a t-distribution with n−1 degrees of freedom. The example above thus becomes

$$m_Y \pm \text{t-percentile } (s_Y /\sqrt{n}) = 278.6033 \pm 2.045\ (53.8656 /\sqrt{30}) = 278.6033 \pm 20.111 = (258.492, 298.715).$$

To interpret the confidence interval, loosely speaking we can say that "we're 95% confident that the mean single-family home sale price in this housing market is between $258,000 and $299,000." To provide a more precise interpretation we have to revisit the notion of hypothetical repeated samples. If we were to take a large number of random samples of size 30 from our population of sale prices and calculate a 95% confidence interval for each, then 95% of those confidence intervals would contain the (unknown) population mean. We do not know (nor will we ever know) whether the 95% confidence interval for our particular sample contains the population mean—thus, strictly speaking, we cannot say "the probability that the population mean is in our interval is 0.95." All we know is that the procedure that we have used to calculate the 95% confidence interval tends to produce intervals that under repeated sampling contain the population mean 95% of the time.

Before moving on to Section 1.6, which describes another way to make statistical inferences about population means—hypothesis testing—let us consider whether we can now forget the normal distribution. The calculations in this section are based on the central limit theorem, which does not require the population to be normal. We have also

seen that t-distributions are more useful than normal distributions for calculating confidence intervals. For large samples, it doesn't make much difference (the percentiles for t-distributions get closer to the percentiles for the standard normal distribution as the degrees of freedom get larger), but for smaller samples it can make a large difference. So for this type of calculation we always use a t-distribution from now on. However, we can't completely forget about the normal distribution yet; it will come into play again in a different context in later lessons.

Degrees of freedom

When using a t-distribution, how do we know how many degrees of freedom to use? One way to think about degrees of freedom is in terms of the information provided by the data we are analyzing. Roughly speaking, each data observation provides one degree of freedom (this is where the n in the degrees of freedom formula comes in), but we lose a degree of freedom for each population parameter that we have to estimate. So, in this chapter, when we are estimating the population mean, the degrees of freedom formula is $n-1$. In Lesson 2, when we will be estimating two population parameters (the intercept and the slope of a regression line), the degrees of freedom formula will be $n-2$. For the remainder of the book, the general formula for the degrees of freedom in a multiple linear regression model will be $n-(k+1)$ or $n-k-1$, where k is the number of predictor variables in the model. Note that this general formula actually also works for Chapter 2 (where $k = 1$) and even this chapter (where $k = 0$, since a linear regression model with zero predictors is equivalent to estimating the population mean for a univariate dataset).

‹ 1.4 - Random Sampling (/stat462/node/251) up (/stat462/node/78) 1.6 - Hypothesis Testing › (/stat462/node/253)

STAT 462

Applied Regression Analysis

1.6 - Hypothesis Testing

Another way to make statistical inferences about a population parameter such as the mean is to use hypothesis testing to make decisions about the parameter's value. Suppose that we are interested in a particular value of the mean single-family home sale price, for example, a claim from a realtor that the mean sale price in this market is $255,000. Does the information in our sample support this claim, or does it favor an alternative claim?

The rejection region method

To decide between two competing claims, we can conduct a hypothesis test as follows.

- Express the claim about a specific value for the population parameter of interest as a *null hypothesis*, denoted NH. [More traditional notation uses H0.] The null hypothesis needs to be in the form "parameter = some hypothesized value," for example, NH: E(Y) = 255. A frequently used legal analogy is that the null hypothesis is equivalent to a presumption of innocence in a trial before any evidence has been presented.
- Express the alternative claim as an *alternative hypothesis*, denoted AH. [More traditional notation uses Ha or H1.]. The alternative hypothesis can be in a *lower-tail* form, for example, AH: E(Y) < 255, or an *upper-tail* form, for example, AH: E(Y) > 255, or a *two-tail* form, for example, AH: E(Y) ≠ 255. The alternative hypothesis, also sometimes called the research hypothesis, is what we would like to demonstrate to be the case, and needs to be stated before looking at the data. To continue the legal analogy, the alternative hypothesis is guilt, and we will only reject the null hypothesis (innocence) if we favor the alternative hypothesis (guilt) beyond a reasonable doubt. To illustrate, we will presume for the home prices example that we have some reason to suspect that the mean sale price is higher than claimed by the realtor (perhaps a political organization is campaigning on the issue of high housing costs and has employed us to investigate whether sale prices are "too high" in this housing market). Thus, our upper-tail alternative hypothesis is AH: E(Y) > 255.
- Calculate a *test statistic* based on the assumption that the null hypothesis is true. For hypothesis tests for a univariate population mean the relevant test statistic is

$$\text{t-statistic} = \frac{m_Y - \text{E}(Y)}{s_Y/\sqrt{n}},$$

  where m_Y is the sample mean, E(Y) is the value of the population mean in the null hypothesis, s_Y is the sample standard deviation, and n is the sample size.
- Under the assumption that the null hypothesis is true, this test statistic will have a particular probability distribution. For testing a univariate population mean, this t-statistic has a t-distribution with n−1 degrees of freedom. We would therefore expect it to be "close" to zero (if the null hypothesis is true). Conversely, if it is far from zero, then we might begin to doubt the null hypothesis:
  - For an upper-tail test, a t-statistic that is positive and far from zero would then lead us to favor the alternative hypothesis (a t-statistic that was far from zero but negative would favor neither hypothesis and

the test would be inconclusive).
    - For a lower-tail test, a t-statistic that is negative and far from zero would then lead us to favor the alternative hypothesis (a t-statistic that was far from zero but positive would favor neither hypothesis and the test would be inconclusive).
    - For a two-tail test, any t-statistic that is far from zero (positive or negative) would lead us to favor the alternative hypothesis.
- To decide how far from zero a t-statistic would have to be before we reject the null hypothesis in favor of the alternative, recall the legal analogy. To deliver a guilty verdict (the alternative hypothesis), the jury must establish guilt beyond a reasonable doubt. In other words, a jury rejects the presumption of innocence (the null hypothesis) only if there is compelling evidence of guilt. In statistical terms, compelling evidence of guilt is found only in the tails of the t-distribution density curve. For example, in conducting an upper-tail test, if the t-statistic is way out in the upper tail, then it seems unlikely that the null hypothesis could have been true—so we reject it in favor of the alternative. Otherwise, the t-statistic could well have arisen while the null hypothesis held true—so we do not reject it in favor of the alternative. How far out in the tail does the t-statistic have to be to favor the alternative hypothesis rather than the null? Here we must make a decision about how much evidence we will require before rejecting a null hypothesis. There is always a chance that we might mistakenly reject a null hypothesis when it is actually true (the equivalent of pronouncing an innocent defendant guilty). Often, this chance—called the *significance level*—will be set at 5%, but more stringent tests (such as in clinical trials of new pharmaceutical drugs) might set this at 1%, while less stringent tests (such as in sociological studies) might set this at 10%. For the sake of argument, we use 5% as a default value for hypothesis tests in this course (unless stated otherwise).
- The significance level dictates the *critical value(s)* for the test, beyond which an observed t-statistic leads to rejection of the null hypothesis in favor of the alternative. This region, which leads to rejection of the null hypothesis, is called the rejection region. For example, for a significance level of 5%:
    - For an upper-tail test, the critical value is the 95th percentile of the t-distribution with n−1 degrees of freedom; reject the null in favor of the alternative if the t-statistic is greater than this.
    - For a lower-tail test, the critical value is the 5th percentile of the t-distribution with n−1 degrees of freedom; reject the null in favor of the alternative if the t-statistic is less than this.
    - For a two-tail test, the two critical values are the 2.5th and the 97.5th percentiles of the t-distribution with n−1 degrees of freedom; reject the null in favor of the alternative if the t-statistic is less than the 2.5th percentile or greater than the 97.5th percentile.

It is best to lay out hypothesis tests in a series of steps, so for the house prices example:

- State null hypothesis: NH: E(Y) = 255.
- State alternative hypothesis: AH: E(Y) > 255.
- Calculate test statistic: t-statistic = $m_Y - \mathrm{E}(Y)/(s_Y/\sqrt{n}) = (278.6033 - 255)/(53.8656/\sqrt{30}) = 2.40$.
- Set significance level: 5%.
- Look up critical value: The 95th percentile of the t-distribution with 29 degrees of freedom is 1.699; the rejection region is therefore any t-statistic greater than 1.699.
- Make decision: Since the t-statistic of 2.40 falls in the rejection region, we reject the null hypothesis in favor of the alternative.
- Interpret in the context of the situation: The 30 sample sale prices suggest that a population mean of $255,000 seems implausible—the sample data favor a value greater than this (at a significance level of 5%).

The p-value method

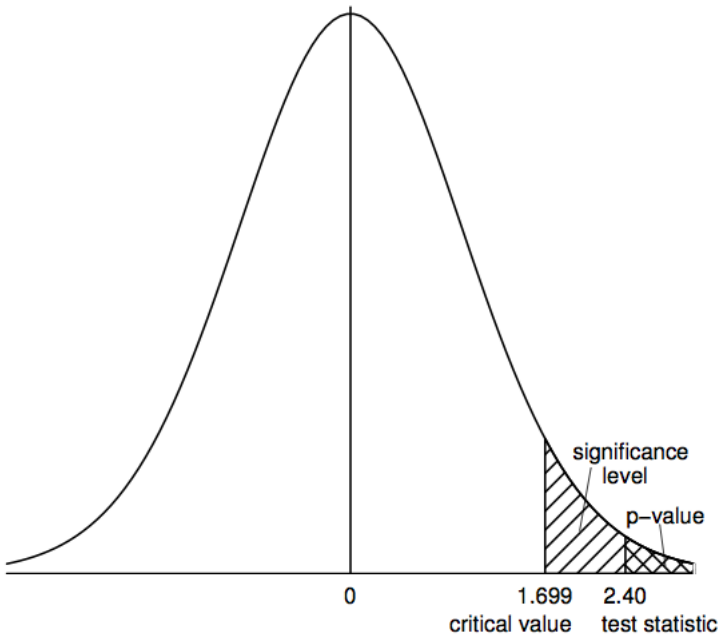
An alternative way to conduct a hypothesis test is to again assume initially that the null hypothesis is true, but then to calculate the probability of observing a t-statistic as extreme as the one observed or even more extreme (in the direction that favors the alternative hypothesis). This is known as the *p-value* (sometimes also called the observed significance level):

- For an upper-tail test, the p-value is the area under the curve of the t-distribution (with n−1 degrees of freedom) to the right of the observed t-statistic.
- For a lower-tail test, the p-value is the area under the curve of the t-distribution (with n−1 degrees of freedom) to the left of the observed t-statistic.
- For a two-tail test, the p-value is the sum of the areas under the curve of the t-distribution (with n−1 degrees of freedom) beyond both the observed t-statistic and the negative of the observed t-statistic.

If the p-value is too "small," then this suggests that it seems unlikely that the null hypothesis could have been true—so we reject it in favor of the alternative. Otherwise, the t-statistic could well have arisen while the null hypothesis held true—so we do not reject it in favor of the alternative. Again, the significance level chosen tells us how small is small: If the p-value is less than the significance level, then reject the null in favor of the alternative; otherwise, do not reject it. For the home prices example:

- State null hypothesis: NH: E(Y) = 255.
- State alternative hypothesis: AH: E(Y) > 255.
- Calculate test statistic: t-statistic = $m_Y - \mathrm{E}(Y)/(s_Y/\sqrt{n}) = (278.6033 - 255)/(53.8656/\sqrt{30}) = 2.40$.
- Set significance level: 5%.
- Look up p-value: The area to the right of the t-statistic (2.40) for the t-distribution with 29 degrees of freedom is less than 0.025 but greater than 0.01 (since the 97.5th percentile of this t-distribution is 2.045 and the 99th percentile is 2.462); thus the upper-tail p-value is between 0.01 and 0.025.
- Make decision: Since the p-value is between 0.01 and 0.025, it must be less than the significance level (0.05), so we reject the null hypothesis in favor of the alternative.
- Interpret in the context of the situation: The 30 sample sale prices suggest that a population mean of $255,000 seems implausible—the sample data favor a value greater than this (at a significance level of 5%).

The following figure shows why the rejection region method and the p-value method will always lead to the same decision (since if the t-statistic is in the rejection region, then the p-value must be smaller than the significance level, and vice versa).

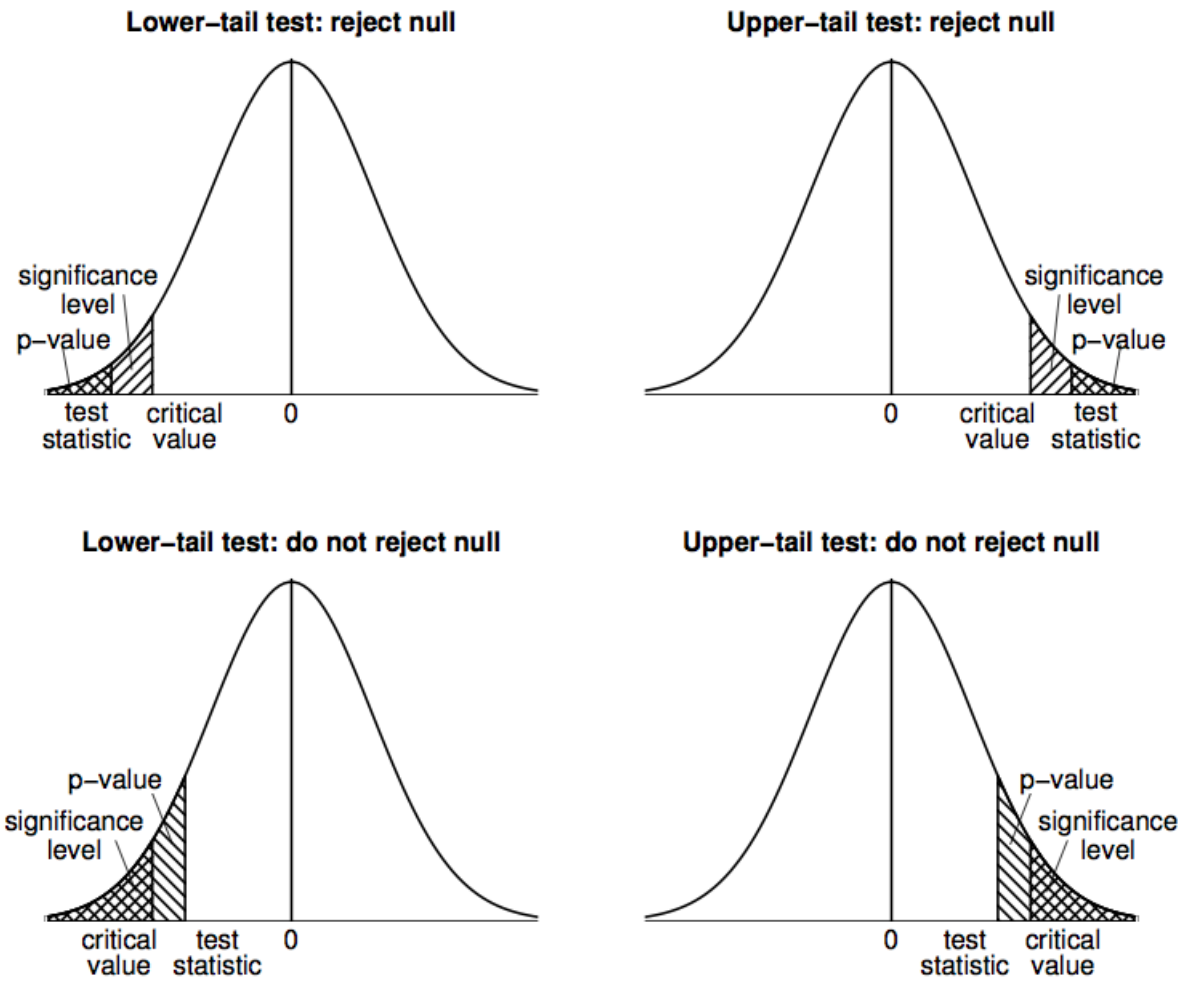


Why do we need two methods if they will always lead to the same decision? Well, when learning about hypothesis tests and becoming comfortable with their logic, many people find the rejection region method a little easier to apply. However, when we start to rely on statistical software for conducting hypothesis tests in later chapters of the

book, we will find the p-value method easier to use. At this stage, when doing hypothesis test calculations by hand, it is helpful to use both the rejection region method and the p-value method to reinforce learning of the general concepts. This also provides a useful way to check our calculations since if we reach a different conclusion with each method we will know that we have made a mistake.

Lower-tail tests

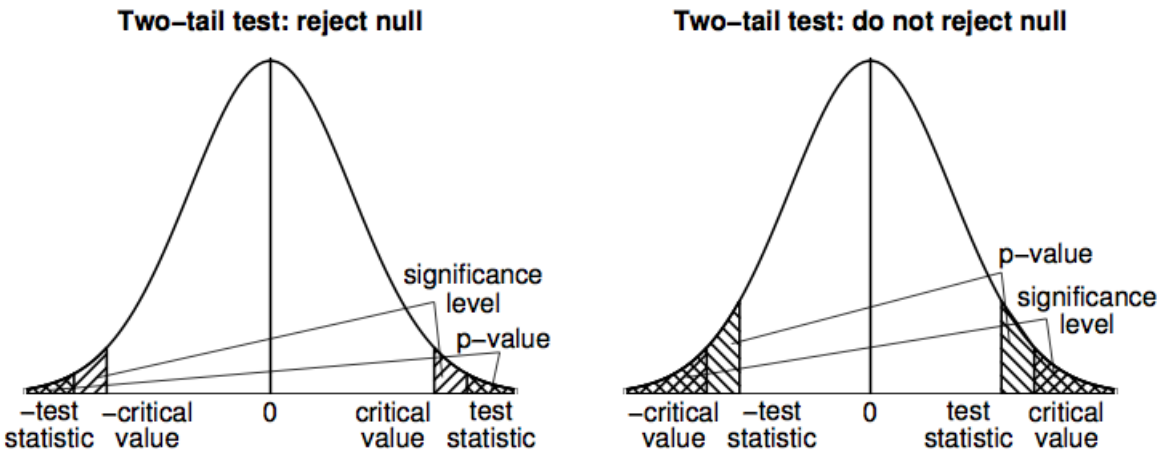
Lower-tail tests work in a similar way to upper-tail tests, but all the calculations are performed in the negative (left-hand) tail of the t-distribution density curve; the following figure illustrates.



A lower-tail test would result in an inconclusive result for the home prices example (since the large, positive t-statistic means that the data favor neither the null hypothesis, NH: $E(Y) = 255$, nor the alternative hypothesis, AH: $E(Y) < 255$).

Two-tail tests

Two-tail tests work similarly, but we have to be careful to work with both tails of the t-distribution; the following figure illustrates.



For the home prices example, we might want to do a two-tail hypothesis test if we had no prior expectation about how large or small sale prices are, but just wanted to see whether or not the realtor's claim of $255,000 was plausible. The steps involved are as follows.

- State null hypothesis: NH: E(Y) = 255.
- State alternative hypothesis: AH: E(Y) ≠ 255.
- Calculate test statistic: t-statistic = $m_Y - \mathrm{E}(Y)/(s_Y/\sqrt{n}) = (278.6033 - 255)/(53.8656/\sqrt{30}) = 2.40.$
- Set significance level: 5%.
- Look up t-table:
  - critical value: The 97.5th percentile of the t-distribution with 29 degrees of freedom is 2.045; the rejection region is therefore any t-statistic greater than 2.045 or less than −2.045 (we need the 97.5th percentile in this case because this is a two-tail test, so we need half the significance level in each tail).
  - p-value: The area to the right of the t-statistic (2.40) for the t-distribution with 29 degrees of freedom is less than 0.025 but greater than 0.01 (since the 97.5th percentile of this t-distribution is 2.045 and the 99th percentile is 2.462); thus the upper-tail area is between 0.01 and 0.025 and the two-tail p-value is twice as big as this, that is, between 0.02 and 0.05.
- Make decision:
  - Since the t-statistic of 2.40 falls in the rejection region, we reject the null hypothesis in favor of the alternative.
  - Since the p-value is between 0.02 and 0.05, it must be less than the significance level (0.05), so we reject the null hypothesis in favor of the alternative.
- Interpret in the context of the situation: The 30 sample sale prices suggest that a population mean of $255,000 seems implausible—the sample data favor a value different from this (at a significance level of 5%).

Hypothesis test errors

When we introduced the significance level above, we saw that the person conducting the hypothesis test gets to choose this value. We now explore this notion a little more fully. Whenever we conduct a hypothesis test, either we reject the null hypothesis in favor of the alternative or we do not reject the null hypothesis. "Not rejecting" a null hypothesis isn't quite the same as "accepting" it. All we can say in such a situation is that we do not have enough evidence to reject the null—recall the legal analogy where defendants are not found "innocent" but rather are found "not guilty." Anyway, regardless of the precise terminology we use, we hope to reject the null when it really is false and to "fail to reject it" when it really is true. Anything else will result in a hypothesis test error. There are two types of error that can occur, as illustrated in the following table:

	Decision	
	Do not reject NH in favor or AH	Reject NH in favor of AH

Reality	NH true	Correct decision	Type 1 error
	NH false	Type 2 error	Correct decision

A type 1 error can occur if we reject the null hypothesis when it is really true—the probability of this happening is precisely the significance level. If we set the significance level lower, then we lessen the chance of a type 1 error occurring. Unfortunately, lowering the significance level increases the chance of a type 2 error occurring—when we fail to reject the null hypothesis but we should have rejected it because it was false. Thus, we need to make a trade-off and set the significance level low enough that type 1 errors have a low chance of happening, but not so low that we greatly increase the chance of a type 2 error happening. The default value of 5% tends to work reasonably well in many applications at balancing both goals. However, other factors also affect the chance of a type 2 error happening for a specific significance level. For example, the chance of a type 2 error tends to decrease the greater the sample size.

‹ 1.5 - Interval Estimation (/stat462/node/252) up (/stat462/node/78) 1.7 - Random Errors and Prediction › (/stat462/node/254)

STAT 462

Applied Regression Analysis

1.7 - Random Errors and Prediction

So far, we have focused on estimating a univariate population mean, E(Y), and quantifying our uncertainty about the estimate via confidence intervals or hypothesis tests. In this section, we consider a different problem, that of "prediction." In particular, rather than estimating the mean of a population of Y-values based on a sample, $Y_1,...,Y_n$, consider predicting an individual Y-value picked at random from the population.

Intuitively, this sounds like a more difficult problem. Imagine that rather than just estimating the mean sale price of single-family homes in the housing market based on our sample of 30 homes, we have to predict the sale price of an individual single-family home that has just come onto the market. Presumably, we'll be less certain about our prediction than we were about our estimate of the population mean (since it seems likely that we could be farther from the truth with our prediction than when we estimated the mean—for example, there is a chance that the new home could be a real bargain or totally overpriced). Statistically speaking, there is "extra uncertainty" that arises with prediction—the population distribution of data values, Y (more relevant to prediction problems), is much more variable than the sampling distribution of sample means, M_Y (more relevant to mean estimation problems).

We can tackle prediction problems with a similar process to that of using a confidence interval to tackle estimating a population mean. In particular, we can calculate a prediction interval of the form "point estimate ± uncertainty" or " (point estimate − uncertainty, point estimate + uncertainty)." The point estimate is the same one that we used for estimating the population mean, that is, the observed sample mean, m_Y. This is because m_Y is an unbiased estimate of the population mean, E(Y), and we assume that the individual Y-value we are predicting is a member of this population. As discussed in the preceding paragraph, however, the "uncertainty" is larger for prediction intervals than for confidence intervals. To see how much larger, we need to return to the notion of a model that we introduced in Section 1.2.

We can express the model we've been using to estimate the population mean, E(Y), as Y-value = deterministic part + random error or $Y_i = E(Y) + e_i$ (i = 1,...,n). In other words, each sample Y_i-value (the index i keeps track of the sample observations) can be decomposed into two pieces, a deterministic part that is the same for all values, and a random error part that varies from observation to observation. A convenient choice for the deterministic part is the population mean, E(Y), since then the random errors have a (population) mean of zero. Since E(Y) is the same for all Y-values, the random errors, e, have the same standard deviation as the Y-values themselves, that is, SD(Y). We can use this decomposition to derive the confidence interval and hypothesis test results of Sections 1.5 and 1.6 (although it would take more mathematics than we really need for our purposes in this course). Moreover, we can also use this decomposition to motivate the precise form of the uncertainty needed for prediction intervals (without having to get into too much mathematical detail).

In particular, write the Y-value to be predicted as Y^*, and decompose this into two pieces as above: $Y^* = E(Y) + e^*$. Then subtract M_Y, which represents potential values of repeated sample means, from both sides of this equation: $Y^* - M_Y = (E(Y) - M_Y) + e^*$, which defines prediction error = estimation error + random error. Thus, whereas in estimating the population mean the only error we have to worry about is estimation error, in predicting an individual Y-value we have to worry about both estimation error and random error.

Prediction interval for an individual Y-value

Recall from Section 1.5 that the form of a confidence interval for the population mean is $m_Y \pm$ t-percentile($s_Y/\sqrt{n}$). The term $s_Y/\sqrt{n}$ in this formula is an estimate of the standard deviation of the sampling distribution of sample means, M_Y, and is called the *standard error of estimation*. The square of this quantity, s_Y^2/n, is the estimated variance of the sampling distribution of sample means, M_Y. Then, thinking of E(Y) as some fixed, unknown constant, s_Y^2/n is also the estimated variance of the estimation error, $E(Y)-M_Y$.

The estimated variance of the random error, e^*, is s_Y^2. It can then be shown that the estimated variance of the prediction error, $Y^* - M_Y$, is $s_Y^2/n + s_Y^2 = s_Y^2(1/n+1) = s_Y^2(1+1/n)$. Then $s_Y\sqrt{(1+1/n)}$ is called the *standard error of prediction* and leads to the formula for a prediction interval for an individual Y-value as $m_Y \pm$ t-percentile($s_Y\sqrt{(1+1/n)}$).

As with confidence intervals for the mean, the t-percentile used in the calculation comes from a t-distribution with n−1 degrees of freedom. For example, for a 95% interval (i.e., with 2.5% in each tail), the 97.5th percentile would be needed, whereas for a 90% interval (i.e., with 5% in each tail), the 95th percentile would be needed. For example, the 95% prediction interval for an individual value of Price picked at random from the population of single-family home sale prices is calculated as

$$m_Y \pm \text{t-percentile}(s_Y\sqrt{(1+1/n)}) = 278.6033 \pm 2.045(53.8656\sqrt{(1+1/30)})$$

$$= 278.6033 \pm 111.976 = (166.627, 390.579).$$

To interpret the prediction interval, loosely speaking we can say that "we're 95% confident that the sale price for an individual home picked at random from all single-family homes in this housing market will be between $167,000 and $391,000." More precisely, if we were to take a large number of random samples of size 30 from our population of sale prices and calculate a 95% prediction interval for each, then 95% of those prediction intervals would contain the (unknown) sale price for an individual home picked at random from the population.

As discussed at the beginning of this section, this interval is much wider than the 95% confidence interval for the population mean single-family home sale price, which was calculated as

$$m_Y \pm \text{t-percentile}(s_Y/\sqrt{n}) = 278.6033 \pm 2.045(53.8656/\sqrt{30})$$

$$= 278.6033 \pm 20.111 = (258.492, 298.715).$$

Unlike for confidence intervals for the population mean, statistical software does not generally provide an automated method to calculate prediction intervals for an individual Y-value. Thus they have to be calculated by hand using the sample statistics, m_Y and s_Y. However, there is a trick that can get around this (although it makes use of simple linear regression, which we cover in Lesson 2). First, create a variable that consists only of the value 1 for all observations. Then, fit a simple linear regression model using this variable as the predictor variable and Y as the response variable, and restrict the model to fit "without an intercept." The estimated regression line for this model

will be a horizontal line at a value equal to the sample mean of the response variable. Prediction intervals for this model will be the same for each value of the predictor variable, and will be the same as a prediction interval for an individual Y-value.

We derived the formula for a confidence interval for a univariate population mean from the t-version of the central limit theorem, which does not require the data Y-values to be normally distributed. However, the formula for a prediction interval for an individual univariate Y-value tends to work better for datasets in which the Y-values are at least approximately normally distributed.

‹ 1.6 - Hypothesis Testing (/stat462/node/253) up (/stat462/node/78)
