Print

Lesson 1: Introduction and Review of Concepts

Introduction

Overview

Suppose you're interested in what causes the spread of a disease, or how likely an applicant is to default on a bank loan. But how do you answer these questions from your data? We're going to start off in Chapter 1 by exploring our data and learning about the SAS tools that can help us answer our questions.

It's often impossible to gather data on an entire population, such as every single person who gets sick or defaults on a loan, so we'll learn how to make inferences from our data samples of those populations. These inferences will help us answer our questions so we can make decisions about future research or business strategies.

We'll begin by briefly discussing the models required to analyze different types of data and the difference between explanatory vs predictive modeling. We'll then move into a review of fundamental statistical concepts, such as the sampling distribution of a mean, hypothesis testing, p-values, and confidence intervals.

After reviewing these concepts, we'll apply one-sample and two-sample t tests to our Ames Housing data to confirm or reject preconceived hypotheses.

Statistical Modeling Overview

Statistical Modeling: Types of Variables

Let's start with a review of statistical modeling. As you might recall, the type of modeling depends on the level of measurement of two types of variables: response and predictors.

Response variables are also known as the outcome, target, or, in designed experiments, dependent variables. They are typically the focus of your business or research, and the variables that you seek to predict. Predictor variables are also known as input, explanatory, or, in designed experiments, independent variables. They are the measures associated with the response variables and therefore can be used to predict the value of the response variables.

In our Ames housing example, square footage, number of garages, and quality of heating system are predictor variables that we'll use to predict the value of the response variable, SalePrice.

Both the response and predictor variables can be either continuous, categorical, or ordinal. A continuous variable can take on any numeric measurement, whereas a categorical variable is associated with specific non-numeric levels. For example, a common categorical variable is status, with the levels on and off. An ordinal variable is similar to a categorical variable, but the levels have a natural hierarchy such as low, medium, and high.

Overview of Models

So let's take a look at the models that we use in this course. We'll study the general linear model, an umbrella term for several different analyses, and also logistic regression. In each case, there's a response variable and predictor variables. When the response variable is continuous and you can assume a normal distribution of errors, you can use a general linear model to model the relationship between predictor variables and the response variable.

The formula is Y, the response, equal to a linear function of the predictors considered, the Xs, and the unknown parameters, the Betas, which we estimate from the data. In this case, we're attempting to predict the response, with k predictors. ε accounts for the unexplained variation in our model.

When will you use ANOVA, or analysis of variance? If you have a response variable that's continuous and all of your predictor variables are categorical, your best approach in terms of a statistical method is going to be ANOVA. With ANOVA, you're looking at how changing the level of your predictors can affect Y, your response variable. For example, what if you wanted to know how sale price relates to the heating quality of a home. This predictor, heating quality, has four levels: excellent, good, average, and fair. Our ANOVA model can explain how the sale price changes from one level of heating quality to another.

If your response is continuous and all of your predictors are continuous, then you're going to use ordinary least squares regression. In the simplest case, imagine you want to predict the sale price of a home using the size of the home in square feet. Ideally, the sampled data exhibit some kind of linear relationship. The regression model indicates what the expected response, or sale price, would be for each value of square feet.

What if you have a categorical response variable that's binary? Well, then regardless of your predictors, logistic regression is going to the be the optimum choice in terms of modeling. In logistic regression, you model the probability of an event given a set of predictors. Here's the formula. The logit of Y is the logit transformation, or the log odds transformation, of the probability of the event. β_0 is the intercept of the equation, and β_1 is the slope parameter for X_1.

So imagine that homes in Iowa selling for more than $175,000 are eligible for a tax incentive, and you want to estimate the probability of this event given the square footage of the home. Here, a value of 0 indicates that the home is not eligible for a tax incentive, and a value of 1 indicates that it is. The logistic regression model predicts the probability of an incentive-eligible home with a sigmoidal curve. You can see that the estimated probability of being eligible increases as the size of the home increases.

Explanatory versus Predictive Modeling

Regardless of the statistical model you use, you need to distinguish between explanatory and predictive modeling. In explanatory models, or inferential statistics, you make conclusions or inferences about a population from the analysis of a random sample drawn from that population. So you generalize from the data you observe to the population that you haven't observed. The goal is to develop a model that answers the question, "How is X related to Y?" That is, how does the outcome change as I change the predictor value? In explanatory modeling, you're concerned with accurately estimating model parameters, and you assess this using p-values and confidence intervals. You typically have small sample sizes and few variables.

Predictive modeling, on the other hand, predicts future values of a response variable based on the existing values of predictor variables. It's focused on making accurate predictions. That is, regardless of the parameter estimates, can I still make good model predictions? You assess the prediction's accuracy using a holdout or validation data set,

and the model usually has many variables and a large sample size.

In this course, we'll focus predominately on explanatory modeling. However, when you're comfortable with creating these models, we'll show you a few techniques to transition into the world of predictive modeling.

Quick Review of Statistical Concepts

Population Parameters and Sample Statistics

In inferential statistics, the focus is on learning about populations. Examples of populations are all people with a certain disease, all drivers with a certain level of insurance, or all customers, both current and potential, at a bank.

Parameters are evaluations of characteristics of populations. They are generally unknown and must be estimated through the use of samples.

A sample is a group of measurements from a population. In order for inferences to be valid, the sample should be representative of the population. A sample statistic is a measurement from our sample. You infer information about population parameters through the use of sample statistics.

A point estimate is a single, best estimate of a population parameter.

Statisticians use Greek letters to represent population parameters (for example, μ, σ, and ρ) and letters from the English alphabet to represent sample statistics (for example, $\bar{x}$, r, and s). You can use $\bar{x}$, the sample mean, to estimate μ, the population mean. Similarly, you can use s, the sample standard deviation, to estimate σ, the population standard deviation.

In this course, our population of interest is all homes in Ames, Iowa. Unfortunately, we can't measure the attributes of each home, so we infer real estate attributes for the region with a sample. We sampled 300 homes sold between 2006 and 2010 and found that the average sale price in our sample is $137,525, the standard deviation is $37,623, and the standard error is $2,172. These values are referred to as point estimates.

Read about It: Parameters and Statistics

Statisticians use Greek letters, for example, μ, σ, and ρ, to represent population parameters, and letters from the English alphabet, for example, $\bar{x}$, $\hat{p}$, r, and s, to represent sample statistics. You can use $\bar{x}$, the sample mean, to estimate μ, the population mean. Similarly, you can use σ, the sample standard deviation, to estimate s, the population standard deviation.

	Sample Statistics	Population Parameters
Mean	$\bar{x}$	μ
Standard Deviation	s	σ
Variance	s^2	σ^2
Correlation	r	ρ

Let's look at these statistics in more detail and see how you can use them to estimate parameters. Suppose you have a sample x_1, x_2, through x_n from some population. You can

calculate the mean for that sample using the formula shown here.

$$\bar{x} = \frac{1}{n}\sum x_i$$

The sample variance, s^2, is a measure of the variability of your sample around the mean. Sample variance gives you a specific measurement indicating how much your data values vary in comparison with the average value. You can calculate sample variance using the formula shown here.

$$s^2 = \frac{1}{n-1}\sum \left(x_i - \bar{x}\right)^2$$

You can use this statistic to estimate the population variance, σ^2. The sample standard deviation, another common measure of variability, is simply the square root of the variance. You can use this statistic to estimate the standard deviation for the population. You calculate the sample standard deviation using the formula shown here.

$$s = \sqrt{\frac{1}{n-1}\sum \left(x_i - \bar{x}\right)^2}$$

Because it's the square root of the variance, the resulting measure of variability will be in the same units as the data, and therefore, the same units as the mean.

For example, suppose that you're interested in knowing the average dollar amount people spend in a store. The unit of measurement is dollars. The data you gather and the mean you calculate from the data will be in dollars. The sample variance will be a measure of the spread in your data in dollars squared. Because the standard deviation is the square root of the variance, it puts the measure of spread back on the original dollar scale.

Normal (Gaussian) Distribution

Because sampling involves variability, parameter estimates also have variability. Often, the variability of sample statistics is approximately normal.

Perhaps the most famous statistical theorem, the central limit theorem, indicates that the sample mean of ANY distribution approximates a normal distribution. That is, if we could sample all subsets of our population of size n, and create a histogram, the sample means will be bell shaped with the population mean and standard error measuring its variability.

For example, imagine our population distribution is bell shaped, with a population mean, μ=50, and standard deviation, σ=5. Suppose we sample n=30 observations and the sample mean is 48.3 Then we take another 30 observation sample, and the sample mean is 52.1. If we repeat this process many times, the distribution of the sample mean will follow a bell shaped normal distribution with population mean μ=50 and standard error of σ divided by the square root of n=30.

Another name for the normal distribution is the Gaussian distribution. The Gaussian distribution is bell-shaped, symmetric, and defined by two parameters, μ (the mean) and σ (the standard deviation). The mean locates the midpoint of the distribution, or the peak of the bell, and the standard deviation describes the variability, or spread of the distribution. The area under the curve between two points is the probability of getting values between those two points.

Some well-known probabilities are associated with the mean and standard deviation of the distribution. For example, approximately 68% of the normal distribution lies within 1 standard deviation of the mean. Approximately 95% lies within 2 standard deviations of the

mean, and approximately 99% lies within 3 standard deviations.

Statisticians often consider values that are more than 2 standard deviations from the mean as unusual. And now you can see why. Only about 5% of all values are that far away from the mean. Depending on the context, some statisticians treat only values more than 3 standard deviations away from the mean as unusual.

Because the normal distribution has many useful mathematical properties, statistical procedures for data based on a random sample often assume the normal distribution. So it's important to know how to check this assumption for your data.

Read about It: Normal Distribution

The formula for a normal distribution of x around a mean, μ, with standard deviation, σ, is

$$f(x, \mu, \sigma) = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}$$

The standard normal curve has μ=0 and σ=1. The area under the curve between any two values can be calculated. In statistics, think about probabilities related to the normal curve. Given the variability around the center (the mean, or point estimate of the parameter), you can think about the probability of sampling a value within some distance, zσ, from the mean. It's the area under the normal probability density curve in an area ranging from -zσ to zσ.

Standard Error of the Mean

Sample statistics are a direct function of the sample collected. This means that if a different sample was collected, the sample statistics would change. However, we collect only one sample due to time and resources.

The variability associated with the sample mean, $\bar{x}$, is measured by the standard error. The standard error of our estimate is the standard deviation of the sample data divided by the square root of the total number of sampled data points. Thus, the larger the sample, the closer we get to measuring all the data of the population, and the smaller the standard error will be. The smaller the standard error, the more precise our estimate, and the more confident we are that the sample mean is a good estimate of the population mean.

Confidence Intervals

Confidence intervals for the mean are interval estimators of the population mean. They take into account the variability of the sample statistic, in this case, the sample mean. The confidence interval shows a range of plausible values for the unknown population mean by reporting an upper and lower bound.

Here's a good way to think about confidence intervals. If you were to draw infinitely many samples and estimate your confidence interval exactly the same way each time, your confidence interval would represent the percentage of those intervals that would contain the true population mean. Of course, nobody draws infinitely many samples. It's rare that you would draw more than one. An important thing to remember is that the confidence interval calculated for a particular sample might or might not contain the value of the true population mean.

You get to choose your desired degree of confidence, or confidence level. A typical confidence level is 95%, meaning that 95% of those theoretically infinite number of intervals would contain the true population mean, and 5% would not.

Here's the formula for calculating the confidence interval for the mean. $\bar{x}$ is the sample mean. t is the t quantile value that is determined by the confidence level and the sample size. $S_{\bar{x}}$ is the standard error of the mean.

To construct a confidence interval for the mean, you must first select a significance level so that the appropriate value of t is used in the formula. Higher confidence levels are associated with larger t values, which, in turn, result in wider intervals. The formula calculates both the upper and lower bounds, equidistant from the sample mean estimate in the middle.

Why not raise the confidence level to 99.9%, so that any confidence interval you calculate contains the true value of the population mean? As you increase the confidence level, the width of the interval increases, making it less informative. In the extreme, a 100% confidence interval for the mean for any sample ranges from negative infinity to positive infinity, which would tell you nothing useful about where the true population mean lies.

Statistical Hypothesis Test

In inferential statistics, you use statistics to infer information about population parameters. But inferences aren't exact. As you've seen, there's variability in parameter estimates. In a hypothesis test, you phrase questions as tests of hypotheses about population parameters. Your null hypothesis is usually one of equality, whereas the alternative hypothesis is one of inequality. The alternative hypothesis is typically what you suspect, or what you're attempting to demonstrate.

The conclusions reached from your hypothesis test are usually phrased in reference to the p-value, or the probability of obtaining a test statistic as extreme or more extreme than the one observed in your data given that the null hypothesis is true. When the p-value is low, it provides doubt about the truth of the null hypothesis. But how low does the p-value need to be before you reject the null hypothesis completely? That depends on you. A common significance level is 0.05 (1 chance in 20). If you require a stricter cutoff, you might consider lowering your significance level when planning your analysis. Let's look at an example.

To begin each game of a worldwide soccer tournament, one of the teams chooses heads or tails, and the match referee tosses a coin. The team that wins the toss decides which goal it will attack, and the team that loses the toss takes the kick-off to start the match. But how do you know the coin is fair? You might suspect that the coin is not fair, but you begin by assuming that it is fair.

Next, you select a significance level: if you observe five heads in a row or five tails in a row, you conclude that the coin is not fair. Otherwise, you decide that there is not enough evidence to show that the coin is biased. With a fair coin, a true null hypothesis, the probability of observing 5 heads or 5 tails in a row in five trials is 1 out of 16.

Why 1 out of 16? There are 5 tosses, and each has a 50% probability of being heads. Tosses are independent, and therefore, the probability of 5 heads is $(1/2)^5$ or 1 out of 32. The probability of 5 tails is also 1 out of 32. These probabilities can be added together to give the probability of 5 heads or 5 tails as 2 out of 32, or 1 out of 16. So the significance level is 1/16, or 0.0625.

To collect evidence, you flip the coin five times and count the number of heads and tails. Finally, you decide either that there is enough evidence to reject the assumption that the coin is fair (either all trials are heads or all trials are tails), or that there is not enough

evidence to reject the assumption that the coin is fair (meaning not all trials are either heads or tails).

So, you performed a hypothesis test and used a decision rule to decide whether the coin was fair or not. But was your decision correct? You began by assuming that the null hypothesis is true: that the coin is fair. But what if you're wrong?

If you reject the null hypothesis when it's actually true, you've made a Type I error. The probability of committing a Type I error is α. α is the significance level of the hypothesis test. In the coin example, it's the probability that you conclude that the coin is not fair when it is fair.

A Type II error, often referred to as β, is when you fail to reject the null hypothesis and it's actually false. In the coin example, it's the probability that you fail to find that the coin is not fair when it is in fact biased. Type I and Type II errors are inversely related. As one type increases, the other decreases.

Power is the probability that you correctly reject the null hypothesis. The power of a statistical test is equal to (1 - β), where again, β is the Type II error rate.

p-Value: Effect Size and Sample Size Influence

If you flip a coin 100 times and observe 50 heads, you wouldn't doubt that the coin is fair. But you might be skeptical if you observe 40 or 60 heads. You'd be more skeptical if you observe 37 or 63 heads, and you'd be highly skeptical if you observe 15 or 85 heads. As the difference between the number of heads and tails increases, you have more evidence that the coin is not fair.

Statisticians refer to the difference between the observed statistic and the hypothesized value as the effect size. The null hypothesis of a fair coin suggests 50% heads and 50% tails. If the coin was actually weighted to give 55% heads, the effect size would be 5%.

A p-value measures the probability of observing a value as extreme as the one observed or more extreme, assuming the null hypothesis is true. Suppose you flip the coin 100 times, and you observe 55 heads and 45 tails. The difference of 10 is associated with a p-value of .3682. p-values this large are often seen in experiments with a fair coin.

As the difference between heads and tails gets larger (for example, 20, 26, or as high as 70), the corresponding p-values get smaller. You would rarely see a small p-value (for example, less than .0001) with a fair coin. In the case of 15 heads and 85 tails, you have evidence that the coin is not fair. So the p-value is used to determine statistical significance. It helps you assess whether you should reject the null hypothesis.

A p-value is not only affected by the effect size (in this case, the observed proportion of heads). It's also affected by the sample size (in this case, the number of coin flips). For a fair coin, you'd expect 50% of the flips to be heads. What if you get 40% heads instead of the 50% you expect? Is it a fair coin? Let's say that you flip the coin 10 times and observe 40%, or 4 heads. What if you flip the coin 400 times and you observe 40% heads?

The evidence becomes stronger and the p-values become smaller as the number of trials increases. As you saw when we talked about confidence intervals, the variability around the mean estimate gets smaller as the sample size gets larger. For larger sample sizes, you can measure means more precisely. Therefore, 40% heads out of 400 flips makes you more confident that this was not just a chance difference, when compared to 40% heads out of only 10 flips.

The smaller p-values reflect this confidence. The p-value here, less than .0001, assesses the probability that this difference from 50% occurred purely by chance. Remember, as you saw earlier, you'd rarely see a p-value less than .0001 with a fair coin.

One-Sample t Tests

Scenario

There are many houses in Ames, Iowa, and you saw in a local news article that the mean home sale price is $135,000, and you'd like to verify this. You can't obtain details on every home, so you decide to estimate the mean sale price using your sample of 300 homes. Earlier, you calculated a sample mean of $137,525 and a standard error of $2,172. Given that there's variability in your sample, is it safe to assume that the population parameter could in fact be $135,000, or is this value unreasonable given the data? You can use a one-sample t test to answer this question.

Performing a t Test

A one-sample t test compares the mean calculated from a sample to a hypothesized mean. You're testing the null hypothesis, μ is equal to μ0, against the alternative hypothesis, μ is not equal to μ0. When you don't know the true population standard deviation, σ, you must estimate it from the sample, and you must use Student's t distribution, rather than the normal distribution, for calculating p-values and confidence limits.

Student's t distribution is similar to the normal distribution, but it has more probability in the tails and is not as peaked as the normal distribution. Student's t distribution approaches the normal distribution as the sample size increases. You calculate the value of Student's t statistic using the equation $t = (\bar{x}-\mu 0)$ all divided by the standard error. Let's calculate Student's t statistic for the Ames housing data to test our null hypothesis.

μ0 is the hypothesized value of $135,000, $\bar{x}$ is the sample mean of SalePrice, $137,525, and $s_{\bar{x}}$ is the standard error of the mean, $2,172.1. The resulting t value is 1.16.

What does this t value tell us? It measures how far the sample mean is from the hypothesized mean, in standard error units. The t value is positive when the sample mean is larger than the hypothesized mean, and negative when the sample mean is less than the hypothesized mean. If your null hypothesis is true, you'd expect $\bar{x}$ to be relatively close to μ0, and the corresponding t value to be close to zero, providing evidence in favor of the null hypothesis. If $\bar{x}$ is far from μ0 and the t value is large, you have evidence against the null hypothesis in favor of the alternative.

Before you can make a decision about your hypothesis, you need to know the probability of observing a test statistic of t=1.16 or more extreme, given that the null hypothesis is true. This probability is the p-value. It quantifies how likely we are to obtain the sample mean. If the p-value is less than α, you reject the null hypothesis in favor of the alternative. On the other hand, if your p-value is greater than α, evidence suggests that your null hypothesized value is statistically reasonable, so you fail to reject the null hypothesis.

Consider the Ames housing example. A p-value less than 0.05, our α, means there's less than a 5% chance you would have a sample mean of $137,525 if your population mean was in fact $135,000.

Let's look at the t distribution. Our alternative hypothesis is an inequality, so this is a two-sided test. The shaded area in the graph is the rejection region. For a two-sided test, the rejection region is contained in both tails, and the area in each tail corresponds to α/2, or in this case, 2.5%. If the t value falls in the shaded region, then you reject the null hypothesis. Otherwise, you fail to reject it. Our t value, 1.16, falls outside the rejection region, so we fail to reject our null hypothesis.

The α and t distribution mentioned here are directly related to those in confidence intervals. In our example, α is 0.05, or 5%. The area outside the confidence interval corresponds to the shaded region in the t distribution, or the rejection region.You can calculate the t value and p-value using t distribution tables, or with PROC TTEST in SAS.

Demo: Performing a One-Sample t Test Using PROC TTEST

Filename: **st101d02.sas**

We'll use the TTEST procedure to determine whether the population mean sale price of homes in Ames, Iowa, is $135,000 given our sample. This procedure performs t tests and computes confidence limits. It can also use ODS Graphics to produce histograms, quantile-quantile plots, box plots, and confidence limit plots. Let's look at program st101d02.sas. The PROC TTEST step analyzes the SalePrice variable. The H0= option specifies our null hypothesis value of 135,000. The INTERVAL option requests confidence interval plots for the means, and the SHOWNULL option displays a vertical reference line at the null value of 135,000.

Let's submit the code, and look at the output.

The first table provides descriptive statistics of our sample, including sample size, mean, standard deviation, standard error, and minimum and maximum values of SalePrice.

The second table provides confidence limits for μ and σ. The default level is 95%, but you can change it with the ALPHA= option in the PROC TTEST statement. Set alpha equal to 1 minus the confidence level.

The last table provides the t test information, including the degrees of freedom, the t value and the p-value, 0.2460. Recall that if the t statistic is close to zero, and the p-value is greater than α, evidence suggests the hypothesized population parameter is statistically reasonable, and we can fail to reject the null hypothesis. Our t value is 1.16 and the p-value is greater than our α, 0.05, so we conclude that the mean sale price of homes in Ames, Iowa, is not statistically different from $135,000.

This confidence interval plot shows the confidence interval around the mean estimate of sale price, and the vertical line references the null hypothesis value. Because the vertical reference line is within the bounds of the confidence interval, we conclude that the mean sale price of homes in Ames, Iowa, is not statistically different from $135,000. Finally, we need to verify the validity of the test by checking that the distribution of the prices of houses is normal. Let's look at the histogram and Q-Q plot to verify this assumption.

The histogram appears to be bell shaped, like a normal distribution. The normal and kernel density estimates are nearly overlapping, indicating that the estimated data distribution from our sample is nearly equivalent to a normal distribution.

If the data are normal, a Q-Q plot produces a relatively straight line with some deviations due to random sampling. In our Q-Q plot, the sorted sale prices are plotted against quantiles from a standard normal distribution. The tail ends seem to be skewed due to possible outliers, but, overall, the plot fails to show departures from normality.

Based on the t test results, we can assume that the Student's t test is valid, and we can conclude that the mean sale price of homes in Ames, Iowa, is not statistically different from $135,000.

Two-Sample t Tests

Scenario

Recall in the one-sample t test, we tested a hypothesized value to make a claim about the population mean. Based on the sampled data, we calculated our p-value and either rejected or failed to reject that hypothesis. Now, instead of claiming that a population mean is equal to some number, you want to test whether two populations' means are equal.

Specifically, you want to test the population means of sale price for homes with and without masonry veneer. Is there a difference in sale price? Can you claim that the population means are statistically different from one another? To answer this question, let's explore the two-sample t test.

The null hypothesis for the two-sample t test is that the means for the two groups are equal, which is μ1=μ2, or μ1-μ2=0. The alternative hypothesis is that the means for the two groups are not equal, or $\mu_1 - \mu_2$ does not equal 0.

Assumptions for the Two-Sample t Test

Before using a two-sample t test, you need to verify three statistical assumptions. Otherwise, the test might be invalid.

The first assumption is that the observations are independent, meaning that when you sampled the data, you collected each unit of information independently from one another.

For example, you can't have repeated measurements on the same observations weighing the results toward a specific result. As you can probably tell, you verify this assumption during the design stage of your analysis.

The second assumption is that you have normally distributed populations. If the populations from which you obtained your samples are normally distributed, then your sample data will most likely be normal too. For large samples, the two-sample t test is fairly robust to deviations from the assumption of normality. However, for small samples, it's important to verify this assumption by visually examining plots of the data. You should confirm that the histogram appears normal and QQ-plots follow a straight line, and we do this in the upcoming demonstration.

Finally, for the last assumption, you need to verify that you have equal population variances. To do this, you can use the folded F Test to test for equality of variances.

The null hypothesis is that the population variances are equal. The formula is sigma 1 squared equals sigma 2 squared. Remember that sigma squared represents variance.

The alternative is that the variances are not equal.

Your test statistic, or F value, is simply the ratio of the maximum sample variance of the two groups to the minimum sample variance of the two groups.

By construction, theFstatistic is always greater than or equal to 1.

If the null hypothesis is true and the variances in the two populations are equal, then the F value is close to 1 and the p-value for F is statistically nonsignificant, a p-value most likely greater than 0.05.

Consequently, a large value for the F statistic is evidence against the assumption of equality.

Testing for Equal and Unequal Variances

In SAS, the equality of variance tests and the t test are all going to be grouped together in the output. You'll look at the F test for equal variances. Remember, this probability value is

saying that this is the probability of getting a statistic at least this extreme, assuming the null hypothesis is true. And the null hypothesis is that the variances of the two groups are equal.

In this case, your p-value is 0.7446. Let's say α was set ahead of time to be 0.05. This p-value is certainly greater than α, so you would fail to reject the null hypothesis. You don't have enough evidence to say that the variances are unequal. Therefore, you look at the equal variance t test, or pooled t test, in terms of the means. By default, SAS shows the 95% intervals for both the pooled method, assuming equal variances for group 1 and group 2, and the Satterthwaite method, assuming unequal variances. SAS calculates a pooled t test that uses a weighted average of the two sample variances.

Here the p-value is 0.0003, less than our α, 0.05, so the means of the two groups are not equal, and you would reject the null hypothesis.

If your F statistic has the p-value 0.0185 for your equality of variance test, which is less than your α, then you would reject the null hypothesis. You have reason to believe that the variances are not equal. Therefore, you would look at the unequal variance t test, which is referred to as the Satterthwaite approximation, and look at the p-value 0.0320 to test the group means. SAS calculates a Satterthwaite t test that compensates for unequal variances and enables you to move forward with the equality of means test when the variances are not equal.

In this case, again, the conclusion for the mean is that the means are not equal, especially if α is 0.05.

As an important side note, if you were to choose the equal variance t test in this case, you would not reject the null hypothesis at the 0.05 level. This shows the importance of choosing the appropriate t test. We also verify this assumption, as it pertains to our data, in the demonstration.

Demo: Performing a Two-Sample t Test Using PROC TTEST

Filename: **st101d03.sas**

Let's now use PROC TTEST to perform a two-sample t test, and test whether the mean of SalePrice is the same for homes with masonry veneer as for those without. I'll open the program st101d03.sas.

This PROC TTEST step doesn't use the null hypothesis option like you saw before, because we're testing the equality of means. The step includes a CLASS statement with Masonry_Veneer as the grouping variable. The CLASS statement is required in a two-sample t test. The classification variable can be numeric or character, but must have exactly two levels, because PROC TTEST divides the observations into the two groups using the values of this variable. Classification levels are determined from the formatted values of the CLASS variable, so if necessary, you can apply a format to collapse the data into two levels. The FORMAT statement here applies the $NoYes format to display Yes and No in the output instead of Y and N.

Ill submit the code.

In the results, let's start by verifying our assumption of normality of the distribution of each group by looking at the histograms and Q-Q plots.

In the summary panel, the histogram of each subgroup appears to be normally distributed, with, of course, a different center of location. Both histograms have a blue normal reference curve superimposed on the plots to help determine whether the distributions are normal.

How about the Q-Q plots? If the data in a Q-Q plot come from a normal distribution, the points cluster tightly around the reference line. The first plot shows homes without masonry veneer, or the Nos, and the second shows homes with, the Yess. The Q-Q plots both exhibit relatively straight lines. Theres slight curvature on both, but nothing too extreme. From these four plots, its safe to say both populations are normally distributed.

Next, to test the equality of variances, we look at the Equality of Variances table. The F statistic is 1.36 and the p-value is relatively large at 0.1039. Based on this, do we reject or fail to reject the null hypothesis of equal variances? The p-value is greater than alpha, so we do not reject the null hypothesis. We dont have enough evidence to say the variances are unequal.

So based on the results of the F test, we now look in the t tests table at the t test for the hypothesis of equal means, the pooled t test. Here we see that the p-value is less than .0001, which is less than 0.05, so we can reject the null hypothesis that the group means are equal. We can conclude that the sale price of homes with masonry veneer is significantly different from homes without it.

Before we go on, notice that the t statistic values for both tests are almost equal, -5.38 and -5.72. When the population variances are equal, the t values are equivalent mathematically. The slight difference here is due to random sampling differences when calculating the variances.

We can make the same conclusion about the means from the confidence interval plot. We can see both pooled and Satterthwaite 95% confidence intervals. Notice for the pooled variance method, the confidence interval for the difference in means is between about -$33,000 and -$15,000. It doesn't include 0, which is our hypothesized value. In other words, we have enough evidence to say that the difference of the means is significantly different from 0 at the 95% confidence level.

Now lets go back up to the descriptive statistics table. From our sample of 300 homes, 89 homes have masonry and 209 do not, and there are 2 homes with missing data, so well remove these from the analysis. For the 209 homes without masonry veneer, the sale price sample mean is $130,172, and for the 89 homes with it, the sample mean is $154,705. The difference mean value between no masonry veneer and masonry veneer is -$24,533. From the sample data, its clear that homes with masonry veneer tend to have a statistically significant higher value.