

Delve Deeper into Survey Data with Minitab

2-Sample t-Tests, Proportion Tests, ANOVA and Regression

In a previous article, we explored several basic survey analysis tools in Minitab. Now we'll use more sophisticated techniques, including 2-sample t-tests, proportion tests, ANOVA and regression, to dig deeper into our data.

Hypothesis Tests

In [Analyzing Survey Data with Minitab](#), we began looking at hypothesis testing by using a chi-square test to confirm an association between gender and pet preference. Many other hypothesis tests are commonly used to analyze survey data, including t tests to evaluate means and proportion tests to evaluate percentages. These types of tests can be used to compare averages or proportions to a target value, or to compare averages or proportions to each other.

Most hypothesis tests in Minitab are located in the **Stat > Basic Statistics** menu.

2-Sample t-Test

We can use the 2-sample t-test to compare the averages between two groups and determine if there is a significant difference between them or if the observed difference is due instead to random chance.

Suppose our local pet adoption and rescue agency surveys a random sample of 100 people who adopted pets to investigate the potential economic implications on pet

adoption. We want to see if the amount of money people spend on their pets annually for food, supplies, medical costs, etc. varies by the type of pet people adopt. We can use a 2-sample t-test to see whether people who adopted cats spent more, less or the same amount of money as those who adopted dogs. The null hypothesis would be that the average money spent by dog and cat owners are equal, and the alternative hypothesis is that the average expenses are not equal.

We select **Stat > Basic Statistics > 2-Sample t** to run this test on our data, and Minitab gives the following output:

Two-Sample T-Test and CI: Cost, Pet					
Two-sample T for Cost					
Pet	N	Mean	StDev	SE Mean	
Dog	50	501.4	40.3	5.7	
Cat	50	443.5	36.3	5.1	
Difference = mu (Dog) - mu (Cat)					
Estimate for difference: 57.95					
95% CI for difference: (42.74, 73.17)					
T-Test of difference = 0 (vs not =): T-Value = 7.56					
P-Value = 0.000 DF = 98					

We select a value called the α -level before conducting a hypothesis test, and we compare the p-value of the t-test to the α -level. The p-value tells us how likely it is that we would obtain our results if the null hypothesis is true. If the p-value is less than or equal to the α -level, we reject the null hypothesis and conclude that there is a difference in average annual expenses for cats and dogs.

For our test, we used an α -level of 0.05, which is very common. Since our data yielded a p-value of 0.000, which is less than our 0.05 α -level, we can reject the null hypothesis. Our data support the conclusion that cat owners do not spend the same amount of money on their pets, on average, as dog owners. In fact, we can conclude that cat owners spend significantly less than dog owners.

The 2-sample t-test also constructs a confidence interval that gives us more detail about the difference between groups. Our data were analyzed with an α -level of 0.05, so Minitab gives us a 95% (or 0.95) confidence interval. This interval tells us that, based on the sample data, we can be 95% confident that the true mean difference between expenses for the two populations is between 42.74 and 73.17 dollars. Note that this confidence interval does not contain 0, which indicates that the difference between our group means is significant, or not equal to 0.

Proportion Tests

What if we want to make inferences about a population proportion? We can use Minitab's one proportion test to do this.

Suppose we have survey data for 1,000 randomly selected local pet owners. We wish to determine if the population proportion of ferret owners is different from the national average of 6.5%.

When we use one proportion procedures, we are really trying to decide which of two opposing hypotheses is true, based on our data:

- *There is no difference between local ferret ownership rates and the national average. We call this the "null hypothesis."*

- *There is a significant difference between local ferret rates and the national average. We call this the "alternative hypothesis." We also can make the alternative hypothesis directional, to see if our rate is higher or lower than average.*

To run a one proportion test, we use **Stat > Basic Statistics > 1 Proportion**. Minitab gives the following output:

Test and CI for One Proportion					
Test of p = 0.065 vs p not = 0.065					
Sample	X	N	Sample p	95% CI	Exact P-Value
1	87	1000	0.087000	(0.070268, 0.106208)	0.008

The analysis provides a p-value of 0.008, which indicates that there is only a 0.8% chance that we would have obtained this sample proportion (or a more extreme sample proportion) if the population proportion was actually equal to our reference value of 0.065, the national average.

The proportion test also gives us a confidence interval, which tells us that we can be 95% confident that the local ferret owner population proportion is greater than or equal to 0.070268, or 7.02%, and less than or equal to 0.106208, or 10.6%. Since the confidence interval does not contain our reference value of 0.065, and the p-value is below 0.05, we can reject the null hypothesis and conclude that the population proportion is not 0.065. The proportion is significantly greater than 6.5%.

Minitab also lets us perform two proportions tests to make inferences about the difference between two population proportions. Let's say we want to know whether the proportion of visitors to an animal rescue shelter who adopt a pet could be increased by providing

an incentive such as free pet food. We might offer the incentive to half of our visitors, survey visitors who did and did not receive the offer, and use the two proportions test to see if the results suggest offering an incentive would encourage more of the overall population of visitors to adopt.

ANOVA

What if we want to use our annual pet expenses survey data to understand information about three or more groups? In this case we can use Minitab's ANOVA (analysis of variance) tools. An ANOVA is similar to a t-test in that both analyses compare group means for a continuous Y (e.g. pet expenses), but in addition to using ANOVA to test if three or more means differ for a single grouping variable, it can also be used to compare group means for multiple variables.

There are different types of ANOVAs. To test if the group averages for a single categorical factor are equal, you can use one-way ANOVA. For instance, our pet agency could use a one-way ANOVA to determine if pet expenditures differ for three different education levels.

An ANOVA procedure could also be used to determine whether the average amount of money spent on pets differed between multiple categorical factors, such as three education levels, and between the type of pet. ANOVA also can be used to determine if there are interactions between two or more variables. If an interaction between two factors is present, then the effect of one factor on the response depends on the level of another factor.

We can use **Stat > ANOVA > General Linear Model** for this type of analysis. For our pet ownership survey, we could use this

tool to determine (1) whether pet expenses differed across educational levels, (2) whether pet expenses differed between dogs and cats, and (3) whether there was an interaction between educational level and pet type.

General Linear Model: Cost versus Pet, Education						
Analysis of Variance for Cost, using Adjusted SS for Tests						
Source	DF	Seq SS	Adj SS	Adj MS	F	P
Pet	1	83965	81743	81743	54.23	0.000
Education	2	895	580	290	0.19	0.825
Pet*Education	2	1488	1488	744	0.49	0.612
Error	94	141689	141689	1507		
Total	99	228037				

In the output above, the Pet variable has a p-value of 0.000, while Education and the Pet*Education interaction both have p-values well above 0.05. Based on this analysis, our data do not support the conclusion that pet expenses differ significantly by education level, or an interaction between educational level and the type of pet owned.

There is also a type of ANOVA called an ANCOVA. If you have a combination of both categorical and continuous factors, you can use Minitab's General Linear Model tool to perform an ANCOVA as well.

Regression

Minitab includes a wide variety of regression analyses, which can be used to examine or predict how particular continuous variables affect a particular outcome—for example, how a person's household income correlates with pet expenses. It can be used for:

- Determining whether a relationship exists between dependent (Y) and independent variables (Xs).
- Determining the strength and structure of the relationship, if one exists.

- Predicting values of the dependent variable based on values of the independent variables.

Simple linear regression tells us about the relationship between one Y variable and one X variable. We use multiple regression to tell us about relationships between one Y variable and several X variables. In multiple regression, we are still showing how Y depends on X, but now Y may depend on many different X's or even the interaction between the X's.

Let's say we want to use simple linear regression to assess whether household income can be used to predict pet expenses. We can go to **Stat > Regression** to run our analysis in Minitab.

Regression Analysis: Cost versus Income				
The regression equation is				
Cost = 439 + 0.511 Income				
Predictor	Coef	SE Coef	T	P
Constant	438.525	3.341	131.25	0.000
Income	0.51145	0.02325	22.00	0.000
S = 12.2225 R-Sq = 91.0% R-Sq(adj) = 90.8%				

The Income p-value of 0.000 is less than $\alpha=0.05$, which indicates that there is a significant linear relationship, or correlation, between a pet owner's income and how much they spend on their pet.

It's very important to remember that correlation tells us about the nature and degree of association between variables, but cannot tell us that a cause-and-effect relationship exists. An association between an independent and dependent variable does not mean that X *causes* Y—only that as X increases or decreases, so does Y. This is what statistics professors mean when tell students that “correlation does not imply

causation.” It is particularly important to be clear about this when you are communicating the results of your analysis to people who may not be well versed in statistics.

The good news is that you can still use regression analysis to make predictions, because prediction does not *require* causation. Regression analysis describes the observed relationship between one or more variables and a response, and we can use that relationship for prediction without worrying about causation as long as the patterns found in the data continue to hold true.

Powerful Tools for Survey Analysis

This article provides only a brief overview of the kinds of information you can glean by analyzing your data with Minitab Statistical Software. Minitab has the tools you need to analyze survey data and make sound conclusions about markets, customers, or whatever you're trying to assess. For more information and additional examples detailing how to use these and other useful tools, Minitab offers an extensive Help system and free Technical Support.

Eston Martz
Senior Creative Services Specialist,
Minitab Inc.

Michelle Paret
Product Marketing Manager, Minitab Inc.

Visit www.minitab.com for more information about statistics.