Spring 2011, Volume 1, Issue 1

THE STATSWHISPERER

The StatsWhisperer Newsletter is published by staff at StatsWhisperer.™ Visit us at: www.statswhisperer.com

Introduction to this Issue

The current issue of this newsletter is geared to provide some foundational support to those approaching data analysis.

Please be forewarned, there may not be much in this issue that is of great interest to individuals experienced in analysis. My hope is to lend a hand to those who have been tasked with analysis, but are at the stage of just first sitting down at the computer and trying to determine their first steps.

At this juncture, I have found that a bit of orientation to the data analysis process (as well as conducting and interpreting statistical tests) is very helpful. Therefore, I hope this issue will provide the beginning analyst with some idea of how to approach this process.

INSIDE THIS ISSUE

This issue addresses the process of data analysis at the univariate and bivariate levels (Please reference in a text or even Google any terms that I mention but do not discuss).

If readers find this issue useful I would be happy to write a subsequent issue regarding multivariate analysis as a next step. Also, I used some practice data (that I made up) to generate results to present in this issue. Anyone who would like these data (in SPSS) to use for practice should feel free to email me.

Doing Your Data Analysis

For our purposes here, I will assume that the data to be used have been entered into a database and cleaned (i.e., checked for mistakes/accuracy). At this point, one must prepare the data for analysis, identify the appropriate statistical test, and conduct and interpret the results of the tests.

In this issue I will focus on the last two processes. However, the importance of correctly preparing the data for analysis cannot be overstated. Checks should be made concerning issues such as multivariate and univariate normality, missing data, multicollinearity, homo/heteroscedasity, outliers, relative variances, and scale reliability.

I have found that it is often easier to understand such data preparation techniques after one understands how the data will be used in statistical tests. Therefore, in the current newsletter I will lead with an overview of some of the most widely applied statistical tests and will discuss data preparation techniques in subsequent issues.

An Orientation to Statistical Analysis

The purpose of a newsletter is to be brief, to the point, and as useful as possible. Therefore, I will present an overview of some statistical techniques. The reader should consider that there are various points of view regarding how tests are chosen and even how they are conducted. Also, there are statistical details that are not explained here simply because this is a brief overview.

Essentially, our discussion here is a jumping off point. What are presented here are the initial considerations that one may consider when pursuing analysis. My view has always been that a little bit of something is better than a lot of nothing. Thus, this overview is a little bit of something. It presents some of the most common tests applied to the analysis of data, especially among individuals that are first approaching data analysis.

However, the reader should absolutely feel the need to delve deeper into the details associated with each statistical test mentioned here. Also, the reader should consider that other statistical tests may be more appropriate for their analysis in particular situations. Therefore, it is always a good idea to get consultation of some sort from an authority (e.g., staff statisticians, etc.) if only for verification purposes.

That being said, I would like to start off by noting that most dissertation studies are cross-sectional in nature. Furthermore, most dissertation studies use statistical methods that do not advance beyond the level of linear or binary logistic regression. Therefore, I will limit my discussion to these study characteristics.

Basically, when doing statistical analysis for a study, there is a 3-step crescendo whereby, variables are first examined individually (called univariate analysis) and then on a one to one basis (called bivariate analysis). The final step usually involves examining how a group of variables relates to one outcome variable (called multivariate analysis). These steps can very well encompass all you need to answer your research question.

The univariate statistics (3) to be examined here include the mean, standard deviation, and range. The bivariate tests (4) discussed here include the t-test, chi-square, ANOVA, and correlation. The multivariate tests (2) to be described include linear and binary logistic regression. Again, I will leave the last set of tests for the next issue.

It is often useful to have an example study when describing analysis. Therefore, I will manufacture a study. The study question will be: *Are people who live with dogs happier than people who live with cats?* The independent variable is does the respondent live with a dog or cat (a categorical predictor). The dependent variable is the level of happiness of the respondent (a continuous outcome variable with a possible range of 1-100 with higher scores indicating more happiness).

Data on respondent age, gender, and race will also be presented in this study. There will be 100 respondents involved in this research (50 who live with dogs/50 who live with cats).

Univariate Analysis

Univariate analysis examines each variable in a data set, separately. It looks at the range of values and measures of what are known as *central tendency*. It describes the pattern of response to each variable. One user friendly way to get the numbers you need for this analysis using a program such as SPSS is to use the function:

Analyze ⟶ Descriptive Statistics ⟶ Frequencies

When the dialog box appears click on all of the study variables (respondent age, gender, race, dog/cat status, and happiness) one at a time using the button in between the columns to bring the variables over from the column on the left to the column on the right. In the bottom center section of the dialog box, click on statistics. In this dialog box, in the upper right hand corner under *central tendency*, check mean. In the lower left hand corner under *dispersion*, click standard deviation, range, minimum, and maximum (Please use a statistical text book to look up the definitions of these terms if needed). Then click continue. Then when the original dialog box appears, click ok.

Subsequently, out output should appear. At the top of our out put we see the box presented under this text. The measures of dispersion we used will be useful for our continuous variables (i.e., subject age and level of happiness). For example, under subject age we see the average person (i.e., the mean) is 35.74 years old with a standard deviation of 13.33 years. The range of respondent age is 48 years, with the youngest person being 18 (minimum value) and the oldest being 66 (maximum value) years of age.

Likewise for our outcome variable, we see the average person scored a 40.46 with a standard deviation of 26.66 years. The score range is 78 points with a low score of 12 and a high score of 90. Thus, within the sample, there are respondents who are very happy (with a score of 78), not very happy (with a score of 12), and the average person (with the mean score of 40.46) is below the middle score (50) along the scale of happiness measure (with a continuum of 1-100).

Statistics

		Subject age	Subject gender	Subject race/ethnicity	Does the subject live with a dog or a cat?	What is the level of the subject's happiness?
N	Valid	100	100	100	100	100
	Missing	0	0	0	0	0
Mean		35.7400	.5100	2.2100	1.5000	40.4600
Std. Deviation		13.32547	.50242	1.27363	.50252	26.65925
Range		48.00	1.00	4.00	1.00	78.00
Minimum		18.00	.00	1.00	1.00	12.00
Maximum		66.00	1.00	5.00	2.00	90.00

Univariate Analysis (continued)

Next, as we scroll down the SPSS output, we notice the below boxes. We can see that within our sample, there are 49 males (49%) and 51 females (51%). The racial ethnic composition incorporates 37 White respondents (37%), 32 African-American/Black respondents (32%), 12 Hispanic/Latino respondents (12%), 11 Asian/Pacific Islander (11%), and 8 (8%) respondents within the other category.

Additionally, as we stated before, we see that there are 50 respondents (50%) that live with dogs and 50 (50%) that live with cats.

Thus, we have now examined the characteristics of each variable separately. We can now describe and comment intelligently on the characteristics of the respondents in our sample.

For example, we can say that men and women seem to be represented equally, over two-thirds (69%) of the sample are either of a White or African-American/Black racial/ethnic background, and the average respondent is about 35 years old. We can verify that half the sample either lives with dogs or cats. Lastly, we can say that the typical respondent scored below the middle score on the happiness measure.

Subject gender

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Male	49	49.0	49.0	49.0
	Female	51	51.0	51.0	100.0
	Total	100	100.0	100.0	

Subject race/ethnicity

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	White	37	37.0	37.0	37.0
	African-American/Black	32	32.0	32.0	69.0
	Hispanic/Latino	12	12.0	12.0	81.0
	Asian/Pacific Islander	11	11.0	11.0	92.0
	Other	8	8.0	8.0	100.0
	Total	100	100.0	100.0	

Does the subject live with a dog or a cat?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Dog	50	50.0	50.0	50.0
	Cat	50	50.0	50.0	100.0
	Total	100	100.0	100.0	

An Orientation to Bivariate Analysis

Next, we may want to look at how the variables relate to one another. A very interesting use of bivariate tests concerns how the various respondent characteristics (respondent age, race, and gender), also known as demographic characteristics, and the independent variable relate to the dependent (a.k.a., outcome) variable.

It is also important to consider how the demographic and independent variables relate to one another. However, for the sake of brevity, I will examine how the respondent demographic characteristics and the independent variable relate to our outcome variable.

We will first was to see if any demographic characteristics are significantly related to the outcome variable. This is a very important step. One of the huge benefits of multivariate analysis is the option to "control" for the influence of variables on the outcome variable besides the independent variable.

For example, what if we conduct a bivariate statistical test and find that living with either a dog or a cat is related to our happiness outcome variable. However, another bivariate test indicates that gender is also related to happiness. How do we establish if living with a dog or a cat is really related to happiness among our sample? Or how about which is the stronger predictor? That is where multivariate models come in. We can enter several variables in a model and note the relative strength of each in explaining our outcome variable.

Thus, it is said in the case of a multivariate model such as ours, which examines living with a dog or a cat as a predictor of happiness, that if we include gender as a predictor that we are "controlling" for the effect of gender in our analysis. Generally, when we examine a study relationship we want to control for any salient influence on the outcome variable outside of our independent variable.

Researchers choose what variables they would like to "control" for in a series of ways. For example, some researchers may choose these "control" variables based on a theory or prior research studies. Some prefer (including myself) to use the variables that are identified within a particular sample as being related to the outcome variable (I use any variable associated with the outcome variable at the $p<.10$ level).

Thus, I would conduct bivariate analysis between the outcome variable with the demographic and independent variables. Afterwards, I would use a multivariate model to control for the demographic variables significantly associated with the outcome variable (with the independent variable) to augment my confidence in my findings.

So here is an important point. Bivariate analysis is not just a way to see how variables relate to one another on a one-to-one basis. Bivariate analysis can also be an important step to see which variables are significantly related to the outcome variable and therefore may warrant inclusion in the final multivariate model.

Bivariate Analysis

Below is a very abbreviated guide to which bivariate tests may be considered for four different types of variable combinations:

Categorical ⟶ Continuous ⟶ **Ind samples t-test**

(2 categories)

Categorical ⟶ Continuous ⟶ **ANOVA**

(>2 categories)

Continuous ⟶ Continuous ⟶ **Correlation**

Categorical ⟶ Categorical ⟶ **Chi-square**

Independent samples t-test

Our first combination recommends the use of an independent samples t-test. Of course, this should only be considered when a researcher believes the subjects in each category are from independent samples. In our study here, this test may be employed using the two-group categorical variable *gender* with the continuous variable *happiness*. Using SPSS, we can use the function:

Analyze ⟶ Compare means ⟶ Ind samples t-test

We put the two-group categorical variable into the box labeled *Grouping Variable* (after highlighting and clicking the variable over using the arrow to the left). Next, click define groups box, where you will have the option to specify values for groups 1 and 2. Here you enter the actual values to code the groups in the database. For example, I coded males as 0 and females as 1, so I enter 1 and 0 for each group. Then click *continue*. Lastly, click our continuous variable (happiness) from the column on the left to the column on the right labeled *test variable(s)*. Then click *OK*.

Subsequently, the below output is generated. Under the **Group Statistics** box we see that the mean score for happiness for females is 45.22 and for males is 35.51. We also see under the **Independent Samples Test** that Levene's test is significant (p<.001), so we use the value of t and significance under *Equal variances not* assumed, which tell us the value of t is 1.852 and p<.10. In short we find a difference between the mean score of happiness at the p<.10 level (i.e., sig.=.067) for males and females. Thus, we will control for this significant effect in our multivariate model.

Group Statistics

	Subject gender	N	Mean	Std. Deviation	Std. Error Mean
What is the level of the subject's happiness?	Female	51	45.2157	29.62048	4.14770
	Male	49	35.5102	22.42499	3.20357

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means		
		F	Sig.	t	df	Sig. (2-tailed)
What is the level of the subject's happiness?	Equal variances assumed	13.965	.000	1.842	98	.069
	Equal variances not assumed			1.852	92.981	.067

Bivariate Analysis (continued)

ANOVA

Our second combination recommends the use of an ANOVA. In our study here, this test may be employed using the categorical variable with more than two groups *race* with the continuous variable *happiness*. Using SPSS, we can use the function:

Analyze ⟶ Compare means ⟶ One-way ANOVA

We put the categorical variable *race* into the box labeled *Factor* (after highlighting and clicking the variable over using the arrow to the left). Next, we move the variable *happiness* into the box labeled *Dependent List*. Also, click the box labeled *Options* under in the lower right corner, check the option for *Descriptive* and then *Continue*. Then click the box labeled *Post hoc, which is* the box to the left of the *Options* box. From here check the option labeled *Bonferroni*, and then *Continue*. Then click *OK*.

Subsequently, the below output is generated. Under the **Descriptives** box we see the mean score for happiness by race. Although some may seem largely different, under the table labeled **ANOVA**, we see the significance level for the test is .599. This indicates that there is no significant difference in the mean score of happiness by race. Thus, we will not need to control for the effect of race in our multivariate model explaining happiness scores.

It is also important to note that if the **ANOVA** table had indicated a significant test, we would need to be able to see which means are significantly different from one another. Through checking the option for the *Bonferroni* function our output would include a **Post Hoc Tests** of **Multiple Comparisons**. This additional table would indicate just which mean scores differ significantly from one another.

Descriptives

What is the level of the Subject's happiness?	N	Mean	Std. Deviation
White	37	38.7027	26.23808
African-American/Black	32	40.0313	25.69358
Hispanic/Latino	12	34.6667	26.61624
Asian/Pacific Islander	11	51.7273	31.79022
Other	8	43.5000	27.30777
Total	100	40.4600	26.65925

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1993.293	4	498.323	.692	.599
Within Groups	68367.547	95	719.658		
Total	70360.840	99			

Bivariate Analysis (continued)

Correlation

Our third combination recommends the use of a correlation. In our study here, this test may be employed using the continuous variable *age* with the continuous variable *happiness*. Using SPSS, we can use the function:

Analyze ⟶ Correlate ⟶ Bivariate

We put the continuous variables *age* and *happiness* into the box labeled *Variables* (after highlighting and clicking the variables over using the arrow to the left). Next, we click *OK*.

Subsequently, the below output is generated. Under the **Correlations** box we see that the value of the Pearson Correlation is -.188 and the significance value is <.10 (i.e., sig.=.061). Thus, for our purposes we will consider this relationship significant and will control for *age* in our multivariate model.

It is also important to note the Pearson Correlation is negative (i.e., -.188), which indicates that among our sample, as *age* goes up *happiness* goes down.

Correlations

		Subject age	What is the level of the subject's happiness?
Subject age	Pearson Correlation	1	-.188
	Sig. (2-tailed)		.061
	N	100	100
What is the level of the subject's happiness?	Pearson Correlation	-.188	1
	Sig. (2-tailed)	.061	
	N	100	100

Bivariate Analysis (continued)

Chi-square

Our forth combination recommends the use of a chi-square. A chi-square should be used in the case of two categorical variables. However, in our study here, our outcome (i.e., *happiness*) is continuous. Thus, for didactic purposes, let's dichotomize the outcome into a categorical variable. Specifically, our variable will be, is the respondent happy yes or no. Our second variable for or chi-square will be our independent variable (i.e., lives with a dog or a cat). Using SPSS, we use the following function:

Analyze ⟶ Descriptive Statistics ⟶ Crosstabs

(People also use: Analyze ⟶ Non-Parametric Tests ⟶ Chi-Square. I prefer the first function)

We put the variables happy (yes/no) in the column box and lives with a dog or cat (yes/no) in the row. In the lower middle section of the dialog box, click *Statistics*, then check *Chi-square*, and then click *Continue*. The click the box marked *Cells* (next to the *Statistics* box). Under *Counts* check *Observed* and *Expected* and under *Percentages*, click *Row*, then click *Continue*.

Subsequently, our output is generated. The first table generated (below) describes the observed and expected counts of the distribution. The second table presents a value for a **Pearson Chi-square** that will describe if there is a statistically significant difference between the observed and expected counts. Our test yield a p value <.01, indicating a significant difference. But where is this difference?

Below you will note, expected counts of 25.5 dog owners and 25.5 cat owners under the No (i.e., not happy) category. However, what we observed is 39 dog owners (78%) and 12 cat owners (24%) under this category. Furthermore, we expected counts of 24.5 dog owners and 24.5 cat owners under the Yes (i.e., happy) category. However, what we observed is 11 dog owners (22%) and 38 cat owners (76%) under this category.

Thus, the data indicate significantly more than expected respondents who own cats that are in the Yes (i.e., happy) category. Additionally, our **Pearson Chi-square** indicated that this is a statistically significant difference. This would indicate that preliminarily, respondents that live with cats may be more likely to be happy.

			Is the subject happy?		Total
			No	Yes	
Does the subject live with a dog or a cat?	Dog	Count	39	11	50
		Expected Count	25.5	24.5	50.0
		% within Does the subject live with a dog or a cat?	78.0%	22.0%	100.0%
	Cat	Count	12	38	50
		Expected Count	25.5	24.5	50.0
		% within Does the subject live with a dog or a cat?	24.0%	76.0%	100.0%
Total		Count	51	49	100
		Expected Count	51.0	49.0	100.0
		% within Does the subject live with a dog or a cat?	51.0%	49.0%	100.0%

Final Comments

It is important to note that we had little room to discuss each test. However, I hope that that the material helped people to see how these tests all work together to examine the outcome variable and support a study. It has been my experience that usually there are materials available on how to conduct these tests, but there are limited resources that discuss how to combine tests to form a study.

For further reference, each test can be of course examined in a textbook or through searching for materials using the Google search engine.

If readers find this newsletter useful, I will write a concluding issue examining the use of logistic and linear regression in answering our very important research question:

Are people who live with dogs happier than people who live with cats?

Feel free to visit statswhisperer on the web at:

www.statswhisperer.com

Dr. William M. O'Bannon, Jr., the founder and CEO can be contacted at:

wb@statswhisperer.com

Information on our upcoming seminars can be viewed at:

http://www.statswhisperer.com/seminars/

Thanks for your interest in our newsletter!