

## Paper 155-2012

### How to Perform and Interpret Chi-Square and T-Tests

Jennifer L. Waller

Georgia Health Sciences University, Augusta, Georgia

#### ABSTRACT

For both statisticians and non-statisticians, knowing what data look like before more rigorous analyses is key to understanding what analyses can and should be performed. After all data have been cleaned up, descriptive statistics have been calculated and before more rigorous statistical analysis begins, it is a good idea to perform some basic inferential statistical tests such as chi-square and t-tests. This workshop concentrates on how to perform and interpret basic chi-square, and one- and two-sample t-tests. Additionally, how to plot your data using some of the statistical graphics options in SAS® 9.2 will be introduced.

#### INTRODUCTION

Millions of dollars each year are given to researchers to collect various types of data to aid in advancing science just a little more. Data is collected, entered, cleaned, and a statistician is told it is ready for analysis. When a statistician receives data, there are some basic statistical analyses that are performed first so that the statistician understands what the data look like. Statisticians examine distributions of categorical and continuous data to look for small frequency of occurrence, amount of missing data, skewness, variability and potential relationships. Not understanding what the data look like in their basic form can cause incorrect assumptions to be made and an incorrect statistical analysis could be performed later on.

The first look at a data set includes plotting the data, determining appropriate descriptive statistics, and performing some basic inferential statistics like t-tests and chi-square tests. Knowing what descriptive statistic or inferential statistical analysis is appropriate for the type of variable or variables in the data, how to get SAS® to calculate the appropriate statistics, and what are the necessary things to report off the output is essential. SAS® has a whole host of statistical analysis tools for both descriptive and inferential statistical analyses.

#### DESCRIPTIVE STATISTICS

What are descriptive statistics? These are numbers that describe your data and the type of descriptive statistic that should be calculated depends on the type of variable being analyzed: categorical, ordinal, or continuous.

##### Categorical Variables

Categorical data is data that can take on a discrete number of values or categories with no inherent order to the categories. Examples of categorical variables are sex (male or female), race (Black, White, Asian, Hispanic), disease or no disease, and yes or no variables. The types of descriptive statistics that are calculated for categorical variables

include frequencies and proportions or percentages in the various categories of the variable.

### Ordinal Variables

Ordinal variables are another type of variable where there are a discrete number of values but the values have some inherent order to them. For example, Likert scale variables (strongly disagree, disagree, agree to strongly agree) are ordinal variables. There is an inherent knowledge that strongly disagree is “worse” than disagree. Several types of descriptive statistics can be calculated for these types of variables including frequencies and proportions or percentages, medians, modes, inter-quartile range. Depending on the number of values an ordinal variable can take on, a mean and standard deviation may also be calculated.

### Continuous Variables

Continuous variables are those for which the values can take on an infinite number of values in a given range. While we may not be able to actually measure the variable as precisely as we would wish, the potential number of values is infinite. For example, think about measuring height. We record height in inches or meters and we measure height with a ruler of some sort. But we are limited in how precise height is measured due to our measuring device. Is someone really 5 feet 7 inches or are they really 5.655214328754 inches. We know that height is measured in a given range and that there really are an infinite number of values that height can take on, but the precision of our measurement is at the mercy of our measuring device. Descriptive statistics that are appropriate for a continuous measure include means, medians, modes, quartiles, variances, standard deviations, coefficients of variation, ranges, minimums, maximums, kurtosis, skewness, inter-quartile ranges, and the list goes on.

### **Inferential Statistics**

Inferential statistics are used to examine data for differences, associations, and relationships to answer hypotheses. The types of inferential statistics that should be used depend on the nature of the variables that will be used in the analysis. The most basic inferential statistics tests that are used include chi-square tests and one- and two-sample t-tests.

### Chi-Square Tests

A chi-square test is used to examine the association between two categorical variables. While there are many different types of chi-square tests, the two most often used as a beginning look at potential associations between categorical variables are a chi-square test of independence or a chi-square test of homogeneity. A chi-square test of independence is used to determine if two variables are related. A chi-square test of homogeneity is used to determine if the distribution of one categorical variable is similar or different across the levels of a second categorical variable.

### One- and Two-Sample T-tests

T-tests are used to examine differences between means. A one-sample t-test is used to examine whether the sample mean of a single continuous variable in a single group of individuals is different from a particular hypothesized population value. A two-sample t-

test is used to examine whether the sample mean of a single continuous variable is different between two different groups of individuals.

## THE DATA SET

Before introducing the statistical procedures in SAS® which are used to calculate descriptive statistics and perform chi-square and t-tests, the data set that will be used throughout the rest of the paper will be discussed. The data set comes from the book, "A Handbook of Small Data Sets" by Hand et al. (1994) page 266 dataset number 328. The description from the book is as follows:

"The data come from the 1990 Pilot Surf/Health Study of NSW Water Board. The first column takes on values 1 or 2 according to the recruit's perception of whether (s)he is a Frequent Ocean Swimmer, the second column has values 1 or 4 according to the recruit's usually chosen swimming location (1 for non-beach 4 for beach), the third column has values 2 (aged 15-19), 3 (aged 20-25) or 4 (aged 25-29), the fourth column has values 1 (male) or 2 (female), and, finally, the fifth column has the number of self-diagnosed ear infections that were reported by the recruit.

The objective of the study was to determine, in particular, whether beach swimmers run a greater risk of contracting ear infections than non-beach swimmers."

The data set consists of five variables. Three of these variables are categorical, frequent ocean swimmer status, location, and sex, and two are ordinal, age group and number of ear infections. While the number of ear infections is ordinal in that you can't have half of an ear infection, the range in the data set is such that number of ear infections can be treated as a continuous variable. Below are the variables, variable names, type of variable, and values of the variables. It should be noted that all variables are numeric variables in the dataset.

Variable	Variable Name	Variable Type	Variable Values
Frequent Ocean Swimmer	fos	Categorical	1=Yes 2=No
Location	location	Categorical	1=Non-beach 4=Beach
Age Group	agegroup	Ordinal	2=15 to 19 years 3=20 to 24 years 4=25 to 29 years
Sex	sex	Categorical	1=Male 2=Female
Number of Ear Infections	numearinfections	Continuous	0-17

A SAS® data set was created and named earinfection containing 5 variables and 287 observations. Format statements were used to format the levels of each variable except number of ear infections so that output is easier to understand.

## SAS PROCEDURES

### Descriptive Statistics for Categorical and Ordinal Variables

In SAS®, there are several procedures that can be used to determine frequencies and percentages for categorical and ordinal variables including PROC FREQ, PROC TABULATE and PROC SUMMARY. From a statistical standpoint, I most often use PROC FREQ to provide descriptive information as it not only provides the descriptive statistics that are needed, but it also provides various statistical tests. The basic syntax of PROC FREQ is

```
proc freq data=datasetname;  
  tables var1 var2 var3 / options;  
run;
```

The above code will give a frequency distribution table for each categorical variable var1, var2 and var3. A frequency distribution table gives the frequency of occurrence at each level of a variable, the percentage of individuals at that level of the variable, the cumulative frequency of individuals at that level and below that level of the variable, and the cumulative percentage of individuals at that level and below that level of the variable. Note that because PROC FREQ determines the number of individuals at a particular value or level of the variable you should only run a PROC FREQ on categorical or ordinal variables. Running a PROC FREQ on a continuous variable gives little information about what is happening with that continuous variable.

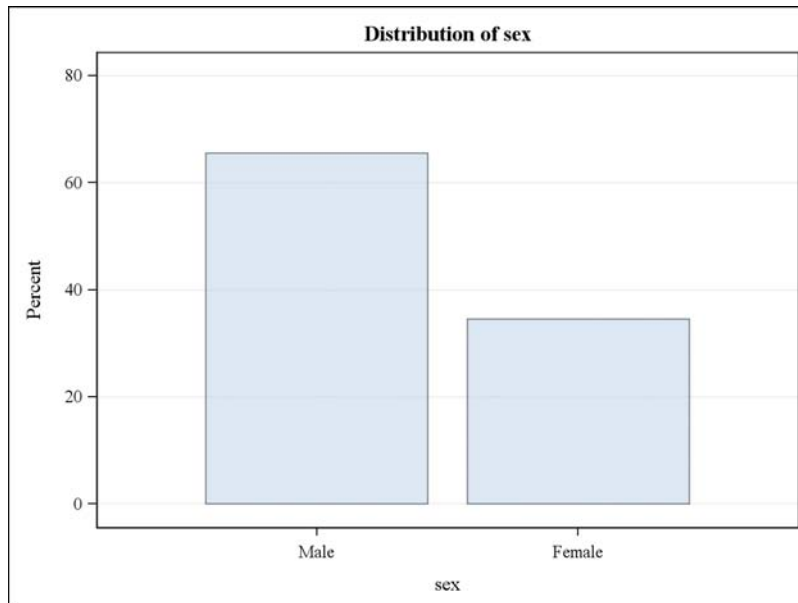
With SAS® 9.2 and ODS statistical graphics, specifying the ODS GRAPHICS ON and ODS GRAPHICS OFF statements around the PROC FREQ statements will produce a bar graph of each variable in the TABLES statement for a visual representation of the percentage of subjects at each level of the variable. Several different types of plots are available. To produce a bar graph of the percent at each level of the categorical variable, include the option `plots=(freqplot(scale=percent))` in the TABLES statement after the “/”.

Example 1: Using the earinfection data set, to get a frequency distribution of sex the following code would be used

```
ods graphics on;  
proc freq data=earinfection;  
  tables sex / plots=(freqplot(scale=percent));  
run;  
ods graphics off;
```

giving following output

Sex				
sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Male	188	65.51	188	65.51
Female	99	34.49	287	100.00



Notice that the number of male subjects, 188, is given under the Frequency column and the percent of male subjects out of the total, 65.51%, is given under the Percent column. The bar graph produced

### Descriptive Statistics for Continuous (and Sometimes Ordinal) Variables

To calculate different measures of location and variation in SAS®, such as means and medians, variances and standard deviations, one can use PROC MEANS, PROC UNIVARIATE, PROC SUMMARY, or PROC TABULATE. Most often, I use PROC MEANS and PROC UNIVARIATE to produce descriptive statistics for continuous or ordinal variables that have a large range in values. The basic syntax of PROC MEANS is as follows

```
proc means data=datasetname options;
  var contvar1 contvar2 contvar3;
run;
```

The above code will produce the default descriptive statistics which include the sample size used in calculation of other statistics, mean, standard deviation, minimum and maximum values for the variables listed in the VAR statement. There are many other descriptive statistics that can be calculated using PROC MEANS and these are requested in the PROC MEANS statement. The SAS® Online Documentation gives the keyword to use to request a specific descriptive statistic. The keywords for the

default and other optional descriptive statistics that I most often request are given below. If you request an optional descriptive and want the default descriptive statistics, you must also request the default descriptive statistic keywords in addition to the optional keywords.

Keyword	Statistic Calculated
n	Number of non-missing observations for the variable.
mean	Mean of the variable
std	Standard deviation of the variable
var	Variance of the variable
median	Median of the variable
min	Minimum value of the variable
max	Maximum value of the variable
sum	Sum of the values of the variable
qrange	Interquartile range of the variable

It should be noted that ODS graphics have not been implemented in PROC MEANS in version 9.2. To obtain a graphical representation of the data in the form of a box plot, PROC BOXPLOT or PROC SGPLOT can be used. Additionally, we can produce descriptive statistics for the continuous variable within the levels of other categorical variables. To do this we add a CLASS statement to the set of PROC MEANS statements:

```
proc means data=datasetname options;
  class catvar;
  var contvar1 contvar2 contvar3;
run;
```

Listing two or more categorical variables causes descriptive statistics to be calculated for all possible combinations of the levels of the two categorical variables. For example, if race (black and white) and sex (male and female) were both listed in the CLASS statement (i.e. CLASS race sex;) the resulting descriptive statistics would be calculated for black males, black females, white males, and white females.

PROC UNIVARIATE gives many of the optional descriptive statistics in PROC MEANS by default. As well, PROC UNIVARIATE will also produce a stem-and-leaf plot, box plot and Normal Probability plot if the PLOTS option is requested in the PROC UNIVARIATE statement. The basic syntax for PROC UNIVARIATE including the PLOTS option is

```
proc univariate data=datasetname plots;
  var contvar1 contvar2 contvar3;
run;
```

The code above will produce tons of descriptive statistics for each continuous variable listed in the VAR statement.

As indicated above, ODS graphics have not been implemented in PROC MEANS and are experimental in PROC UNIVARIATE. However, SAS® has implemented many different types of statistical graphics in version 9.2. Scatter plots, bar graphs, pie charts, box plots, histograms, density plot, LOESS plots, and needle plots are just some of the types of plots that can be produced in the new Statistical Graphics procedures in SAS®. For continuous variables, one of the most useful types of descriptive graphs is a box plot and within the SAS SGPLOT procedure there are two different types of box plots that can be produced: a vertical box plot using the vbox statement, or a horizontal box plot using the hbox statement. Box plots, whether vertical or horizontal are useful in examining variability, skewness, and potential outliers of a variable. To produce a horizontal box plot using PROC SGPLOT the following statements would be used:

```
proc sgplot data=datasetname;
  hbox contvar;
run;
```

If a box plot within levels of a continuous variable are needed, the category=catvar option is given in the hbox statement

```
proc sgplot data=datasetname;
  hbox contvar / category=catvar;
run;
```

Examples of PROC MEANS, PROC UNIVARIATE, and PROC SGPLOT are given below.

Example 2: Using the earinfection dataset, to calculate the default descriptive statistics for number of ear infections we would use the following PROC MEANS and PROC UNIVARIATE statements

```
proc means data=earinfection;
  var numearinfections;
run;

proc univariate data=earinfection plots;
  var numearinfections;
run;
```

and the output that would be produced is

Analysis Variable : numearinfections				
numearinfections				
N	Mean	Std Dev	Minimum	Maximum
287	1.3867596	2.3385412	0	17.0000000

The output from PROC MEANS indicates that the mean number of ear infections reported by the 287 individuals (under the N column) is 1.39 (under the Mean column)



with a standard deviation of 2.34 (under the Std Dev column). The minimum number of ear infections was 0 (under the Minimum column) and the maximum number was 17 (under the Maximum column).

Below is the output from PROC UNIVARIATE. It is divided into several different sections with the first section (named Moments) being the different statistical moments that can be calculated. Included in the Moments section are the mean, standard deviation, variance, and coefficient of variation. The next section is comprised of "Basics Statistical Measures" and includes the mean, median and mode as measures of location, and the standard deviation, variance, range and inter-quartile range as measures of variability. The next section included basic tests of location including a one-sample t-test for testing whether the mean is different from 0, and the Sign test and Wilcoxon Signed Rank test for testing whether the median is different from 0. Quantiles are given in the next section and include the maximum, median, and minimum. Finally, the last section includes Extreme Values with the 5 observations having the lowest values of the continuous variable and the 5 observations having the highest values of the continuous variable being listed. The plots option gives the Stem-and-Leaf plot and Box plot and the Normal Probability Plot.

Moments			
<b>N</b>	287	<b>Sum Weights</b>	287
<b>Mean</b>	1.38675958	<b>Sum Observations</b>	398
<b>Std Deviation</b>	2.33854124	<b>Variance</b>	5.46877513
<b>Skewness</b>	3.2015866	<b>Kurtosis</b>	14.180533
<b>Uncorrected SS</b>	2116	<b>Corrected SS</b>	1564.06969
<b>Coeff Variation</b>	168.633501	<b>Std Error Mean</b>	0.13803972

Basic Statistical Measures			
Location		Variability	
<b>Mean</b>	1.386760	<b>Std Deviation</b>	2.33854
<b>Median</b>	0.000000	<b>Variance</b>	5.46878
<b>Mode</b>	0.000000	<b>Range</b>	17.00000
		<b>Interquartile Range</b>	2.00000

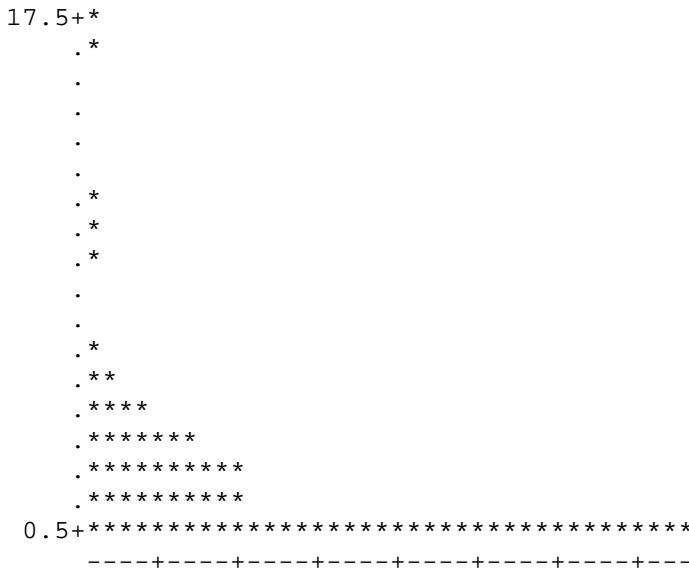
Tests for Location: Mu0=0				
Test	Statistic		p Value	
<b>Student's t</b>	<b>T</b>	10.04609	<b>Pr &gt;  t </b>	<.0001
<b>Sign</b>	<b>M</b>	68	<b>Pr &gt;=  M </b>	<.0001
<b>Signed Rank</b>	<b>S</b>	4658	<b>Pr &gt;=  S </b>	<.0001



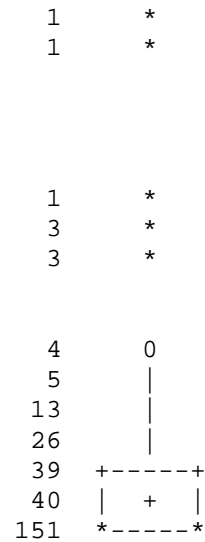
Quantiles (Definition 5)	
Quantile	Estimate
100% Max	17
99%	11
95%	5
90%	4
75% Q3	2
50% Median	0
25% Q1	0
10%	0
5%	0
1%	0
0% Min	0

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	284	10	65
0	283	10	249
0	282	11	30
0	278	16	31
0	277	17	47

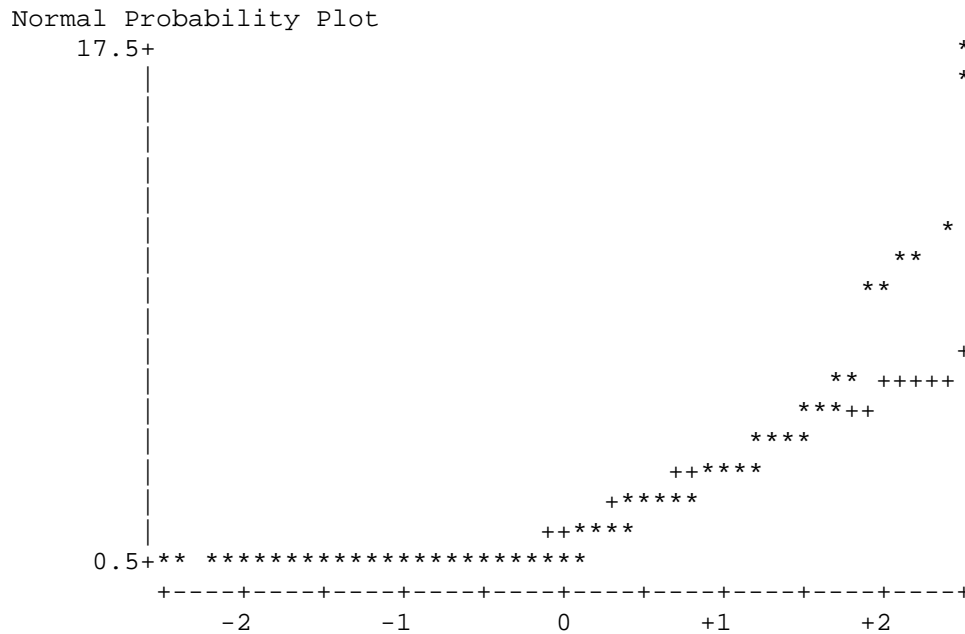
Histogram



# Boxplot



\* may represent up to 4 counts



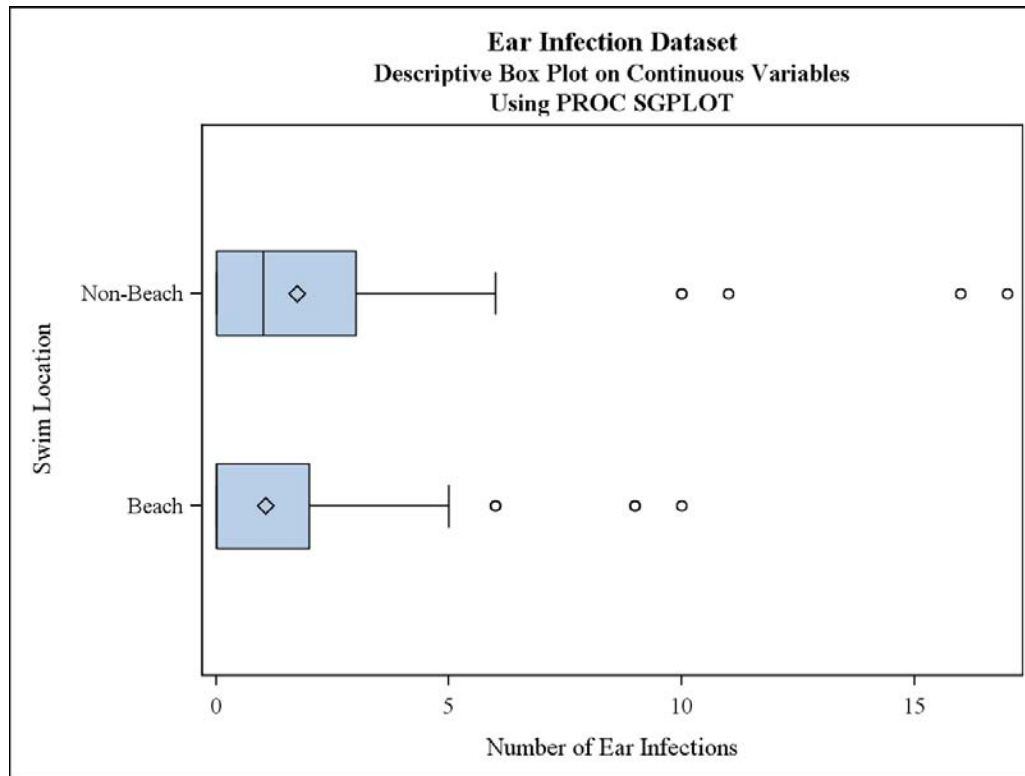
Example 3: To obtain the median in addition to the sample size, mean and standard deviation within each swim location (non-beach, beach) and to produce a box plot of the number of ear infections within each location we would use the following code:

```
proc means data=earinfection n mean std median;
  class location;
  var numearinfections;
run;

proc sgplot data=earinfection;
  hbox numearinfections / category=location;
  label numearinfections='Number of Ear Infections'
        location='Swim Location';
run;
```

Notice that the keywords for the specific statistics you are interested in calculating are given in the PROC MEANS statement. The CLASS statement tells SAS to calculate the requested statistics within the levels of the location variable. In the PROC SGPLOT procedure, a horizontal box plot was selected (hbox numearinfections) and the category=location option was used to ask for the horizontal box plot within the levels of the location variable. The output that is produced is as follows

Analysis Variable : numearinfections					
Numearinfections					
location	N Obs	N	Mean	Std Dev	Median
Non-Beach	140	140	1.7357143	2.6704100	1.0000000
Beach	147	147	1.0544218	1.9224069	0



Here the columns that are produced are the levels of the variable in the CLASS statement in the first column, the number of observations at each level of the CLASS variable in the second column, the number of non-missing observations for the continuous variable at each level of the CLASS variable in the third column, the mean, standard deviation and the median in the fourth, fifth and sixth columns, respectively. A description of the statistics from this output would be that the 140 non-beach swimmer had a mean number of ear infections of 1.73 (sd=2.67) and a median number of ear infections of 1. The 147 beach swimmers had a mean number of ear infections of 1.05 (SD=1.92) and a median of 0 ear infections.

Box plots are boxes made of a rectangle beginning at the first quartile (Q1) and ending at the third quartile (Q3), with the second quartile (or median, Q2) indicated with a vertical line in the middle of the box. The lines extending out of the box are the point just inside  $Q3+IQR$  (inter-quartile range) and  $Q1-IQR$ . The circles indicate potential outliers that are beyond the  $Q1-IQR$  and  $Q3+IQR$  lines. Examining the box plots, the non-beach swimmers tend to have more variability in the number of ear infections since the box is somewhat longer. The diamond located within each box is the mean number of ear infections in that particular swim location. The open circles within each swim

location indicate that there are four potential points for non-beach swimmers and 3 potential points for beach swimmers that may be potential outliers and could influence other inferential statistical analyses that may be performed. For non-beach swimmer, the median is shown. However, because the median and the first quartile for beach swimmers are both 0, there is no vertical line within the box. Both distributions of number of ear infections are positively skewed because the median (the vertical line) is either close to the first quartile (Q1) or at the first quartile.

### Inferential Statistics for Categorical Variables

To examine the relationship or association between two categorical variables, we use a chi-square test. Chi-square tests are performed using PROC FREQ and the basic SAS® code used is

```
proc freq data=datasetname;
  tables catvarrow*catvarcol / chisq
         plots=(freqplot(twoway=groupvertical
                        scale=percent));
run;
```

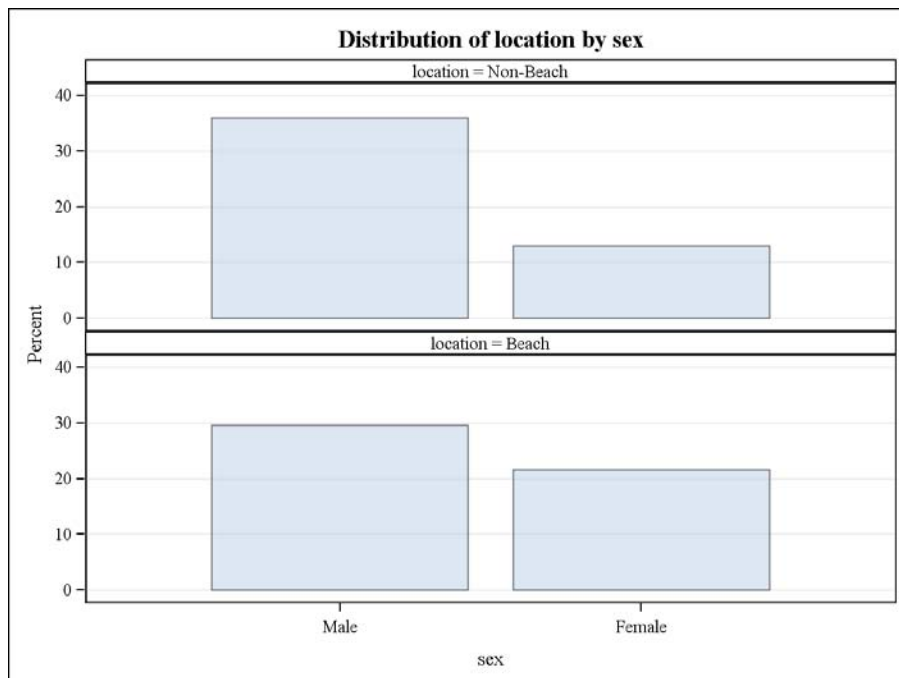
In the TABLES statement, we indicate that we want a two-way table to be calculated with the variable that will determine the rows of the two-way table being listed first (catvarrow) followed by an asterisk (\*) and then the variable that will determine the columns of the two-way table listed second (catvarcol). The option to calculate a chi-square test, chisq, is listed in the TABLES statement following the "/". The test that is produced by the chisq option compares the row or column percentages in the two-way table. The option plots=(freqplot(twoway=groupvertical scale=percent)) is used to plot the overall percentages, and not the row percentages, across the levels of the row variable. If there are several row and column variables you want a chi-square test performed for, you can have multiple row variables listed within parentheses followed by an asterisk and multiple column variables listed within parentheses and PROC FREQ will perform a chi-square test on all possible combinations of the row and column variables listed

```
ods graphics on;
proc freq data=datasetname;
  tables (catvarrow1 catvarrow2)*(catvarcol1 catvarcol2)
         / chisq plots=(freqplot(twoway=groupvertical
                        scale=percent));
run;
ods graphics off;
```

Example 4: For the earinfection data a series of chi-square tests were performed to examine the relationship between frequent ocean swimmers with age group and sex and between swim location with age group and sex. The SAS® code listed is given below. For presentation purposes, only the two-way table for location by sex will be presented.

```
ods graphics on;
proc freq data=earinfection;
  tables (fos location)*(agegroup sex) / chisq
        plots=(freqplot(twoway=grouphorizontal
                        scale=percent));
run;
ods graphics off;
```

Table of location by sex			
location(location)	sex(sex)		
Frequency Percent Row Pct Col Pct	Male	Female	Total
<b>Non-Beach</b>	103 35.89 73.57 54.79	37 12.89 26.43 37.37	140 48.78
<b>Beach</b>	85 29.62 57.82 45.21	62 21.60 42.18 62.63	147 51.22
<b>Total</b>	188 65.51	99 34.49	287 100.00



### Statistics for Table of location by sex

Statistic	DF	Value	Prob
Chi-Square	1	7.8705	0.0050
Likelihood Ratio Chi-Square	1	7.9373	0.0048
Continuity Adj. Chi-Square	1	7.1890	0.0073
Mantel-Haenszel Chi-Square	1	7.8431	0.0051
Phi Coefficient		0.1656	
Contingency Coefficient		0.1634	
Cramer's V		0.1656	

The two-way table for location by sex is printed first followed by a bar graph of the overall percent at each location by sex level in the total sample of 287 individuals. Each cell of the table lists four numbers, the frequency occurring in each cell, the overall percentage of number of observations in that cell over the total sample size, the row percentage of the number of observation in that cell over the total number in that particular row of the table, and the column percentage of the number of observations in that cell over the total number in that particular column of the table. If we are interested in whether the distribution of sex is different for beach and non-beach swimmers (i.e. do beach or non-beach swimmers have a different proportion of males), the correct percentage to examine in the two-way table is the row percentage, which is the third number listed in each cell of the table. The percent of non-beach swimmers who are male is 73.57% (103/140) and the number of beach swimmers who are male is 57.82% (85/147). The question now is whether these two percentages are statistically different and the answer can be found by looking at the chi-square test in the "Statistics Table of location by sex". Looking at row labeled Chi-Square, we find the chi-square test statistic value to be 7.8705 and the associated p-value is 0.0050. If the alpha level is 0.05, then we would conclude that there is a statistically significant difference between the proportion of males among beach and non-beach swimmers with non-beach swimmers being more likely to male than beach swimmers. Why are non-beach swimmers more likely to be male? Because the percentage of males among non-beach swimmers (73.57%) is higher than the percentage of males among beach swimmers (57.82%).

### Inferential Statistics for Continuous Variables

The most basic statistical test to examine differences in a continuous variable is a t-test. The type of t-test that is performed depends on the number of groups of individuals, a single group or two groups, in the data set. If there is a single group (i.e. the entire sample) and you want to test whether the sample mean of a continuous variable, `contvar`, is different from a particular null value, `h0=nullvalue`, a one-sample t-test is performed in SAS® using PROC TTEST as follows

```
proc ttest data=datasetname h0=nullvalue plots=summary;
  var contvar;
run;
```

The `plots(unpack)=summary` produces a histogram and a box plot of the data using the built in SAS® ODS graphics that are available. The default null value in SAS® for `h0=nullvalue` is 0.

If you are interested in testing whether the mean of the continuous variable, `contvar`, is different for a categorical variable having only two groups, `catvar`, then a two-sample `ttest` is performed

```
proc ttest data=datasetname plots=summary;
  class catvar;
  var contvar1 contvar2 contvar3;
run;
```

Here we add a `CLASS` statement with the name of the categorical variable containing only two levels. The `plots(unpack)=summary` will produce the histogram and box plot of the continuous variable for each level of the categorical variable. Note that in `PROC TTEST` the `CLASS` statement can only contain one continuous variable. If you want to run multiple two-sample t-tests on the same continuous variable for different categorical variables, you must specify multiple `PROC TTEST` statement sets. Multiple continuous variables can be specified in the `VAR` statement and the result is a two-sample t-test between the two groups in the `CLASS` statement for each continuous variable in the `VAR` statement.

Example 5: To test whether the number of ear infections is different from 2, a one-sample t-test using `PROC TTEST` is performed. The `h0=2` in the `PROC TTEST` statement indicates that we want to test whether the mean in the `VAR` statement (`var numearinfections`) is different from a null value of 2.

```
proc ttest data=earinfection h0=2 plots=summary;
  var numearinfections;
run;
```

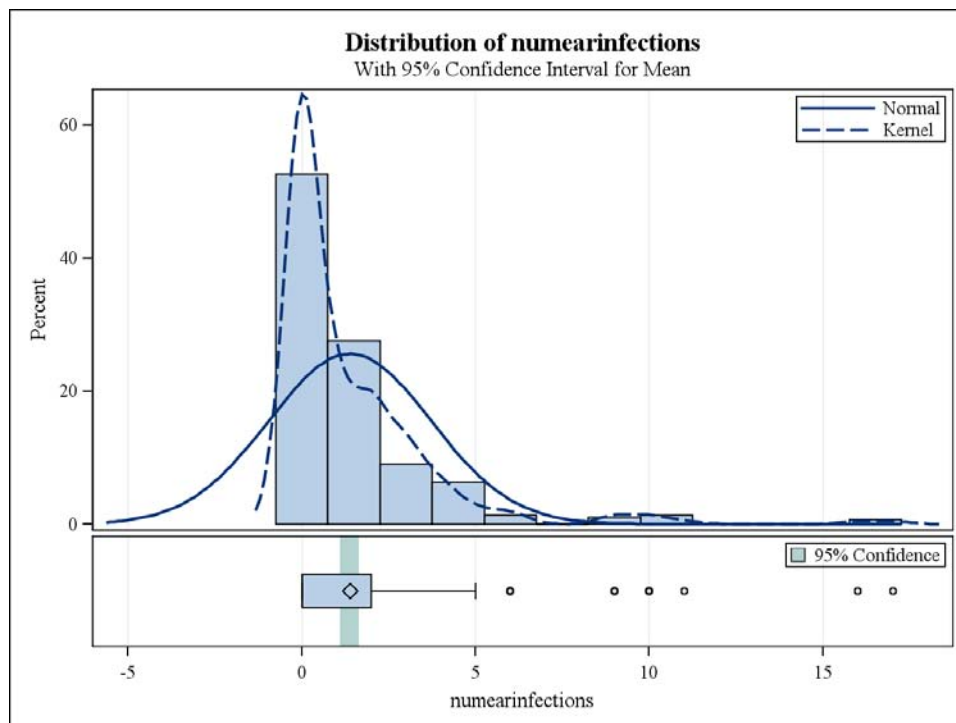
The output which is produced from the above SAS® code is

N	Mean	Std Dev	Std Err	Minimum	Maximum
287	1.3868	2.3385	0.1380	0	17.0000

Mean	95% CL Mean	Std Dev	95% CL Std Dev
1.3868	1.1151 1.6585	2.3385	2.1616 2.5473

DF	t Value	Pr >  t
286	-4.44	<.0001





Here the descriptive statistics for the number of ear infections in the 287 individuals in the sample including the mean, standard deviation, standard error, minimum and maximum are given in the first portion of the output. The next portion of the output contains the mean and a 95% confidence interval for the mean, followed by the standard deviation and a 95% confidence interval for the standard deviation. Finally, the one-sample t-test is given in the final table portion of the output with the t-value (the test statistic) as -4.44 and the associated p-value ( $Pr>|t|$ ) of <.0001. If the alpha level was 0.05, since the p-value is less than the alpha level we would find that the mean number of ear infections in the sample was significantly less than 2. We conclude that the mean is less than the null value of 2 because the mean of 1.3868 is less than 2. The plots that are produced by the plots option in the PROC TTEST statement are shown as well.

Example 6: As a final example using the ear infection data, we want to examine whether the mean number of ear infections is different between the two frequent ocean swimmer groups and between non-beach and beach swimmers. The SAS code to perform the two-sample t-test is

```
proc ttest data=earinfection plots(unpack)=summary;
  class fos;
  var numearinfections;
run;

proc ttest data=earinfection plots(unpack)=summary;
  class location;
  var numearinfections;
run;
```

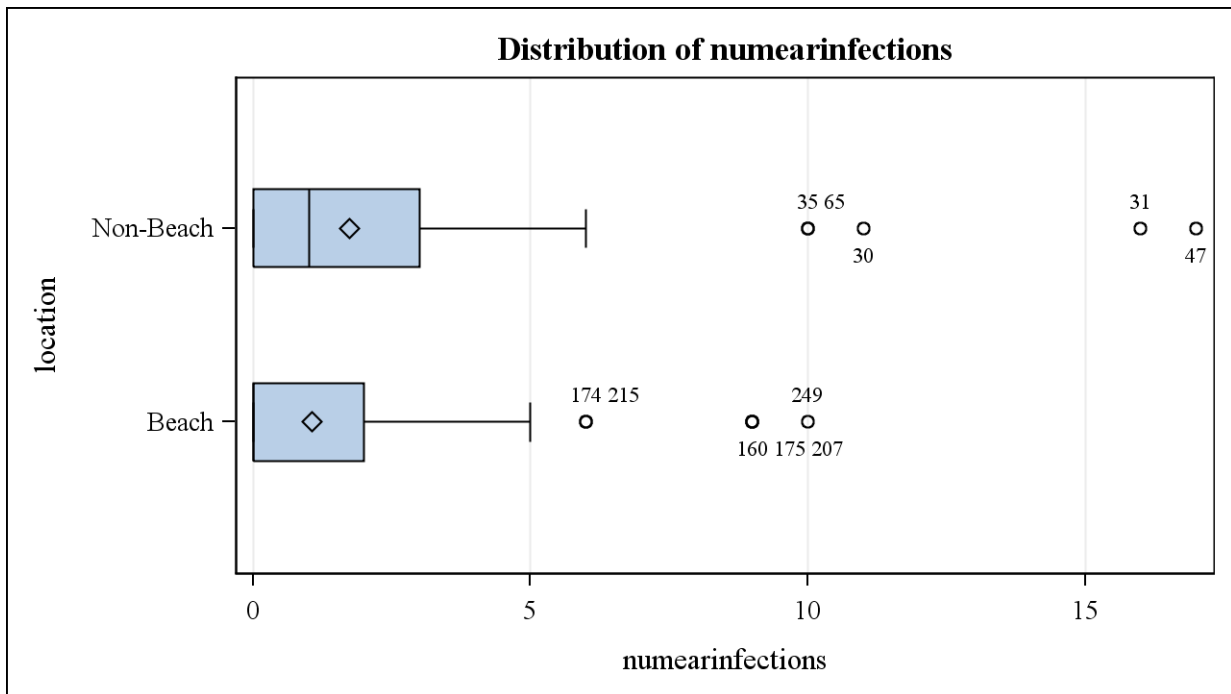
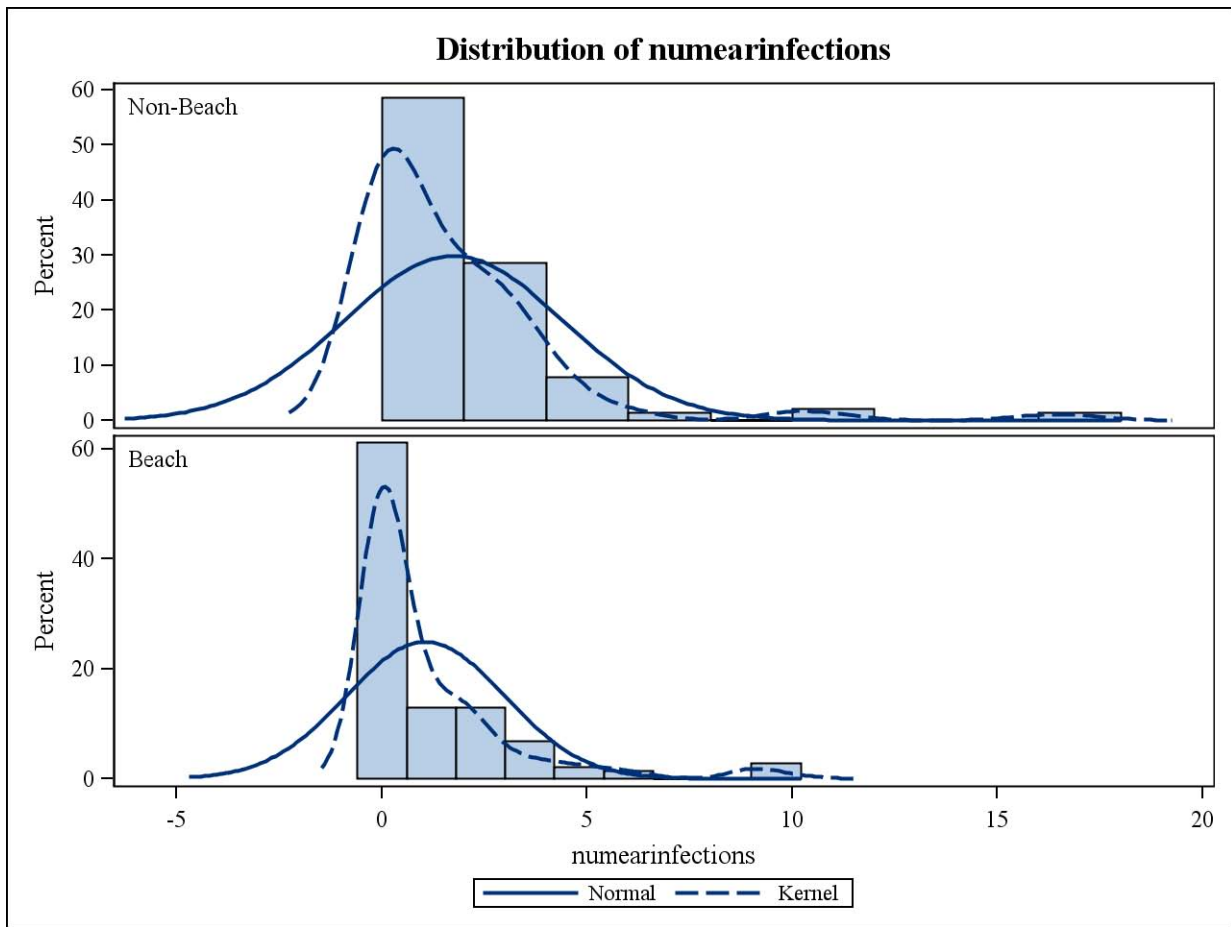
Notice that because there are two different categorical variables that I wish to examine difference between the mean number of ear infections, two different sets of PROC TTEST statements are given with the variable in the CLASS statement changing for each PROC TTEST statement set. The VAR statement contains the continuous variable name, and if there were other continuous variables that we wanted to examine for differences between frequent ocean swimmer status or between swim locations we could have listed them in the VAR statement. The output for differences between swim locations is given below

location	N	Mean	Std Dev	Std Err	Minimum	Maximum
Non-Beach	140	1.7357	2.6704	0.2257	0	17.0000
Beach	147	1.0544	1.9224	0.1586	0	10.0000
Diff (1-2)		0.6813	2.3176	0.2737		

location	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Non-Beach		1.7357	1.2895	2.1819	2.6704	2.3900	3.0260
Beach		1.0544	0.7411	1.3678	1.9224	1.7249	2.1714
Diff (1-2)	Pooled	0.6813	0.1426	1.2200	2.3176	2.1419	2.5248
Diff (1-2)	Satterthwaite	0.6813	0.1381	1.2245			

Method	Variances	DF	t Value	Pr >  t
Pooled	Equal	285	2.49	0.0134
Satterthwaite	Unequal	251.7	2.47	0.0142

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	139	146	1.93	<.0001



Descriptive statistics within the levels of the categorical variable listed in the CLASS statement, location, as well as the difference between the locations (non-beach minus beach) are given in the first section of the output. The next section gives the mean and 95% confidence interval for the mean and the standard deviation and 95% confidence interval for the standard deviation for each level of the CLASS variable and for the difference between the two levels (non-beach minus beach) assuming equal variances (Pooled row) or assuming unequal variances (Satterthwaite row). The next section gives the results of the two-sample t-tests assuming equal variance (Pooled) or unequal variance (Satterthwaite) with the test statistic in the "t value" column and the p-value under the "Pr>|t|" column. The last numeric section is a test for the Equality of Variances and whether you can assume that the variances in the two levels of the categorical variable are equal and use the Pooled t-test or whether you should assume that the variances in the two levels of the categorical variable are unequal and use the Satterthwaite t-test. The test statistic for the Equality of Variances test is given under the "F value" column and the corresponding p-value is under the "Pr>|F|" column. If the p-value for the Equality of Variances test is less than the alpha level, we assume unequal variances and perform the Satterthwaite t-test. If the p-value for the Equality of Variances test is greater than or equal to the alpha level we assume equal variances and perform a Pooled t-test. The last portion of the output is the histograms and box plots for each level of the categorical variable in the CLASS statement.

To know which t-test is appropriate to report, either the Pooled or Satterthwaite t-test, we examine the results of the Equality of Variances test first. In this instance the F test statistic is 1.93 and the corresponding p-value is <.0001 indicating we should assume unequal variances, since the p-value is less than the alpha level of 0.05. There has been a lot of debate in the statistical community as to which two-sample t-test is the most appropriate to present given that we rarely know whether populations variances are equal. At this date, many statisticians feel that it is always appropriate to report the unequal variance, Satterthwaite, t-test. While a little power is gained if variances are equal in the population and an equal variance, or pooled, t-test is performed, it is not a significant increase in the power and the conclusions for both the pooled and Satterthwaite tests are often the same. Very infrequently will the ultimate conclusion that there is a difference or that a difference failed to be detected using a Satterthwaite t-test be different than that of the pooled t-test.

So we finally get to the results of the two-sample t-tests. Since our Equality of Variance test indicated that we should perform the Satterthwaite t-test, the t-value we would report is 2.47 with a corresponding p-value of 0.0142. Since the p-value is less than the alpha level of 0.05, we conclude that the mean number of ear infections among non-beach swimmers is significantly higher than the mean number of ear infections among beach swimmers.

## CONCLUSIONS

Calculation of the descriptive statistics and inferential statistics presented for categorical variables and continuous variables are just the first steps in any statistical analysis. As well, knowing what the data look like using graphical methods can aid in determining whether and what additional and more rigorous statistical analyses should be performed.

The SAS® System provides many different procedures to produce descriptive statistics, chi-square tests, and t-tests as the first steps in a statistical analysis of the data set. Additionally, with the addition of ODS graphic and the Statistical Graphics with version 9.2, SAS® has made it easier to have a graphical representation of the variables in your data set.

## REFERENCES

1. Hand DJ, Daly F, Lunn AD, McConway KJ, Ostrowski E. The Handbook of Small Data Sets, Chapman and Hall, 1994.
2. SAS Institute Inc., SAS 9.2 Help and Documentation, Cary, NC: SAS Institute Inc., 2008.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

## CONTACT INFORMATION

Jennifer L. Waller, Ph.D.  
Georgia Health Sciences University  
Department of Biostatistics  
AE-1012  
Augusta, GA 30912-4900  
jwaller@georgiahealth.edu