Paper 2380-2015

Chi-Square and T-Tests Using SAS®: Performance and Interpretation

Jennifer L. Waller and Maribeth H. Johnson
Georgia Regents University, Augusta, Georgia

ABSTRACT

Data analysis begins with data clean up, calculation of descriptive statistics and the examination of variable distributions. Before more rigorous statistical analysis begins, many statisticians perform basic inferential statistical tests such as chi-square and t-tests to assess unadjusted associations. These tests help to guide the direction of the more rigorous analysis. How to perform chi-square and t-tests will be presented. We will examine how to interpret the output, where to look for the association or difference based on the hypothesis being tested, and propose next steps for further analysis using example data.

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, distributional shape, 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 is necessary to report from the output is essential. SAS has a whole host of statistical analysis tools for both descriptive and inferential statistical analyses.

DESCRIPTIVE STATISTICS

Descriptive statistics 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, 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 assume, 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 feet 6.55214328754 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 to determine the middle of the distribution include means, medians, and modes and to examine the spread of the distribution we can use quartiles, variances, standard deviations, coefficients of variation, ranges, inter-quartile ranges, minimums and maximums. Statistics such as kurtosis and skewness can be used to describe the shape of the distribution.

INFERENTIAL STATISTICS

Inferential statistics are used to examine data for differences, associations, and relationships to answer research questions (or 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 categorical variables. The levels of categories for each variable can be two or more. The types of research questions that can be addressed are:

- Are two (or more) proportions in a single categorical variable different from hypothesized population values?
- Is there an association or dependence between two categorical variables?
- Are two (or more) proportions different from each other?
- Is there a difference of risk (or odds) of an event between two groups?

One- and Two-Sample T-tests

T-tests are used to examine differences between means. The types of research questions that can be addressed are:

- Is the sample mean of a single continuous variable in a single group of individuals different from a particular hypothesized population value?
- Are the sample means of a single continuous variable different between two different groups of individuals?

SAS PROCEDURES

Inferential Statistics for Categorical Variables

Chi-Square Test for Goodness of Fit (One Sample Test of Proportions)

To examine whether two or more proportions for a single categorical variable fit a specified set of values, a chi-square Goodness of Fit (GOF) test is performed. Using PROC FREQ, the SAS code used is

```
proc freq data=datasetname;
  tables catvar / chisq testp=(p1 p2 p3…);
run;
```

In the TABLES statement we indicate that we want a one-way table by listing the categorical variable, catvar. Following the "/" the options to use chisq to perform the chi-square GOF test and testp=(p1 p2 p3….) where p1, p2, p3 and so on are the population percentages that are assumed for each level of the categorical variable. The population percentages p1, p2, p3 and so on must add up to 100 percent.

Chi-Square Tests for Equality of Two Proportions or Association of Two Categorical Variables

To examine whether the equality of two proportions or the association between two categorical variables exists, 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 measures
         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 measures option produces the odds ratio or relative risk (column dependent) and corresponding 95% confidence intervals. The option plots=(freqplot(twoway=groupvertical scale=percent)) is used to plot the *overall percentages,* and not the row percentages, across the column levels of the row variable.

In SAS 9.4, to obtain plots of column percentages by groups you can use the following code

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

For row percentages change the groupby= option to groupby=row. This will not work in SAS 9.3.

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 1: Data were collected on 287 swimmers (Hand et al., 1994). 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. To illustrate a chi-square test of goodness of fit, we will use the age group variable. We want to examine whether the sample frequency distribution fits a hypothesized population distribution of 50% in the youngest age group and 25% in the two older age groups. The SAS® code is listed below and the output (Output 1) follows.

```
proc freq data=earinfection;
  tables agegroup / nocum chisq testp=(50 25 25);
run;
```

agegroup			
agegroup	**Frequency**	**Percent**	**Test Percent**
15-19 yrs	140	48.78	50.00
20-24 yrs	79	27.53	25.00
25-29 yrs	68	23.69	25.00

Chi-Square Test for Specified Proportions	
Chi-Square	1.0139
DF	2
Pr > ChiSq	0.6023

Output 1: Chi-square Test of Goodness of Fit on Age Group

In the first table of Output 1, the sample frequency distribution is given. For any frequency distribution table, the table lists the levels of the categorical variable in the first column, the second column contains the frequency or number of individuals in that specific category, and the third column contains the percent of individuals in that specific category. In the first table the sample percentages are highlighted in blue and then the population distribution we are testing against is highlighted in green. The GOF test statistic and p-value are shown in the next table of Output 1 and are highlighted in yellow. The p-value for the GOF test indicates that the sample distribution is not significantly different at an alpha level of 0.05, since p=0.6023, than the assumed population distribution of 50% in the youngest age group followed by 25% in the two older age groups.

Example 2: Using the ear infection data, to test that the distribution of age groups is different in non-beach and beach swimmers, i.e. is there an association between swim location and age, the SAS® code below was used.

```
proc freq data=earinfection;
   tables location*agegroup / chisq;
run;
```

Table of location by agegroup				
location(location)	agegroup(agegroup)			
Frequency Percent Row Pct Col Pct	15-19 yrs	20-24 yrs	25-29 yrs	Total
Non-Beach	71 24.74 50.71 50.71	46 16.03 32.86 58.23	23 8.01 16.43 33.82	140 48.78
Beach	69 24.04 46.94 49.29	33 11.50 22.45 41.77	45 15.68 30.61 66.18	147 51.22
Total	140 48.78	79 27.53	68 23.69	287 100.00

Statistic	DF	Value	Prob
Chi-Square	2	9.1202	0.0105
Likelihood Ratio Chi-Square	2	9.2542	0.0098
Mantel-Haenszel Chi-Square	1	3.4825	0.0620
Phi Coefficient		0.1783	
Contingency Coefficient		0.1755	
Cramer's V		0.1783	

Fisher's Exact Test	
Table Probability (P)	1.292E-04
Pr <= P	0.0109

Output 2: Chi-square Test of Differences of Association Between Age Groups and Beach Versus Non-Beach Swimmers

Each cell of the first table in the output (Output 2) 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 age is different for beach and non-beach swimmers (i.e. is there an association between swim location and age), 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 row percentages are highlighted in yellow. The chi-square test statistic and p-value are given in the second table. The chi-square test statistic is highlighted in green and is in the column labeled "Value". The p-value is highlighted in blue and is in the column labeled "Prob". The p-value=0.0105 indicates that the association is statistically significant at the 0.05 alpha level. To determine what the association is, we look at the row percentages. Notice that the largest differences between percentages in non-beach and beach swimmers occurs in the 25-29 year age group. Beach swimmers (30.6%) were more likely to be older than non-beach swimmers (16.3%). While we could have produced the plot, in this instance the plot that is produced by PROC FREQ is not helpful as it plots overall percentages and not row percentages. The table with Fisher's Exact test is not needed as a Warning message is not given right above the Fisher's Exact table results.

Example 3: What if our question was whether the proportion of beach swimmers in each age group was different? The SAS code to use is exactly the same as what is used in Example 2. As well, the output produced (Output 3) is exactly the same. What is different is what you use from the output.

Table of location by agegroup

location(location)	agegroup(agegroup)			
Frequency Percent Row Pct Col Pct	15-19 yrs	20-24 yrs	25-29 yrs	Total
Non-Beach	71 24.74 50.71 50.71	46 16.03 32.86 58.23	23 8.01 16.43 33.82	140 48.78
Beach	69 24.04 46.94 49.29	33 11.50 22.45 41.77	45 15.68 30.61 66.18	147 51.22
Total	140 48.78	79 27.53	68 23.69	287 100.00

Statistic	DF	Value	Prob
Chi-Square	2	9.1202	0.0105
Likelihood Ratio Chi-Square	2	9.2542	0.0098
Mantel-Haenszel Chi-Square	1	3.4825	0.0620
Phi Coefficient		0.1783	
Contingency Coefficient		0.1755	
Cramer's V		0.1783	

Output 3: Chi-Square Test for Differences of Proportions in Age Groups between Beach Swimmers.

Because we are now interested in showing that the age groups have different percentages of beach swimmers, we use the column percentages, which are highlighted in yellow in the two-way table. The test statistic and p-value do not change, but the interpretation does. Again, the test is statistically significant at the 0.05 alpha level indicating that

at least one proportion is not equal to the others. The expectation is that the percent of beach swimmers should be the same in each of the three age groups. Looking at the column percentages, 49.2% were beach swimmers in the 16-19 year age group, 41.8% were beach swimmers in the 21-24 year age group, and 66.2% were beach swimmers in the 25-29 year age group. Once again we see that the oldest age group has more beach swimmers than the two younger age groups. The result that you present, row percentages or column percentages, will depend on the research question.

Example 4: A study is run in which 900 individuals are sampled and each is classified as to whether they had contracted the flu during the last year and whether they had been inoculated for the flu. The research question is whether inoculation status and contracting the flu are associated and what is the magnitude of the association. To estimate independent sample risk ratios and odds ratios the data should be arranged with the "disease" status as the columns and the "exposed" status as the rows. "Disease" is the event of interest and for this example it is flu status.

Many times aggregated data are presented in a two-way table (Table 1). Data can be entered in aggregate form similar to the DATA step below. The number of individuals in each cell of the two-way table is the information that is entered (ifcount) and that number is used in the PROC FREQ with a WEIGHT statement.

	Flu Status	
Inoculation Status	Flu	No Flu
Inoculated	150	200
Not Inoculated	300	250

Table 1: Innoculation and Flu Data

```
data flu;
  input inoculation $ 1-14 flu $ 17-22 ifcount 25-27;
datalines;
Inoculated          Flu      150
Inoculated          No Flu   200
Not Inoculated      Flu      300
Not Inoculated      No Flu   250
;
run;


proc freq data=flu;
  tables inoculation*flu / chisq measures;
  weight ifcount;
title 'Chi-Square Test of Association of Innoculation and Flu';
run;
```

Table of inoculation by flu			
inoculation	flu		
Frequency Percent Row Pct Col Pct	Flu	No Flu	Total
Inoculated	150 16.67 42.86 33.33	200 22.22 57.14 44.44	350 38.89
Not Inoculated	300 33.33 54.55 66.67	250 27.78 45.45 55.56	550 61.11
Total	450 50.00	450 50.00	900 100.00

Statistic	DF	Value	Prob
Chi-Square	1	11.6883	0.0006
Likelihood Ratio Chi-Square	1	11.7191	0.0006
Continuity Adj. Chi-Square	1	11.2255	0.0008
Mantel-Haenszel Chi-Square	1	11.6753	0.0006
Phi Coefficient		-0.1140	
Contingency Coefficient		0.1132	
Cramer's V		-0.1140	

Estimates of the Relative Risk (Row1/Row2)			
Type of Study	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	0.6250	0.4770	0.8189
Cohort (Col1 Risk)	0.7857	0.6810	0.9065
Cohort (Col2 Risk)	1.2571	1.1051	1.4301

Output 4: Chi-Square Test of Association between Inoculation and Flu with Measures of Association

In the Output 4 above, the chi-square test statistic (highlighted in green) and corresponding p-value (highlighted in blue) indicate that there is a statistically significant association between inoculation and contracting the flu. The odds ratio (OR) and relative risk (RR) estimates and corresponding 95% confidence interval are given in the final table in the output "Estimates of the Relative Risk (Row1/Row2)". A 95% confidence interval for an estimate that does not contain 1 also indicates a significant association at the 0.05 alpha level.

Whether you report an odds ratio or a relative risk depends on the nature of your study. If you sampled on whether or not someone had the flu or did not have the flu and asked them to recall whether they had been inoculated or not, you would report the Case-Control (Odds Ratio) highlighted in yellow. The interpretation of the odds ratio would be that the odds of contracting the flu are 0.625 times as likely for those who were inoculated than those who did not receive the inoculation. An odds ratio less than one indicates that the inoculation was protective for the disease.

If you sampled those who were inoculated or not inoculated and then followed them to determine whether they contracted the flu or not you would report the Cohort (Col1 Risk) or Cohort (Col2 Risk) relative risk estimates highlighted in purple and grey. The relative risk you report is dependent on whether column 1 of the two-way table is

what you are at risk for or column 2 of the two-way table is what you are at risk for. In this example, we are interested in the risk of contracting the flu, which is in column 1 of the two-way table so we would report the Cohort (Col1 Risk) relative risk. The interpretation of the relative risk would be that those who were inoculated were 0.7856 times as likely to contract the flu than those who were not inoculated. The effect is protective because exposure to the inoculation implies reduced risk of the flu.

Future Steps for Categorical Variables

Determining the magnitude of the association and which measure of association to present, the OR or RR, is dependent on the study design. Odds ratios can also be calculated for cross-sectional (prevalence) studies or for prospective cohort studies. The two-way table set up depends on the study design, i.e. whether the disease status is the row or column variable; the calculation of the odds ratio is the same, but the interpretation is different. Risk ratios can be calculated directly only for cohort studies. Odds ratios, as already discussed, can be calculated not only for cohort studies but also for case control studies. There may be an assumption that the OR gives a close approximation of the RR in all situations. The potential problem is that in some situations the OR may exaggerate the measure of association that would be determined by a RR.

In observational studies, certain factors associated with both the outcome and the exposure can distort the association between the exposure and the outcome. When investigators are aware of and measure these factors – called confounders – they can use certain analytic techniques to adjust for their effects to provide a better estimate of the effect of the exposure itself. Techniques (such as Cochran-Mantel-Haenszel chi-square test or logistic regression) that are commonly used to adjust for confounders yield odds ratios (rather than risk ratios) between each confounder variable and the outcome as well as between the exposure of interest and the outcome. Logistic regression has an added benefit of allowing one to assess the association of multiple independent variables, either categorical or continuous, at the same time on a single, dichotomous outcome variable. A Cochran-Mantel-Haenszel test can be run using the CMH option in the TABLES statement in PROC FREQ.

Additionally, if you are interested in agreement of categorical variables measured at two different time points, a McNemar's test of symmetry for dichotomous variables and Bowker's test of symmetry for categorical variables with 3 or more levels can be used. The AGREE option can be used in the TABLES statement in PROC FREQ to obtain these tests.

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. There is an assumption that the population or populations follow a normal distribution for the variables of interest.

One Sample t-test

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=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.

Two Sample t-test

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 t-test 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=summary will produce the histogram and box plot of the continuous variable for each level of the categorical variable. If you want the plots of the histogram and box plots to be produced in separate boxes, you can use

plots(unpack)=summary, where (unpack) is used to separate the plots. Note that in PROC TTEST the CLASS statement can only contain one dichotomous, categorical 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: Data from 200 births were collected to examine the differences in birth weight (in grams) between mothers who smoked and mothers who did not. We first want to test whether the birth weight of babies born to non-smokers in this sample was different from the population norm that defines low versus normal birth weight, 2500 grams. A one-sample t-test using PROC TTEST is performed. The null value, h0=2500, in the PROC TTEST statement indicates that we want to test whether the mean of the variable listed in the VAR statement is different from a null value of 2500.

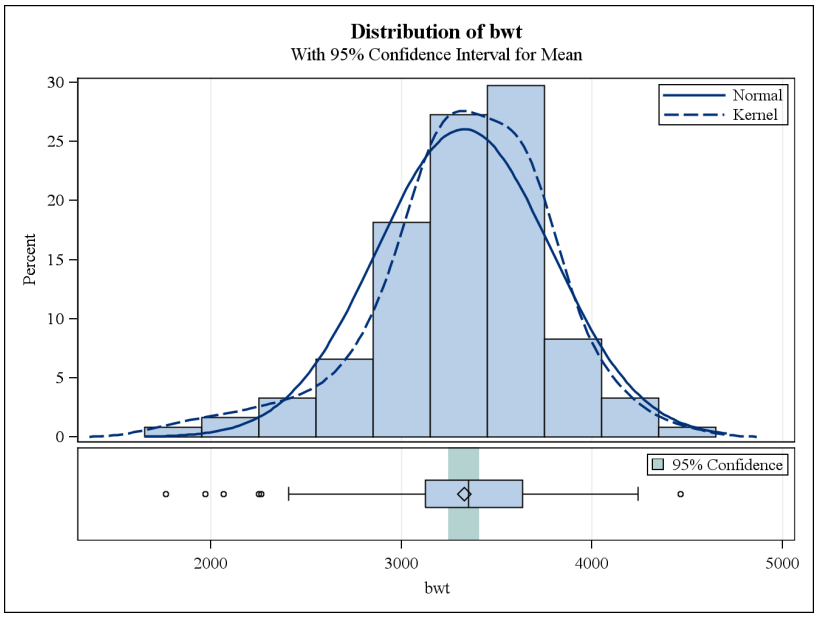
```
proc ttest data=babies h0=2500 plots=summary;
  var bwt;
run;
```

The output (Output 5) which is produced from the above SAS® code is

N	Mean	Std Dev	Std Err	Minimum	Maximum
121	3327.5	459.6	41.7837	1762.0	4463.0

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
3327.5	3244.7	3410.2	459.6	408.1	526.1

DF	t Value	Pr > |t|
120	19.80	<.0001



Output 5: One-Sample T-test for the Mean Birth Weight of Babies of Non-Smoking Mothers Being Different from 2500g.

In Output 5 the descriptive statistics for birth weight in the 200 individuals in the sample including the mean (highlighted in yellow), standard deviation (highlighted in green), standard error, minimum and maximum are given in the first portion of the output (Output 5). The next portion of the output contains the mean (in yellow) and a 95% confidence interval for the mean, followed by the standard deviation (in green) 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 highlighted in blue) as 19.80 and the associated p-value (Pr>|t| highlighted in pink) of <.0001. If the alpha level is 0.05, since the p-value is less than the alpha level we would find that the mean birth weight of babies born to non-smoking mothers in the sample was significantly greater than 2500 grams. We conclude that the mean is greater than the null value of 2500 because the mean of 3327.5 is greater than 2500 grams. The plots that are produced by the plots option in the PROC TTEST statement are shown as well. The plots of the data are fairly symmetric and seem to follow a normal distribution so a t-test is appropriate.

Example 6: As a final example using the babies data, we want to examine whether the mean birth weight is different between babies born to smoking mothers versus babies born to non-smoking mothers. The SAS code to perform the two-sample t-test is

```
proc ttest data=babies plots(unpack)=summary;
  class momsmoker;
  var bwt;
run;
```

The VAR statement contains the continuous variable name, and if there were other continuous variables that we wanted to examine for differences between smoking and non-smoking mothers we could have listed them in the VAR statement. The output for differences between swim locations is given below (Output 6).

Descriptive statistics within the levels of the categorical variable, momsmoker, listed in the CLASS statement as well as the difference between the momsoker (smokers minus non-smokers) are given in the first section of the output (Output 6). Means are highlighted in yellow and standard deviations are highlighted in blue. 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 (smoker minus non-smoker) 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 to determine 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, highlighted in pink, 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 Sattherthwaite 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.

The plots show that the data are fairly symmetric in each group (smoker and non-smoker) and appear to follow a normal distribution. As well, the kernel density (the dotted line) is similar to the normal density (the solid line), which is another indication that the data follow a normal distribution. So a two-sample t-test is appropriate to perform.

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 1.05 and the corresponding p-value is 0.8078, highlighted in pink, indicating we should assume equal variances, since the p-value is greater 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, Sattherthwaite, 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 Pooled t-test, the t-value we would report is -3.50 with a corresponding p-value of 0.0006, highlighted in

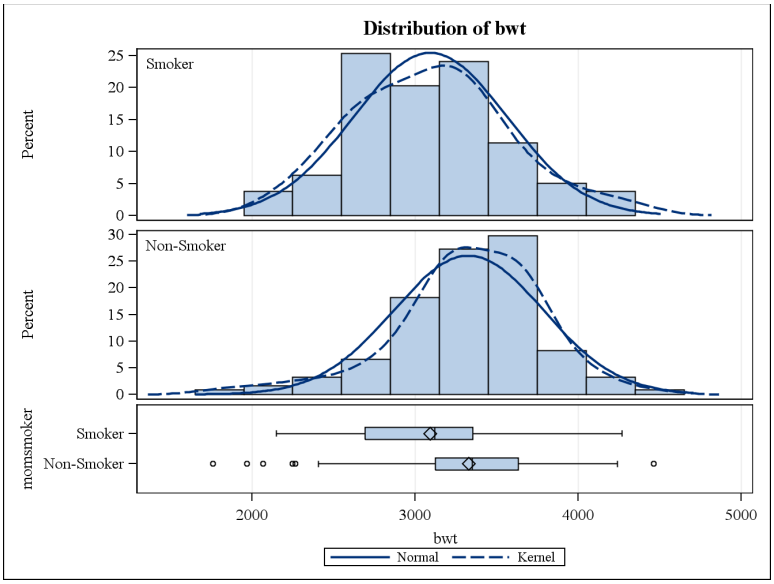
green. Since the p-value is less than the alpha level of 0.05, we conclude that the mean birth weight among babies born to smokers (3092.5g) is significantly lower than the mean birth weight among babies born to non-smokers (3327.5g).

momsmoker	N	Mean	Std Dev	Std Err	Minimum	Maximum
Smoker	79	3092.5	470.6	52.9429	2150.0	4270.0
Non-Smoker	121	3327.5	459.6	41.7837	1762.0	4463.0
Diff (1-2)		-234.9	464.0	67.1108		

momsmoker	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Smoker		3092.5	2987.1	3197.9	470.6	406.9	558.0
Non-Smoker		3327.5	3244.7	3410.2	459.6	408.1	526.1
Diff (1-2)	Pooled	-234.9	-367.3	-102.6	464.0	422.4	514.6
Diff (1-2)	Satterthwaite	-234.9	-368.1	-101.8			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	198	-3.50	0.0006
Satterthwaite	Unequal	164.06	-3.48	0.0006

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	78	120	1.05	0.8078



Output 6: Two-Sample T-test for Differences in Birth Weight of Babies Born to Smoking and Non-Smoking Mothers.

Future Steps for Continuous Variables

One- and two-sample t-tests are certainly not an exhaustive list of the statistical techniques that can be used for continuous variables. The type of study design including the number of groups, number of measurement times and question being asked all guide the type of statistical analysis to be performed. For example if a single continuous variable was measured at two different time points and we were interested in whether the mean changed from time point 1 to time point 2, a paired t-test can be used using the PAIRED statement in PROC TTEST.

If your categorical variable has three or more levels (e.g. the age group variable in the ear infection data set) and it is of interest to test that one or more means is different a one-way analysis of variance (ANOVA) is the appropriate statistical technique to use. PROC GLM can be used to examine whether differences between the groups exist.

If the data violate any assumptions for a t-test (e.g. non-normal distribution of the continuous data, not bell shaped) or the study design contains several categorical factors of interest or measurements taken at several different time points, we would suggest that you consult with an experienced statistician on the most appropriate analysis to use for your study design. The analysis can become very complex very quickly and applying the most appropriate statistical technique will allow you to make the most appropriate conclusions of your data.

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 Graphics and the Statistical Graphics with version 9.3 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.

CONTACT INFORMATION

Jennifer L. Waller, Ph.D.
Georgia Regents University
Department of Biostatistics & Epidemiology
AE-1012
Augusta, GA 30912-4900
jwaller@gru.edu

Maribeth H. Johnson, M.S.
Georgia Regents University
Department of Biostatistics & Epidemiology
AE-1011
Augusta, GA 30912-4900
majohnso@gru.edu