

STAT 515 - SAS Templates

This page contains sample code for using SAS in the course STAT 515 with the text *Statistics, 9th Edition* by McClave and Sincich. It is designed so that you can read along with material below, and copy and paste the SAS code (usually looking like type writer font) directly from the web page into the SAS program editor window. The introduction section is written assuming you have no previous experience using SAS. In the later sections the examples usually correspond to portions of the text book, and it would be helpful to have the book with you. In general, you will be able to use these templates to do your homework assignments after modifying them by entering your own data and making sure that all of the names match up.

Originally created by B. Habing - Last updated: 1/2/04

Introduction to SAS

- [Normal, t, Chi², and F Tables](#) - Section 5.3 and Supplement 6
- [CIs for Means and Variances](#) - Section 7.2 and Supplement 7
- [CI for One Proportions](#) - Section 7.3
- [One Sample t-test](#) - Section 8.4
- [Test for One Proportion](#) - Section 8.5
- [Two Sample t-test](#) - Section 9.1
- [Paired t-test](#) - Section 9.2
- [For Two Proportions](#) - Section 9.3
- [One-way ANOVA](#) - Section 10.2
- [Simple Linear Regression](#) - Chapter 11 and Supplements
- [Goodness of Fit Test](#) - Section 13.2
- [Two-Way Contingency Table](#) - Section 13.3 and Supplement 13.3

[Text Book Data Sets on CD](#)

[Computer Trouble?](#)

[SAS Won't Start?](#)

[Graphs Printing Small in Word?](#)

Introduction to SAS:

To begin with you should open up the program SAS. If there is a SAS icon on the screen already, you can just double click that to start the program. If there is not a SAS icon on the screen already, you need to use the start menu. Click the `Start` rectangle in the bottom left of the screen with your left mouse button. This should open up a list of options above the `Start` rectangle, and you should slide the mouse arrow up to the one labeled `Programs`. This should open up a list of more options to the right and you should keep sliding along until you get to the one labeled `The SAS System for Windows`. Once you get there click on it. This should start up SAS.

There are three main windows that are used in SAS. The Log window, the Program Editor window, and the Output window. The Log and Program Editor window are the two on the screen when you start up the program. The Output window isn't visible yet because you haven't created anything to output. If you happen to lose one of these windows they usually have a bar at the bottom of the SAS window. You can also find them

under the `View` menu.

The Program Editor is where you tell SAS what you want done. The Output window is where it puts the results, and the Log window is where it tells you what it did and if there are any errors. It is important to note that the Output window often gets very long! You usually want to copy the parts you want to print into MS-Word and print from there. It is also important to note that you should check the Log window everytime you run anything. The errors will appear in maroon. Successful runs appear in Blue.

Hitting the [F3] key at the top of your keyboard will run the program currently in the Program Editor window. You can also run programs by clicking on the little image of the runner in the list of symbols near the top of the SAS program screen.

In older editions of SAS, running the program will erase whatever was written in the Program Editor window. To recall whatever was there, make sure you are in that window, and hit the [F4] key.

If you keep running more programs it will keep adding it all to the Output window. To clear the Output window, make sure you are in that window, and choose `Clear text` under the `Edit` menu.

If you happen to lose a window, try looking under the `View` menu.

The following is the SAS code for entering data about the starting salaries of a group of bank employees. The data consists of the beginning salaries of all 32 male and 61 female entry level clerical workers hired between 1969 and 1977 by a bank. (Yes, they really are annual salaries!!!) The data is reported in the book *The Statistical Sleuth* by Ramsey and Schafer, and is originally from: H.V. Roberts, "Harris Trust and Savings Bank: An Analysis of Employee Compensation" (1979), Report 7946, Center for Mathematical Studies in Business and Economics, University of Chicago Graduate School of Business.

The data is formatted in two columns, the first is the starting salary, the second is an id code, `m` for male, and `f` for female. (You can just cut and paste the code below starting with `OPTIONS` and ending with the `;` after all the numbers into the Program Editor window.)

```

OPTIONS pagesize=50 linesize=64;

DATA bankdata;

INPUT salary gender $ @@;

LABEL salary = "Starting Salary"
       gender = "m=male, f=female";

CARDS;

3900 f 4020 f 4290 f 4380 f 4380 f 4380 f
4380 f 4380 f 4440 f 4500 f 4500 f 4620 f
4800 f 4800 f 4800 f 4800 f 4800 f 4800 f
4800 f 4800 f 4800 f 4800 f 4980 f 5100 f
5100 f 5100 f 5100 f 5100 f 5100 f 5160 f
5220 f 5220 f 5280 f 5280 f 5280 f 5400 f

```

```

5400 f 5400 f 5400 f 5400 f 5400 f 5400 f
5400 f 5400 f 5400 f 5400 f 5400 f 5520 f
5520 f 5580 f 5640 f 5700 f 5700 f 5700 f
5700 f 5700 f 6000 f 6000 f 6120 f 6300 f
6300 f 4620 m 5040 m 5100 m 5100 m 5220 m
5400 m 5400 m 5400 m 5400 m 5400 m 5700 m
6000 m 6000 m 6000 m 6000 m 6000 m 6000 m
6000 m 6000 m 6000 m 6000 m 6000 m 6000 m
6000 m 6300 m 6600 m 6600 m 6600 m 6840 m
6900 m 6900 m 8100 m
;

```

Note that `_most_` lines end with a semi-colon, but not all. SAS will crash if you miss one, but usually the log window will tell you where the problem is. An extra/missing semi-colon is probably the single largest reason for a SAS program crashing.

The `OPTIONS`

line only needs to be used once during a session. It sets the length of the page and the length of the lines for viewing on the screen and printing. The font can be set by using the `options` choice under the `Tools` menu along the top of the screen. When you cut and paste from SAS to a word processor, the font 10 point Courier New works well.

The `DATA`

line defines what the name of the data set is. The name should start with a letter, have no spaces, and only letters, numbers, and underscores. The `INPUT` line gives the names of the variables, and they must be in the order that the data will be entered. The `$` after `gender` on the `INPUT` line means that the variable `gender` is qualitative instead of quantitative. The `@@` at the end of the `INPUT` line means that the variables will be entered right after each other on the same line with no returns. (Instead of needing one row for each person.)

If we hit F3 at this point to enter what we put above, nothing new will appear on the output screen. This is no big surprise however once we realize that we haven't told SAS to return any output! The code below simply tells SAS to print back out the data we entered.

```

PROC PRINT DATA=bankdata;
TITLE "Gender Equity in Salaries";
RUN;

```

The only difficulty we have now, is that it would be nice to look at both the men and women separately, so we need to be able to split the data up based on what's in the second column. The following lines will make two separate data sets `male` and `female`, and then print out the second one to make sure it is working right:

```

DATA male;
SET bankdata;

```

```
KEEP salary;

WHERE gender='m';

RUN;
```

```
DATA female;

SET bankdata;

KEEP salary;

WHERE gender='f';

RUN;
```

```
PROC PRINT DATA=female;

TITLE "Female Salaries";

RUN;
```

Whenever you have a `DATA` line, that means you are creating a new dataset with that name. The `SET` line tells it that we are making this new data set from an old one. The `KEEP` line says the only variables we want in this new data set are the ones on that line. The lines after that say any special commands that go into the making of the new data set. In this case the `WHERE` command is used to make sure we only keep one gender or the other. Later we will see examples of making datasets that involve using mathematical functions. In any case, it should be pretty straight-forward when you just stop and read through what the lines say.

The most basic procedure to give out some actual graphs and statistics is `PROC UNIVARIATE`:

```
PROC UNIVARIATE DATA=female PLOT FREQ ;

VAR salary;

TITLE 'Summary of the Female Salaries';

RUN;
```

The `VAR` line says which of the variables you want a summary of. Also note that the graphs here are pretty awful. The `INSIGHT` procedure will do most of the things that the `UNIVARIATE` procedure will, and a lot more. The one thing it won't do is be open to programming it to do new things. Later in the semester we'll see how some of the other procedures in SAS can be used to do things that aren't already programmed in.

```
PROC INSIGHT;

OPEN female;

DIST salary;

RUN;
```

You can cut and paste the graphs from `PROC INSIGHT` right into Microsoft Word. Simply click on the border of the box you want to copy with the left mouse button to select it. You can then cut and paste like normal.

(Before printing the graphs, you might want to **adjust** them so that they print at the correct size.)

While in PROC INSIGHT, clicking on the arrow in the bottom corner of each of the boxes gives you options for adjusting the graphs format. To quit PROC INSIGHT, click on the X in the upper right portion of the spreadsheet.

The graph at the top is a *Box Plot* or a *Box and Whisker Plot*. The wide line in the middle is the median. The edges of the box are the 25th percentile = 1st Quartile = Q1 and the 75th percentile = 3rd Quartile = Q3. A percentile is the point where at least that percent of the observations are less than the percentile and (100-that percent) are greater.

The distance between the 75th percentile and 25th percentile is called the Interquartile Range = IQR = Q3-Q1. It is a measure of how spread out the data is, the larger the IQR the more spread out the middle of the data is.

Extending beyond the edge of the box are the whiskers. The whiskers are allowed to go up to 1.5 IQRs from the edge of the box, and must end at a data point. The values beyond that are displayed as dots. They are possible outliers. We can see from this box plot that about 1/4 of the data is less than 4800, about a 1/4 is between 4800 and 5220, about a 1/4 is between 5220 and 5400, and the last 1/4 is between 5400 and 6300. (Click on the boxes in the box plot to see the numbers!) We also see that there are no really extreme points.

The box plot has the advantage that it is always drawn the same way (unlike a histogram), but the disadvantage that it doesn't show as much detail.

One thing that we can notice from the histogram and box plot is that the data does not look very symmetric (a.k.a. balanced), instead it looks slightly skewed to the left (that is it looks like it collapsed a bit farther out in that direction). We can add a curve over the histogram to make it easier to compare to a bell-shaped (or normal) curve. Under the *Curves* menu, choose *Parametric Density...* Just hit ok on the box that pops up.

As mentioned before, one of the problems with the histogram is that the way it looks can be affected a lot by how the width of the bars is selected, and where they start and begin. The box with the arrow in it, at the lower left side of the histogram lets you control that. Click on that box, and then select *Ticks...* Change the 3800 to 3600, and the 6600 to 6400, and then click ok. Now try 3700 and 6700, and set the Tick Increment to 600.

Unlike the histogram, the Q-Q plot is not subject to options chosen by the user. Under the *Curves* menu select *QQ Ref Line...*

and then click ok in the window that comes up. The idea of the Q-Q plot is that it plots the actual data along the y-axis, and the values that the data would have if they were exactly the percentiles of a normal curve (bell curve). So if the data is approximately like that of a bell curve, the line should look fairly close to straight. If not, it should be off. Notice that this looks very close to a straight line!!! The data isn't actually that far from a bell-curve, and could be made to look even closer in the histogram if we set the bars up just right. Because the data is close to being bell-curved, we will find that some very nice properties will hold. For example it will be close to being symmetric (the mean and median differ by less than 100). [We will see that the bell-curve or normal curve will be very important throughout the semester, and so it will be useful to tell if the data follows a bell curve.]

Lets change one of the values though, so that the data appears less normal. Click on the spreadsheet, and change the 3900 to 8900, the 4020 to 8020, the 4290 to the 8290, and the first 4380 to 8380. Now note that those four points are rather extreme as indicated on the box plot. You can also see the change in the Q-Q plot. And, note that the mean is over 100 larger than the median (we only changed 4 out of 61 values).

To compare both men and women, we could open PROC INSIGHT with the original data set. (It is probably quickest to close it using the X in the spreadsheet, then go to the *Solutions* menu, the *Analysis* submenu, and then *Interactive Data Analysis*. Select *Work* and then *Bank Data* and then hit *Open*.) Under the

Analyze menu choose Box Plot/Mosaic Plot (Y). Pick Salary for Y and Gender for X.

Section 5.3 and Supplement 6 - Normal, t, Chi², and F Tables

SAS has built in functions that can calculate the values you find in the tables. Each distribution has one function that solves $P(X < x_0)=?$, and is called the probability function. Each distribution also has another function that solves $P(X < ?)=p$, and is called the quantile function.

Distribution	Quantile	Probability
Standard Normal	PROBIT(pct)	PROBNORM(val)
chi ²	CINV(pct,df)	PROBCHI(val,df)
t	TINV(pct,df)	PROBT(val,df)
F	FINV(pct,dfx,dfy)	PROBF(val,dfx,dfy)

In each case, the `pct` is the probability (percent of the area) that is less than the `val`, and `df` are the degrees of freedom. Notice that this is the opposite of what the chi², t, and F tables in the book report (the tables in the book give the probability greater than the value).

The following code will obtain the answers for examples 5.2 to 5.10 in the text book:

```
DATA normanswers;
e5p2 = PROBNORM(1.33)-PROBNORM(-1.33);
e5p3 = 1 - PROBNORM(1.64);
e5p4 = PROBNORM(0.67);
e5p5 = PROBNORM(-1.96) + (1-PROBNORM(1.96));
e5p6 = PROBNORM((12-10)/1.5) - PROBNORM((8-10)/1.5);
e5p7 = PROBNORM((20-27)/3);
e5p8 = PROBIT(1-0.1);
e5p9 = PROBIT(0.975);
e5p10 = 550 + 100*PROBIT(0.90);
;
PROC PRINT DATA=normanswers;
RUN;
```

Note that the answers differ slightly from the text because the text rounded.

The code works similarly for other distributions. Say the sample size was 10 and we were asked to find: $P(\text{chi}^2 > 4.16816)$, $P(2 < \text{chi}^2 < 4)$, and the x_0 such that $P(\text{chi}^2 > x_0) = 0.005$. The code to give these three answers would be as follows.

```
DATA chianswers;
a1 = 1 - PROBCHI(4.16816,9);
a2 = PROBCHI(4,9) - PROBCHI(2,9);
a3 = CINV((1-0.005),9);
;
PROC PRINT DATA=chianswers;
```

```
RUN;
```

You can compare the results of this code to what you would get using Table XI in the text. (Drawing the pictures should help.) While the Table can give us the answers for the first and third questions, for the second, the best it can do is to say between (.100 and .050) minus between (.010 and .005), so that the final answer is somewhere between (.095 and .040).

Section 7.2 and Supplement 7 - Confidence Intervals for Means and Variances

The following instructions and code will construct confidence intervals for the population mean, variance, and standard deviation from the data in Example 7.2 on page 301.

The first step is to enter the data.

```
DATA examp7p2;
INPUT numb_chars @@;
CARDS;
1.13    1.55    1.43    0.92    1.25    1.36    1.32    0.85
1.07    1.48    1.20    1.33    1.18    1.22    1.29
;
```

We could then use PROC INSIGHT to analyze this data. You could start PROC INSIGHT by going to the Solutions menu, and then the Analysis submenu, then choose Interactive Data Analysis. In the box that comes up, choose WORK and examp7p2 and hit Open. You could also start PROC INSIGHT by using the following:

```
PROC INSIGHT;
OPEN examp7p2;
DIST numb_chars;
RUN;
```

We can add a Q-Q plot to the output by going under the Curves menu and selecting QQ Ref Line.... Notice in this case that most of the points are fairly near the straight line. We would say "The data seems to approximate a normal distribution and so we can trust the results of the confidence intervals for the mean and variance." If it was less clear that it was approximately normal, then we could still have a reasonable amount of faith in the confidence interval for the mean, but not for the standard deviation. See supplement 7.3 for more on when we can trust the results of a confidence interval. See problem 5.45 on page 241 for three sample Q-Q plots (a.k.a. normal probability plots) where one looks good and two do not.

To construct the confidence intervals, go to the Tables menu and select Basic Confidence Intervals. If you choose your desired percent, all three will be added to the output window.

To report the results of this example, the easiest thing to do would be to open Microsoft word and copy into it the code from the Program Editor window, the box containing the Q-Q plot, and the box containing the confidence intervals.

Section 7.3 - Confidence Interval for One Proportion

To calculate a confidence interval for a proportion we can use SAS's ability to find **probabilities for a normal distribution** and simply type the formulas into a data step.

The following code will calculate the Agresti and Coull corrected confidence interval in example 7.5 (page 312). It is set up so that you can enter the x, n, and confidence level after the cards (in that order). Notice that we have simply entered the formulas exactly as they are in the book and remembered to put in the

semi-colons.

```
DATA pinterval;
INPUT x n confcoeff;
phat = x/n;
pstar = (x+2)/(n+4);
alpha = 1-confcoeff;
zalphaover2 = PROBIT(1-alpha/2);
plusorminus=zalphaover2*sqrt(pstar*(1-pstar)/(n+4));
lower=pstar-plusorminus;
upper=pstar+plusorminus;
KEEP x n confcoeff pstar plusorminus lower upper;
CARDS;
3 200 0.95
;
PROC PRINT DATA=pinterval;
RUN;
```

Section 8.4 - One Sample t-test

The following code is an example of how you can get a test of hypotheses for a single mean. It is example 8.4 on page 343. Please note that `PROC INSIGHT` always gives the p-value for the alternate hypothesis "not equals too". If you want the p-value for $<$ or $>$ then you must draw the picture and see how you need to change the p-value. (The text gives instructions for this on page 344 in its *Note*.)

```
DATA examp8p4;
INPUT hstays @@;
CARDS;
2      3      8      6      4      4      6      4      2      5
8      10     4      4      4      2      1      3      2      10
1      3      2      3      4      3      5      2      4      1
2      9      1      7      17     9      9      9      4      4
1      1      1      3      1      6      3      3      2      5
1      3      3      14     2      3      9      6      6      3
5      1      4      6      11     22     1      9      6      5
2      2      5      4      3      6      1      5      1      6
17     1      2      4      5      4      4      3      2      3
3      5      2      3      3      2      10     2      4      2
;

PROC INSIGHT;
OPEN examp8p4;
DIST hstays;
RUN;
```

We can construct the Q-Q plot for this data by choosing `QQ Ref Line...` under the `Curves` Menu. In this case the data doesn't appear normal at all! (It's very skewed to the right.) Because of this, we can't fully trust the results, but with a sample size of 100 and the robustness of this t-test we can still get some idea of whether the mean is $=5$ or is <5 .

To conduct the test of hypotheses, we can simply go to `Tests for Location...` under the `Tables` menu. Select $\mu=5$ (our null hypothesis value in this example) and hit ok. This gives the results of three tests. The first is our t-test for testing the alternate hypothesis of "not equals to". The other two tests (the Sign Test and the Signed Rank test) are discussed in STAT 518.

Looking at the p-value we can see that it is 0.2042 with a t-statistic of 1.28. Following the instructions on page 344 we cut that p-value in half for our hypotheses and get that the p-value is 0.1021. Comparing this to our $\alpha=0.05$ we fail to reject the alternate hypothesis.

Another way to get the one-sided p-values is to use the following code. It returns all three p-values (one for each possible alternate hypothesis) and you have to choose the correct one. The only portions you need to change are the name of the data set, the name of the variable, and the value after the cards line. The value after the cards line should be the value for the null hypothesis.

If you read through the code, you should be able to make out several of the formulas.

```
PROC MEANS NOPRINT DATA=examp8p4;
VAR hstays;
OUTPUT OUT=temp MEAN=xbar STD=sd N=n
RUN;
DATA temp2;
SET temp;
KEEP xbar mu sd n t pgreater pless ptwoside;
INPUT mu;
t = (xbar-mu)/(sd/sqrt(n));
df = n - 1;
pgreater = 1 - PROBT(t,df);
pless = PROBT(t,df);
ptwoside = 2*MIN(1-ABS(PROBT(t,df)),ABS(PROBT(t,df)));
cards;
5
;
PROC PRINT;
RUN;
```

Of course you don't need to do it both ways! They'll always give you the same answer.

Section 8.5 - Test for One Proportion

A test for one proportion can be conducted in the same manner that we made a [confidence interval for one proportion](#). We can simply put all of the needed formulas into a data step. The p-values are calculated using the same ideas as the note on page 344. The code below will conduct the test for example 8.7 and 8.8 on pages 355-357 (the observed x, n, and p from the null hypothesis appear in the cards section.) Note that the values are slightly different because the book rounded. Also, instead of using the more complicated method to see if the sample size is large enough it will calculate np and $np(1-p)$ instead.

```
DATA ptest;
INPUT x n pnull;
np = n*pnull;
nlminusp = n*(1-pnull);
phat=x/n;
z=(phat-pnull)/sqrt(pnull*(1-pnull)/n);
pgreater = 1 - PROBNORM(z);
pless = PROBNORM(z);
```

```
ptwoside = 2*MIN(1-ABS(PROBNORM(z)),ABS(PROBNORM(z)));
CARDS;
10 300 .05
;
PROC PRINT DATA=ptest;
RUN;
```

Section 9.1 - Two Sample t-test

The following sample code will analyze the data in Example 9.4 on page 387. Notice that we have to enter the group and the value for each observation, and that the group name is a word (and so it needs a \$).

```
DATA examp9p4;
INPUT group $ value @@;
CARDS;
new      80      new      80      new      79      new      81
stand    79      stand    62      stand    70      stand    68
new      76      new      66      new      71      new      76
stand    73      stand    76      stand    86      stand    73
new      70      new      85
stand    72      stand    68      stand    75      stand    66
;

PROC TTEST DATA=examp9p4;
CLASS group;
VAR value;
RUN;
```

The order you enter the groups determines which hypothesis is being tested. Because the `new` group was entered first, the procedure will look at the difference `new-stand`. By default the procedure tests the hypothesis that this difference is equal to zero. To test that the difference is equal to some other value, say 5, you would add `H0=5` to the first line between `examp9p4` and the semi-colon (in `PROC TTEST`).

The first three rows of the output contain the means, confidence intervals for the means, standard deviations, and confidence intervals for the standard deviations for the two groups individually, and for the difference of the two groups. Notice that the third row gives the confidence interval (-1.399, 9.5327) that matches what the text found in the middle of page 388... SAS uses 95% for CIs by default. Also notice that we can find s_p^2 on the third line; 6.119 is the square root of 37.45.

The next two rows of the output are for the two t-tests. The `Pooled` line is the case where the variances are equal, and the `Sattertwaite` line is for unequal variances. One important thing to note here is that SAS is doing the two-sided test for the alternate hypothesis "not equals to". If you are testing either `<` or `>`, then you will need to draw the picture and adjust the p-value by hand.

The final line is a test of the null hypothesis that the variances of the two groups are equal or not. The `Pr>F` column is the p-value for this test, so if it is a small value (less than alpha) you reject the null hypothesis that the variances are equal. For this example we see that the p-value is 0.8148, which is very large, and so we fail to reject that the variances are equal.

However, this F test for two-variances is not robust at all and should not be used.

In checking the assumption that the data is normal, we need to make sure we check it for BOTH samples. We can still use PROC INSIGHT for this.

```
PROC INSIGHT;
OPEN examp9p4;
RUN;
```

When you choose `Distribution (Y)` under the `Analyze` menu, choose `value` for `Y` and choose `group` for `group`. This will put the information for both groups in the same window (use the scroll bar at the bottom to switch between them). If you add the q-q plot and q-q line, it will add them for both variables. In this case the second sample appears to look fairly normal, but the first one is a bit questionable. Because of this, I would trust the two sample t-test (it is robust against) slight violations.

A logical next question is "If I can't trust the F-test, then how do I know if I can use the test where we assume the variances are equal?" Because the two-sample t-test for two means is fairly robust (especially when the two samples are about equal sized, see pg. 383) we could just compare the standard deviations, or we could use two box-plots like they do on page 382 and see if we believe the variances are equal or not. If we have our doubts though, especially since the sample sizes are not equal here, we should use the Satterwaite formula.

NOTE:

There are more robust tests available to see if two variances are equal than the F test. One of these that SAS can be made to perform in PROC GLM is called the "Modified Levene's Test" or the "Brown and Forsythe Test". Similarly SAS can be made to perform tests of the null hypothesis "the data comes from a normally distributed population" vs. the alternate hypothesis that the population is not normal. Perhaps the best of these tests is the "Anderson-Darling Test". Because these tests are somewhat complicated we will not be covering them in STAT 515... but if you ever need a test of either of these hypotheses a statistical consultant can show you how to conduct them.

Section 9.2 - Paired t-test

The following code will conduct the paired t-test shown on pages 399-400 (the data is in table 9.4). Remember that the p-value is for the two-sided alternate hypothesis and would need to be adjusted for testing either $>$ or $<$.

```
DATA learner;
INPUT new standard;
CARDS;
77      72
74      68
82      76
73      68
87      84
```

```

69      68
66      61
80      76
;

```

```

PROC TTEST DATA=learner;
PAIRED new:standard;
RUN;

```

Another way to conduct this t-test is to remember that it is simply a one-sample t-test on the differences. In PROC INSIGHT you could choose Variables under the Edit menu. Select the Other... option. We can now make a new variable that is equal to new-standard. Click on Y-X in the transformation box, then select new for the Y value, standard for the X value, and click on OK. You can now choose Distribution (Y) under the Analyze menu and simply do a one sample t-test on the new variable you created.

Section 9.3 - Tests and Confidence Intervals for Two Proportions

Tests and confidence intervals for two proportions can be made in a similar fashion to how we made [tests](#) and [confidence intervals](#) for one proportion.

The following code will calculate the confidence interval discussed on pages 511 and 512.

```

DATA ci2p;
INPUT x1 n1 x2 n2 confcoeff;
plh = x1/n1;
p2h = x2/n2;
alpha = 1-confcoeff;
zalphaover2 = PROBIT(1-alpha/2);
diff=plh-p2h;
plusorminus=zalphaover2*sqrt(plh*(1-plh)/n1 + p2h*(1-p2h)/n2);
lower=diff-plusorminus;
upper=diff+plusorminus;
KEEP confcoeff diff plusorminus lower upper;
CARDS;
546 1000 475 1000 0.95
;
PROC PRINT DATA=ci2p;
RUN;

```

This code performs the test of hypotheses for the data in examples 9.6 and 9.7 on pages 413-415.

```

DATA test2p;
INPUT x1 n1 x2 n2;
plh = x1/n1;
p2h = x2/n2;
ph = (x1+x2)/(n1+n2);
diff=plh-p2h;
z=(plh-p2h)/sqrt(ph*(1-ph)/n1 + ph*(1-ph)/n2);
pgreater = 1 - PROBNORM(z);
pless = PROBNORM(z);
ptwoside = 2*MIN(1-ABS(PROBNORM(z)),ABS(PROBNORM(z)));
KEEP ph z pgreater pless ptwoside;
CARDS;
555 1500 578 1750
;

```

```
PROC PRINT DATA=test2p;
RUN;
```

Section 10.2 - One-way Analysis of Variance

The following code will analyze the data in examples 10.3 and 10.4. Notice that each observation must have both a group identifier and a measurement. Recall that the \$ tells SAS that the variable before it is characters (not numbers) and the @@ means that more than one observation occurs on a line.

```
DATA iron;
INPUT brand $ dist @@;
CARDS;
A      251.2   B      263.2   C      269.7   D      251.6
A      245.1   B      262.9   C      263.2   D      248.6
A      248.0   B      265.0   C      277.5   D      249.4
A      251.1   B      254.5   C      267.4   D      242.0
A      260.5   B      264.3   C      270.5   D      246.5
A      250.0   B      257.0   C      265.5   D      251.3
A      253.9   B      262.8   C      270.7   D      261.8
A      244.6   B      264.4   C      272.9   D      249.0
A      254.6   B      260.6   C      275.6   D      247.1
A      248.8   B      255.9   C      266.5   D      245.9
;

PROC INSIGHT;
OPEN iron;
FIT dist = brand;
RUN;
```

This output gives you the ANOVA table automatically. To check the assumptions we need to verify that the variances seem to be approximately equal and that the distributions seem approximately normal. (Remember that you also need for the samples to be random and independent... but you can't really check that from the plots.)

One way to check whether the groups seem to have the same variance is to use the residual versus predicted plot at the bottom left of the PROC INSIGHT window. This is basically like Figure 10.10 on page 457 except that it is turned on its side and all of the columns have been made to have the same mean. If the four (in this example) columns all look equally spread out then we would say it looks like the variances are the same. Another way to check whether the groups seem to have the same variances is to use side by side box plots like we did for the two-sample t-test. (Similar to Figure 10.10, but using Box Plots instead of Dot Plots). Under the Analyze menu choose Box Plot/Mosaic Plot (Y). Select dist for Y and brand for X. This plot shows us both how the means differ (C looks larger than the others) and that all have roughly the same variance (the size of each box plot isn't that much different from the others). Finally, you also could have chosen Distribution (Y) under the Analyze menu with dist for Y and brand for group. This will calculate the standard deviation for each group and we could conclude that since all the sds are between 3.8 and 5.2 that they are close enough that the procedure should work well. Any of these three methods is acceptable, and you do not need to do all of them.

To check if each of the samples looks like they came from normal distributions, we really should make a separate Q-Q plot for each group, just like for the two-sample t-test. Under the Analyze menu choose Distribution (Y). Select dist for Y and brand for group. You can then choose QQ Ref Line... under the curves menu. (To get rid of all the extra information, you could deselect the various "checked" tables and graphs under the Tables and Graphs menus.) For this example, Brands A and C look very close to normal, Brand B looks a little odd, and Brand D appears to have one outlier. Since the F test for ANOVA is robust however, none of these seem bad enough

not to trust the test.

Unfortunately these can be hard to check if there weren't many observations in each sample. An alternative approach is to use the combined q-q plot that will show up at the bottom of the PROC INSIGHT window that has the ANOVA table in it. To add that plot to the output, choose Residual Normal QQ under the Graphs menu.

NOTE: PROC GLM

is another commonly used method of performing both Analysis of Variance and regression. It will construct the ANOVA table and many other statistics that we will use in STAT 516... it won't construct the various graphs however. The code we would use with this data and PROC GLM would be:

```
PROC GLM DATA=iron;
CLASS brand;
MODEL dist=brand;
RUN;
```

The last page of this output is the same as that on the top of page 453, except that it was generated using PROC GLM instead of PROC ANOVA.

Chapter 11 and Supplements - Simple Linear Regression

The following code will calculate all of the statistics for the data in Table 11.1 on page 514 and give the output that is discussed in Sections 11.2 through 11.8.

```
DATA stimulus;
INPUT amount_x reaction_y;
CARDS;
1          1
2          1
3          2
4          2
5          4
;

PROC INSIGHT;
OPEN stimulus;
FIT reaction_y = amount_x;
RUN;
```

To generate the q-q plot for the residuals go to Residual Normal QQ under the Graphs menu. To generate the confidence interval for the slope, select C.I. / C.I. (Wald) for Parameters under the Tables menu.

To get the prediction intervals and confidence interval for the regression line (like Figure 11.24 on page 556) you can simply go to Confidence Curves under the Curves menu and select the type you want. (You can put both on the same graph.) This picture will not let you see the actual values though. To get output like that in Figure 11.19 and 11.20 you would use the following code:

```
PROC GLM DATA=stimulus;
MODEL reaction_y = amount_x / ALPHA=0.05 CLI;
RUN;

PROC GLM DATA=stimulus;
MODEL reaction_y = amount_x / ALPHA=0.05 CLM;
RUN;
```

Where CLI is for the individual prediction interval and CLM is for the mean (or regression line).

If we wanted to get the intervals for a new x value, say $x=4.5$, we would add a new data point to the data set and rerun the `PROC GLM` code. (The `.` for the y -value tells SAS that it shouldn't be included in the calculation of the regression line...)

```
DATA stimulus;
INPUT amount_x  reaction_y;
CARDS;
1          1
2          1
3          2
4          2
4.5       .
5          4
;
```

Section 13.2 - Goodness of Fit Test

The following code will analyze the data in example 13.2 on page 710.

```
DATA ex13p2;
INPUT opinion $  count;
CARDS;
legal          39
decrim         99
exist          336
noopin         26
;
PROC FREQ DATA=ex13p2 ORDER=data;
TABLES opinion / TESTP=(.07,.18,.65,.10);
WEIGHT count;
RUN;
```

Instead of using `TESTP` (test proportion), you also could use `TESTF` (test frequency). In this case you would put the expected values in instead of the proportions. One further complication with `PROC FREQ` in SAS is that it doesn't handle observed values of zero well. If there is a cell that was 0, use the value 0.00001 instead. This way SAS will actually recognize that it is a cell, and it won't throw the test statistic off by very much.

Section 13.3 - Two-way Contingency Tables

The following code will work example 13.3 on pages 719-720.

```
DATA ex13p3;
INPUT rel $  marit $  count;
CARDS;
A  D          39
```

```

B   D           19
C   D           12
D   D           28
None D           18
A   Never       172
B   Never        61
C   Never        44
D   Never        70
None      Never   37
;

```

```

PROC FREQ DATA=ex13p3;

    WEIGHT count;

    TABLES rel*marit / chisq expected nopercnt;

RUN;

```

The Text Book Data

The various data sets used in the text book can be found on the CD in the back of the text in text files. . The key to the various names in this directory can be found in Appendix B of the text on pages 823-827.

Computer Trouble?

In most cases, help with the computers (NOT the programming) can be gained by e-mailing help@stat.sc.edu

For the printers on the first and second floor, printer paper is available in the Stat Department office. For printers on the third floor, paper is available in the Math Department office.

If you are using a PC restarting the machine will fix many problems, but obviously don't try that if you have a file that won't save or the like.

If SAS won't start, one of the things to check is that your computer has loaded the X drive correctly (whatever that means). Go to My Computer and see if the apps on 'lc-nt' (X:) is listed as one of the drives. If it isn't, go to the Tools menu and select Map Network Drive.... Select X for the drive, and enter \\lc-nt\apps for the Folder. Then click Finish. This should connect your computer to the X drive and allow SAS to run. If you already had the X-drive connected, then you will need to e-mail help@stat.sc.edu.

If your graphs print out extremely small

after you copy them to word, you might be able to fix the problem by "opening and closing" the image. In word, left click once on the image, and select Edit Picture or Open Picture Object under the Edit menu. A separate window will open with the image in it. Simply choose Close Picture. It should now print out ok. This will also make the spacing between the characters in the labels look right if they were somewhat off.

If the problem is an emergency requiring immediate attention see the statistics department computer person in room 209D.

If they are not available and it is an emergency see Minna Moore in room 417.

Flagrantly non-emergency cases may result in suspension of computer privileges.
