Helpful Guide to SPLAT

(Statistical Package for Learning and Teaching)

December 1, 2024

Program and Guide written by
Chris Olsen
crolsen@fastmail.com

(More or less accurate)
Table of Contents

(More or less accurate)
Table of Contents
(Cont'd)

(More or less accurate)
Table of Contents
(Cont'd)

First of all…

- **SPLAT (Statistical Package for Learning and Teaching) is totally freeware** and designed to facilitate the learning and teaching of statistics. (Hence the name!) The specific target audience for SPLAT is Advanced Placement Statistics teachers and students, but SPLAT should work well in some other AP classes (e.g. AP Biology and AP Psychology) and also statistics classes other than AP. SPLAT is very easy to install and even easier to use.

- I am happy to provide the source code to SPLAT should your school district IT folks wish to check for malware.

Preface

I wrote SPLAT for mostly two reasons: my profession has involved teaching statistics and I enjoy computer programming. When I began presenting statistics workshops my initial intent was to provide freeware to complement students' use of the TI-calculator in the workshops. I found that many AP Statistics teachers did not have statistical software due to budget constraints; my hope back then (and now) was (is) that teachers and their students can use the same statistical software in their classes and at home without breaking anyone's budget or (after installation) having to go to the internet if they live in places where access is limited and/or expensive.

SPLAT has become a labor of love and I will probably continue to add features and tinker with it. (Everyone must have some hobby after all!)

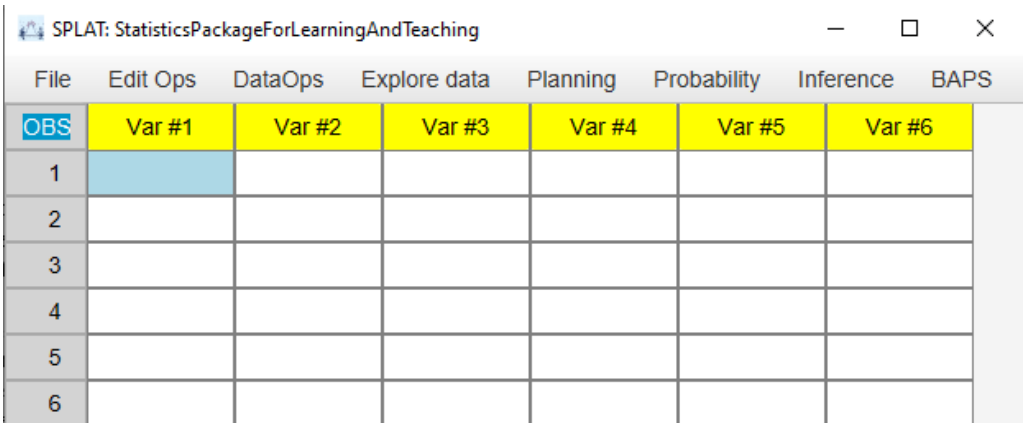
A short bug alert:

After my workshops in the summer of 2023, I made the initial graph sizes smaller for easier use on laptops. This has (for unknown reasons) created a few problems with some of the **initial** presentations of graph panels. The solution seems to be to take that little rodent and click and drag on the corners or edges of the panel. I have not determined what causes this, or how to fix it, or even if it is possible to fix the problems, but I'm working on it. In some cases, the problems seem to be known to the Java programming community and are hardware related. So if a graph does not initially appear, jiggle its window a little bit and it should fix itself.

So, that is the preface. On to the post-preface…

Making SPLAT work

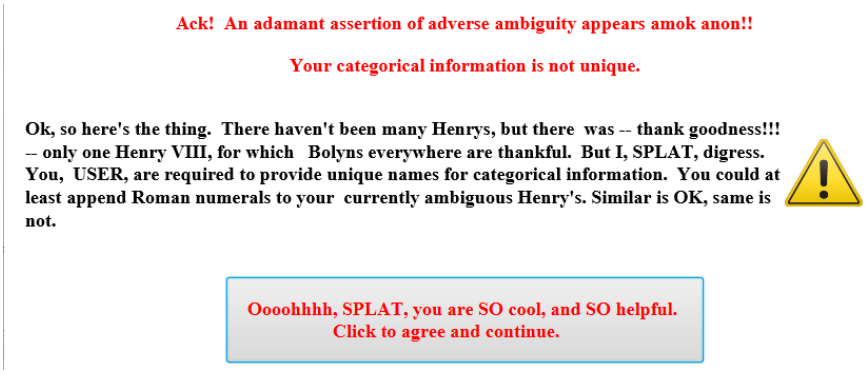
When SPLAT starts up you will see what is known as a Splash Screen. The splash screen in SPLAT has no purpose except to give you my email address so you can contact me about SPLAT if you have questions or suggestions for improvements. Just click the Get-on-with-it button and you will see the spreadsheet below. The menu items across the top reflect the 2019+ AP Statistics course description, plus some stuff that is "BAPS" (Beyond AP Statistics). I have included in SPLAT some BAPS procedures for those teachers with the time and inclination to teach more statistical topics after the AP Statistics exam. Also, sometimes student projects go beyond the AP Syllabus, and one hopes the BAPS choices will provide support for students who can thus proceed with their projects and not search in vain on the internet. Teachers should be aware that I make no attempt to teach the procedures; I do attempt to present the statistical output as "professional" software would.



Students (but, of course, not teachers!) make occasional judgement and/or keystroke errors. SPLAT tries to discover these as soon as they occur and clearly provide feedback to the user about the nature of the error. My past classroom experience with allegedly helpful error messages displayed by statistical software has sensitized me to this aspect of user interaction. SPLAT's error messages and warnings are intended to be informative -- and lighthearted – and then return the student to a reasonable and safe place to continue working. Users should carefully read the error messages to get an idea of what the problem is, and then just click on Agree. (You will discover soon enough that "Agree" is the only option.)

As an example, suppose a student is setting up a chi square Goodness of Fit test. SPLAT is expecting unique names for the values of the variable. Further suppose the variable has these values: Red, Green, and Blue, but the student enters Red, Red and Blue. Kind and gentle statistical software should not just beep loudly in a stern stentorian tone and wait for the student to divine what the error is, it should give the cherub some sort of clue. I have programmed information into SPLAY's error messages. (My students in the past have suggested the error messages are at the level of "Dad Jokes".) The intent is to leave the student with some degree of comfort in the face of an error. Also, the cherubs should know that SPLAT feels their pain and that Users are only human. In some cases, in as the immediately following example, some specific directions are given, but mostly the errors should be read, smiled (or groaned) at, and dispensed with.

Here is what the error messages look like, in this case if labels for variables are not unique in a data file:



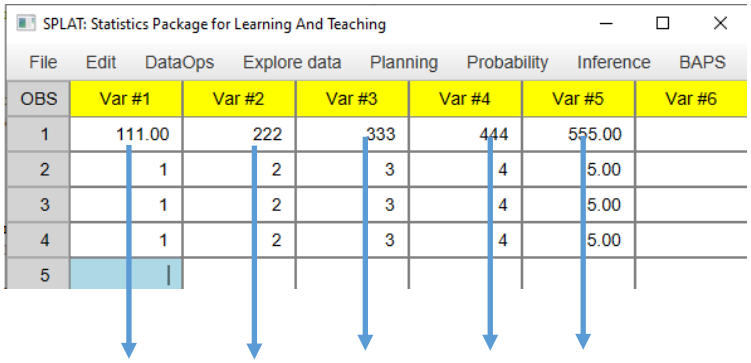
Data entry with SPLAT.

There are two ways one can enter data in the SPLAT spreadsheet: (1) "like-the-Texas Instruments TI-8x calculators" and (2) "like-Excel." As is well known, the TI is set up to enter data into "Lists." Students with TIx's will be experienced with this vertical entry. Some may also be familiar with Excel, where data entry can also be performed horizontally by tabbing. I have programmed both options into SPLAT.

If you want to enter data into SPLAT vertically (the TI way), click on a cell and enter a value, followed by the "Enter" key. If you want to go the horizontal tabbing route, click on a cell, and enter values in the cells, followed by TABs. After entering your data in the last column, press "Enter." Then repeat on successive lines. Pressing Enter while you are entering data horizontally tells SPLAT you are done with horizontal entry and are ready to go to the next line. If -- as I typically do -- you mess up and hit Enter by accident or habit, you can re-establish the horizontal data entry by clicking on a new starting point and tabbing away.

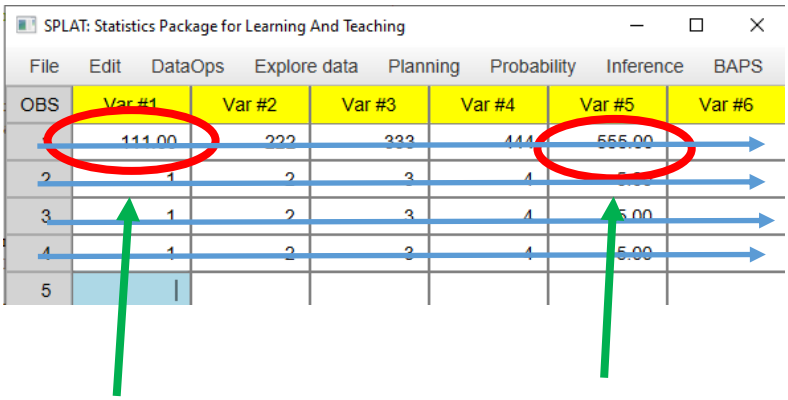
Data entry the TI-8x Calculator way –

#, Enter, Repeat



Data entry the Excel (Tab) way

#, Tab, Repeat;
Enter on last column



Start by entering the number and tabbing across the row

"Enter" here to establish the last column, tab on successive row-entry of data.

I would like to ~~show off~~ gracefully introduce some of the features of SPLAT by demonstrating some common statistical procedures. You will find that the look-and-feel of SPLAT is very similar across procedures, so that the learning curve for SPLAT should be a graceful glide, not an appalling precipice.

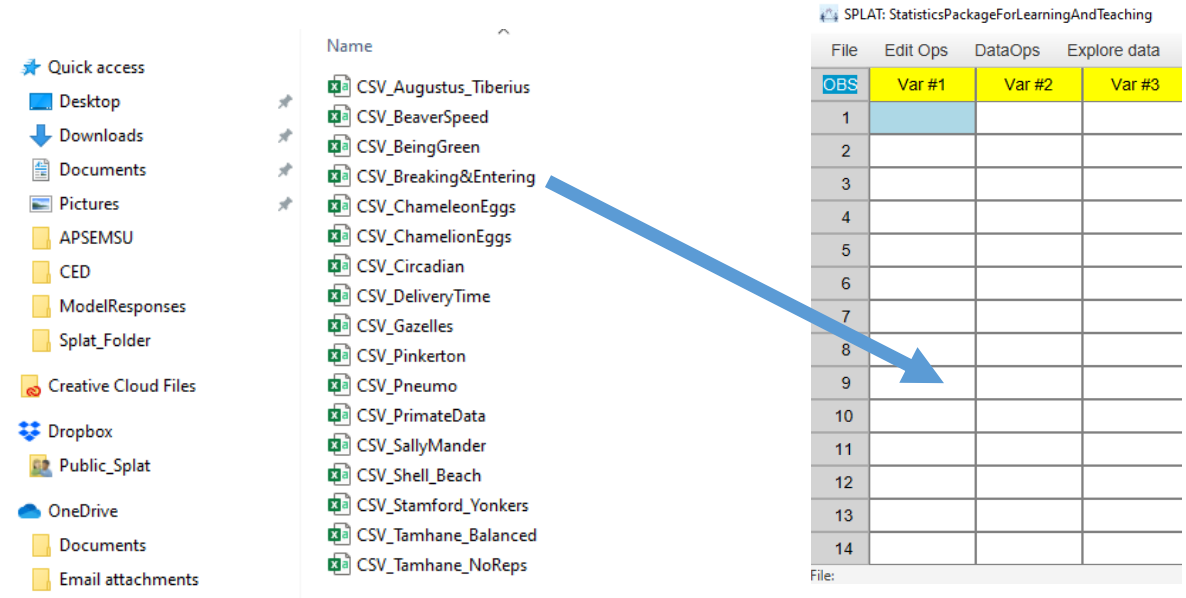
Data files:

I am assuming that you, dear User, do not have lots of time to enter data and might even make errors doing so. To minimize the frustration of getting erroneous results due to these sorts of errors, I have provided data files to use with SPLAT. (They reside in Dropbox, where you downloaded the SPLAT installer.) The data file names have a prefix: CSV and the file names looks like this: "CSV_file_name." SPLAT reads from and writes to files in a standard format known as "comma separated values." CSV files are editable using a text editor or Excel if you prefer to do that. I suggest you copy the "CSV_files" folder to your desktop for easy access while traversing this Guide. (The prepended "CSV" is not necessary, it is just my own personal convention.)

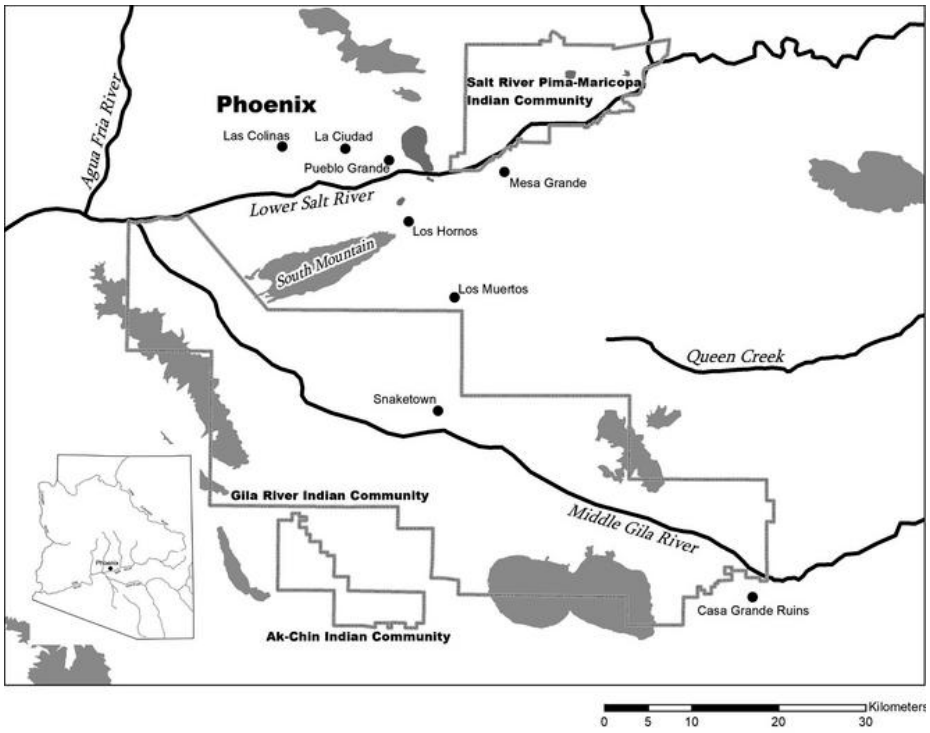
Note: If your intended data analysis requires some sort of weird data manipulation like taking the inverse secant of a variable, or maybe evaluating complex Boolean logic expressions, you would be well-advised to enter data and do those manipulations in Excel. Then save the data in a CSV file. There are things that Excel can do easily but SPLAT is not programmed to do. Be careful to save the Excel file as a CSV-formatted file to make it "SPLAT-readable."

On to Univariate statistics…

In SPLAT you can access a data file in the usual way, by clicking on File→Open. SPLAT also has a Drag and Drop capability for reading CSV files. You can click on a file and drag it over to the SPLAT spreadsheet, as shown below. Open (or Drag) the file, CSV_Breaking&Entering, and we will take a drive with SPLAT. The file contains data reported in, Beck, M. E. (2002). The Ball-On-Three-Ball Test for Tensile Strength: Refined Methodology and Results for Three Hohokam Ceramic Types. *American Antiquity* 67(3):558-569.



Dr. Beck measured the tensile strength of ceramic objects from archeological sites in south-central Arizona. Her unit of analysis was the "sherd," a fragment of pottery. Tensile strength is the maximum load that a material can support without fracturing. Her results are reported in units of force (kg) needed to shatter the sherd.

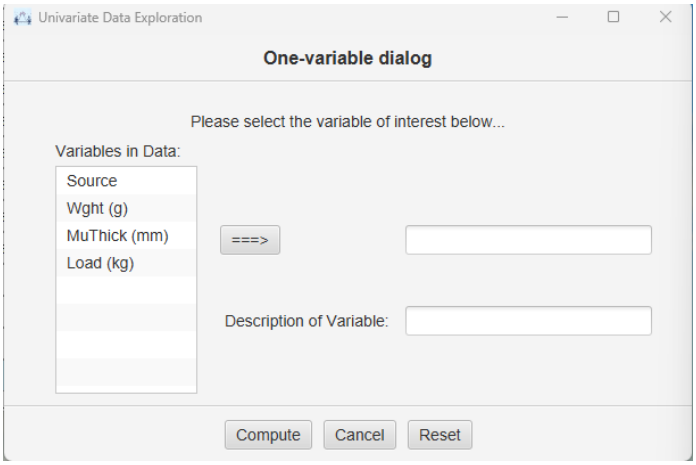


Three different sources of sherds are reported: Sacaton, Casa Grande, and Gila Plain, all sites in the Phoenix Basin of Arizona.

The Sacaton and Gila Plain sherds date from 950-1100 CE, and the Casa Grande sherds from 1100-1300 CE. The weight of the sherd is measured in grams, and the mean thickness of the sherd in mm. (Notice that the data were NOT entered in the TIx format as separate Lists.)

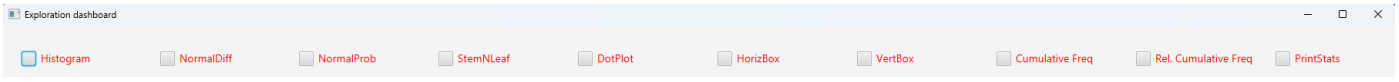
In SPLAT, selecting variables for analysis is consistent in appearance across the different statistical procedures. SPLAT will present a list of variables in the file, one or more arrows to select the variable(s) of interest, and a place for you to optionally prefer and supply label(s) for the variable or variables of interest in your analysis of the data.

To start, click this sequence: Explore data → Univariate data → A single quantitative variable. This is the more-or-less standard presentation of choices when your procedure involves data. The variables are listed on the left, the little arrow is for selection of the variable of interest, and the "Description of Variable" is where you can substitute your own preference for a label. The default label will be what is in the "Variables in Data" part of the dialog.



Scroll down a bit in the "Variables in Data" list and choose the sherd weights by clicking on "Wght (g)" and then clicking on the right-pointing arrow. In the Description of Variable field, enter something you will remember – like "Weight in grams" and click on "Compute."

The standard SPLAT dashboard will appear with a light blue background and a row at the top where your current graphic and statistical options are displayed. This dashboard contains a **breathtakingly long** list of options, **many more options than is usual** in SPLAT. This list of options is long because there is a breathtakingly long list of possible univariate graphs to consider here.



Click on some of the various graphics plot options, Histogram, NormalProb, etc., to see how this works. All SPLAT dashboards will be similar, but (thankfully!) will contain fewer options. One of the design principles programmed into SPLAT is: get the user to the good stuff (the graphs) with a minimum number of mouse clicks. This is really the only dashboard in SPLAT that pays such a heavy price in terms of choices. To close the dashboard, click on the little 'x' in the upper right corner of the panel. Mac people, your x's are those little red circles in the upper left corner.

To view a plot, click on a check box. The plot sizes are initially small enough to accommodate laptops. The plot sizes can be increased (but not, for now, decreased) in size and can be moved about by:

- clicking and dragging on the corners and edges of the panels to resize them.
- clicking and dragging on the tops to move the graph hither and thither across the dashboard as desired.

Hint: Some of the graph panels can be a bit picky about moving and resizing, for reasons I do not (yet!) understand. If you encounter a picky one, resizing by clicking and dragging on the northwest corner of the panels seems to work best.

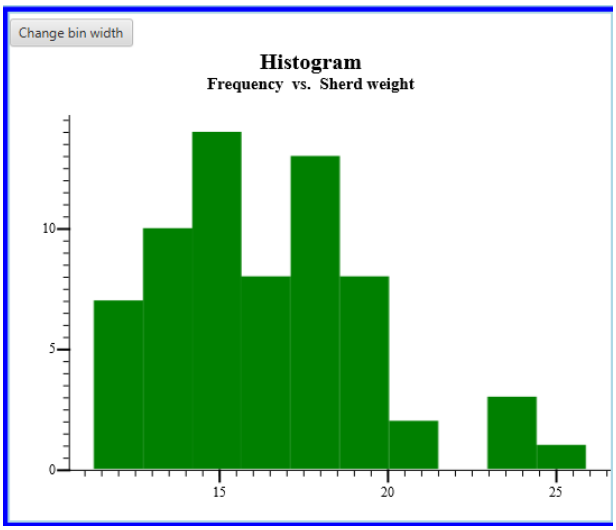
Copy and pasting with SPLAT

If you wish to copy and paste a plot into Word or whatever, click on the plot and press Control-C to copy to the system clipboard. (I believe the Mac equivalent is Command-C.) Then Control-V will paste it to your Word document. If that doesn't work for some reason, you can always use your favorite snipping tool.

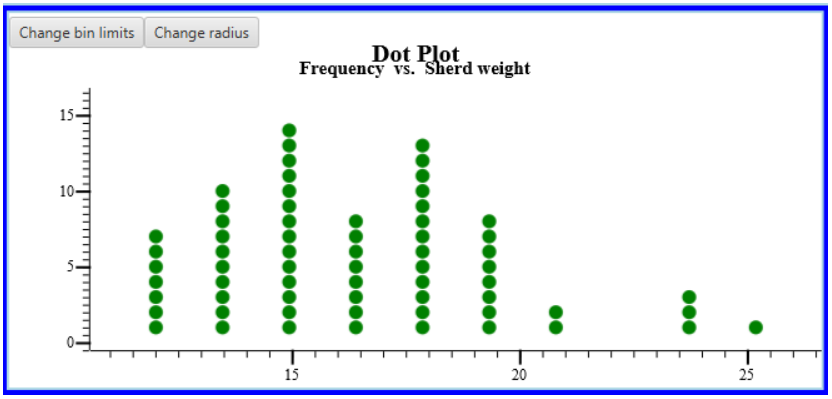
Now I would like to introduce you to some of the plots…

Histogram, Dotplot, Cumulative Frequency:

The bin placements in histograms, dot plots, and cumulative frequency plots can be changed to fit your desires. The initial presentations, especially of the dot plot, are almost never what you want. For example, notice that the tick marks do not line up perfectly with the middles of the bins in the histogram above. The same problem occurs with the dot plots (after changing the radius…)

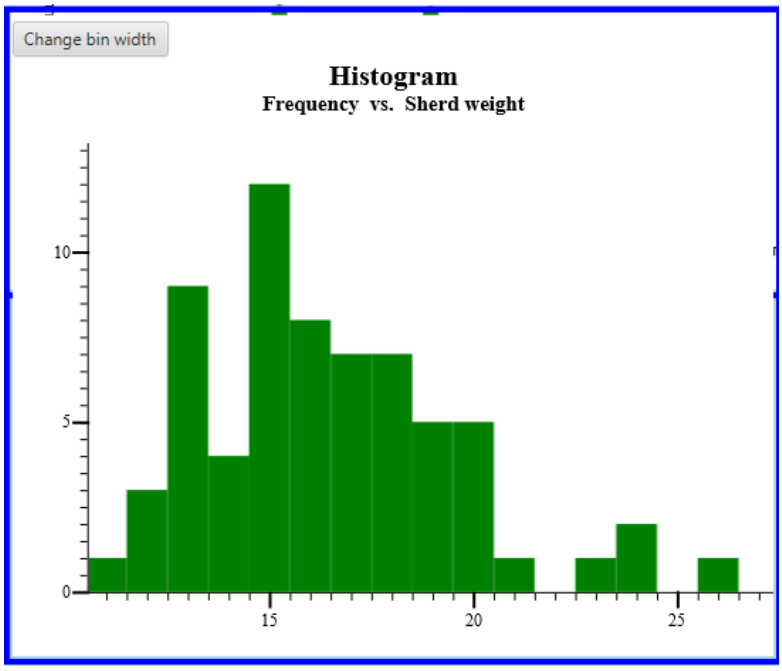


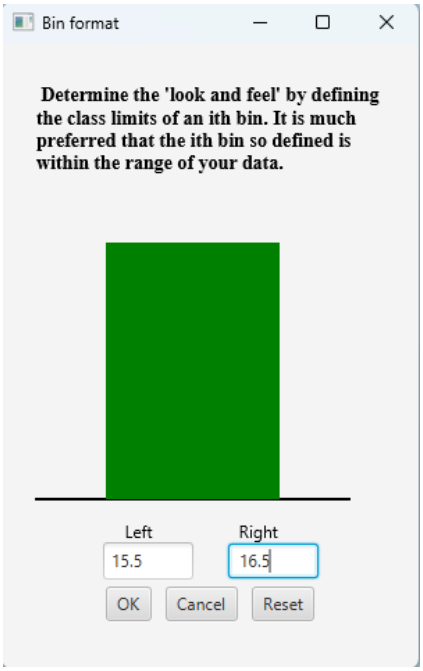
Changing bin placement and sizes – similar to the TI!



You can change the bin width and locations to fit your taste by clicking on…wait for it… "Change bin width."

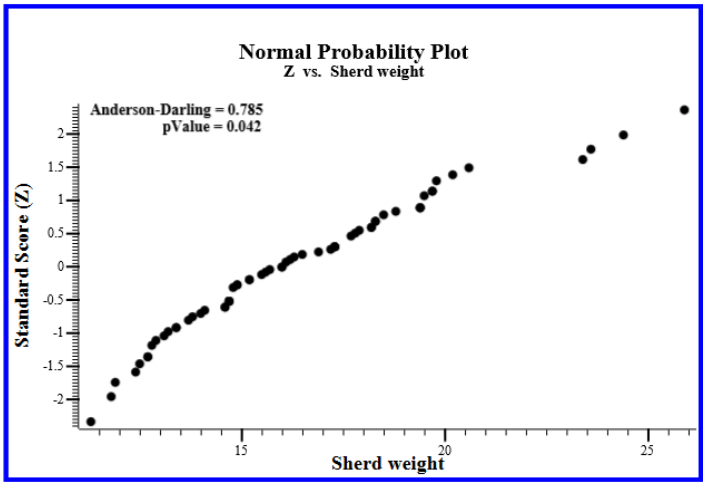
The basic idea here is that you imagine what one of the bins would look like – where does it start, where does I end? – and enter those two values and click "OK." SPLAT will then re-create the graph consistent with your choices and re-display it. As an example, tell SPLAT that you want the left and right sides to be (say) 15.5 and 16.5, respectively. You can choose any values for the sides, but if you choose values outside the range of the data SPLAT will object – and might even crash!





Normal probability plot:

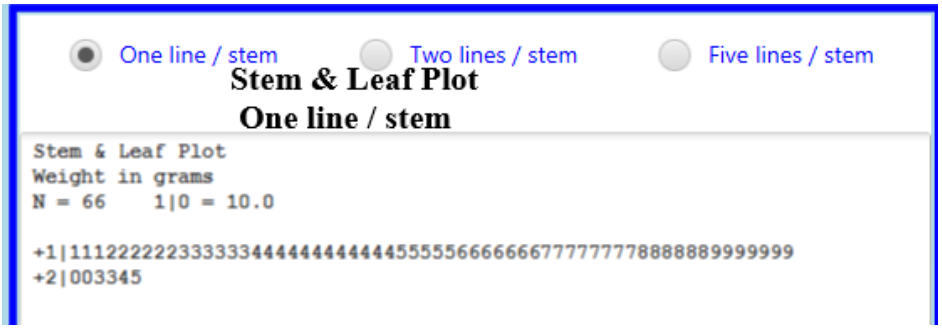
The normal probability plot is that last one in the second row on TI's. The Anderson-Darling (AD) statistic shown in this plot is used to test the hypothesis that the data have been sampled from an approximately normal population. This is one of those places in SPLAT where some BAPS things are reported. These are for those who are familiar with the statistics, and also to suggest to AP Statistics students that there is more way cool statistical stuff to learn in their futures.



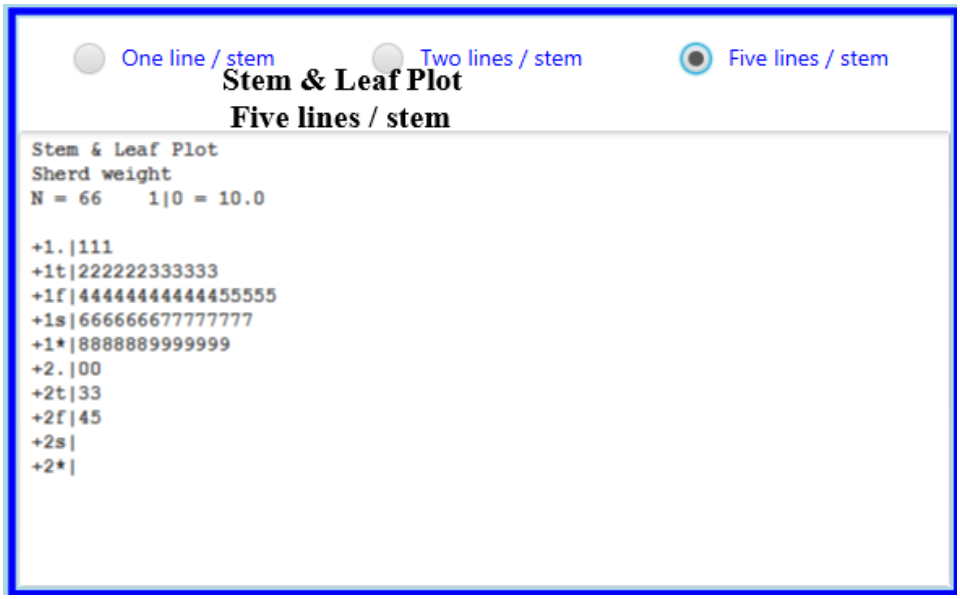
Checking assumptions is an important idea in AP Statistics, and the AD statistic is included here to reinforce that idea. In real classroom life, my recommendation to students is generally to inspect the normal probability plot and only consider the AD statistic if the plot is ambiguous.

Stem and leaf plot(s)

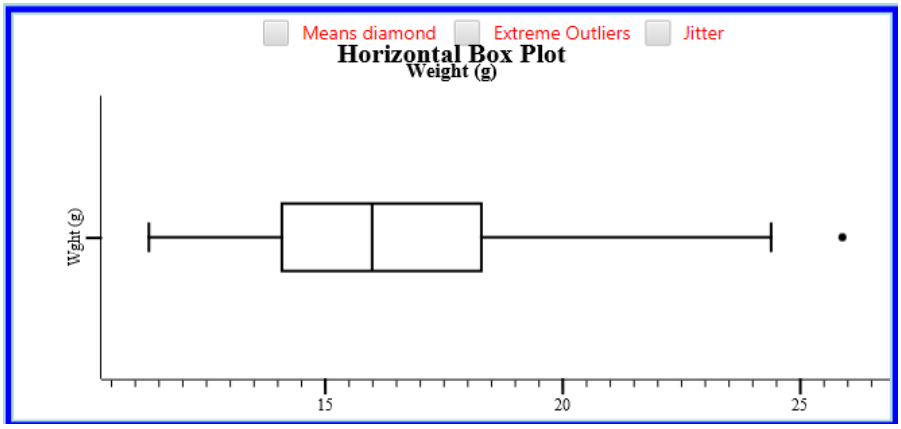
The initial stem and leaf presentation is in a "one line per stem" view:



SPLAT has three stem and leaf options: one, two, and five lines per stem. For this example, that initial presentation may not be best one. If you click on the five lines per stem option to get a more pleasant display. Any guess what the s, f, and t indicate? (Hint: it works in English but not French.)



The box plots



The "Means diamond" displays the mean (top and bottom vertices of the diamond) and plus/minus one standard deviation (left and right vertices of the diamond) of the data. When "Extreme Outliers" is checked, outliers more extreme than 3 IQRs from the nearest quartile are presented as circles (example below); solid dots indicate the usual non-extreme 1.5 IQR outliers. "Jittering" allows you to see where the original data points are located, something that the usual box plot hides from you.

Print stats

These are the usual suspect statistics plus a few. While I'm here I would like to mention the zooming capability of text-based panels in SPLAT. Right click on the Print Statistics panel and a small window appears. This small window can be resized and is intended for use when projecting in a classroom.

Resize the Print Stats panel to a decent width and height as desired; then, using the wheel on your mouse or clicking on the Zoom menu item you can make the text bigger so the cherubs in the back of the room (and those of us of a certain age) can see it more easily.

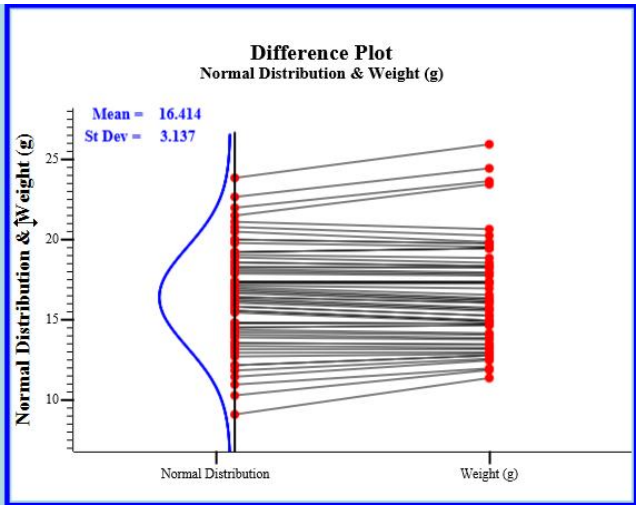
Print univ statistics

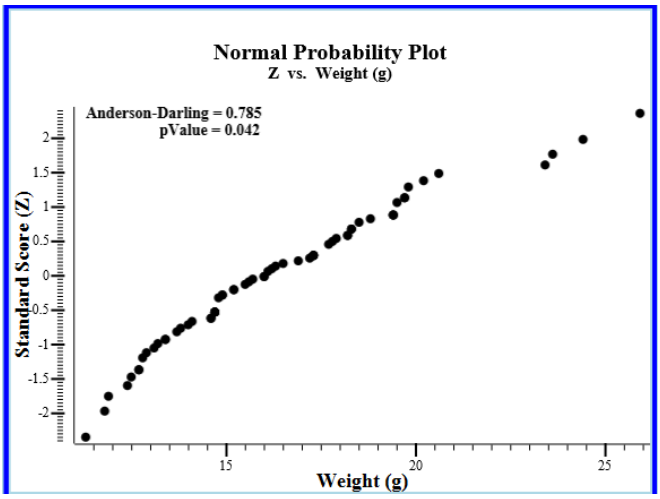
```
       Univariate statistics

*****  File information  *****
     Variable: Something
     N in file:    66
     N missing:     0
       N Legal:    66

*****  Basic mean based statistics  *****
          Mean:       16.414
      Variance:        9.842
     Stand dev:        3.137
          Skew:        0.802

  *****  Other mean based statistics  *****
   Trimmed mean:       16.159
      Kurtosis:        3.536
Excess Kurtosis:       0.536
            CV:        0.191

 *****  Five-number summary  *****
       Minimum:       11.300
            Q1:       14.100
        Median:       16.000
            Q3:       18.300
       Maximum:       25.900

  *****  Other median based statistics  *****
           IQR:        4.200
         Range:       14.600
```

The information in text boxes can also be copy-and-pasted to your favorite text editor, but you will need to choose a mono-spaced font to make the display decently readable.

The normal probability plot (and friend)

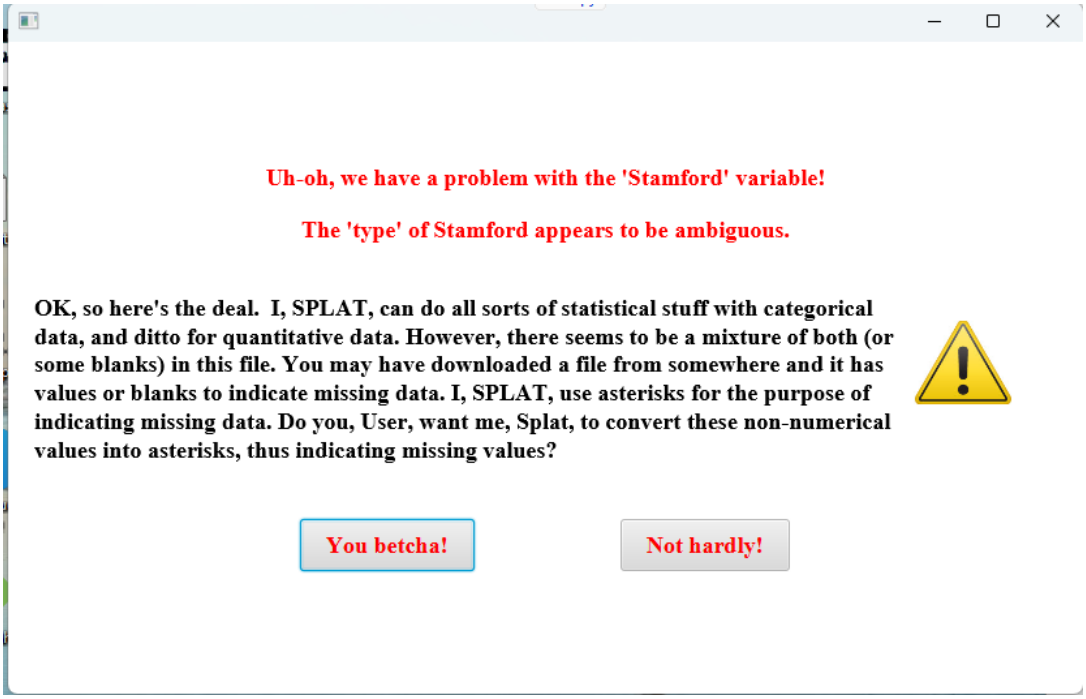
These two plots are used to help students check for skew. The general idea of the normal difference plots is to address the question of whether the data credibly came from a normal distribution. If the data source is credibly normal, one should see a bunch of relatively horizontal lines going from the ideal normal distribution (left) to the data (right).





If you have two quantitative variables…

OK, close the blue dashboard and return to the spreadsheet. Open the file, CSV_Stamford_Yonkers. These data, gathered back in the 1970's, are ozone concentrations in parts per billion from the two northeastern U.S. cities. Open the file, and … Yikes! An error message…



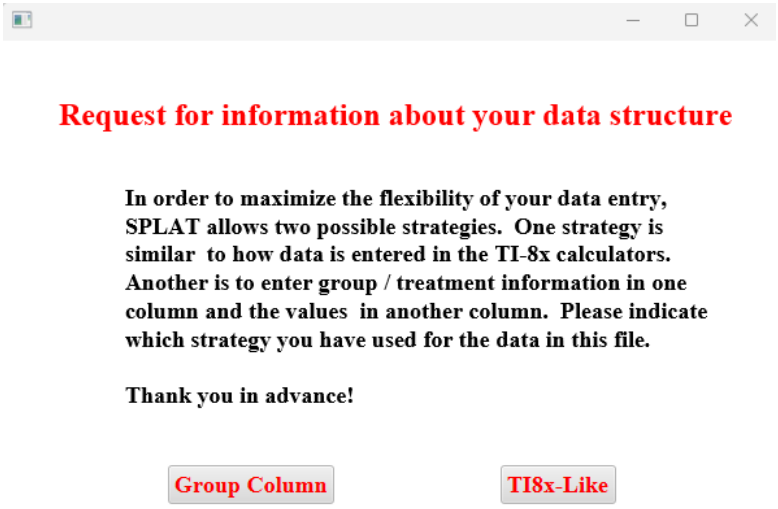
Sacre bleu!!!, as they say on the left bank. Hath something wicked this way comethed, possibly the three witches from Macbeth?? Well, not really. What has happened here is that these data were downloaded from the internet and contain some entries SPLAT is concerned about. In SPLAT, missing data are indicated with an asterisk ("*"). This file may be indicating missing data with a different symbol (e. g. "N/A"), or possibly blanks. SPLAT is therefore worried about the data type, whether the values of this variable are numeric or categorical.

Click on the "You betcha!" button. The entries of concern (periods in this case) will be converted to asterisks in the file. The conversion may not be initially apparent but will appear if you click on the spreadsheet. In real life make sure you have a backup copy of the data!!

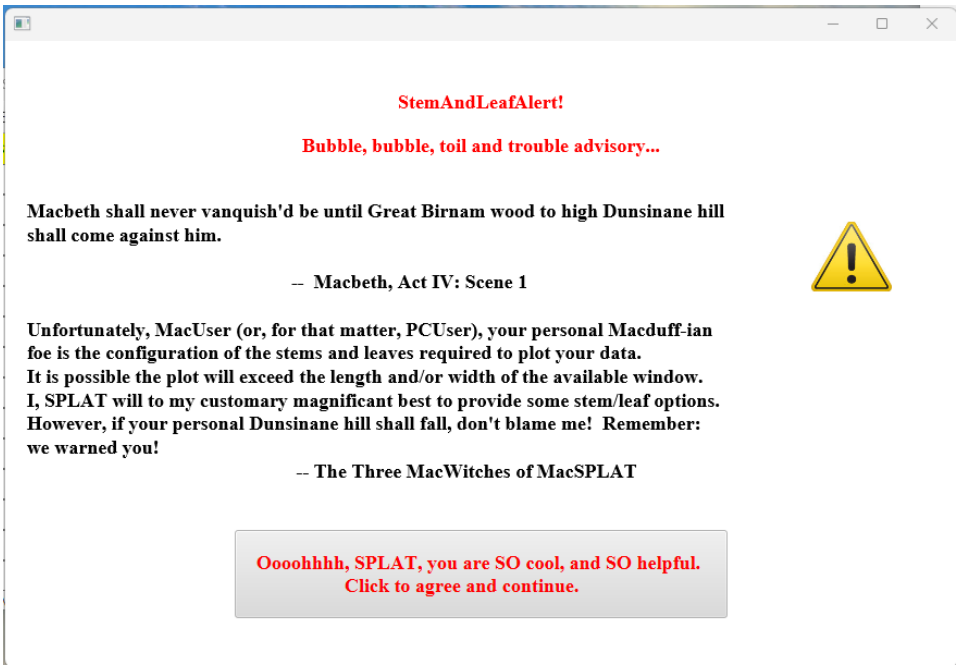
To compare two -- as distinguished from (a) fewer than two, and (b) two or more -- variables, do this: "Explore data → Univariate Data → Compare Exactly Two Quantitative distributions"

For Heaven's sake!!! Yet another request for information! Will no one rid me of these meddlesome requests??? (Answer: No)

There are two different data formats for entering data into SPLAT. (This is not the same thing as discussed above, the TI and Excel stuff.) One format is the List structure that one sees using the TI. The "other" method is to have a single data column (= List) indicate Group membership and then other columns have values for the different variables. The ozone data is "TI8x-Like" so choose that option.



Now select "Stamford" and "Yonkers" (as if you had any other choices) as the variables of interest. The data structure is, again, "TI8x-Like." Fill in the variable names you would like to see in the graphs. I will use Stamford for variable #1 and Yonkers for variable #2, and we will proceed to ANOTHER problem reported by the helpful (??) and perhaps pedantic (!!) SPLAT…



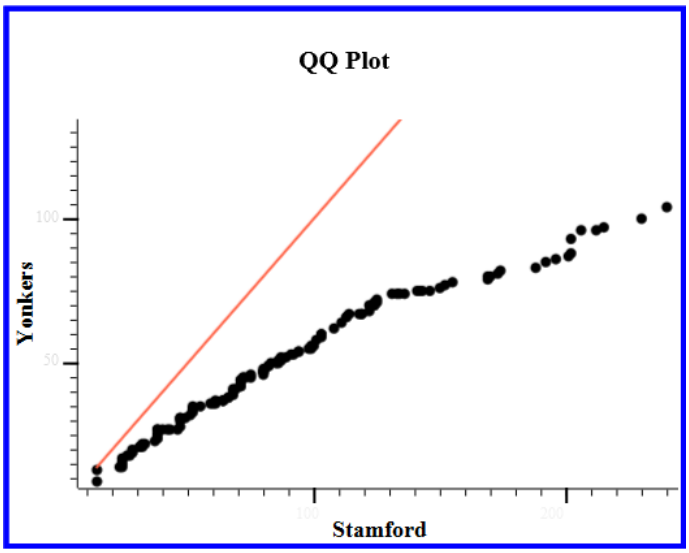
Here SPLAT is warning you about the window size needed to completely display the back to back Stem&Leaf plot. Because the sample size here is very large there may not be enough room to display all the Stem&Leaf options. Clicking your -- by now -- standard agreement and we shall continue.

The QQ Plot

The QQ ("Quantile-Quantile") plot is similar to the Normal Probability Plot but the QQ plot compares two distributions irrespective of their shapes. An approximately straight presentation of the dots indicates the two distributions have approximately similar shapes but may have different means and/or standard deviations. Interpreting QQ plots is something of an art; for more information about the QQPlot, check this site out:

https://en.wikipedia.org/wiki/Q%E2%80%93Q_plot

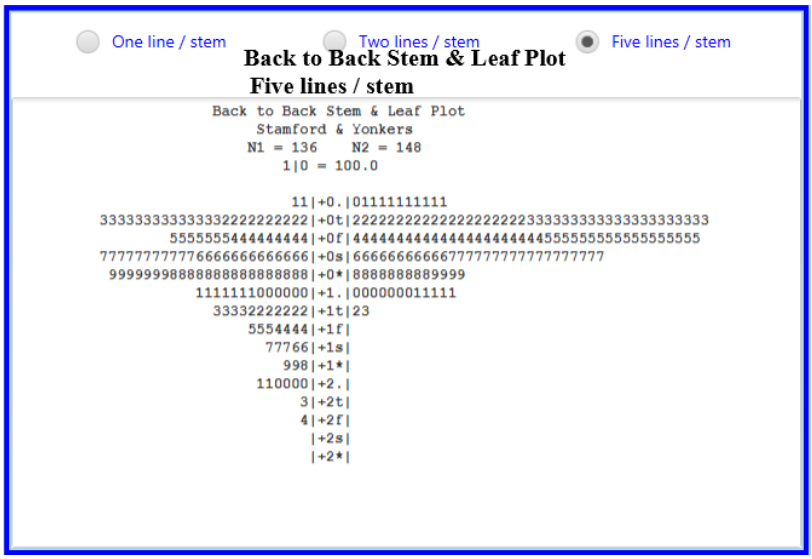
Many statistical procedures assume that both distributions are approximately normal, or at least differ only in location (their means or medians). If the dots in the QQ plot differs from a 45-degree straight line, one might wish to investigate the distributions further.



The Back-to-Back Stem and Leaf plot (BBSLPlot) < **Insert wild applause!!** >

SPLAT and I (well, actually, I) burned many a midnight oil to program the back-to-back stem and leaf plot. (You are welcome.) One, two, or five lines per stem can be chosen to get the best result for your descriptive purposes. The data here is impossible to fit in the window with one line per stem, and the two lines per stem takes up a lot of width in the panel. The five lines per stem option best displays the distributions for comparison IMHO, and in any case is the only one that easily fits. On occasion the stem and leaf plots will require either more rows or more character space than is available in the SPLAT panel – this will happen with particularly large data sets.

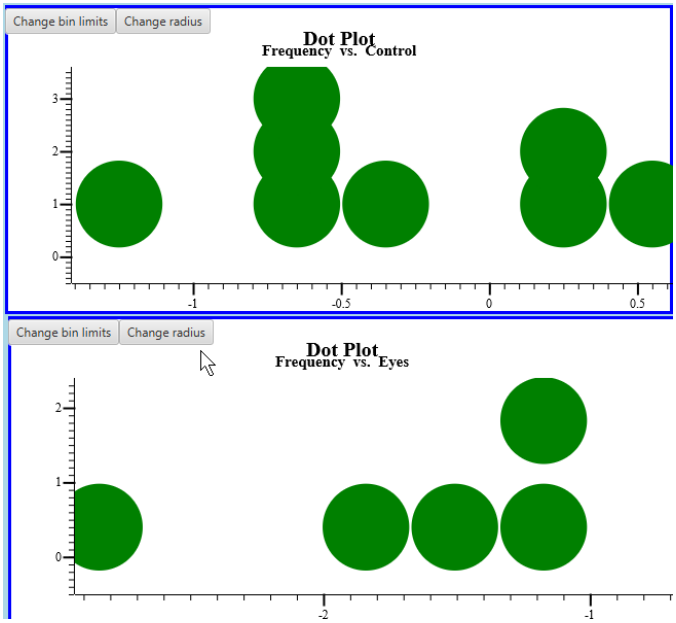
SPLAT will warn you there might be some difficulty in constructing a back-to-back stem and leaf plot for large data sets.



Comparative dot plots

The comparative dot plots work a little differently!!! Hold your breath and deal with it. First of all, the plot may not initially appear; you may have to nudge it by dragging on one of the corners.

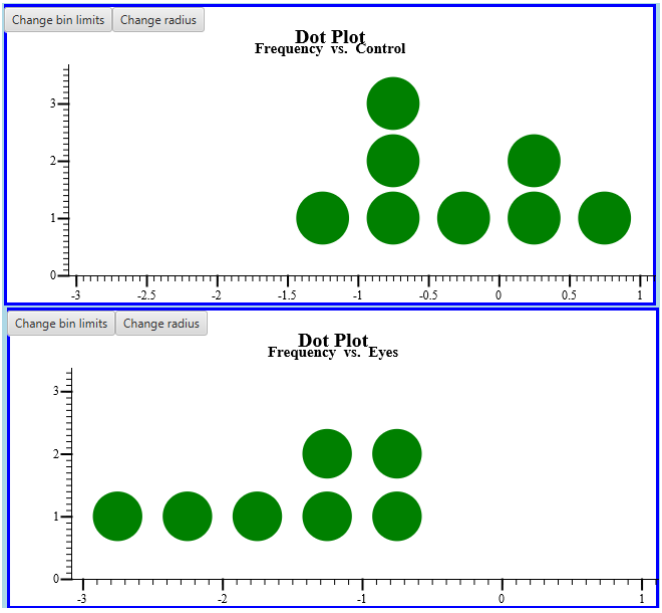
I'm going to cheat a little bit and use a data set on jet lag that will be explained further below. I will consider the Control and Eyes variables. My first look at the dots is not encouraging, dotplot-wise – no dots! But I resized the plots by dragging on the northwest corners…



To get the comparative dot plots above I had to do a bit of messing around:

- Since their original scales differed in their ranges, I needed to fix that. (Click and drag on the horizontal scales.)
- Since their original scales differed, I had to make sure the default bin limits were the same. (I chose -1.5 and -1 for left and right limits.)
- Since the y-axes are constructed separately by SPLAT I had to click and drag to make them approximately the same.
- After that I had to make the dot sizes approximately the same.

That got me to…



I know you are hoping with Polonius that: "Though this be madness, yet there is method in 't." (Hamlet, Act 2, Scene 2). SPLAT is maximizing your choices for displaying the dot plots. Basically, you can move the original-sized dot plots around separately OR move them around together in that panel hiding behind the individual dot plots. (Truth be told, this is a bug, not a feature. It just popped up one day and I decided to keep it.)

In SPLAT it sometimes takes a lot of playing around to get the comparative dot plots to look as you might want. I would make it easier if I knew how; but I don't and thus haven't. In general, the big programming problem with dotplots is how to make the algorithm work with both small and large data sets.

Note that there is another option for displaying multiple-variable dot plots; you will see this in a few pages. If you choose "Two or More Quantitative Distributions" and pick "Dot Plot" in the dashboard you can check this out. It is a relatively new addition to SPLAT and was something requested by an AP Statistics teacher.

If you have more than two quantitative variables…

Circadian rhythms in humans are cycles of about 24 hours. The circadian rhythms provide regularity to a wide range of human biological functions. The cycle is thrown for a bit of a loop when one crosses time zones, resulting in what is known as "jet lag." Over time people adjust as the new time zone gradually resets their internal circadian clock. Campbell & Murphy (1998) reported that in addition to the usual resetting of the clock by simply doing normal stuff in the light of day, the internal clock could be reset by exposing the back of the knee to light. (I am NOT making this up!)

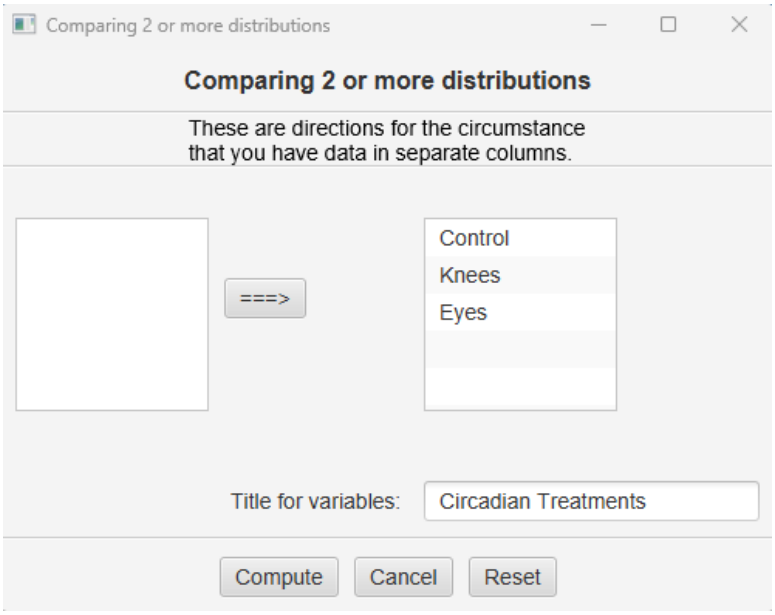
Wright & Czeisler (2002) attempted to replicate the experiment, with more and better controls. Their treatment groups consisted of a Control group, a Knees group, and an Eyes group. During the experimental trials participants were awakened from sleep and a three-hour session of bright lights were applied to the Eyes, Knees, or Neither. The response variable is the magnitude and direction of phase shift in the daily cycle of melatonin production. The data are in the file, CSV_Circadian.

Campbell, S. S., & Murphy, P. J. (1998). Extraocular Circadian Phototransduction in Humans. *Science*, 179:396-398.

Wright, K. P., & Czeisler, C. A. (2002). Absence of Circadian Phase Resetting in Response to Bright Light Behind the Knees. *Science* 297:571.

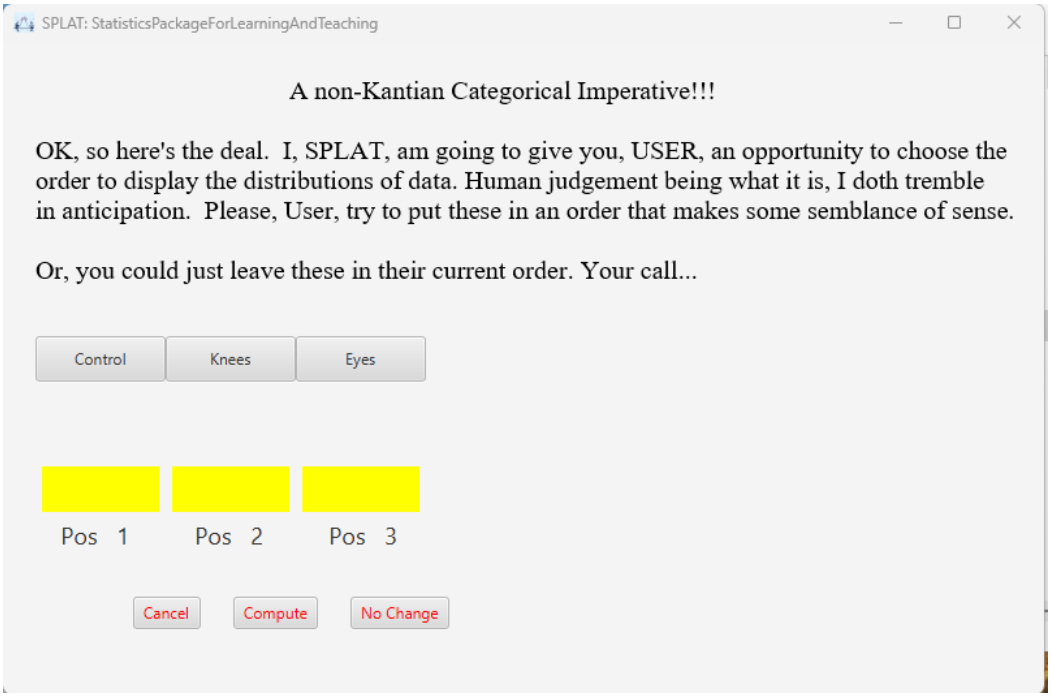
To compare two or more, as distinguished from (a) fewer than two, and (b) exactly two distributions, do this: "Explore data → Univariate Data → Compare Two or More distributions." As usual, SPLAT will ask about your data organization; in this file the data are organized in the "TIx-Like" format.

Clicking on all three variables, provide a title, and then click on "compute." I chose "Circadian Treatments" as my Title.

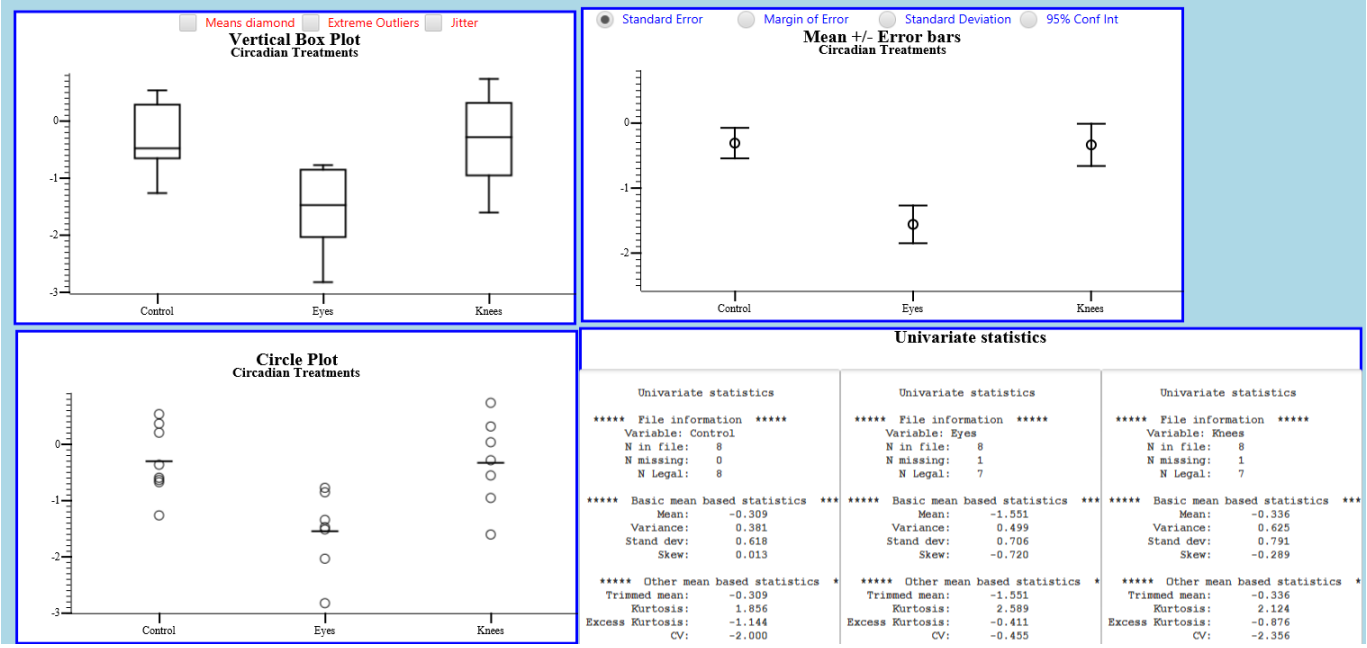


Now you will see a feature that appears now and then with categorical variables in SPLAT: you can choose the order of presentation (or not) in the subsequent graphs. Just click and drag the buttons into positions.

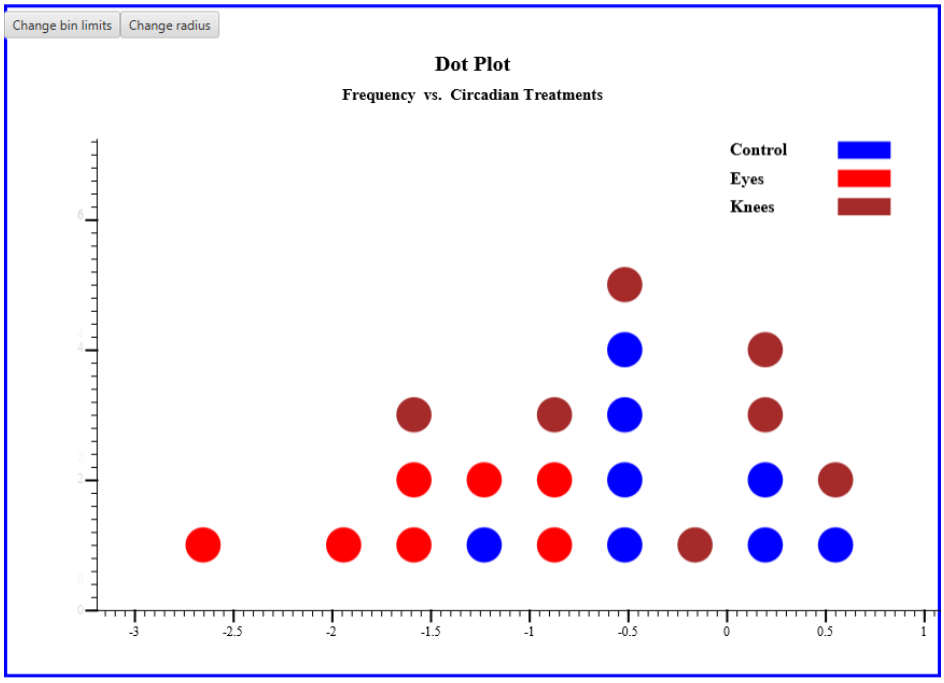
For example, to alphabetize the order, drag Control to Position 1, Eyes to Position 2, and Knees to Position 3, and go for it!



The various plots shown below have been already resized, moved around, and rescaled to be able to present them all together. (Couldn't quite get the complete table in with the other three.)



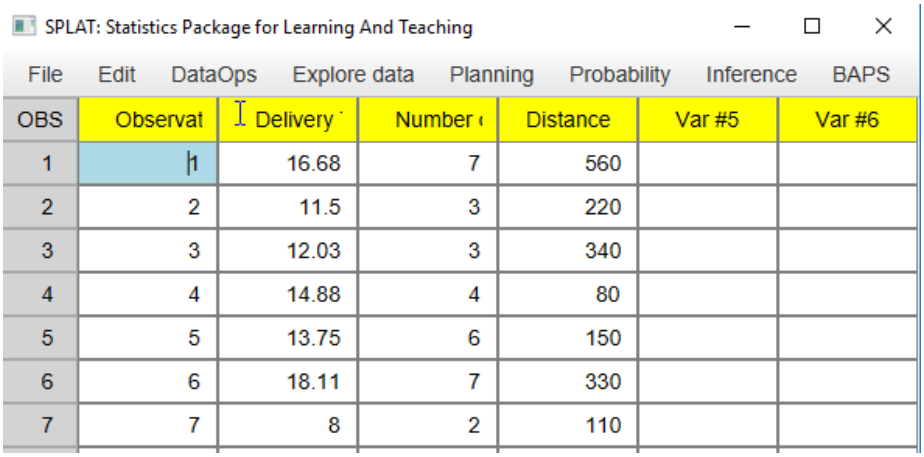
Here is the new multi-variable dot plot mentioned earlier:



I had to click and drag on the y-axis to get the dots to behave initially. (Have to work on this problem…) Also, the stuff on the right was initially hidden so I had to widen the panel.

On to Simple Regression!

OK, close the blue Univariate dashboard. We are once again back at the main menu. Click on "File → Clear Data" to start fresh, and open the file, CSV_DeliveryTime. These data are delivery times from truck to convenience store for several stores. I ~~stole~~ borrowed the data from my fav regression book, Montgomery, Peck, Vining, Introduction to Linear Regression Analysis.

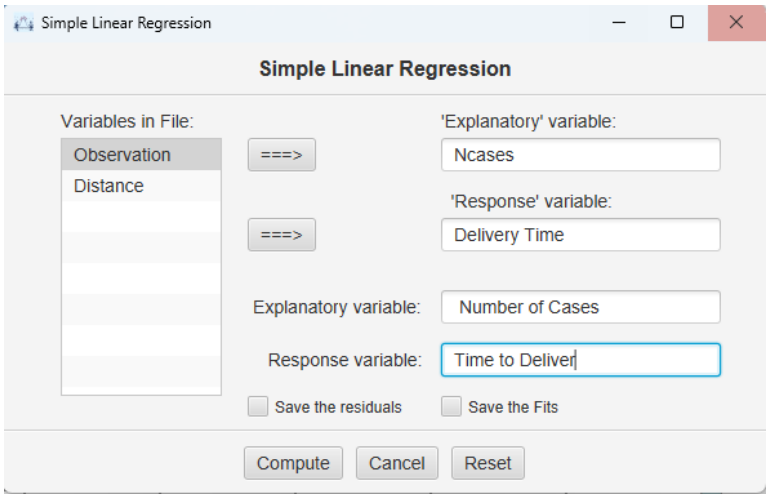


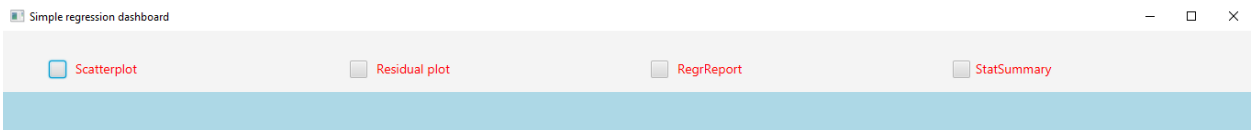
OBS	Observat	Delivery	Number	Distance	Var #5	Var #6
1	1	16.68	7	560		
2	2	11.5	3	220		
3	3	12.03	3	340		
4	4	14.88	4	80		
5	5	13.75	6	150		
6	6	18.11	7	330		
7	7	8	2	110		

In SPLAT there are two paths to linear regression. If you choose the sequence "Explore data → Bivariate data → Linear regression" you will get the "simple" version of regression output. If you choose the sequence "Inference → Regression" you get additional regression goodies of a diagnostic nature. Let us take the road more traveled by early in a statistics course: the "Explore data" route.

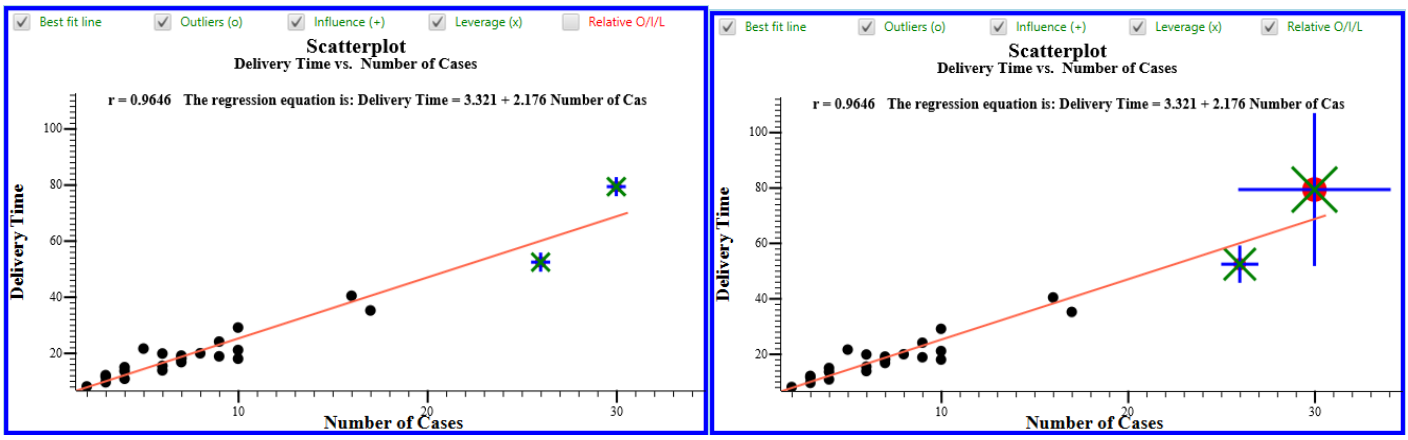
Choose "Delivery Time" as the Response Variable (second arrow), and "NCases" as the explanatory Variable (top arrow). The "Save the *xxx*" options add the residuals and/or the *y*-hats to the SPLAT spreadsheet; you may check those if you wish. Now click on "Compute."

Once Compute is clicked, the regression dashboard will appear:





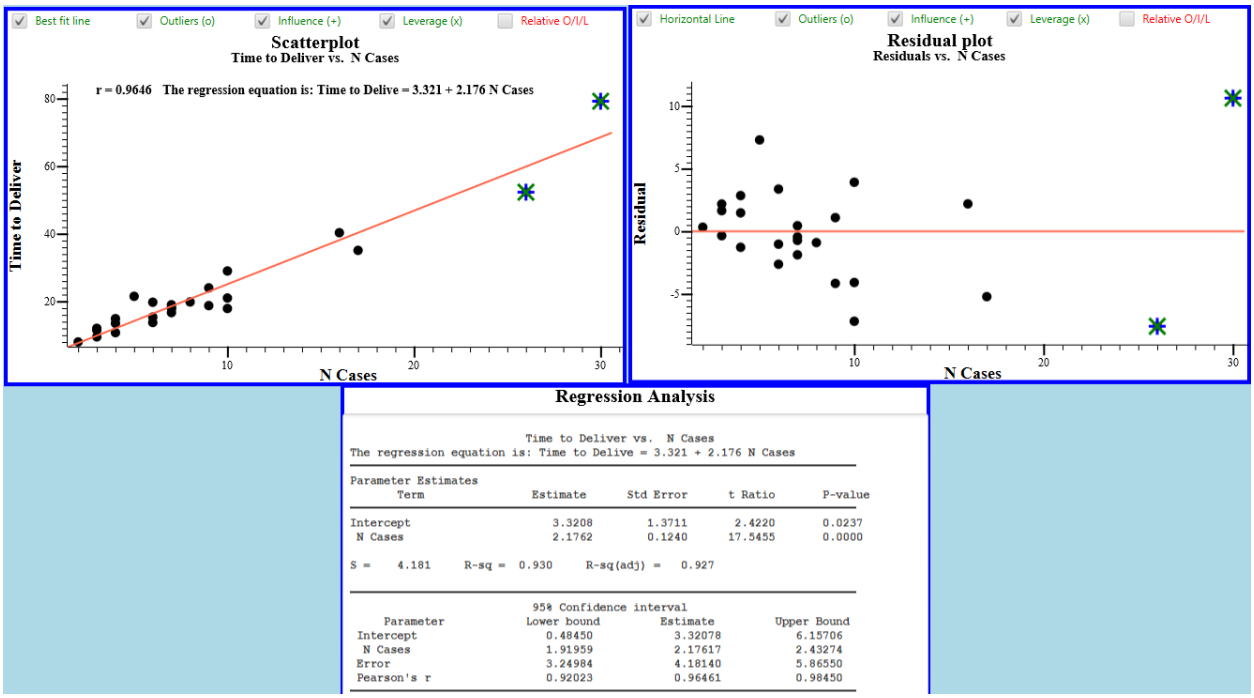
First, click on "Scatterplot." In the panels below I have checked four of the five plot options presented with the Scatterplot option: Best fit line, Outliers, Influential points, and High Leverage points. Outliers are shown as red dots. High leverage points are shown as green x's. Influential points (combination of sufficiently outlying and/or high enough leverage points) are indicated by blue plusses. The correlation will always appear in the scatterplot. The regression equation will be displayed only in the "No inference" version of the regression output and then only if the "Best fit line" is checked. The basic reason for these choices is to provide a "simple" regression output for classes with few statistical needs, like AP Psychology, AP Biology, or perhaps an elementary science class. Note that the northeasterly point -- (30, 79.24) -- is also an outlier, but the influence and leverage indications cover it up in the scatterplot on the left.



If the "Relative O/I/L option is chosen, the "amounts" of outlier-hood, influence-hood, and leverage-hood are represented as proportionally larger circles and crosses and plusses. The calculation formulas for the circles and crosses and plusses are way BAPS – it is the concepts these graphics are intended to illustrate. (If you wish to see the diagnostic statistics that underlie the sizes for the circles, crosses, and plusses, choose the inference version of Linear Regression.)

The different sizing of these symbols portrays the relative "amounts" of influence, outlier-hood, and leverage of each point. The measure of influence is Cook's D, a statistic commonly used in multiple regression, but also appropriate in simple regression. The light blue color indicates a Cook's D greater than 0.5 ("interesting"). A Cook's D greater than 1.0 ("Seriously interesting!") would appear in dark blue. The areas of the graphics are proportional to the respective measures of outlier-hood, leverage-hood and influence-hood. For complete descriptions of these ideas and calculations, I recommend any edition of my above-mentioned fav, Montgomery, Peck, Vining, Introduction to Linear Regression Analysis.

Recall that you can click and drag on SPLAT's numeric axes to change them to fit your taste. In this plot I had to adjust the Y-axis again to show the influential point (30, 79.24) clearly. Recall also that you can click on the graphs and drag them around on the dashboard to show what combinations of graphs you wish and where you want them. So, for example, you might want to see the regression line, the residual plot, and the regression information at the same time.



Comparing Regressions (Descriptive)

In addition to means and proportions it is also possible to compare linear relations. Here we will consider this only as a problem of description; inference for differences among slopes will rear its ugly head in our presentation of the way BAPS topic of Analysis of Covariance.

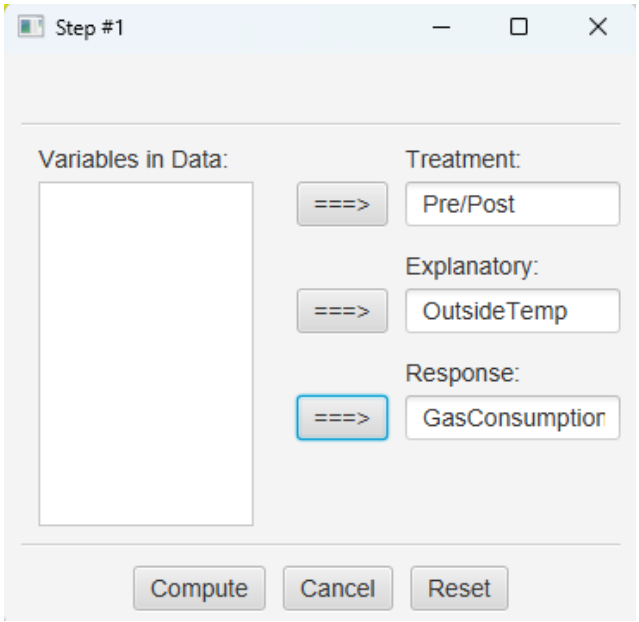
I lifted the data from Tamhane, A. C. (2009). *Statistical Analysis of Designed Experiments: Theory and Applications* and is in the file, CSV_HomeHeating. Of interest is the effect of insulation on home heating. The variables are temperature centigrade, Yes/No insulation, and the consumption of heating gas in thousands of cubic feet. The data are organized as shown in the "Not TIx" format in the file. Treat_A is Pre-insulation, Treat_B is Post-insulation in the same English house.

SPLAT: StatisticsPackageForLearningAndTeaching

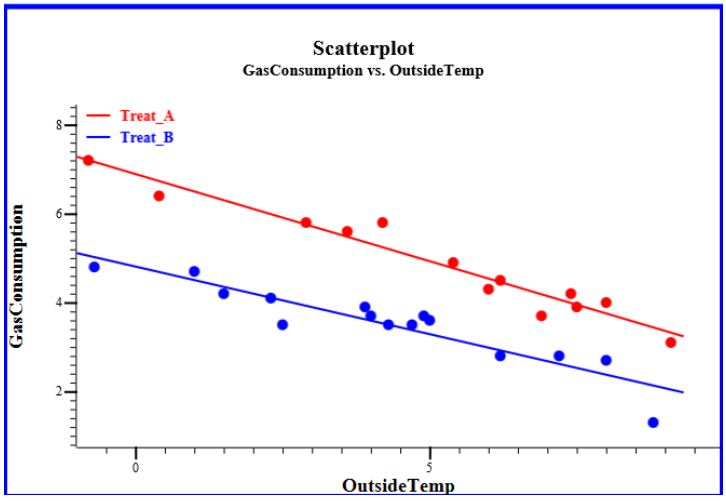
File Edit Ops DataOps Explore data Planning Probability Inference BAPS

OBS	OutsideTemp	GasConsump	Pre/Post	Var #4	Var #5	Var #6
1	-0.8	7.2	Treat_A			
2	0.4	6.4	Treat_A			
3	2.9	5.8	Treat_A			
4	3.6	5.6	Treat_A			
5	4.2	5.8	Treat_A			
6	5.4	4.9	Treat_A			
7	6.0	4.3	Treat_A			

Select Explore data → Bivariate data → Compare regressions, make the choices shown, and click on Compute.



The usual suspect regression plots are available in SPLAT, but I will present only two here.



The slopes of both lines are negative, which is what we would expect – gas consumption is less for higher temperatures. In addition, our reading of the intercepts is that the gas consumption goes down about 2000 cubic feet on average at all the outside temperatures in the rang of the data after putting in insulation.

Regression Comparison

********** Univariate parameter estimates for Groups/Treatments **********

Treatment/ Group	Sample Size	Sample Mean	Sample St Dev	Std Err of mean	Margin of Error	Lower 95PC Bound	Upper 95PC Bound
Treat_A	13	4.877	1.195	0.345	0.711	4.166	5.588
Treat_B	15	3.520	0.874	0.233	0.481	3.039	4.001

********** Bivariate parameter estimates for Groups/Treatments **********

Treatment/ Group	Slope	Intercept	Correlation
Treat_A	-0.392	6.892	-0.973
Treat_B	-0.304	4.810	-0.922

Homogeneity of Slopes

Source of Variation	Sum of Squares	df	Mean Square	F	P-value
Heterogeneity	0.394	1	0.394	3.767	0.0636
Residuals	2.508	24	0.104		
Within Resids	2.901	25			

The statistical stuff presents the usual regression stuff, and in addition a hypothesis test for equal slopes. A hypothesis test for equal intercepts is the stuff of which the Analysis of Covariance is concerned. At this point we will remain mercifully ignorant of these inference procedures and just be thankful for the graphs. ("Please, sir, I want some more" might work in Oliver Twist, but not here.)

On to Still Simple but Nonlinear Regression!

SPLAT performs elementary transformations of data. These transformations can be used when one is confronted with bivariate relationships that are not quite straight-ish when plotted. As an example, consider the problem confronting G. K. Adams and his friends Robin and Lydia in their study of chameleon eggs. [Adams, G. K., et al (2010), Eggs under Pressure: Components of Water Potential of Chameleon Eggs during Incubation. *Physiological and Biochemical Zoology* 83(2): 207-214].

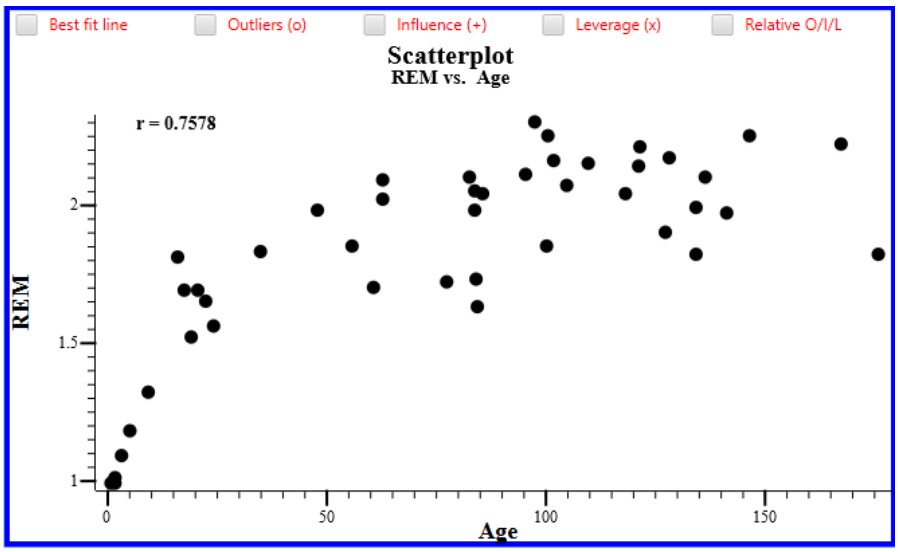
The investigators were interested in the relationship between the relative egg mass (mass on the day the egg was sampled divided by its initial mass) and the age of chameleon eggs. Mother Nature threw Adams, et al. a curve. Their data is in the file, CSV_ChameleonEggs, and the initial cases look like this:

SPLAT: Statistics Package for Learning And Teaching

File Edit DataOps Explore data Plan

OBS	Age (days)	Relative	Var #3
1	0.9	0.99	
2	1.8	0.99	
3	1.8	1.01	
4	3.3	1.09	
5	5.2	1.18	
6	9.4	1.32	

Explore Data → Bivariate Data → Linear Regression brings us to the following scatterplot of Relative Egg Mass vs. Age:



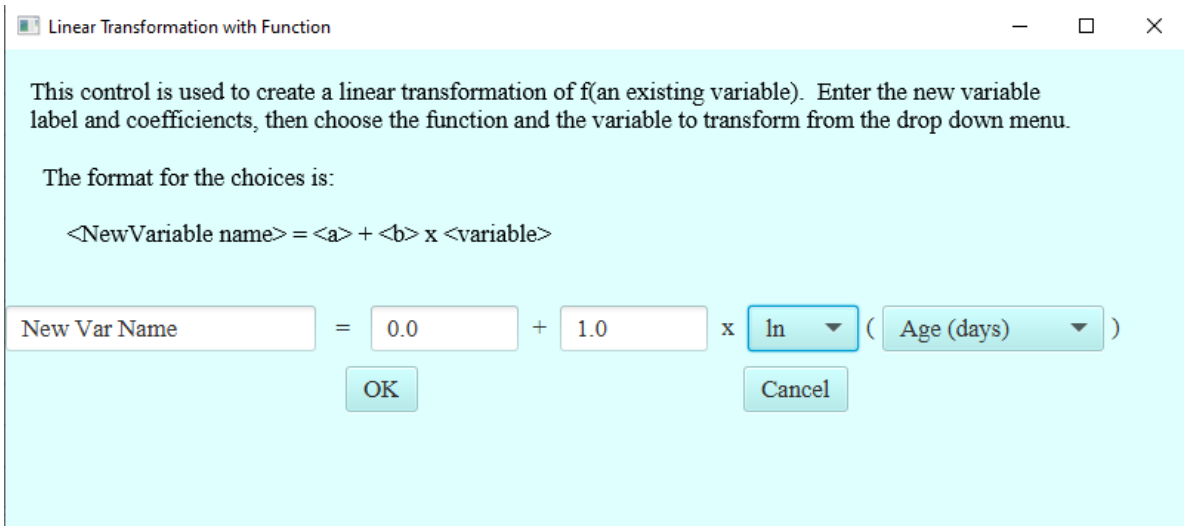
A straight-line model does not seem adequate; a quick mental run through elementary algebra functions suggests a logarithmic model may be better used to capture this relationship over the domain of the data. We will transform our Age variable and construct a best fit line of the form, $REM = a + b * Ln(Age)$. To accomplish this, we need to first navigate back to the spreadsheet and take a different fork in the road. Click on that red X (or red circle if you are on a Mac) to close the simple regression dashboard and return to the SPLAT spreadsheet…

SPLAT: Statistics Package for Learning And Teaching

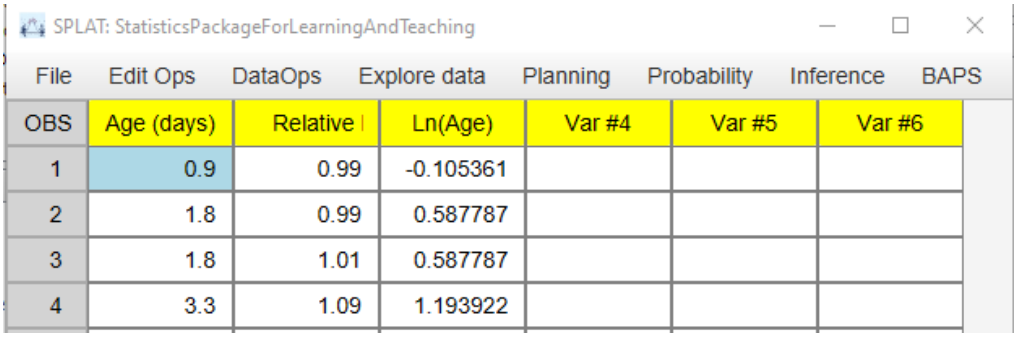
File Edit DataOps Explore data Plan

OBS	Age (days)	Relative	Var #3
1	0.9	0.99	
2	1.8	0.99	
3	1.8	1.01	
4	3.3	1.09	
5	5.2	1.18	
6	9.4	1.32	

Click on DataOps → Nonlinear transformations, and the following panel will appear:

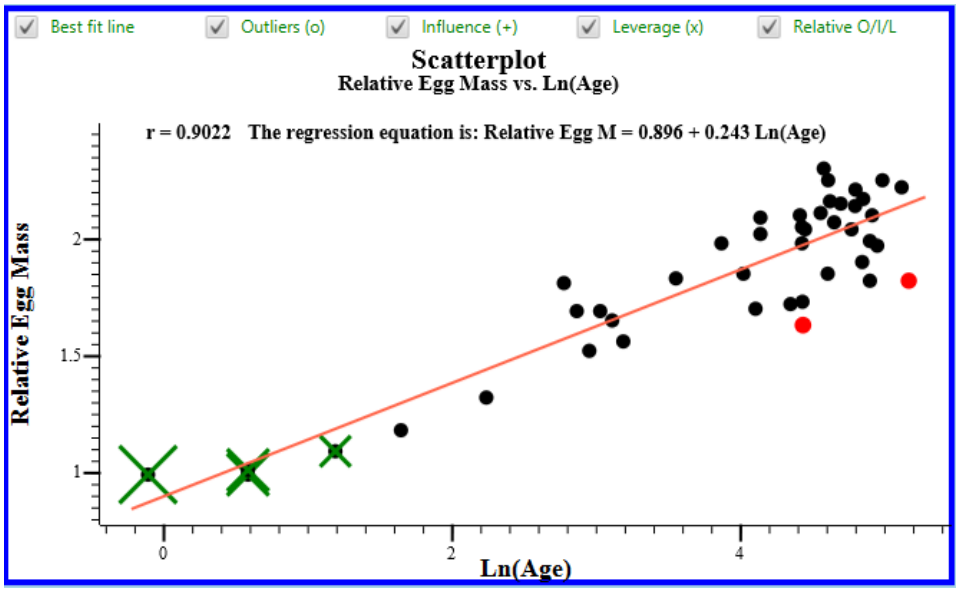


Wow, this is a stroke of luck: the default choice appears to be just what we are looking for. We should provide a better name for the variable, however, and "Ln(Age)" seems to be a reasonable choice. Click on the text field that now has "New Var Name" and make that change. After clicking on OK, our SPLAT spreadsheet reappears with the results of the calculation:



OBS	Age (days)	Relative	Ln(Age)	Var #4	Var #5	Var #6
1	0.9	0.99	-0.105361			
2	1.8	0.99	0.587787			
3	1.8	1.01	0.587787			
4	3.3	1.09	1.193922			

Repeating the SPLAT procedures for linear regression, [x = Ln(Age), y = Relative Egg Mass] we arrive at a straight line fit to the transformed data, this time indicating some high leverage points with low values of *Ln(Age)*. After a bit of clicking and dragging on the axes to bring the points away from the edges, and clicking all the options at the top, the points appear in all their glory within the boundaries of the plot.

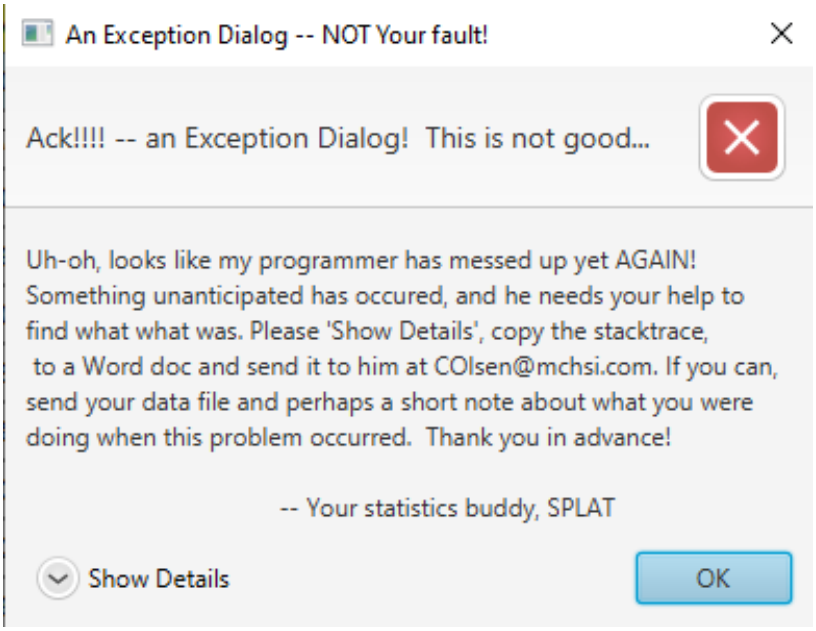


A note on transforming variables in SPLAT

Some problems can occur with data transformations: not all functions and operations are legal with all real numbers. Logs and square roots of negative numbers, division by zero, stuff like that, are frowned upon. SPLAT attempts to warn about these possible problems as they occur.

A note of humility!

SPLAT advisories and error messages are constructed from my long experience helping cherubs try to figure out what statistical software is telling them after they have strayed into error. I have endeavored to trap as many of these sorts of problems as I can, but I suspect SPLAT may fall prey to the occasional seriously Demonic Student User. In that circumstance, SPLAT may report the following Exception dialog:



If this happens, please "Show Details" and send me the (hopefully helpful) diagnostic information presented there. You are welcome to read said diagnostic information but trust me; Agatha Christie mysteries are more pleasant to read than diagnostic mysteries! Or maybe Micheal Connelly for the younger set.

Note that the email should be crolsen@fastmail.com. For reasons unknown I am not able to generate an unanticipated error to show you; must be getting more difficult to be Demonic.

If this appears, it will have the correct email! Since we really don't know why the message has appeared it is a good idea to exit SPLAT and start over. You DID keep a backup of the data, right?

One-parameter models in the AP Statistics CED?!?

The AP Statistics Course and Exam Description, "Essential knowledge" VAR-7.M.2, mentions what is usually referred to as one-parameter or "no-intercept" regression. Within AP Statistics this presumably involves the linear model: $y = \beta_1 x + \varepsilon$. No-intercept models are at best controversial and interpreting them is almost never easy. Consider the research by Brian Stafford and his colleagues. (Stafford, B. J., et al. (2002). Gliding Behavior of Japanese Giant Flying Squirrels (*Petaurista leucogenys*). *Journal of Mammalogy* 83(2):553-562). FYI, "Giant" in this case means a body of 25 – 50 cm with a tail an additional 30 – 40 cm.

Among other things, Brian and colleagues measured the horizontal distances and altitudes lost for a set of glides by these creatures. It stands to reason that in theory, if one of these glides has a horizontal distance of 0.0, or near 0.0, the altitude lost will also be 0.0, or near 0.0. So, the model, $y = \beta_1 x + \varepsilon$ seems like it might be a possible choice.

There are heavy-duty interpretation problems with regression through the origin. A no-intercept model can wreak havoc with the slope (in red below). Does that no-intercept slope look at all like it could be interpreted as the usual average increase in the response variable per one unit change in the explanatory variable? (Hint: No.)

One supposes that the choice between a one-parameter model and the usual two-parameter model could unfold something like this:

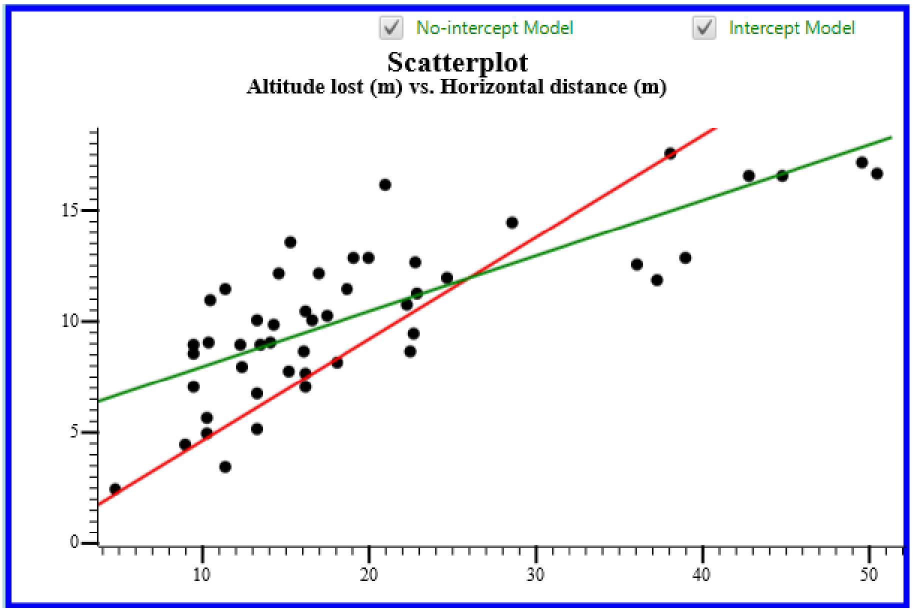
Step 1: In the usual two-parameter model, test the hypothesis, $H_0 : \alpha = 0$.
Step 2: If the hypothesis is not rejected, judge the one-parameter model to be reasonable.

But if the one-parameter model is judged to be reasonable there is the further problem of how to compare it to the usual (and perhaps also reasonable) two-parameter model. The only directly comparable statistic between the intercept and no-intercept models is the **Mean Square for residuals**, circled below with a value of 12.27. The mean square for residuals for the corresponding two-parameter model is 5.2; smaller is better. One cannot fruitfully compare the R-squared, the standard deviation of the residuals, or the correlation. If one is looking for the "best model," – or, in this case, the "better model" -- how does one choose? What information does the one-parameter give you that is superior to that given by the two-parameter model?!?!?

My $0.02 advice in general: Choose no-intercept models at your peril!!

But if you like to walk on the perilous side of the street, here is what SPLAT will deliver…

SPLAT displays the no-intercept best fit (red) and the usual with-intercept best fit line (green).



If you wish to explore the one-parameter option in SPLAT, Explore Data → Bivariate Data → One Parameter Linear Regression is the path. The data are in the file, CSV_GlidingData.

No-intercept Regression Analysis

```
Altitude lost (m) vs. Horizontal distance (m)

The regression equation is: Altitude lost = 0.458 Horizontal dis

Parameter Estimates
        Term                  Estimate     Std Error     t Ratio      P-value

 Horizontal dis                 0.4579       0.0220       20.782       0.0000


   **********************************************************************
   *****     NOTE: No intercept in model.  R-Square is redefined!     *****
   *****     For regression through the origin (the no-intercept      *****
   *****     model), R-square measures the proportion of variability *****
   *****     in the response variable about the ORIGIN that is        *****
   *****     explained by regression.  This CANNOT be compared to     *****
   *****     R-Square for models which include an intercept.          *****
   *****     The ONLY statistic appropriate for comparing this model *****
   *****     to the usual (intercept) simple regression model in the *****
   *****     ANOVA table below is the mean square of the residuals!   *****
   **********************************************************************

                            Analysis of Variance

Source of         Sum of     Degrees of
Variation         Squares     Freedom        Mean Square       F          P-value

 Regression       5300.837        1            5300.84      431.90        0.000
 Residual          589.113       48              12.27
 Total            5889.950       49

                                R-square = 0.052
```

Anticipating inference: A short note about effect sizes in SPLAT

In inferential statistics the P-value gives us a measure of "statistical" significance, the probability that the difference between our expectations and results would occur if chance alone were operating. "Statistical" significance is usually distinguished from "practical" significance.

Measures of "practical" significance are legion and are typically referred to as "effect sizes." Since students typically confuse statistical and practical significance, SPLAT's inference procedures calculate, distinguish, and report effect sizes. In real life statistical software, a researcher might have a choice of effect size to report. SPLAT, on the other hand, reports what I believe to be the most commonly used effect size for means and proportions: Cohen's d and Cohen's H. SPLAT reports Cramer's V for measures of association between categorical variables and, of course, Pearson's r and the Coefficient of Determination for association between quantitative variables. For ANOVA, the omega squared, and Cohen's d are reported.

An understandable (and short) discussion of effect sizes and P-values can be downloaded at:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3444174/pdf/i1949-8357-4-3-279.pdf

For an accessible discussion of P-values and Friends, I recommend Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar (2019) Moving to a World Beyond "$p < 0.05$", The American Statistician, 73:sup1, 1-19, DOI: 10.1080/00031305.2019.1583913.

For the actual calculations of effect sizes, I have relied on these sources:

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd). Mahwah, NJ: Lawrence Erlbaum Associates.

Grissom, R. J., & Kim, J. J. (2012). Effect Sizes for Research: Univariate and Multivariate Applications (2nd). New York: Taylor & Francis.

"Effect size" https://en.wikipedia.org/wiki/Effect_size#:~:text=are%20considered%20large.-,Cohen's%20d,similarly%20for%20the%20other%20group.

You should note that the values of Cohen's d for means reported in SPLAT are unbiased corrections of the "raw" Cohen's d and may differ from reported values if you check SPLAT's calculations with data from research reports. (See Grissom & Kim, p70).

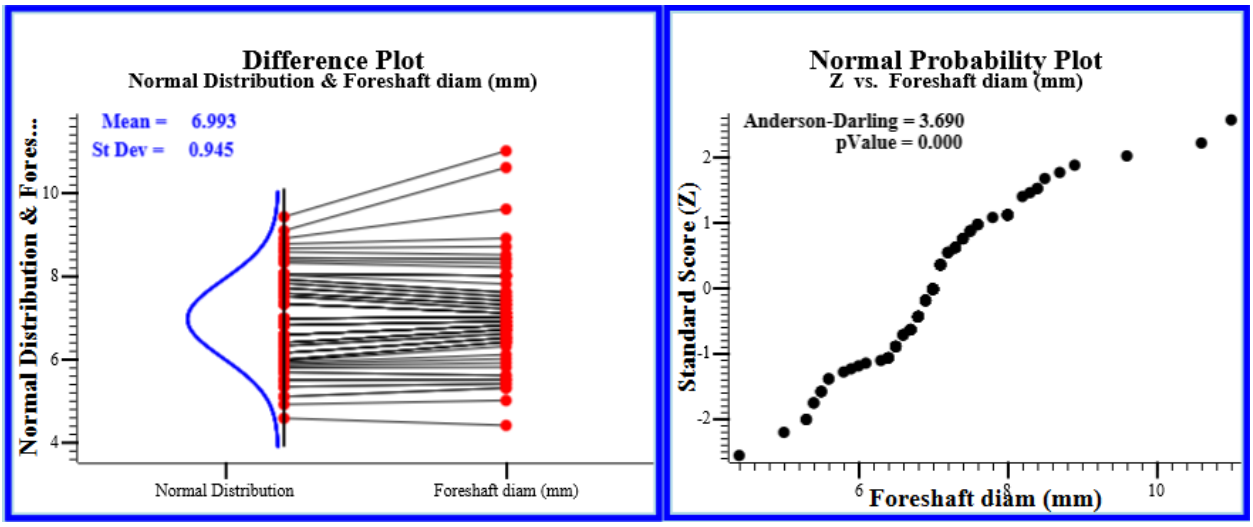
OK, enough of this technical stuff – on to inference!

A Non-Boston t Party – Inference for means

I would like to illustrate the inference procedures in SPLAT, starting with inference for a single population mean. The sampling distributions of the test statistics for independent *t* and matched *t* procedures are *t* distributions, and the sampling distribution for the test statistics for the difference between means is approximated by a *t* distribution. Thus, the presentations of results for these three procedures are similar.

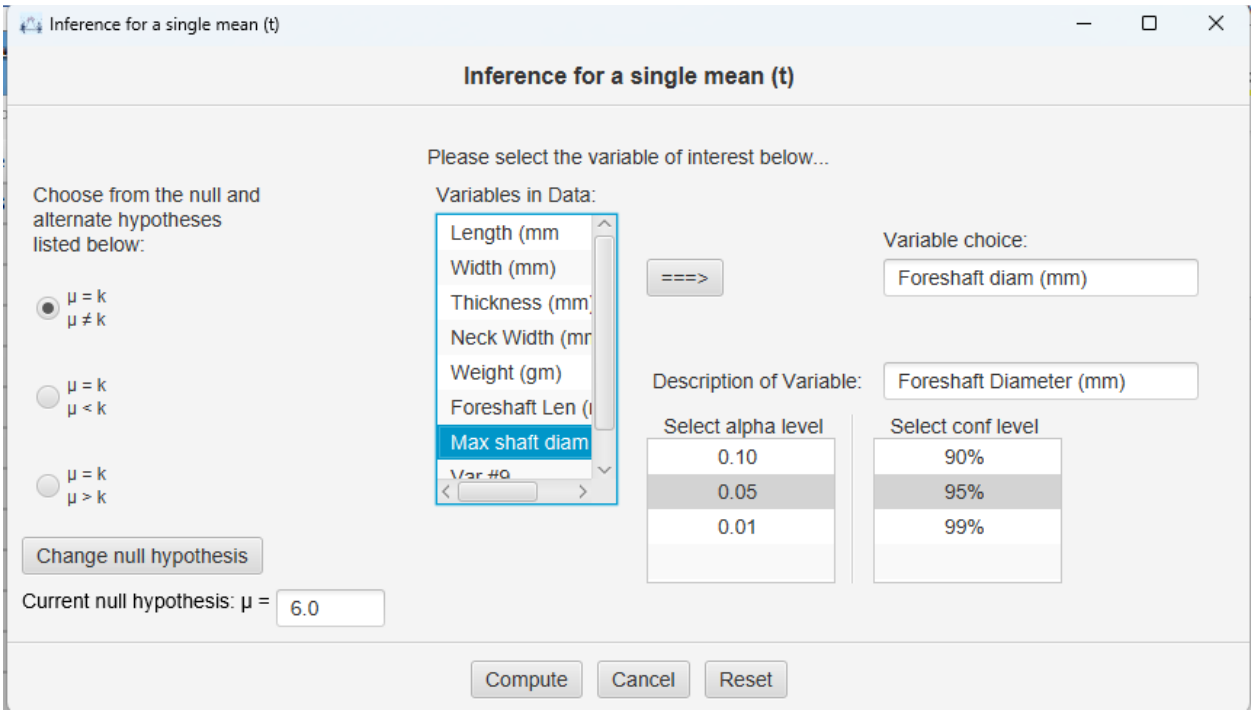
I have dredged up some data on Clovis projectile points from [Thomas, D. H., Arrowheads and Atlatl Darts: How the Stones Got the Shaft. *American Antiquity* 32(3): 461-472.] The measure of interest is the foreshaft diameter (mm). The data are in the file, CSV_Arrowheads.

We are mindful of the presumption of the plausibility the data were drawn from an at least approximately normal population. Taking advantage of SPLAT's univariate graphing capability to "Explore data →Univariate Data →A Single Quantitative Variable, " we will assess that plausibility. Choose the variable "Foreshaft Diam (mm)" in the file and pick your favorite univariate plot; here are the NormalDifference plot and normal probability plot for the foreshaft diameters:



We can see that the distribution wiggles around. The difference plot shows slopes both positive and negative and the normal probability plot exhibits slopes of tangent lines of varying value. There are hints of outliers on both ends of the distribution. The Anderson-Darling P-value is below the .05 level of significance. We note in passing that there are 117 data points, and recall that the t-procedures are very robust, so we will cross our fingers, and proceed. Well, OK, we might institute a bit of whistling in the dark also.

Now that we are OK with the presumption of approximate normality in the population, navigate back to the spreadsheet and execute the sequence, Inference → One mean. Since we have actual data, rather than pre-existing sample statistics, click on "Data" in the information request panel. Here is the single-mean inference panel with my choices already entered. Somewhere on the internet I found out that Clovis diameters range from 3 to 9 mm, so I will use a null hypothesis in the middle for this example: $\mu = 6mm$.



One can choose the options in this panel in any sequence; nothing is etched in cyber-stone until the Compute button is pressed. Click on "Change null hypothesis" and indicate a hypothesized mean of 6, leaving the alternative hypothesis as a "not equal" and the alpha level at .05. Then click on Compute. The resulting "Inference report" for the *t*-test looks like this:

Inference for a single mean report

*** Summary information ***

	NSize	Mean	StDev	StErr
Foreshaft Diame	117	6.993	0.945	0.087

*** Hypothesis Test ***

Null hypothesis: μ = 6.0
Alt hypothesis: μ ≠ 6.0

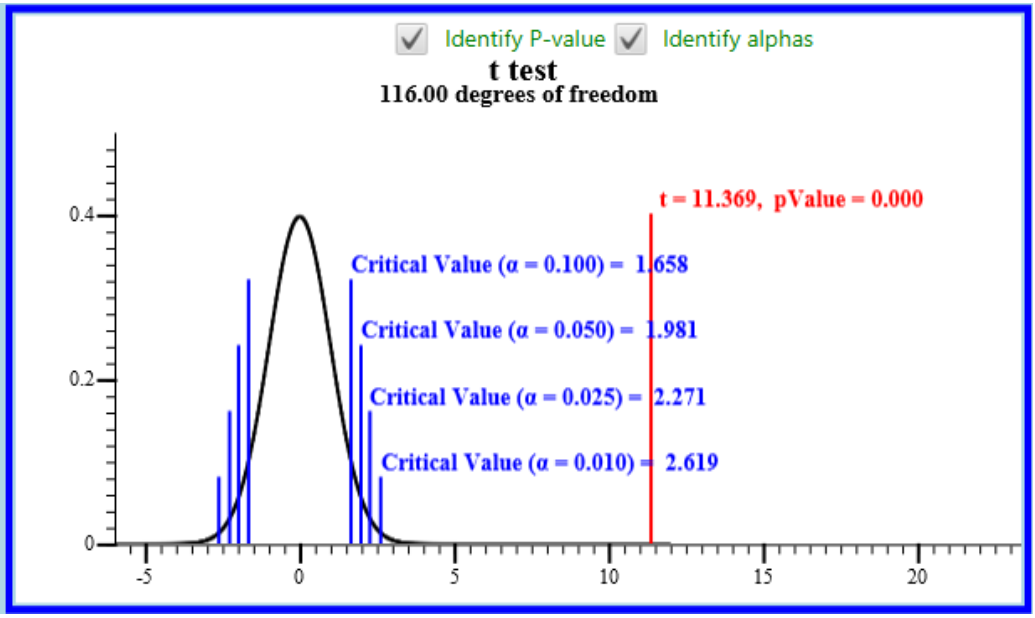
*** Hypothesis Test & 95% Confidence interval ***

Mean	StandErr	df	t_Value	pValue	ciLow	ciHigh
6.993	0.087	116	11.369	0.000	6.820	7.166

Effect size (Cohen's D) = 1.044

You may not be familiar with Cohen's D. It is a measure of "effect size" briefly discussed above in this Guide. For some enlightenment on Cohen's D check out the sources indicated above. Suffice it to say that a Cohen's *d* above 1.0 is huge; my null hypothesis is seriously embarrassed! Now click on the "t-test" button to see where the t-statistic fits in the sampling distribution, and then click on the two buttons to get the P-value and some critical values. You will probably have to click and drag on the right side of the x-axis a bit to see everything. Now click on the "Identify P-value" and "Identify alphas" buttons.

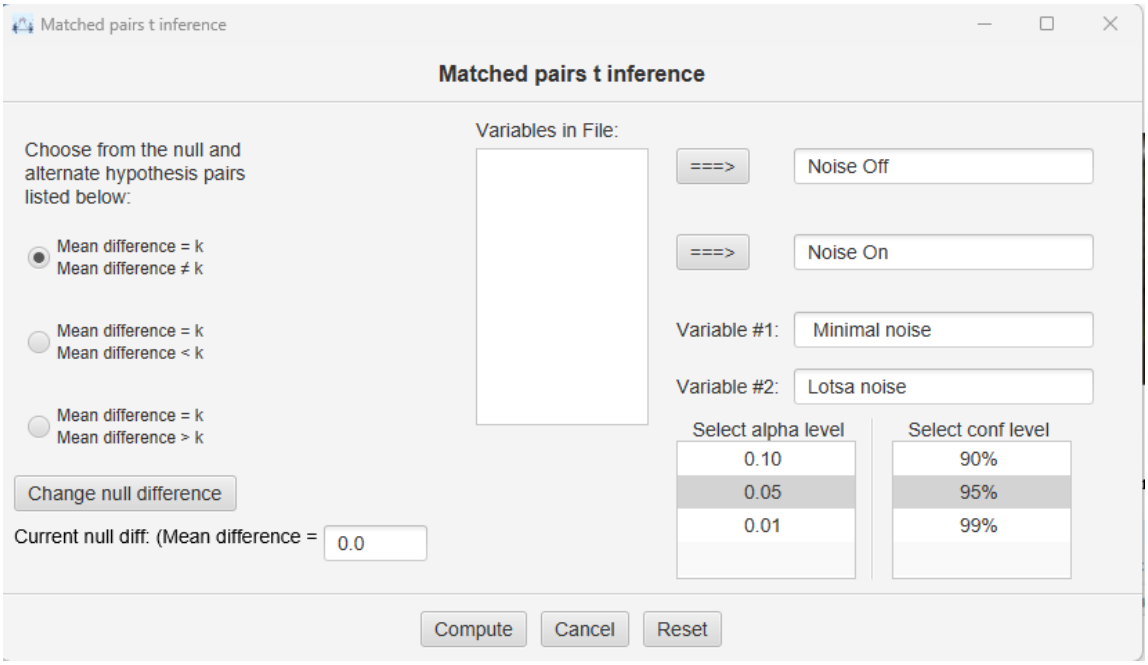
The main reason I added the option of showing all this Critical Value stuff is to assist AP Biology students who may see some of this terminology in their classes but may be a bit hazy about how these terms are related.



Another Non-Boston t Party – Inference for paired means

Inference for paired data is inference for a single population of differences between pairs. For our example we will use data in the file, CSV_Chaffinch. [Quinn, J. L., et al. (2006). Noise, predation risk compensation and vigilance in the chaffinch *Fringilla coelebs*. *Journal of Avian Biology*. 37:601-608.] The investigators were studying the effect of urban noise on the foraging behavior of these birds. It is generally true that because of the location of the eyes on their heads, birds are not able to peck for food on the ground and scan for predators at the same time. The data are mean times in seconds that the birds had their heads "down," between those times they were scanning for predators. There are two samples in this study: observed in "noise" and "no noise" conditions.

Choose Inference → Paired mean, and the familiar choices for inference will be presented.



Click on Noise Off and Noise On as the variables of interest. SPLAT will provide the usual suspect plot choices (Yikes! Lots of them!). Inspect these at your leisure…

Yet Another Non-Boston t Party – Inference for means w/o raw data
The Case of the Mummy's Curse

What may be the most famous mosquito bite in history was delivered to George Herbert (Lord Carnarvon), the man who financed the expedition that unearthed the tomb of the Egyptian Pharaoh, Tutankhamen (c. 1241 BC – c. 1323 BC).

His death (Herbert's not Tut's) was reported in international newspapers, and speculation abounded that his death was due to a "mummy's curse." The source of the (alleged) curse is unknown but possibly can be traced to an 1869 short story by *Little Women*'s author, Louisa May Alcott: "Lost in a Pyramid: the Mummy's Curse." We will grab the result of research by Mark Nelson and address this mummy's curse business. [Nelson, M. R. (2002). The mummy's curse: historical cohort study. *British Medical Journal* 325:1482-4]

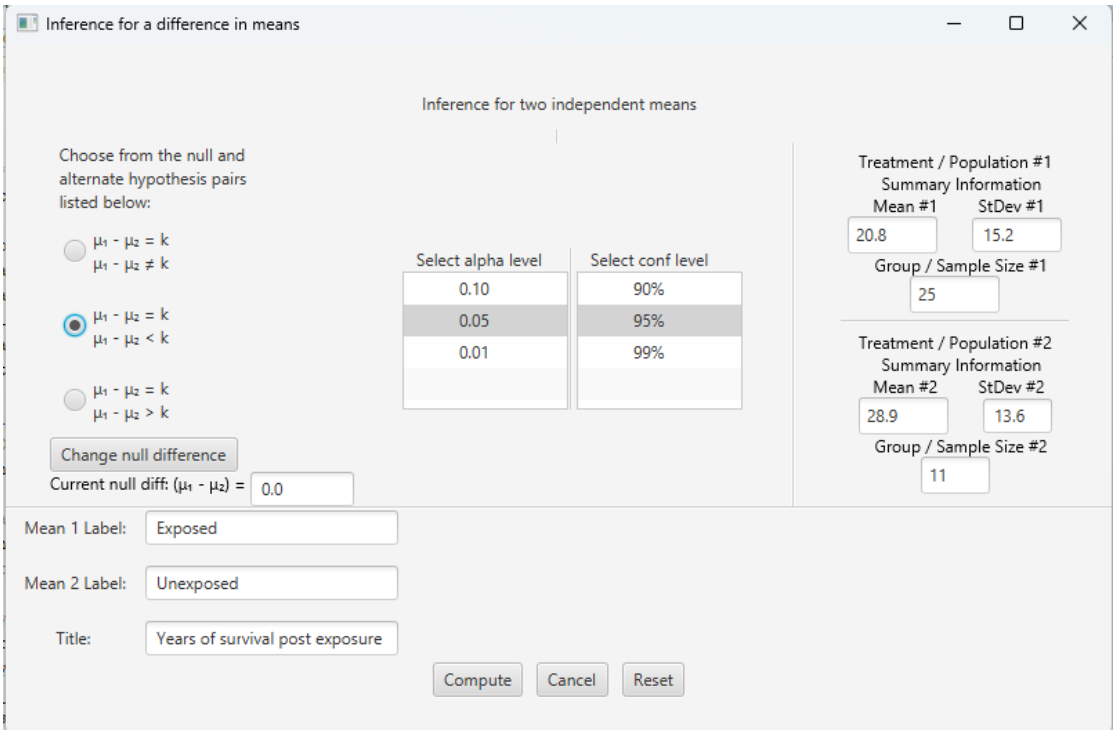
Nelson divided the individuals reported by Carter to be in Egypt at the time of the Big Bite and present (or not) at the opening at the third door (where the sarcophagus was located) and/or the opening of the sarcophagus and/or the opening of the coffins and/or the examination of the mummy. (So the number of "exposures" to the curse was between 1 and 4.) The data on years of survival after exposure are in the following table:

Category	n	$\bar{x}$	s
Exposed	25	20.8	15.2
Unexposed	11	28.9	13.6

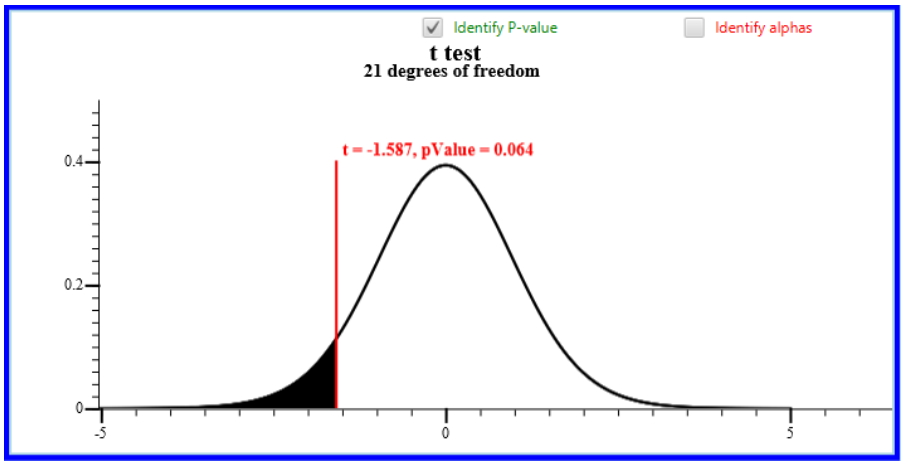
Years of Survival

At the SPLAT main menu select **Inference → Independent Means** and click on the "**Summary**" button. You should see the panel below. I have filled in appropriate values; note that I am conducting a one-sided test:

$$H_a : \mu_{Exposed} < \mu_{Unexposed}$$



The graphic presentation below suggests that we will not reject the null hypothesis. Close, but no below .05 level cigar.



Consistency being a virtue in this context, the Independent Means Report will not contradict the plot; the plot thickens not, as it were.

Inference for independent means report

***** Descriptive statistics and confidence intervals *****

	NSize	Mean	StDev	StErr	CI_Low	CI_High
		*** 95% Confidence intervals for μ_1 , μ_2 ***				
Variable	25	20.800	15.200	3.040	14.526	27.074
Variable	11	28.900	13.600	4.101	11.663	29.937

***** Hypothesis test: *****

Null hypothesis: $\mu_1 - \mu_2 = 0.000$
Alternative hypothesis: $\mu_1 - \mu_2 < 0.000$

*** 95% Confidence intervals for $\mu_1 - \mu_2$ ***

Method	DiffMeans	St Err	df	t-stat	p-Value
Satterthwaite	-8.100	5.105	21.329	-1.587	0.064
Pooled	-8.100	5.336	34	-1.518	0.069

***** Estimation for the difference, Exposed - Unexposed

*** 95% Confidence intervals for $\mu_1 - \mu_2$ ***

Method	DiffMeans	StandErr	df	ciLow	ciHigh
Satterthwaite	-8.100	5.105	21.329	-16.877	+∞
Pooled	-8.100	5.336	34	-17.122	+∞

Effect size (Cohen's D) = -0.537

A Non-Boston Non-t Party – Inference for proportion(s)

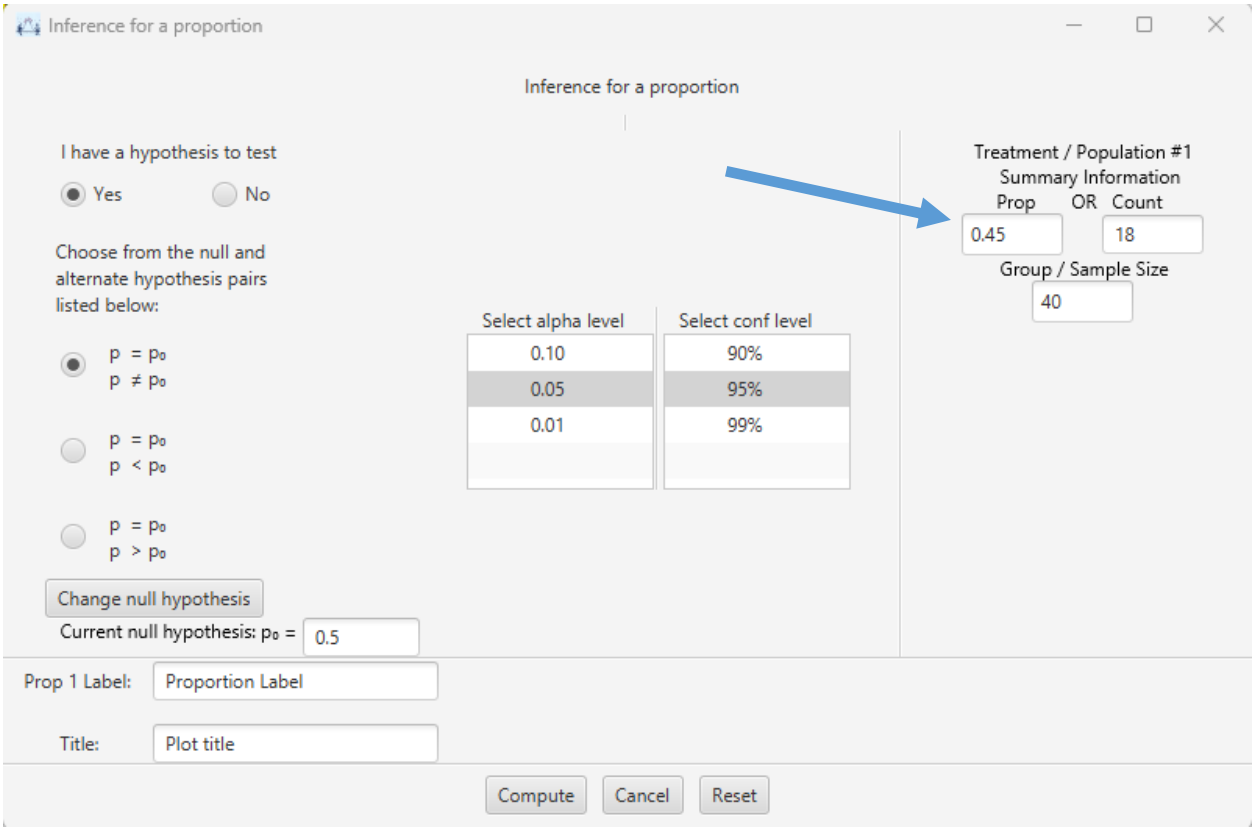
In general, the user inference for proportions follows along that of means, though – of course! – there are some wrinkles unique to proportions. Let's take these one at a time…

Execute the sequence: Inference → One proportion.

Not a great deal new to see here. The little blue arrow reminds me to mention that the cherubs can input either a proportion or a count with the sample size. As is almost always the case in panels in SPLAT, the values can be entered in in any order.... **except**…you must make the hypothesis testing choices in a certain order: <u>You have to say "Yes, I have a hypothesis to test</u>" before you choose either the alternative hypothesis or the value of the null hypothesis. (I don't actually remember why that is the case; some deficiency in my understanding of the Java language, probably,)

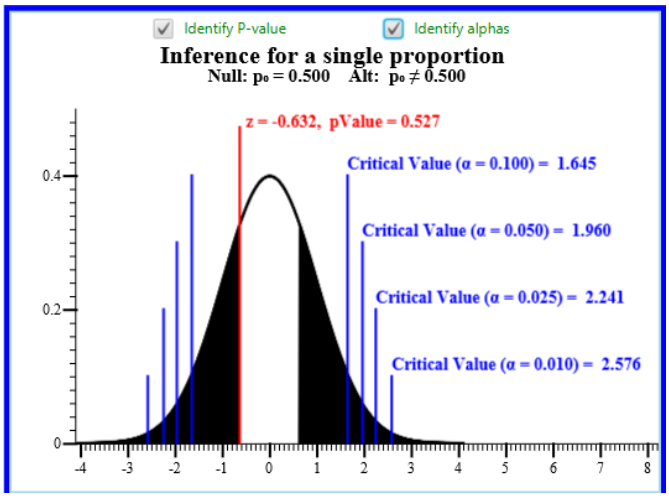
So here I am choosing to do a hypothesis test of a proportion equal to 0.5, and my data are 18 out of 40 successes. (SPLAT kindly filled in the 0.45 when I Entered the count and sample size.)

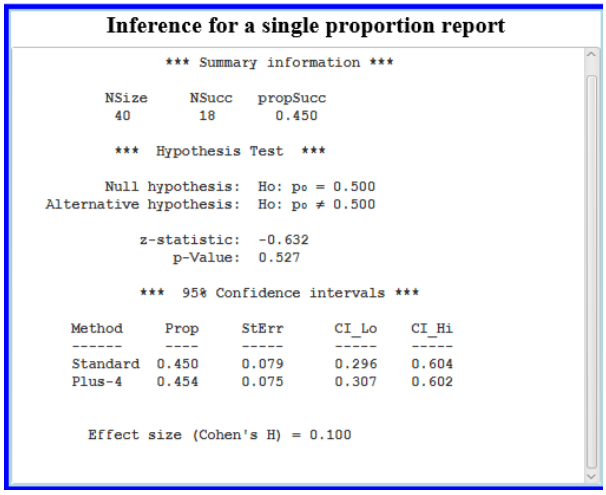
Now, compute!



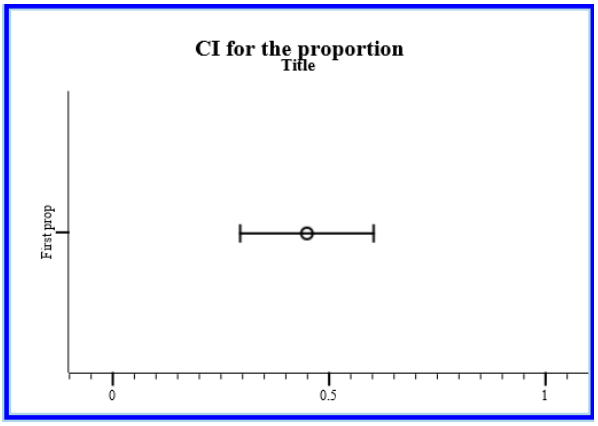
There are four output panels for proportions: the sampling distribution for the statistic, a numeric report, a 95% confidence interval, and finally a choice for "exact" inference. In the "exact" panel, the exact binomial distribution of counts is there for those times when the assumptions for inference are not met (e.g., small sample sizes).

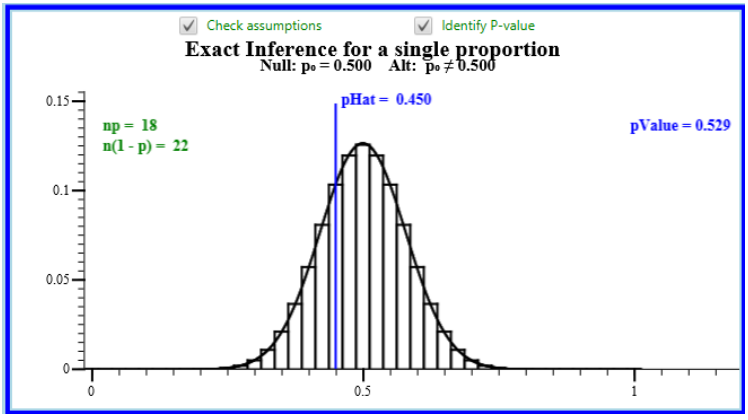
I provide these options even though most of the time in textbook inference problems, assumptions will be met. Here, students will see that there are alternatives to the usual normal curve approximation. Since some statistics books mention the "plus-4" statistic, I have included that as well in the numeric summary. The two-proportion inference works in a similar manner but does not present the "exact" results.





The 95% confidence interval, and the "exact" report for inference about a single proportion, with a superimposed normal curve for comparison:





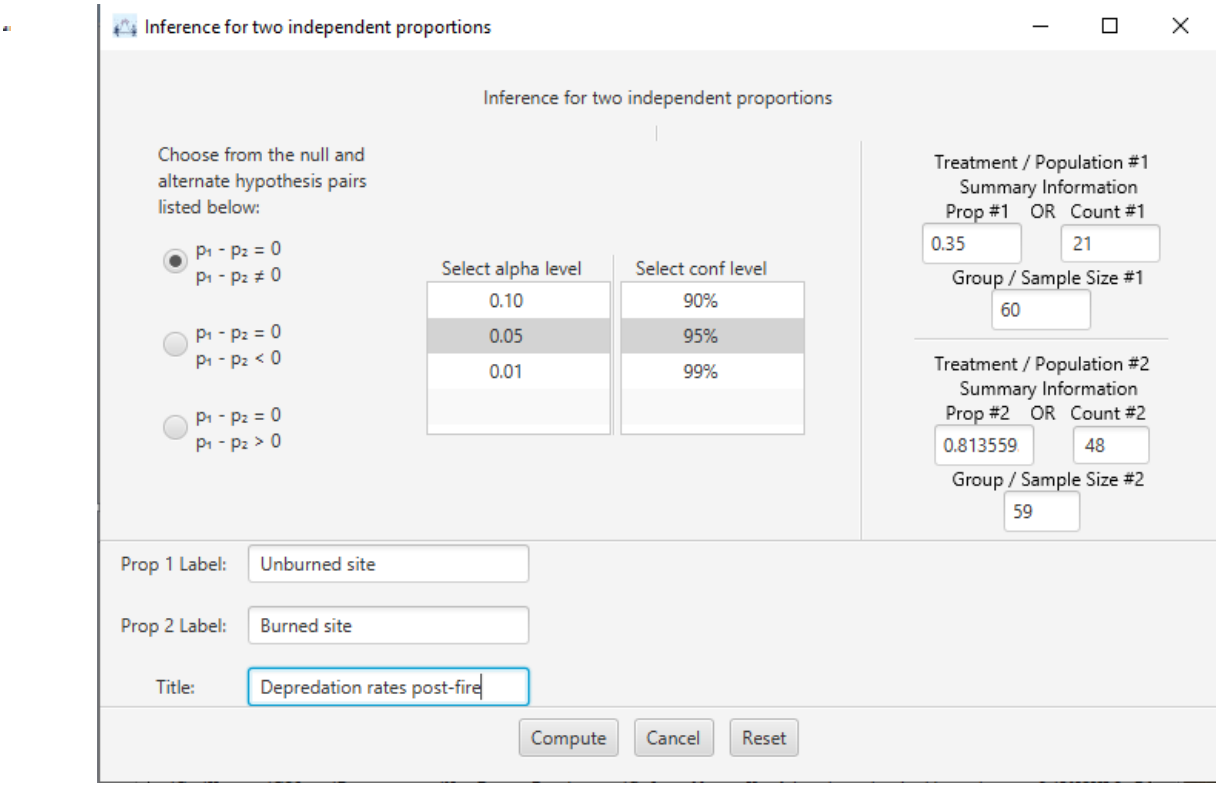
Inference for two independent proportions.

Unlike inference for means, your only hypothesis testing option for two independent proportions is a difference of zero (but confidence intervals for any $p_1 - p_2$ are provided). The hypotheses and levels of significance / confidence levels are selected by clicking on them; the defaults are "not equal to" and $\alpha = 0.05$ / 95% confidence.

For each treatment or population, you can input either the sample proportion and sample size, or the sample count of successes and the sample size. If the count and sample size are given, the proportion will be calculated by SPLAT. If the proportion and sample size are given, an estimate of the count will be calculated.

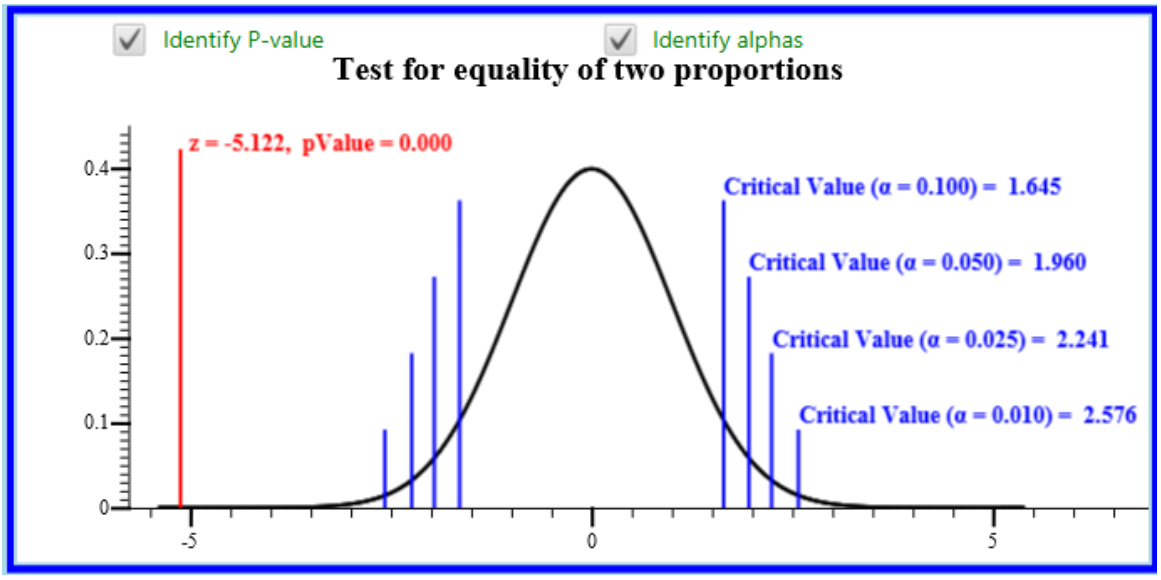
In this example I will use results from Almario, B., et al. (2009). Effects of Prescribed Fire on Depredation Rates of Natural and Artificial Seaside Sparrow Nests. *The Wilson Journal of Ornithology* 121(4): 770-777. The researchers placed artificial nests and eggs on unburned and burned tidal marsh areas in Maryland. They monitored the nests for "depredation" (missing eggs or eggs with teeth or beak marks.) Twenty-one of the 60 artificial nests on unburned sites were depredated compared to 48 of 59 on burned sites.

Click on: Inference → Two Proportions and the data entry panel will appear. The panel, with data entered, looks remarkably like this:



After entering the counts and sample sizes, click on Compute. The dashboard options for the difference between proportions inference are: (a) z-test (using the normal approximation); (b) a TwoPropReport; (c) a confidence interval for the difference in proportions; and (d) confidence intervals for the two proportions.

The output for the z-test, all options selected, is shown below. (Once again, I had to click and drag the right end of the scale a bit to the left to get everything in.)



The statistical summary ("TwoPropReport") below provides the usual hypothesis test and confidence interval information…

Inference for a difference between proportions

*** Summary information ***

Prop	NSize	NSucc	prop	ciLow	ciHigh
Prop #1	60	21	0.350	0.229	0.471
Prop #2	59	48	0.814	0.714	0.913

*** Hypothesis Test ***

Null hypothesis: $H_o: p_1 - p_2 = 0$
Alternative hypothesis: $H_o: p_1 - p_2 \neq 0$

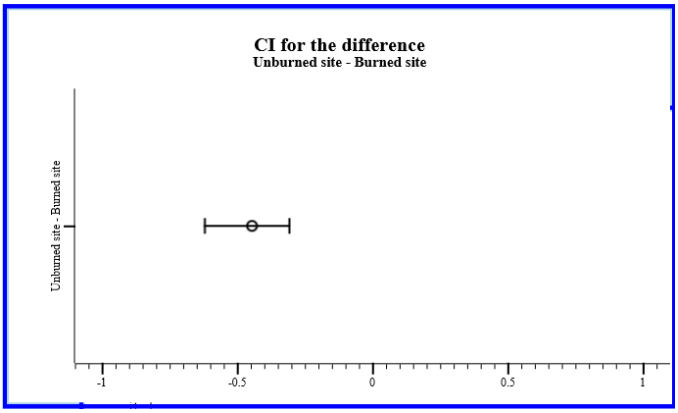
z-statistic: -5.122
p-Value: 0.000

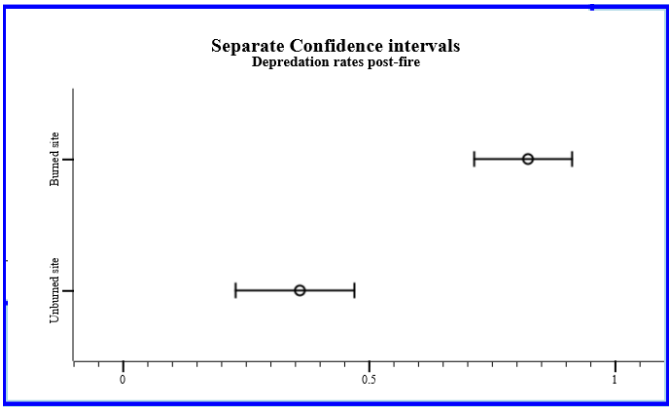
*** Confidence interval for $p_1 - p_2$ ***

p1 - p2	StandErr	ciLow	ciHigh
-0.4636	0.080	-0.620	-0.307

Effect size (Cohen's H) = 0.983

… and the confidence intervals for the individual proportions and their difference are what one would expect.

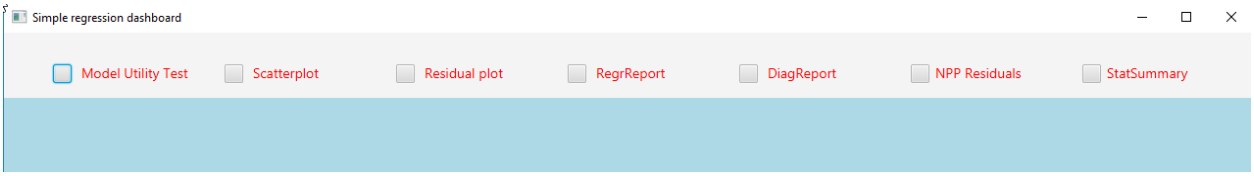




If the assumptions necessary for the normal curve approximation are not met, the P-value from Fisher's Exact test is provided with a warning about the assumption violation. The reason I put this in SPLAT is pedagogical: teachers could then respond to those cherubs who wonder what could be done if the assumptions necessary for the normal approximation to be appropriate (i.e., the "np rules" are not met).

Inference for Regression

Inference for regression works just like the Data Exploration version of Regression discussed earlier; it is over on the right in the menu choices: Inference → Regression slope. You get output about inference for the intercept also, but in real life the intercept is seldom of interest to researchers.



The added options you get with the test for a slope of zero are:

Model Utility Test – similar to other inference presentations
Regression report – 95% confidence intervals for the parameters and ANOVA
Diagnostic Report – the calculated values used for assessing outliers,
high leverage points and influential points (way BAPS!)
Normal probability plot of the residuals
Joint confidence interval for slope and intercept
Statistical summary – univariate and bivariate statistics

Not all these will be meaningful to AP Statistics students; some SPLAT procedures and output panels are intended for those teachers and students who have time and inclination after the AP exam and wish to explore more advanced statistical topics.

Note that the regression report provides much more information than would be seen on an AP Statistics exam, and the Regression Diagnostics panel is probably best ignored unless a unit on multiple regression is anticipated, or you want to get beyond AP Statistics and into the calculations of leverage and influence.

Explore the graphics options if you wish, and then we will move on to chi square…

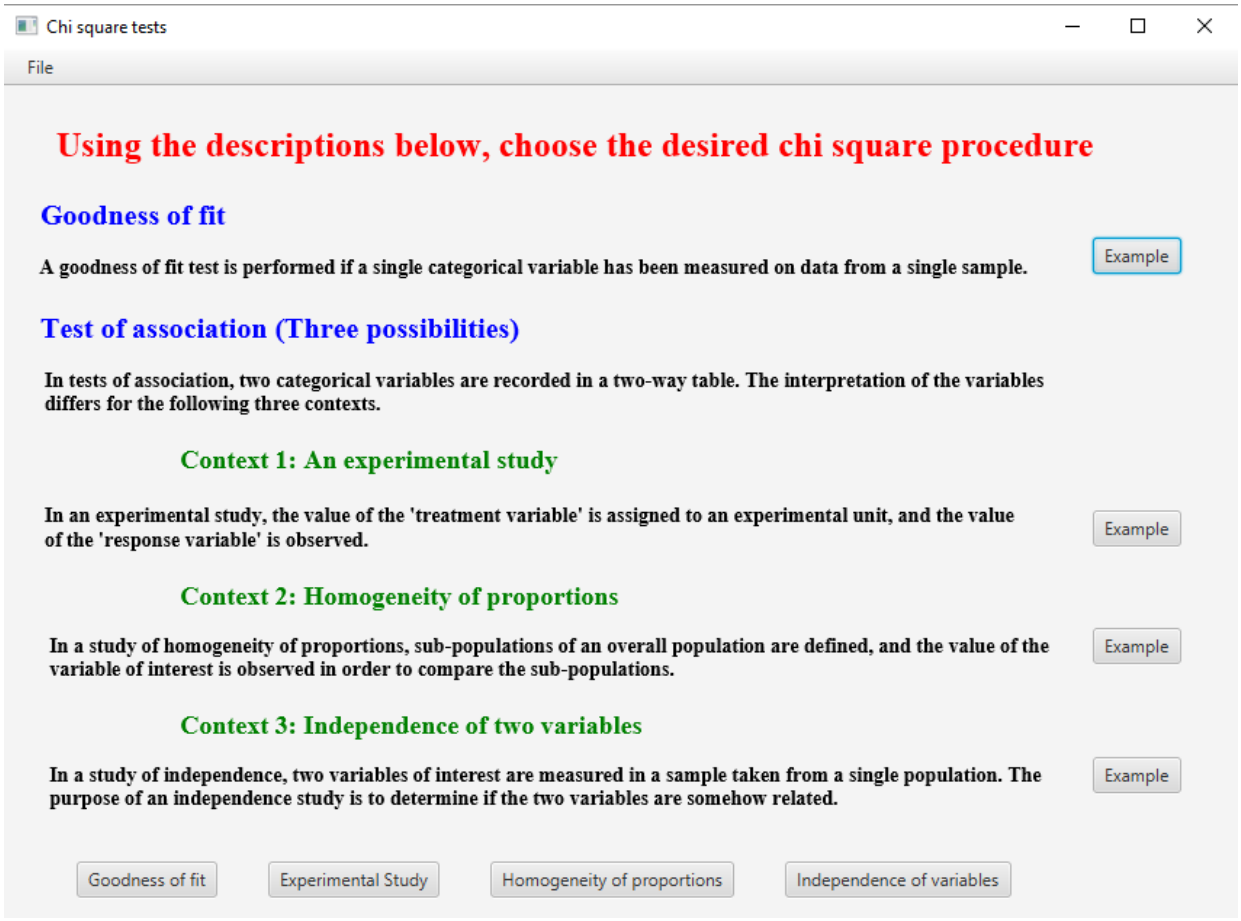
Data entry the Chi Square way

(I.e., entry in formatted panels, not in the SPLAT spreadsheet)

Now, for Chi square… start by executing this sequence in the menu:

Inference → Chi square → I will enter data in a table

Then you will be confronted with more choices…

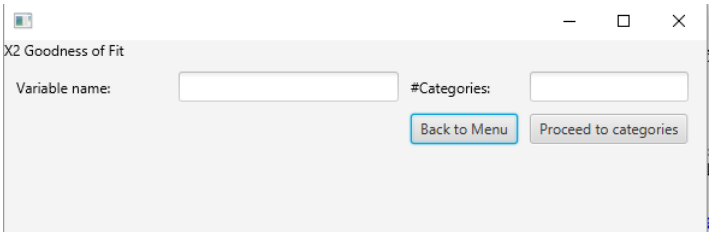


There are two main hypothesis tests that use the chi square procedures in elementary statistics: The Goodness of Fit (GOF) test and the Tests of Association. In AP Statistics the Tests of Association are further divided into a test of "homogeneity of proportions" and test of "independence of variables". Data for chi square problems will usually be a summary of counts in some sort of table, but the table alone will seldom clearly indicate which chi square procedure is appropriate. The panel shown above is intended to help the fledging statistician make the correct decision.

The three contexts of tests of association shown in the panel above perform the same computations, but the verbal presentations of directions, the null hypothesis and the subsequent interpretation of the results differ among the three contexts. The examples on the right side of the panel are taken from the biology literature and are intended to illuminate the descriptions on the menu.

The Chi Square Goodness of Fit Test

The goodness-of-fit test involves values of a single characteristic in a single sample from a single population. The values of the characteristic are partitioned by the investigator into non-overlapping categories, and the numbers of units observed to be in each category make up the "observed counts." To perform the GOF test in SPLAT, click on the Goodness of fit button in the lower left of the menu panel; this panel will be presented:



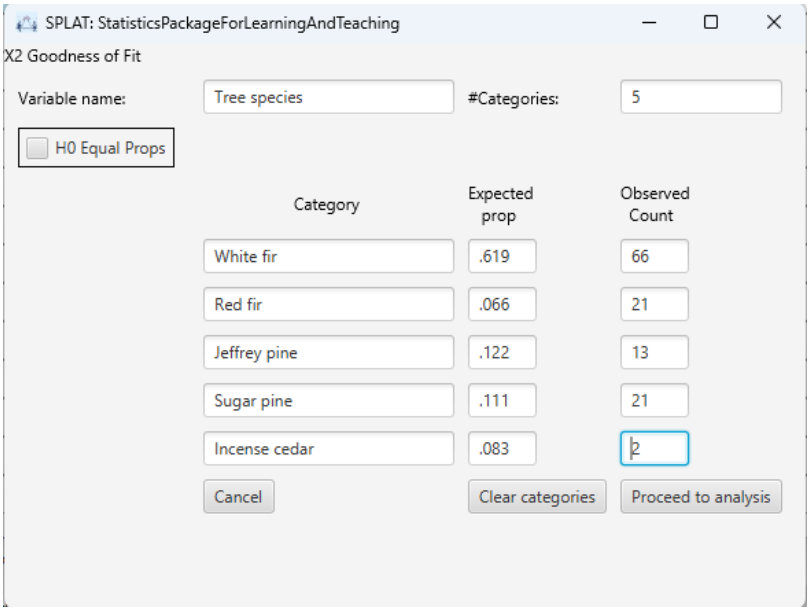
SPLAT designed to keep students out of trouble with data entry as much as possible. If, for example, you enter something other than a positive integer in the #categories below you will get an error message. (In addition to decimals, SPLAT allows common fractions (e.g. 12 / 37) to be entered in the "expected prop[ortions]" field.) The usual traversal methods of tabbing and mouse clicking can also be used to move around the data entry fields, both forward and backward.

As you proceed you will see the following:

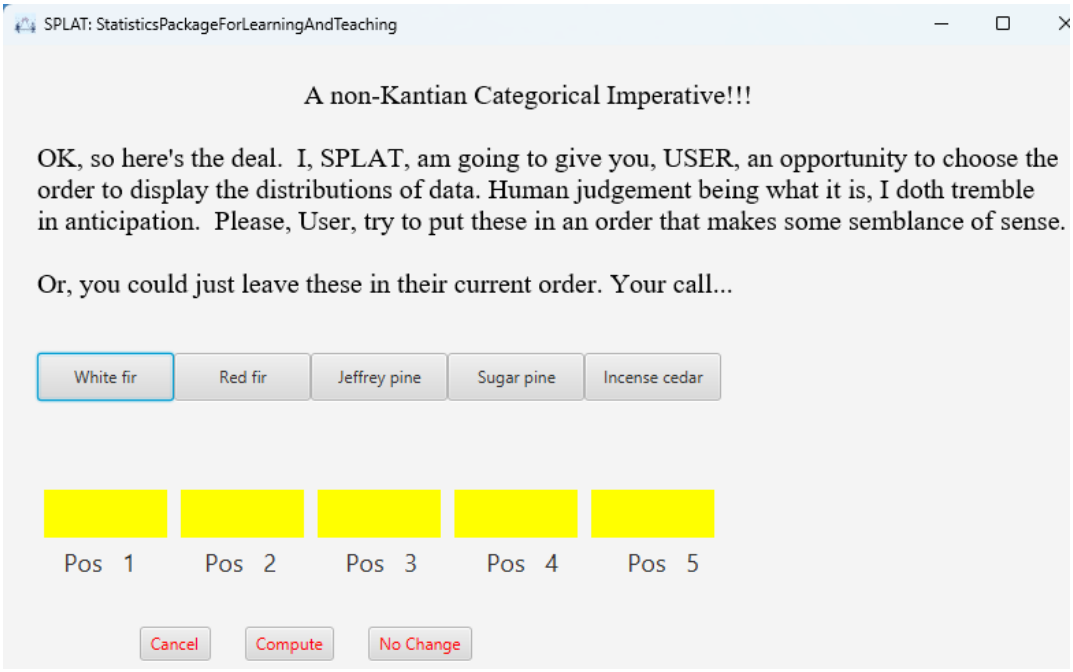
In the example to follow the variable name is "Tree Species" and there are 5 categories. Enter those values in the fields above and press "Proceed to categories."

Enter the species, expected proportions, and observed counts and click on "Proceed to Analysis" as shown below.

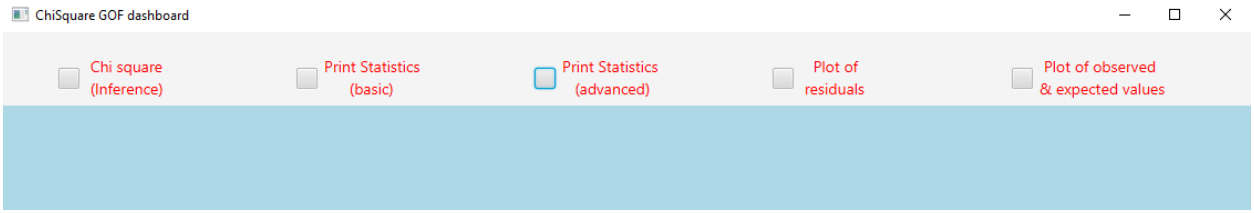
A numerical problem that can occur with chi-square tests is that the sum of the expected proportions add up to a value slightly different from 1.0 due to round-off error. If SPLAT is suspicious (i.e. the total of the proportions is different from 1.0 by .01 or more) you will be given a chance to (a) re-enter the proportions or (b) assure SPLAT that the numbers are correct and the problem is merely due to rounding. If so assured, SPLAT will silently adjust the expected proportions so that they sum to 1.0.



Now click on the Proceed to analysis button. The panel that appears then gives you the option of ordering the presentation of the variables in the graphs associated with the GOF test. I like the alphabet, so I will indicate my preferences to SPLAT by dragging the variables to the positions in alphabetical order.



After dragging the tree species to alphabetical order and clicking on Compute, our display options are presented:

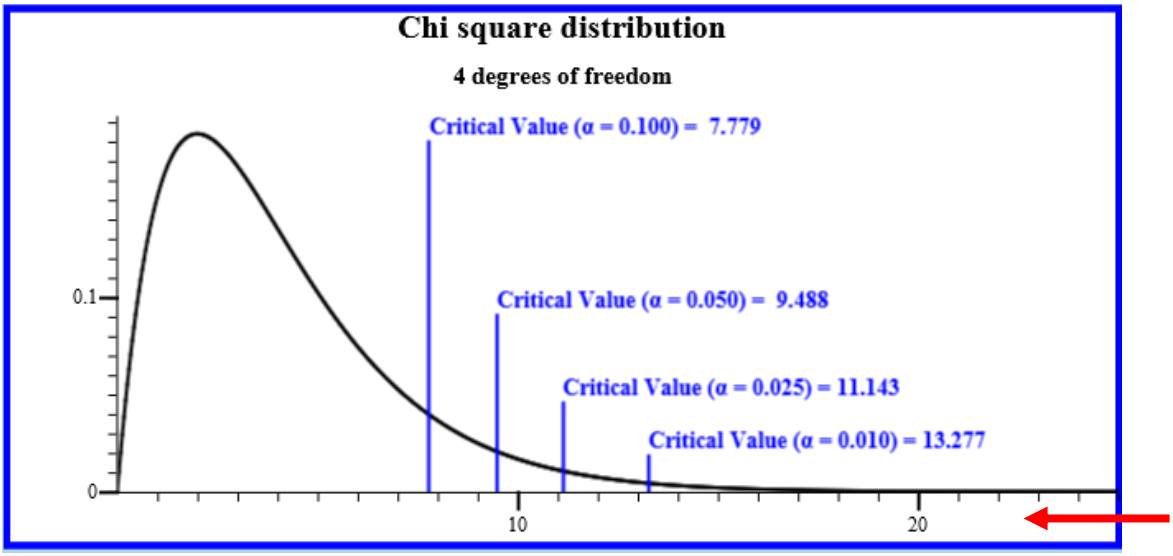


Once your data is entered the Dashboard will appear. Recall that the SPLAT dashboards are the places where you make your decisions about what output will be presented.

The display options are (as usual) available by clicking on the check boxes at the top of the big blue panel. For the chi square goodness of fit, these options are:

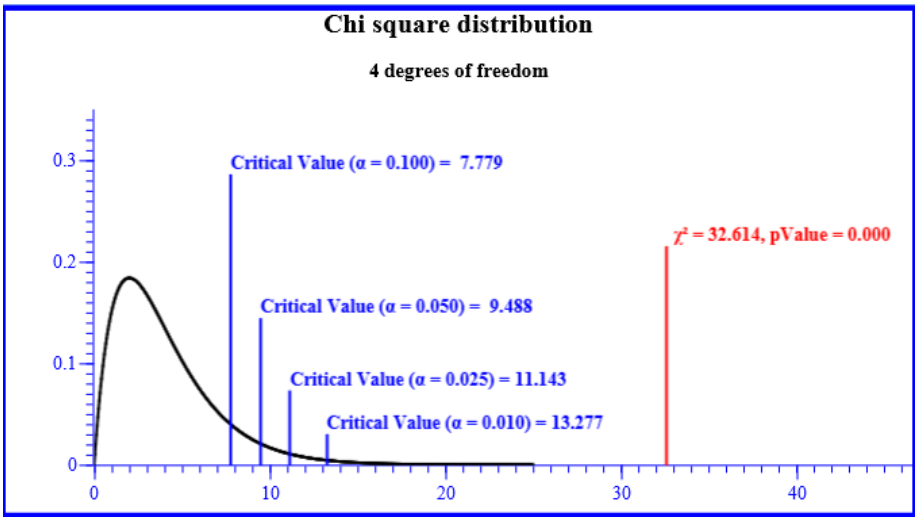
Chi square (inference)
Print Statistics (basic)
Print Statistics (advanced)
Plot of residuals
Plot of observed & expected values

The "Chi square (inference)" display is shown below and has been slightly adapted from one presented in Kamin, L. F. (2010). Using a Five-Step Procedure for Inferential Statistical Analysis. *The American Biology Teacher* 72(3): 186 – 188.



The chi square / (inference) display is intended to give a graphical meaning to the frequently seen terms: "Critical value", "Level of significance" (alpha), and "P-value." For these data the chi square statistic (32.614) is initially "off stage right" as they say on 42nd Street. To bring the chi square value into view, click and drag on the scale, as suggested by the red arrow above.

Almost all numerical scales in SPLAT can be adjusted with a click and drag. As much as possible I want the user – not SPLAT -- to decide what the scale should look like. (This is colloquially known as "Passing The Buck.") You can check this out by clicking and dragging on both the horizontal and vertical scales of the chi square (inference) display.



The size of SPLAT's graphic displays can be manipulated similarly by dragging on edges and corners. Sometimes this is a little tricky; for unknown reasons the edges (especially the right edge) can sometimes be picky. The northwest and southeast corners seem to respond the best to resizing by clicking and dragging. The text display of the results of a GOF analysis can be seen in the "Print Statistics / (basic)" panel. Click on "Chi square (inference)" to hide it, and then on "Print statistics (basic)" to make that panel appear. You should see the panel below.

Elementary chi square statistics

Variable of interest: Tree species

Category	Observed count	Hypothesized proportion	Expected count	Contribution to Chi Square
Incense ced	2	0.083	10.21	6.601
Jeffrey pin	13	0.122	15.01	0.268
Red fir	21	0.066	8.12	20.442
Sugar pine	21	0.111	13.65	3.954
White fir	66	0.619	76.14	1.350

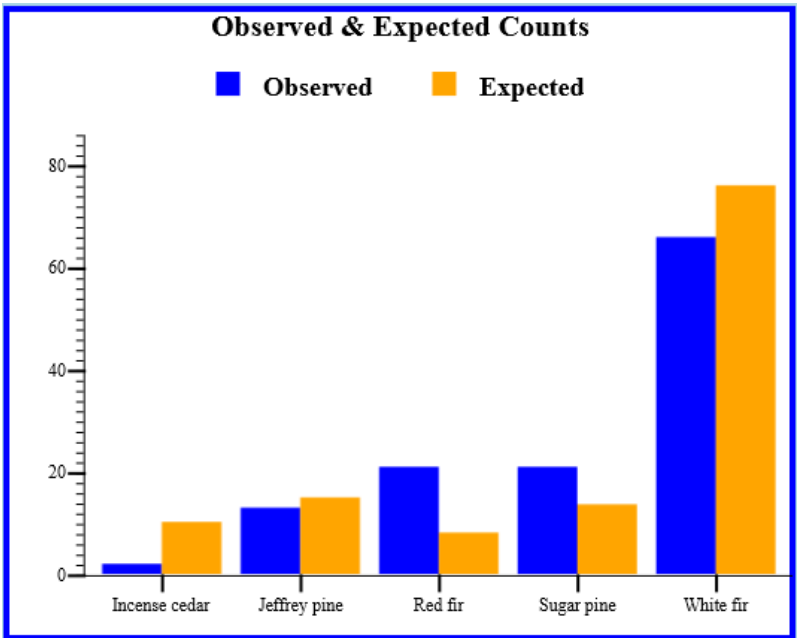
Chi Square = 32.614
df = 4
p-Value = 0.000

The panels in SPLAT are resizable, draggable, and almost all are zoomable. The text and graphics as they initially are presented are intended to be viewed by you at your desktop, but you can also use SPLAT with a projector -- you can zoom the graphs and text panels in and out. SPLAT initially presents graphs in a rectangle of somewhat arbitrary but possibly too small size, in case you have a small screen, e.g. a laptop. You can resize graphs by clicking and dragging

the sides, bottom, and corners of a panel with graphs or text. You can move graphs to different locations in the dashboard by clicking and dragging at the top of a window; The cursor will change to a hand (for moving) and to an arrow (for resizing).

To repeat, for reasons unknown SPLAT will sometimes refuse to resize unless you choose a specific corner. For example, it may only resize from clicking on the upper left corner. In baseball the home team bats last; in ecology Nature bats last; in computer programming the language – Java, in the present case -- bats last…

Click on the " Print statistics (basic)" check box to hide it, and click the "Plot of observed & expected values" check box to display that panel. For these data you will see the following (notice – alphabetized):



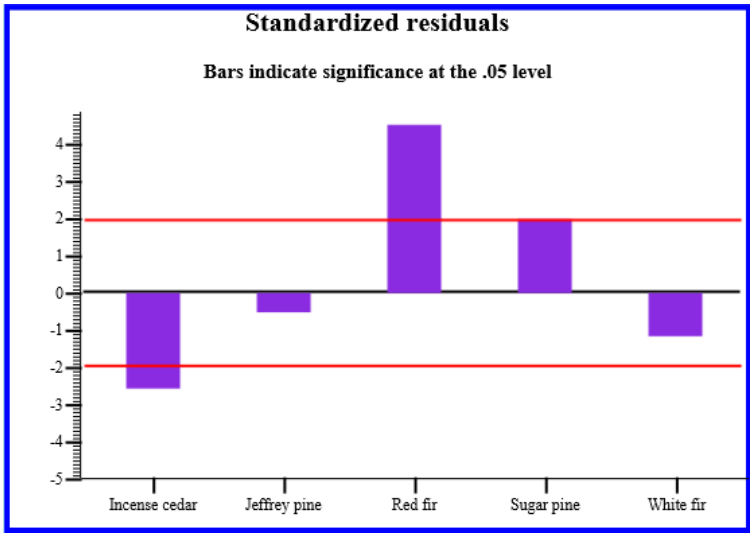
The observed / expected counts show how much the observed counts differ from the expected counts. The general idea is to understand – after rejecting the hypothesis -- which species in this sample differ from what was hypothesized, and in what direction. The observed/expected plot presents the discrepancies in units of "counts."

Residuals, in statistics, are differences between what is observed and what is expected. The residual plot below is perhaps a more informative presentation; click on the "Plot of residuals" check box. These are technically "adjusted" residuals. Adjusted residuals differ from what is presented as "Standardized (Pearson) residuals" in some textbooks. The numeric values are similar, and either can be used. The advantage of the adjusted residuals is that they are asymptotically standard normal, and are therefore interpretable as z-scores. [Ref: Agresti, A. (2018). Categorical Data Analysis (3rd). John Wiley & Sons. Hoboken, NJ.]

The formula for the Pearson residuals is:

$$\text{Standardized (Pearson) residual} = \frac{\text{Raw Residual}}{\sqrt{ExpectedValue}}$$

If your textbook discusses standardizing the residuals and they use the Peason residuals, your textbook values will be slightly larger than what SPLAT presents. The formula for the adjusted residuals is slightly more complicated and can be found in Agresti (2018).



The vertical axis displays the adjusted residuals, and each category has one. It is difficult to interpret "raw" residuals, and these deviations are commonly presented as "Contributions to Chi-Square." The contributions to chi-square are the values acquired from each category in the chi-square calculations. The quantity

$$\frac{(Observed - Expected)^2}{Expected}$$

is calculated for each cell, and these quantities are summed to get the chi square statistic. The square roots of the contributions to the chi square statistic, appropriately positive or negative, are the Pearson residuals discussed above. An adjusted residual will be negative if the observed count is less than expected, and positive if the observed count is greater than expected under the null hypothesis. The red lines in the graph -- the ± 1.96 values of the z-statistic – are boundary values for statistical significance at the .05 level. It is this information about statistical significance that gives added utility over the raw "contributions" in the observed / expected plot. An adjusted residual extending beyond these red lines indicate a statistically significant difference from what is expected by virtue of the null hypotheisis.
Residuals are primarily of interest after a rejection of the null hypothesis, when one seeks to identify specific categories where values differ from what is expected and in what direction. In the residuals plot for these data the red fir and incense cedar deviated most from the hypothesized expectations, and appear to be the categories mainly responsible for the statistically significant chi square value.

The numeric values of the contributions to chi square and the standardized residuals are presented in the "Advanced Statistics" panel.

The Chi square tests of Association

The computations for the two tests of association are exactly the same; it is the null hypothesis and the interpretation of the results that distinguish the two tests. The two tests – or, more accurately, the two interpretations of the tests of association – are tests of (a) independence [of variables], and (b) homogeneity [of proportions]. A test of association is interpreted as a test of independence if the researcher has taken two measures from a single population and is interested in whether the two variables are related, or "associated." A test of association is interpreted as a test of homogeneity if (1) the researcher has randomly assigned experimental treatments to individuals in at least two treatment groups and is interested in whether the distributions of proportions of categorical results differ across the populations or treatment groups, or (2) information was gathered from at least two separate populations.

Consider first the test for independence. Since the statistical and graphical displays are the same for the test of homogeneity, the details of using SPLAT will be presented here and not repeated in detail for the test of homogeneity.

The test for independence of two categorical variables

Come, let us all dine together…

Norovirus **Salmonella** **Staphylococcus**

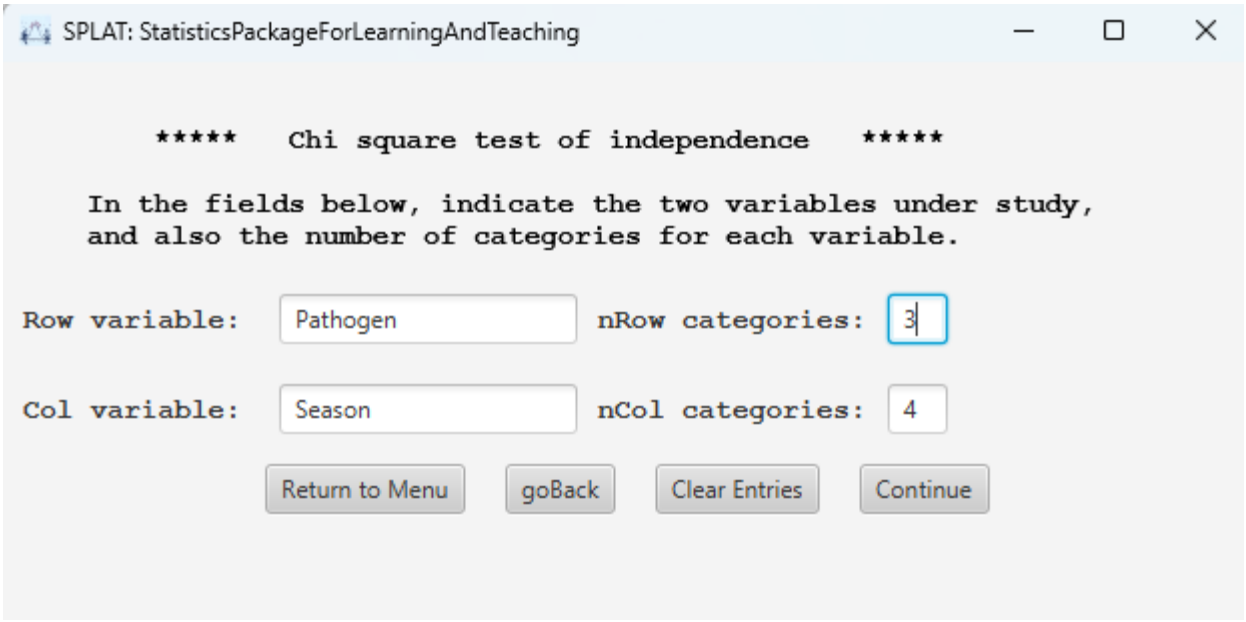
"Georgia is a state on the southeastern coast of the U.S. It has a population of nearly 10 million and an area of 59,441 square miles, with a humid subtropical climate. Archival data on foodborne illness in Georgia between 1998 and 2010 from the [CDC's] Foodborne Outbreak Online Database were examined."

- Wilson, E. (2015). Foodborne Illness and Seasonality Related to Mobile Food Sources at Festivals and Group Gatherings in the State of Georgia. *Journal of Environmental Health* 77(7):8-11.

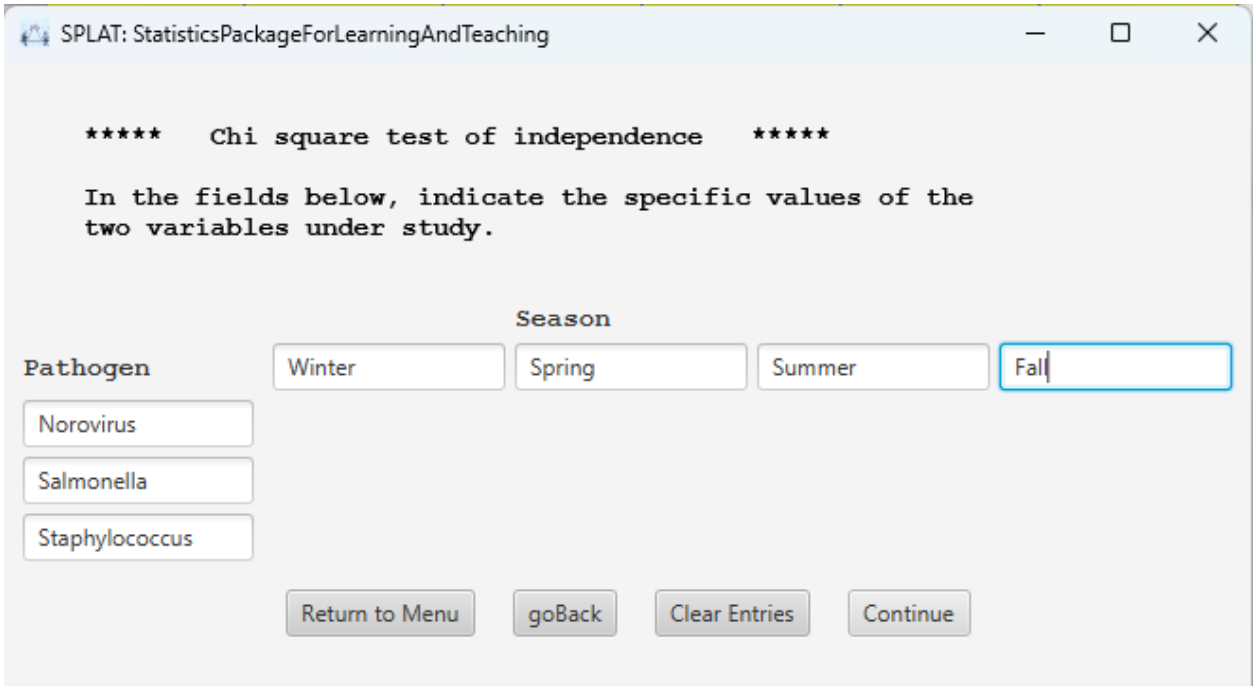
Research question: Is there an association between pathogen and season of the year? Here is a selection of data from the study:

Pathogen	Season				Total
	Winter	Spring	Summer	Autumn	
Norovirus	33	28	20	13	94
Salmonella	14	13	28	18	73
Staphylococcus	7	7	6	7	27

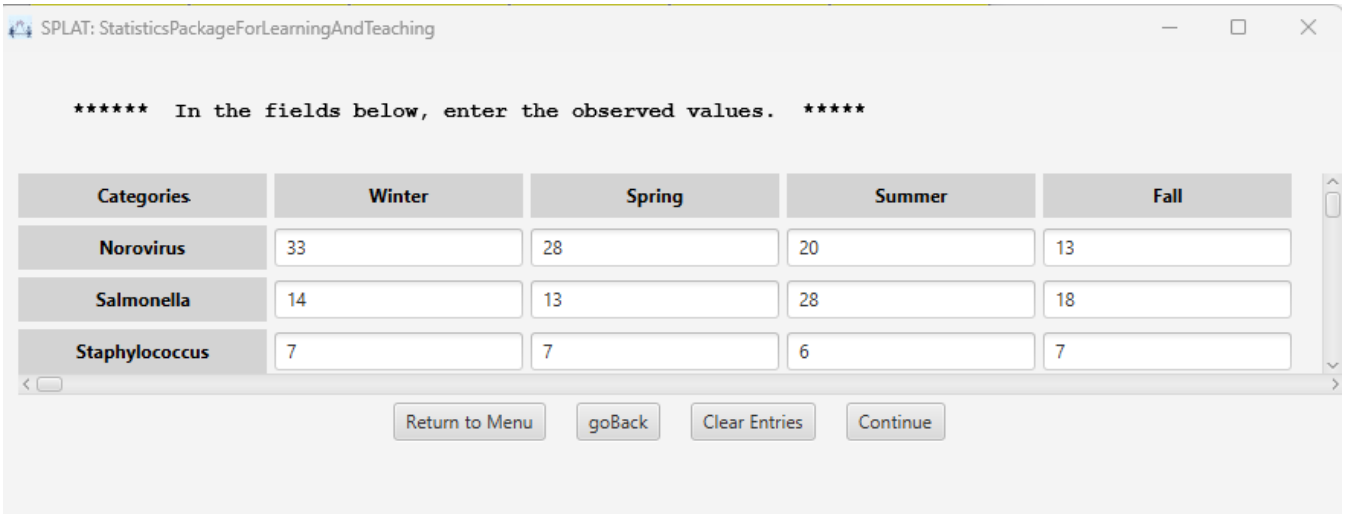
To follow along with this example, click on Inference ➔ chi square ➔ "I will enter data in a table" and finally ➔ "Independence of variables." The variables we are checking for independence in this example are "Pathogen" (Row variable, 3 categories) and "Season" (Col variable, 4 categories). Fill in the fields as indicated below…and then click on "Continue."



Now tell SPLAT the different categories for each of the variables, and again click on Continue:



We are within epsilon of done! Enter the numbers as shown below and (yes, once again!) click on Continue.



The "Chi square distribution" panel is the same as was shown in the GOF discussion. All else, however, is new and different for the tests of association. Here is the Print Statistics / (basic) panel:

Elementary chi square statistics

Association between: Pathogen and Season

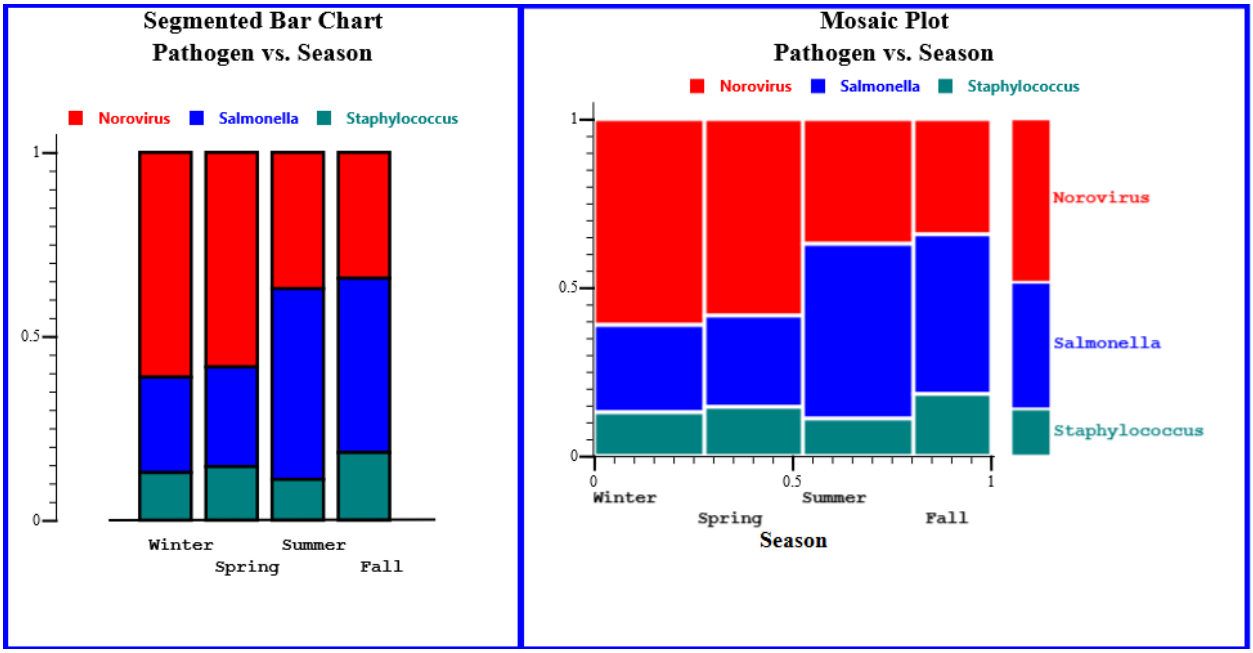
	Winter	Spring	Summer	Fall
Norovirus				
Observed values	33.00	28.00	20.00	13.00
Expected values	26.16	23.26	26.16	18.41
Contrib to X2	1.79	0.97	1.45	1.59
Stand. Resid (z)	2.19	1.58	-1.98	-1.96
Salmonella				
Observed values	14.00	13.00	28.00	18.00
Expected values	20.32	18.06	20.32	14.30
Contrib to X2	1.97	1.42	2.90	0.96
Stand. Resid (z)	-2.09	-1.74	2.54	1.38
Staphylococcus				
Observed values	7.00	7.00	6.00	7.00
Expected values	7.52	6.68	7.52	5.29
Contrib to X2	0.04	0.02	0.31	0.55
Stand. Resid (z)	-0.24	0.15	-0.70	0.89

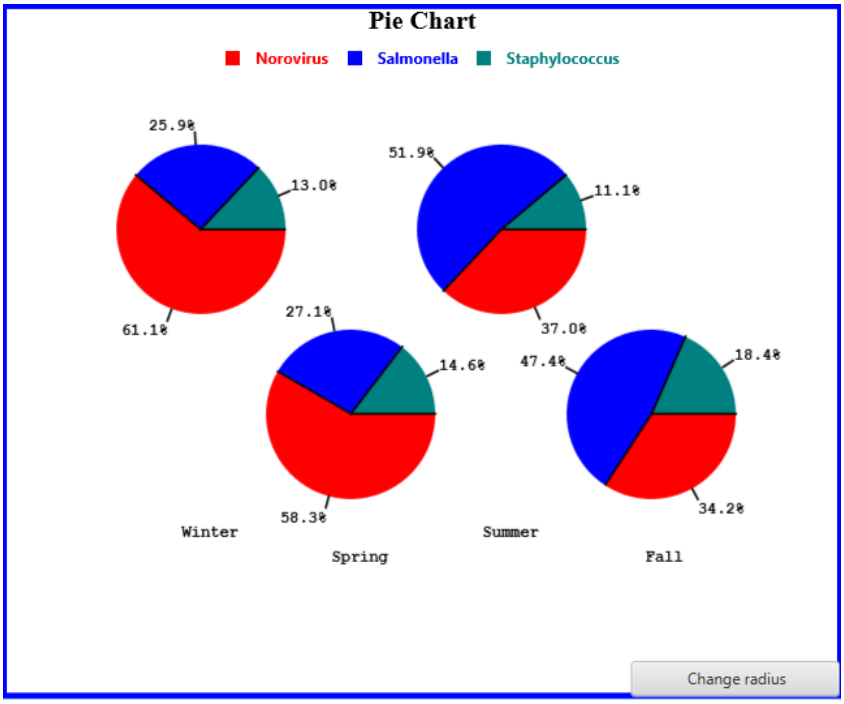
Chi Square = 13.951
df = 6
p-Value = 0.030

This panel presents information that is analogous to the Goodness of Fit test but is somewhat more complicated because there are two variables to consider. The Print Statistics / (advanced) window shows in addition a standardized measure of association (an effect size) known as Cramer's V statistic. That panel is not presented here, but you are very welcome to check it out.

Three additional graphic displays give a visual presentation of the results: the segmented bar chart, the Mosaic plot, and the pie chart. These plots display information about how the proportions of pathogens and seasons are distributed in the sample. The Mosaic plot provides slightly more information than the segmented bar chart: the proportion of pathogens combined over all seasons is indicated in the rightmost column of the Mosaic plot, and the proportions in each season are indicated by the widths of the rectangles. For example, the "Winter" bar is slightly wider in the Mosaic plot since in this sample a higher proportion of pathogens was found in Winter.

These three plots are of interest only if the null hypothesis has been rejected. In that case the researcher focusses on which species and habitat combinations have different proportions than hypothesized, in which direction, and why.



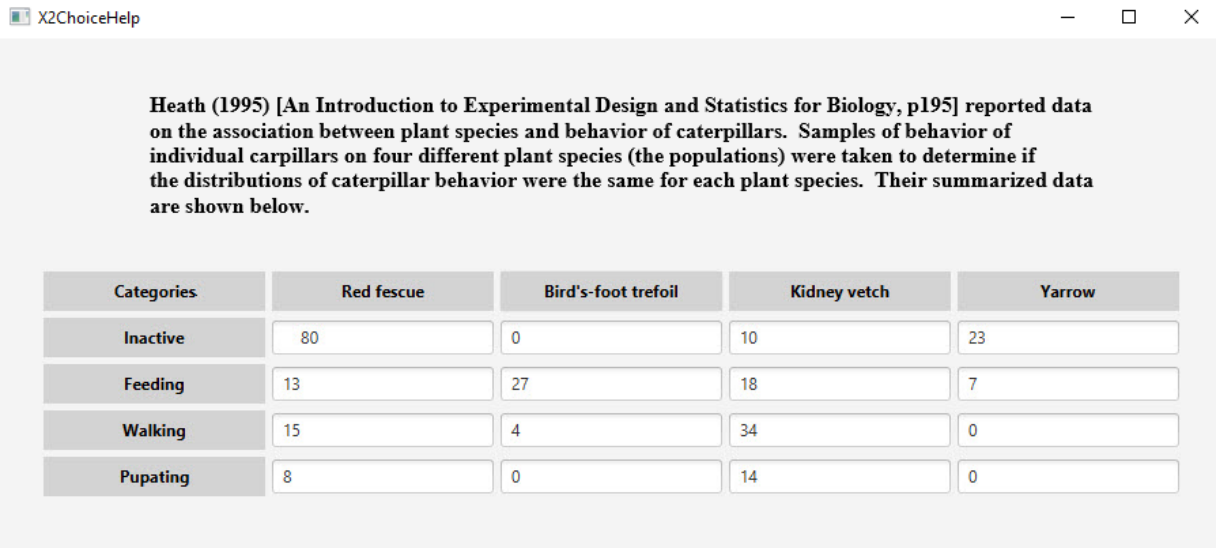


Note: In the initial presentation the "pies" were smaller. I used the "Change radius" capability in SPLAT to blow them up to their presented size.

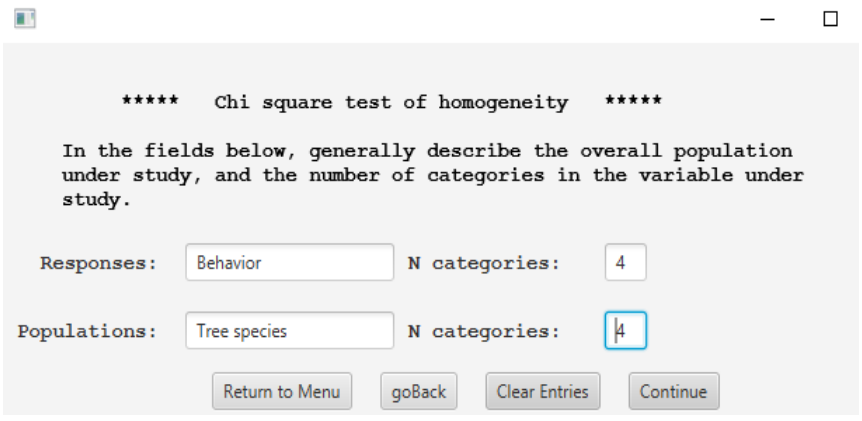
The test for homogeneity of proportions

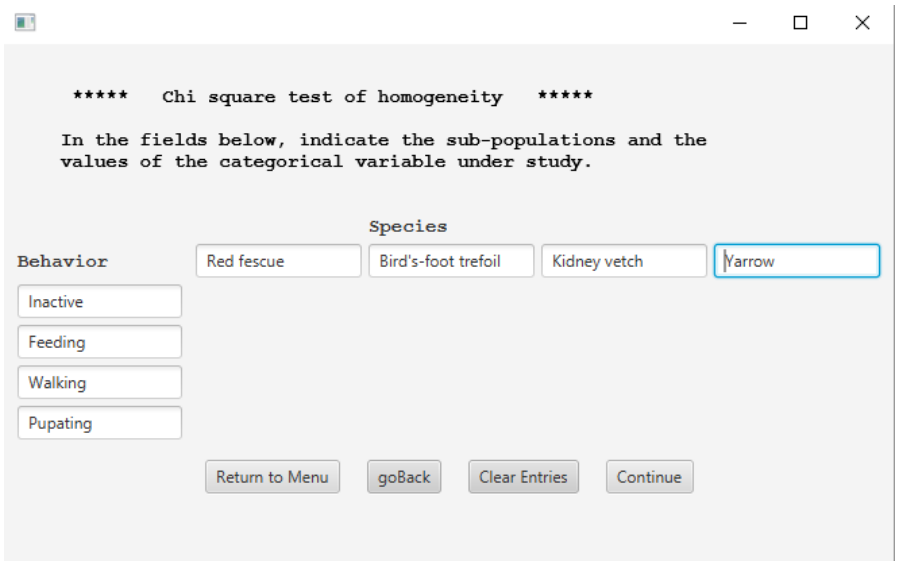
An important thing to remember with the chi square tests of association is that while the rows and columns are interchangeable for the test of independence, they are not interchangeable for tests of homogeneity of proportions across populations or in an experiment. In the test for homogeneity of proportions, the populations (or experimental treatments) are treated as an "independent" or "explanatory" variable and should appear as a row of column categories spread across the top of the table. This is consistent with the idea that the explanatory variable, i.e. the species, may impact the value of the response variable, but not *vice versa*.

Click on Inference ➔ chi square ➔ "I will enter data in a table," but this time click on "Homogeneity of Proportions." For this study, the sub-populations are tree species, and the "dependent" response variable is behavior. The final data entry is…



OK, back to information entry…





Now go back to the actual counts shown about and enter them.

Tick…tick…tick.

OK, take a deep breath and click on the Print Statistics / (basic) output. You may need to make the window larger to see everything. Notice that the program gives a warning if there are expected values less than 5. A chi square test which results in cells with expected values less than 5 should generally be interpreted with a dollop of caution.

Elementary chi square statistics

Association between: Behavior and Tree species

	Red fesc	Bird's-f	Kidney v	Yarrow
Inactive				
Observed values	80.00	0.00	10.00	23.00
Expected values	51.81	13.85	33.94	13.40
Contrib to X2	15.34	13.85	16.89	6.88
Stand. Resid (z)	7.15	-5.34	-6.61	3.76
Feeding				
Observed values	13.00	27.00	18.00	7.00
Expected values	29.80	7.96	19.53	7.71
Contrib to X2	9.47	45.50	0.12	0.06
Stand. Resid (z)	-4.85	8.35	-0.48	-0.31
Walking				
Observed values	15.00	4.00	34.00	0.00
Expected values	24.30	6.49	15.92	6.28
Contrib to X2	3.56	0.96	20.53	6.28
Stand. Resid (z)	-2.88	-1.18	6.09	-3.00
Pupating				
Observed values	8.00	0.00	14.00	0.00
Expected values	10.09	2.70	6.61	2.61
Contrib to X2	0.43	2.70	8.27	2.61
Stand. Resid (z)	-0.93	-1.83	3.60	-1.80

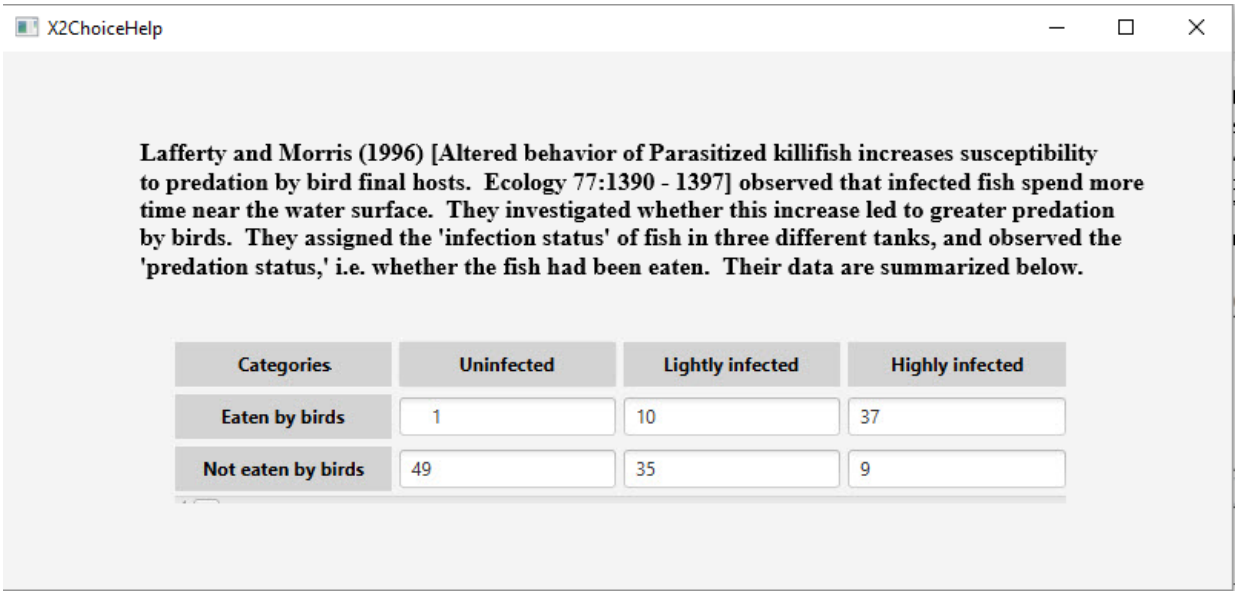
Chi Square = 153.442
df = 9
p-Value = 0.000

********** Warning! **********

*** There are 2 cells with expected values less than 5 ***

The test for homogeneity of proportions (when performing an experiment)

The mechanics of testing for homogeneity of proportions in an experimental study is the same as for an observational study. Again, it makes a difference which variable is the row variable and which is the column variable. Here the "X" or "independent" variable is the experimental treatment, "infection status." The response possibilities, "Fate," are "Eaten" and "Not eaten." Once again, for purposes of checking your progress if you are following along, the Print Statistics / (basic) output is shown here.



Elementary statistics

Association between: Predation status and Infection status

	Uninfect	Lightly	Highly i
Eaten by birds			
Observed values	1.00	10.00	37.00
Expected values	17.02	15.32	15.66
Contrib to X2	15.08	1.85	29.08
Stand. Resid (z)	-5.95	-2.03	8.09
Not eaten by birds			
Observed values	49.00	35.00	9.00
Expected values	32.98	29.68	30.34
Contrib to X2	7.78	0.95	15.01
Stand. Resid (z)	5.95	2.03	-8.09

Chi Square = 69.756
df = 2
p-Value = 0.000

Planning a study: Power

The power of a statistical test is the probability of rejecting the null hypothesis, though we usually only care about the probability of rejecting a false null hypothesis. Power is a function of the sample size, chosen level of significance (α), standard error of the statistic, and the chosen effect size (difference considered to be of practical importance.)

SPLAT presumes that power analyses are conducted before gathering the data. One could toss in values from an experiment already completed (and, as R. A. Fisher famously opined, do a *postmortem* on a failed experiment) but the formulas I am using are for the planning part of a study, before any data are available. Power calculations for the different inference procedures in AP Statistics are straightforward, except for the power associated with the difference between independent proportions. There seems to be no end of different approaches to calculating power for a difference of proportions, including arcsine square root transformations (with continuity correction), Fisher-Irwin tests for two-by-two tables, power functions for Fisher's "exact" test, and for all I know some method the Three Witches put together to drive Macbeth toward his tyrannical desire for power. I turned to Chow, S., et al. (2018). Sample Size Calculations in Clinical Research (3rd ed). CRC Press, Boca Raton, and programmed SPLAT to do the calculations I found there.

Example: Power for a single mean

The information used in this example is taken from p598 in Starnes, D., & Tabor, J. (2020). The Practice of Statistics (6th ed). [Thanks guys!] The context for the example is a company that is developing a new AAA battery ("deluxe") that is supposed to last longer than their existing battery.

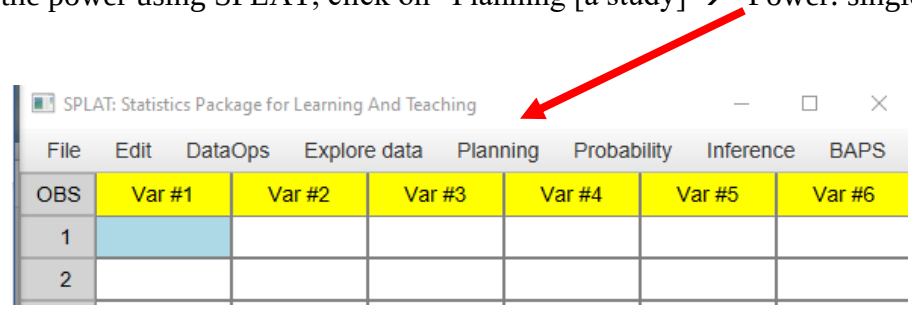
"Based on years of experience, the company's regular AAA batteries last for 30 hours of continuous use on average. The company plans to select an SRS of 50 deluxe AAA batteries and use them continuously until they are completely drained."

Daren and Josh's choices of values for the power calculation are:

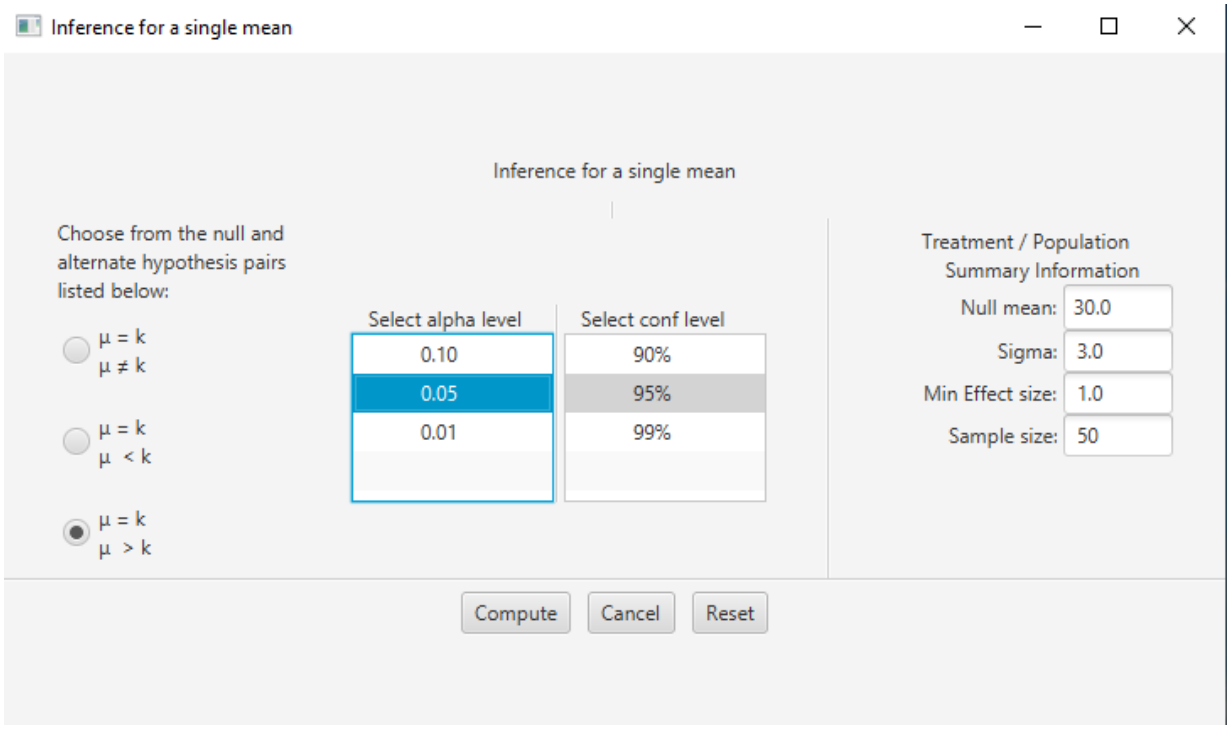
$$H_o : \mu = 30$$
$$H_a : \mu > 30$$

Their level of significance is $\alpha = 0.05$ and they wish to detect a difference as small as 1.0 hours.

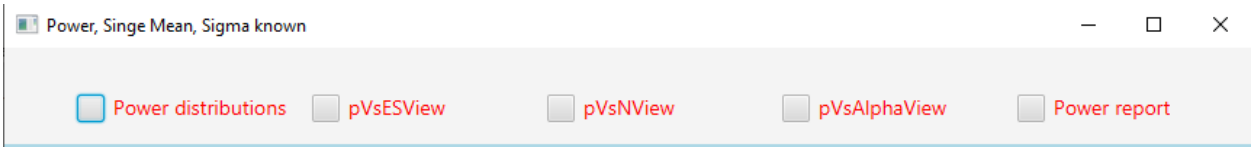
To calculate the power using SPLAT, click on "Planning [a study] → "Power: single mean"



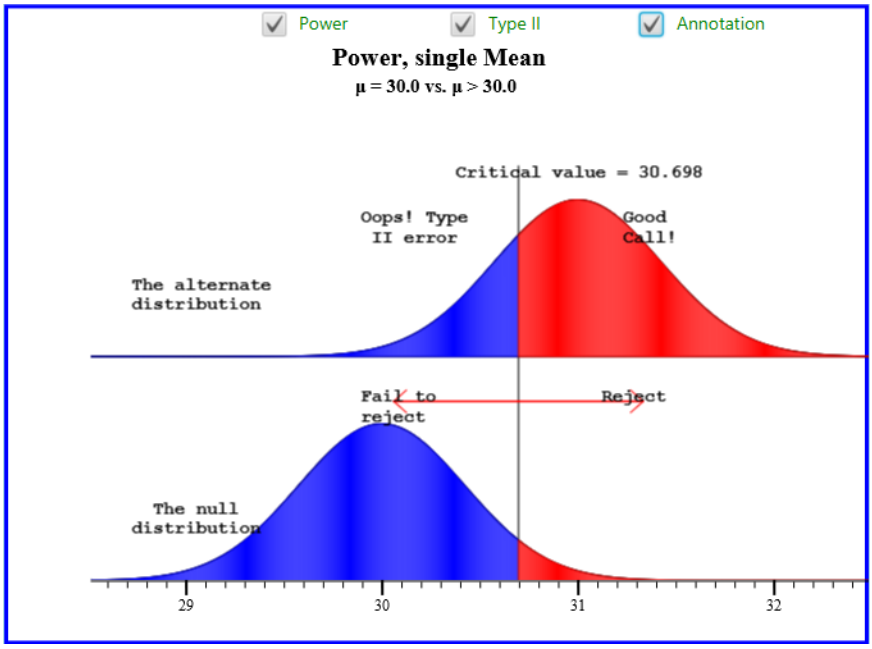
The decisions about the quantities influencing power are all made on a single panel. Here are the values to substitute for this example:



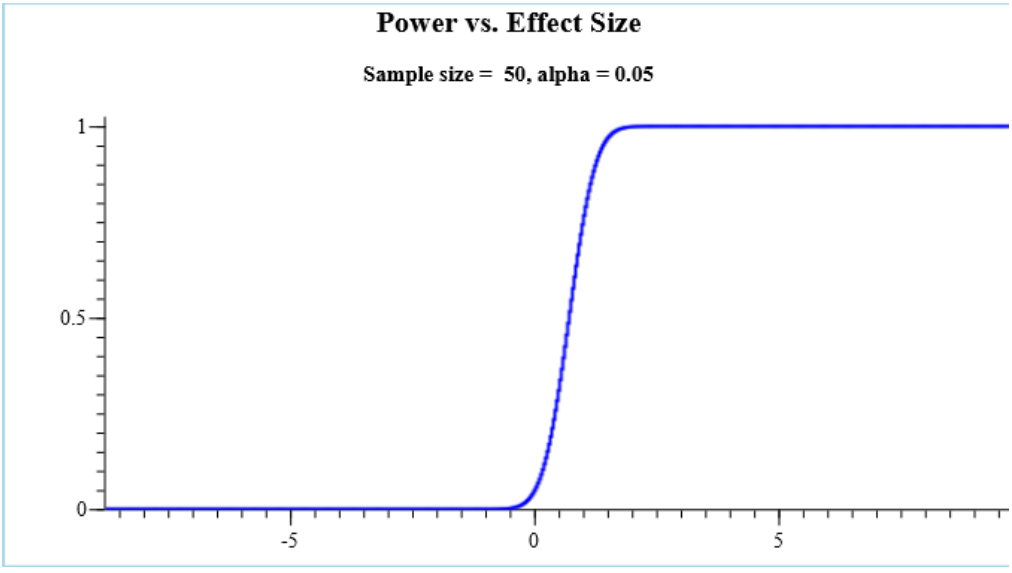
Select "Compute" and you will see the usual dashboard setup:

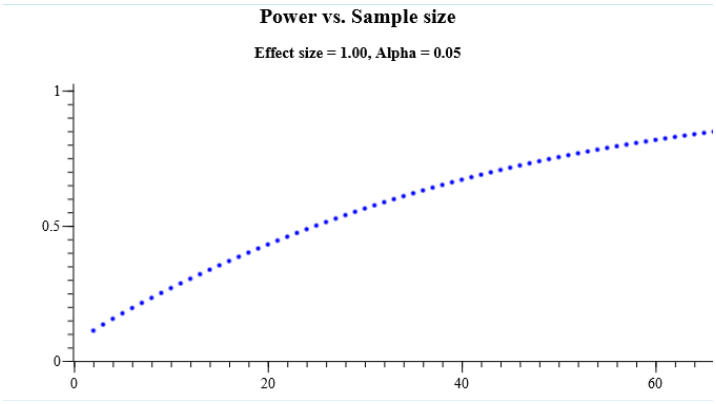


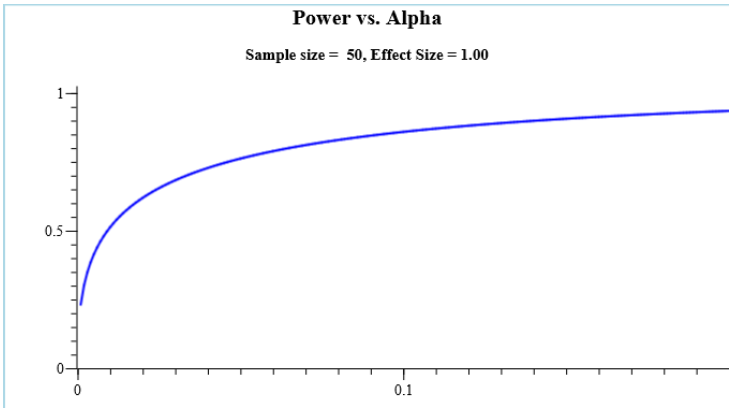
The “Power distributions” option displays the sampling distributions of the sample mean for a true null and an alternative null hypothesis (based on the smallest practically significant effect size. The intent here is to illustrate the ideas of power and Type I and II error.



The power vs Effect Size, Sample Size, and Alpha choices present more traditional graphs, pictures designed to express their respective relationships with power.







The "Power report" summarizes the power calculations:

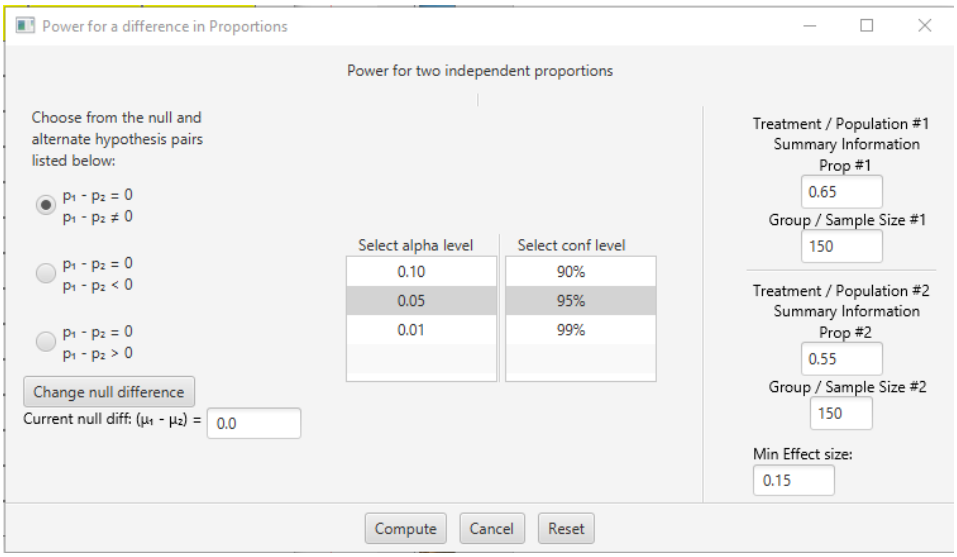
Power Analysis: One Mean

```
       Power Statistics

Single Sample Mean

       Null hypothesis: μ = 30.0
Alternative hypothesis: μ > 30.0
              alpha =    0.050
       Sample size =      50
     Assumed Sigma =    3.000
    Standard error =    0.424
       Effect Size =    1.000
              Power =    0.762
```

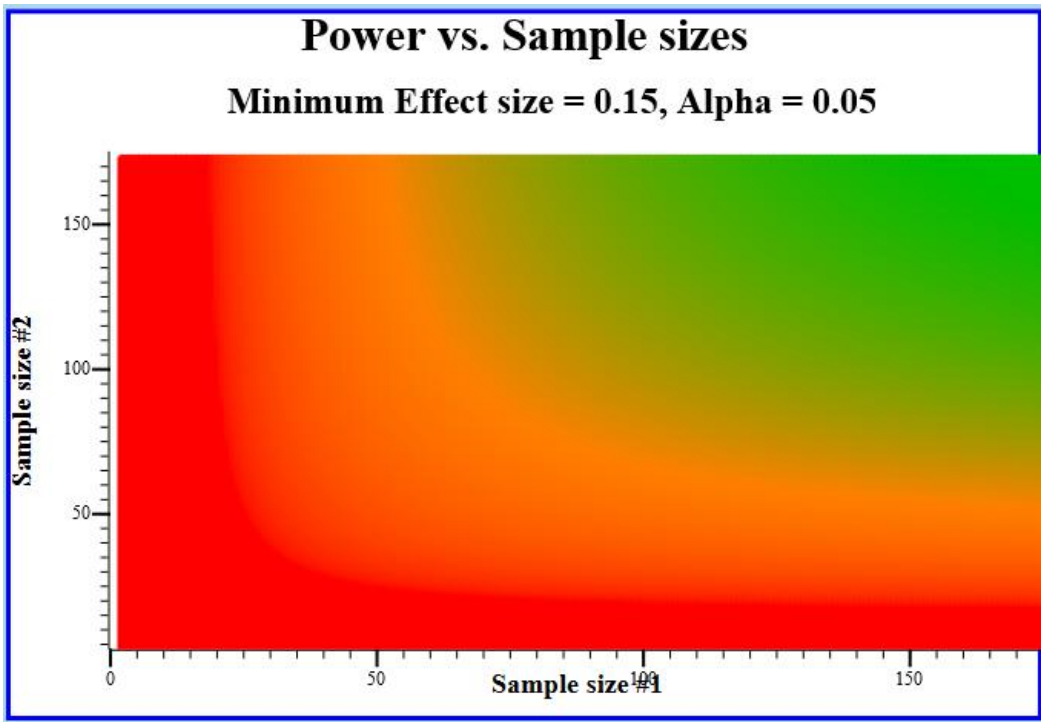
SPLAT does similar power calculations for a single proportion, and for independent means and proportions.

Another example: Power for difference in independent proportions

Power for a difference in proportions and power for a difference in means will be similar to what is presented for a single mean and a single proportion. However, here there are two sample sizes to consider, so Power vs. Sample Size will have a different look. Click on Planning → Power: Two Props and enter these values…

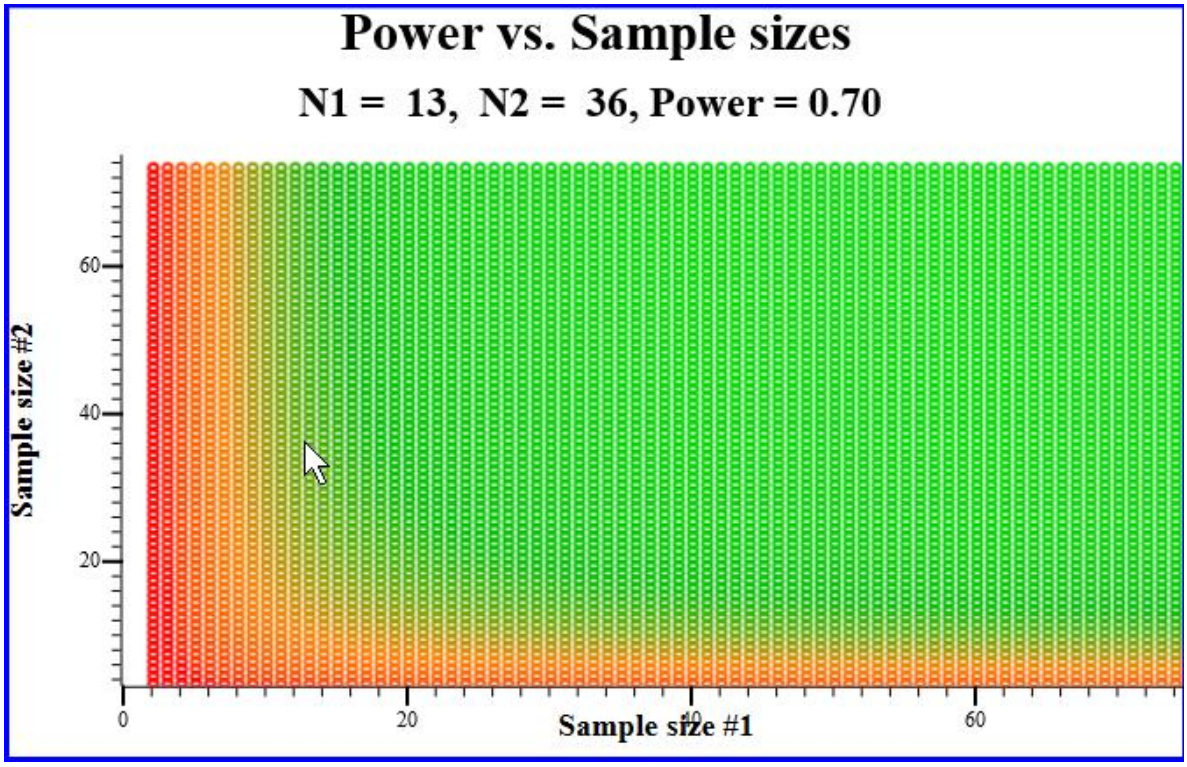


Now click on Compute. The calculations may take a bit of time because of the large sample sizes, but after a while the dashboard should appear. Click on "PVsNView" and you should see…



The power calculations for different combinations of sample sizes are presented as a visual representation of low power to high power, red to green, as the sample sizes get larger. The yellowish transition from red to green occurs where the power is 0.80, the usual standard for acceptable power in statistical studies. If you click and depress on a particular ordered pair of sample sizes, the power will be presented at the top of the display.

With smaller sample sizes you will see discrete circles. The placement (and click) of the arrow determines the sample sizes giving the power above the plot.

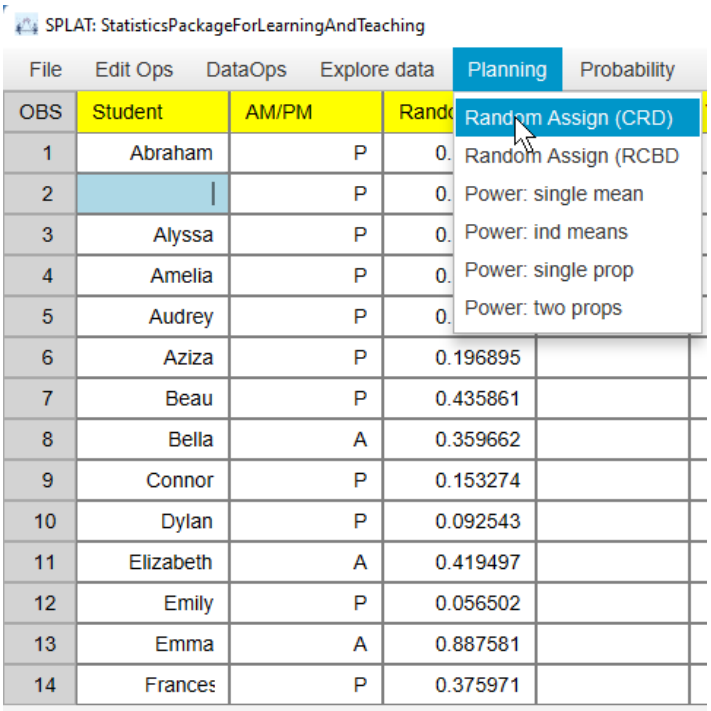


Executing a study: Random Assignment to treatments

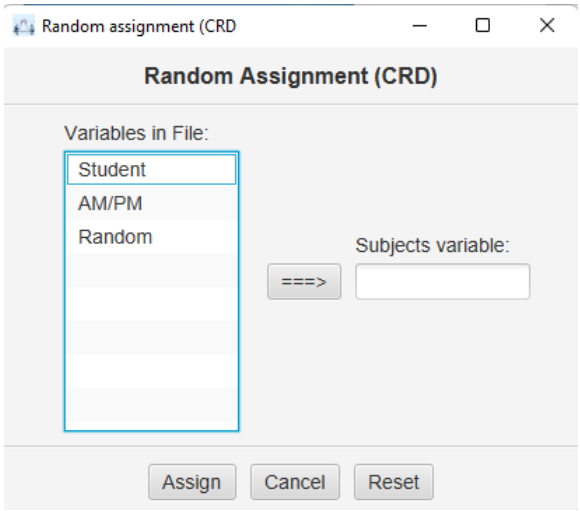
In an experiment the first line of defense against confounding is random assignment to treatments. SPLAT can assist with this important task (and save some time!). The two experimental designs in AP Statistics – the Completely Randomized and Randomized (Complete) Block designs differ in how subjects and treatments are brought together. A quantitative or categorical variable may be used as a blocking variable. SPLAT will handle all these scenarios.

Open the file, CSV_StatClasses and click on Planning. There are three "variables" in this file. The left-most variable is the students just waiting with bated breath [!] to be assigned to treatments. AM/PM and Random are two possible blocking variables, one categorical and one quantitative. (In real life you would supply these possible blocking variables. The quantitative blocking variable in this example file consist of meaningless random numbers. For reasons completely unknown, some of the students seem to be hiding offstage right in the cells. There MUST be some way to fix this, but it has eluded me so far. If this offends you, you can click in the names column and arrow up and down – that seems to bring the values into view. The three assignments to treatments are shown in the Dropdown menu. Two of the treatments are lumped together under the not unreasonable choice, **R**andomized **C**omplete **B**lock **D**esign.)

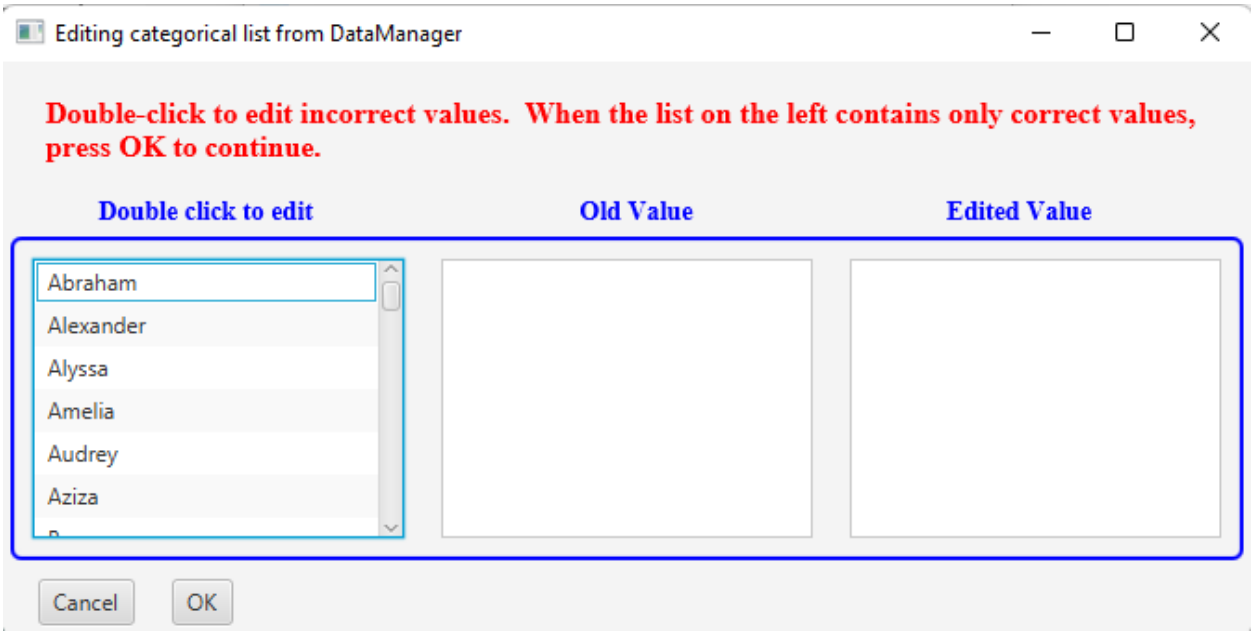
Execution of the assignment to treatment strategies is elementary, Dr. Watson! For the **C**ompletely **R**andomized **D**esign one needs to tell SPLAT which variable includes the subjects, how many treatments there are and what they are. **Select Random Assign (CRD).** A by now familiar panel will appear…



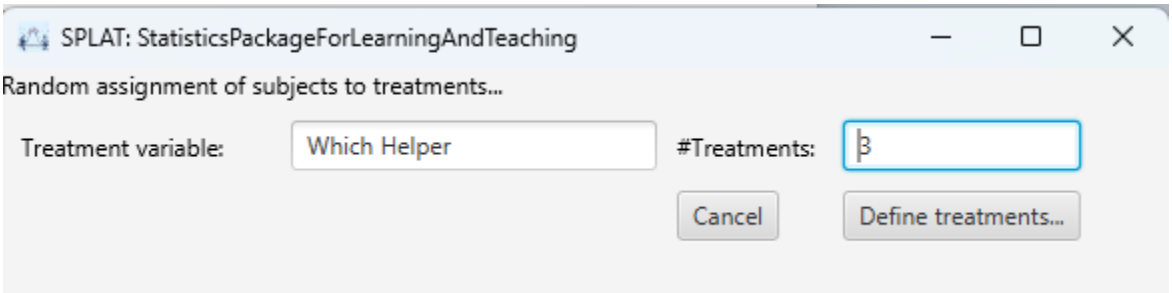
OBS	Student	AM/PM	Rando
1	Abraham	P	0.
2		P	0.
3	Alyssa	P	0.
4	Amelia	P	0.
5	Audrey	P	0.
6	Aziza	P	0.196895
7	Beau	P	0.435861
8	Bella	A	0.359662
9	Connor	P	0.153274
10	Dylan	P	0.092543
11	Elizabeth	A	0.419497
12	Emily	P	0.056502
13	Emma	A	0.887581
14	Frances	P	0.375971



Select Student as the Subjects variable – bet you didn't see THAT coming – and click on **Assign**…



Elsewhere in SPLAT, where categorical values of a variable can appear more than once, you can fix them here. However, as experimental subjects presumably your subjects have unique names. You will not be able to fix those here – bail out and fix the names in the spreadsheet. If all is OK at this point, click on OK.



Here I have defined the variable "Which helper" and informed SPLAT there are 3 treatments. In the case of the CRD, SPLAT will come as close as possible to providing equal treatment group sizes.

On to randomized complete blocks!

The only assignment to treatments SPLAT will do for the Randomized Complete Block design is complete blocks, i.e. each block contains one of every treatment. If the number of subjects divided by the number of treatments is not a whole number, SPLAT will display a typically kind and gentle advisory…

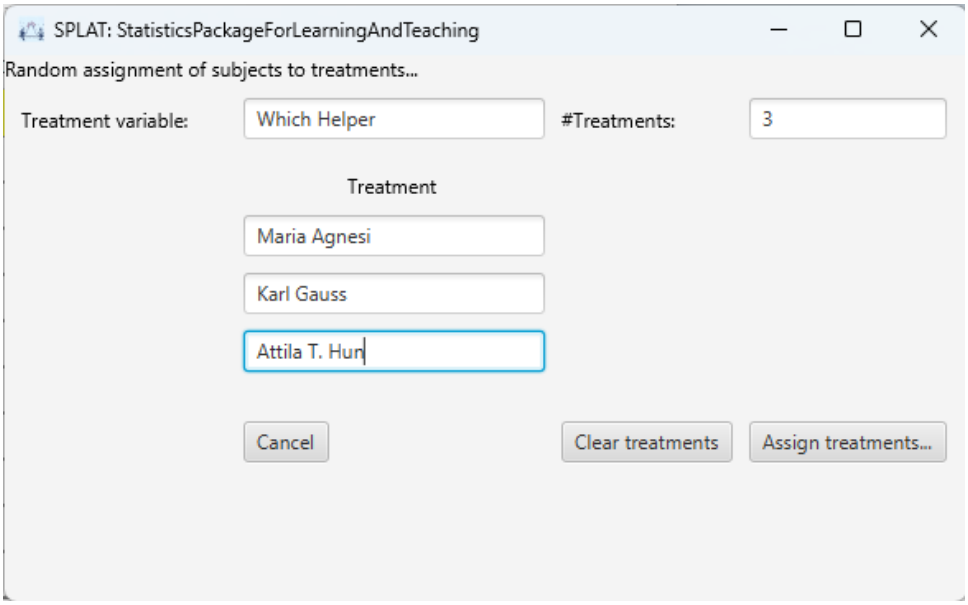
Hey, USER -- This ain't Animal Farm!!

I, SPLAT, only handle COMPLETE blocks!!!

So, what's the deal here?!?!?! You think that some blocks are more equal than other blocks? Think you can skimp on subjects by shorting some of the blocks?!??! SPLAT is shocked, SHOCKED at this perfidious attempt to slip one past your Instituional Review Board!!! You DO have and IRB, do you not?? Or SOMEONE to explain the Randomized Complete Block facts of life? Harumph!!

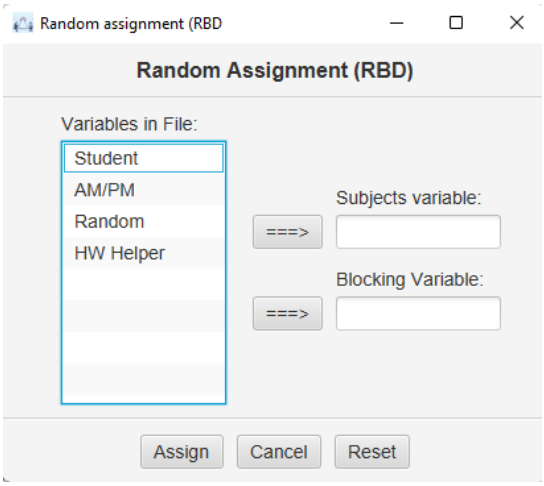
Oooohhhh, SPLAT, you are SO cool, and SO helpful. Click to agree and continue.

For this experiment we will assign the students to one of our homework helpers…

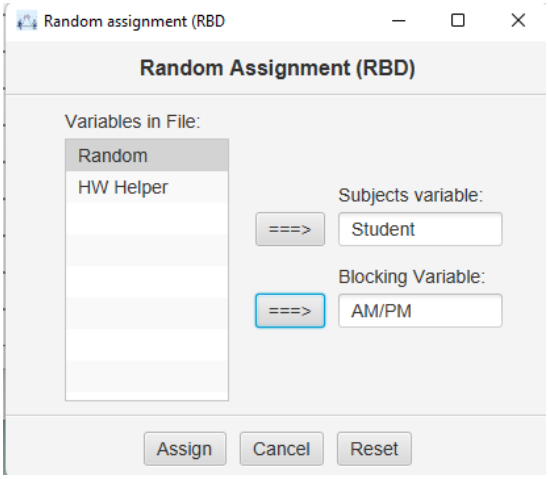


…and click on **Assign treatments**.

Assigning treatments for a randomized complete block design works in a very similar manner to the completely randomized design; the only difference is that you have an additional decision – which is the blocking variable? We will continue from our earlier assignment…



I'm thinking…Class meeting time (AM/PM).



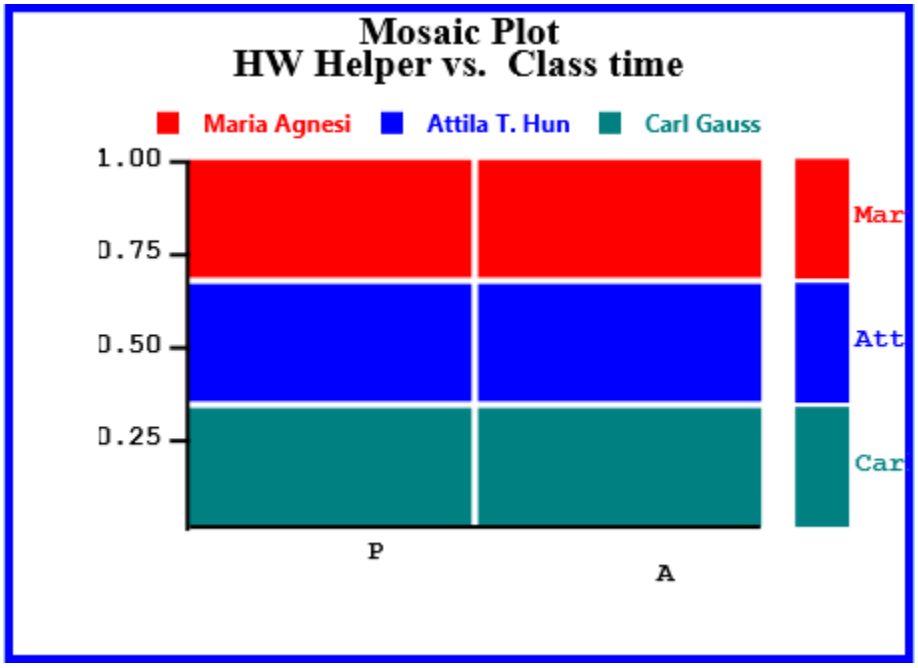
Use the same treatments as before….Ta-daaaaaa!

SPLAT: StatisticsPackageForLearningAndTeaching

File | Edit Ops | DataOps | Explore data | Planning | Probability

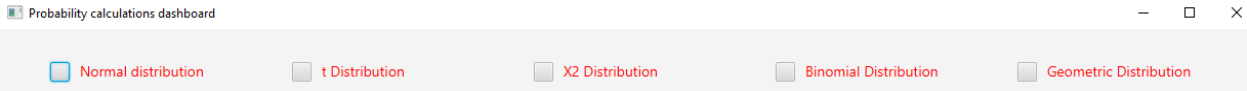
OBS	Student	AM/PM	Random	Treatment
1	Abraham	P	0.335285	Carl Gauss
2	Alexander	P	0.372748	Carl Gauss
3	Alyssa	P	0.796440	Carl Gauss
4	Amelia	P	0.905633	Attila T. Hun
5	Audrey	P	0.526136	Maria Agnesi
6	Aziza	P	0.196895	Attila T. Hun
7	Beau	P	0.435861	Carl Gauss
8	Bella	A	0.359662	Maria Agnesi
9	Connor	P	0.153274	Maria Agnesi
10	Dylan	P	0.092543	Carl Gauss

If we are skeptical, we can verify the "completeness" of the random assignment with a Mosaic plot…

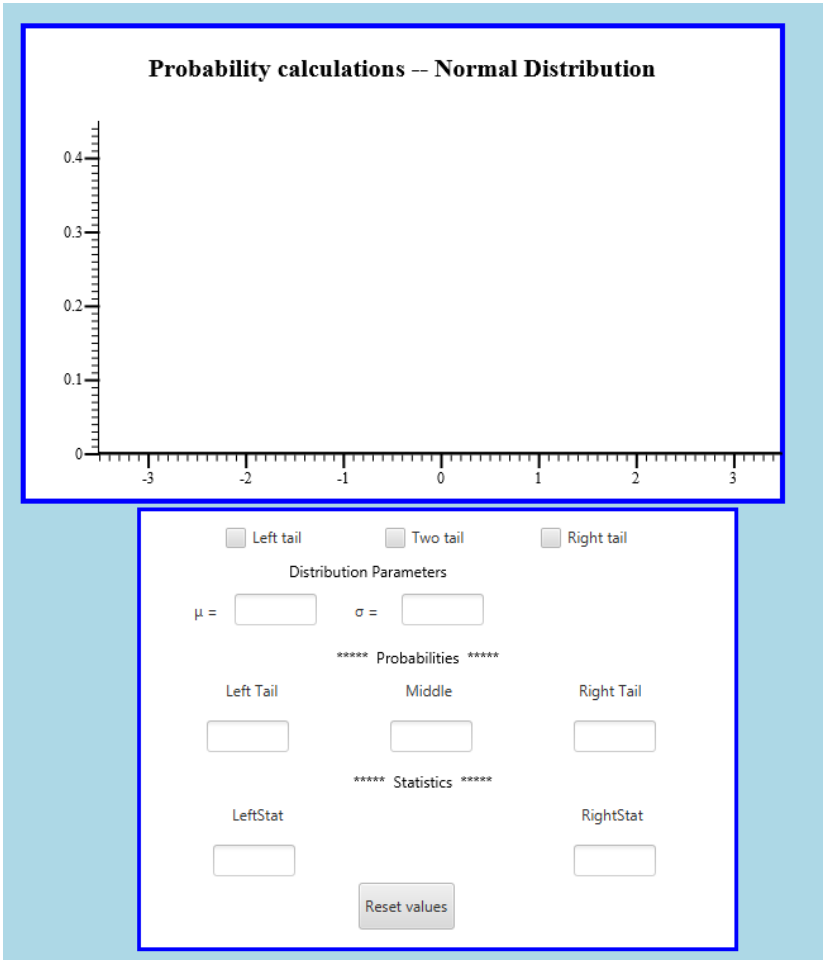


Probabilities – Normal, *t*, Chi square

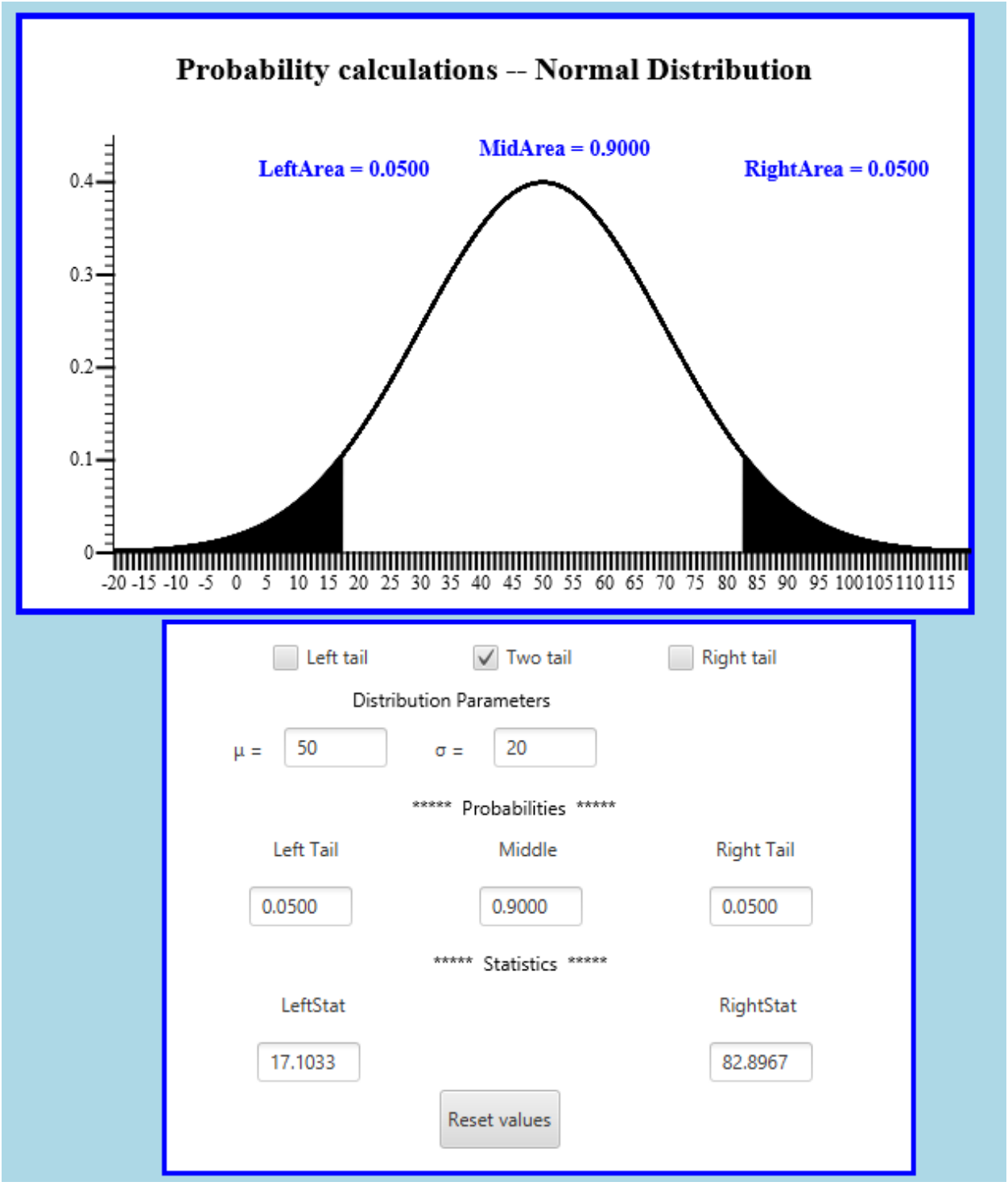
SPLAT can act in place of the customary tables of the statistical persuasion. To capitalize on this capability, click on Probability → "Probability Distributions." What to your wondering eyes should appear, but the usual SPLAT panel of choices?



As you can clearly see – if you have a magnifying glass – SPLAT provides calculations for the random variables in the AP Statistics CED. These random variables all work in a similar manner and present similar information; check the "Normal" box and you should see:

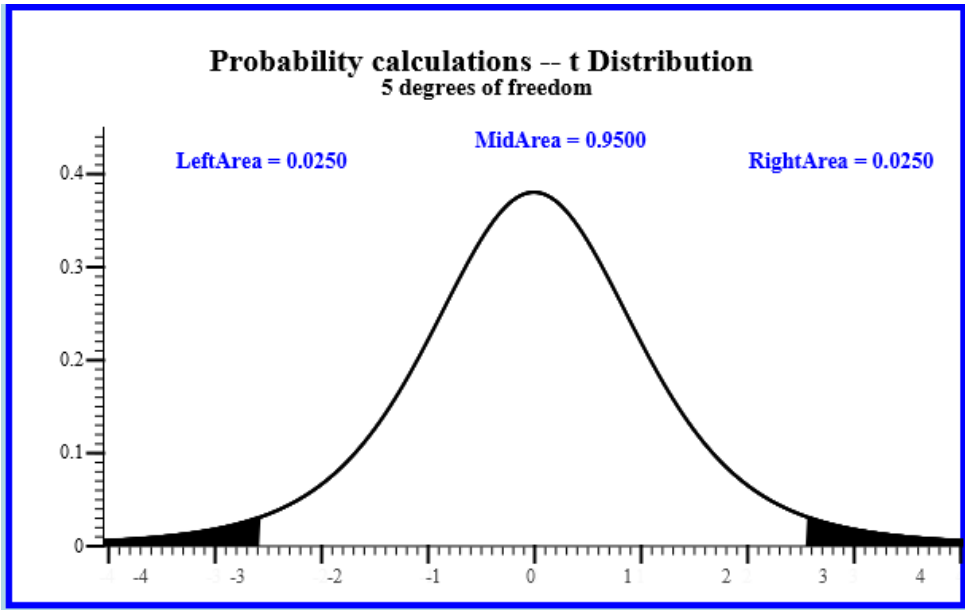


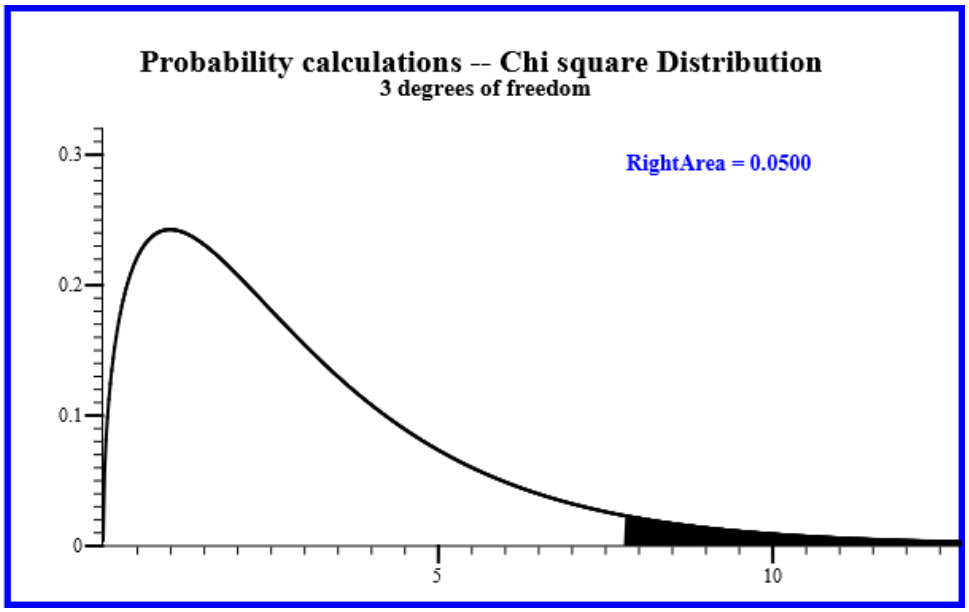
Now click on "Two tail" and fill in a mean and standard deviation (not necessarily 0 and 1) and choose a "Middle" probability, and add a probability, say .90. You should see something like this:


Probability calculations -- Normal Distribution
LeftArea = 0.0500
MidArea = 0.9000
RightArea = 0.0500
0.4
0.3
0.2
0.1
0
-20 -15 -10 -5 0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100 105 110 115
Left tail
Two tail
Right tail
Distribution Parameters
μ = 50
σ = 20
***** Probabilities *****
Left Tail
Middle
Right Tail
0.0500
0.9000
0.0500
***** Statistics *****
LeftStat
RightStat
17.1033
82.8967
Reset values

You may see some stray information presented, like "Left Area = NaN." This is the Java language equivalent of "There is no such number." I am trying to just not print anything when there is no such number, but I am still scratching my head about how to do this. It SHOULD be obvious to any decent programmer but obvious has not worked for me so far. Also, it may be relatively easy for the wantonly destructive student types to mess this up, e.g., with probabilities summing to a value greater than 1.0. As I find these errant possibilities, I will add more error alerts – but keeping up with the young Maxwellian Demons is a constant chore!

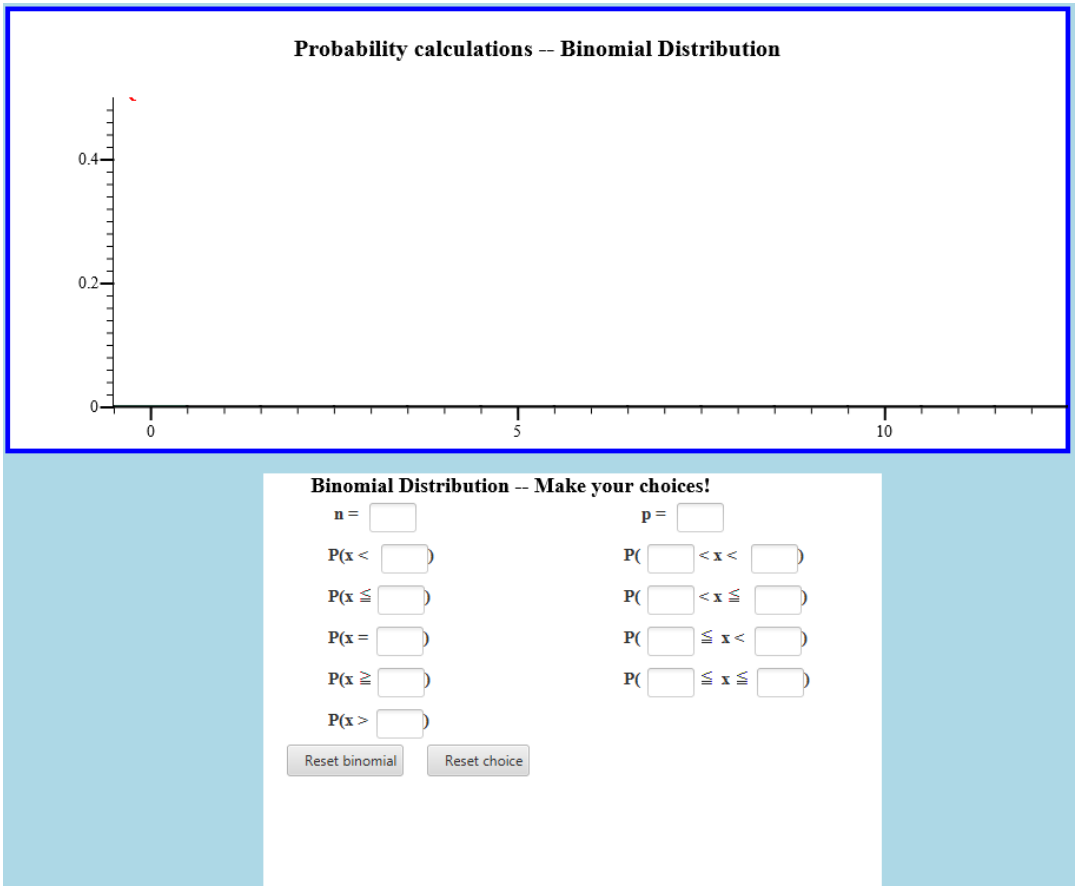
The *t* and χ^2 tables work in a similar manner.





Probabilities – Binomial and Geometric

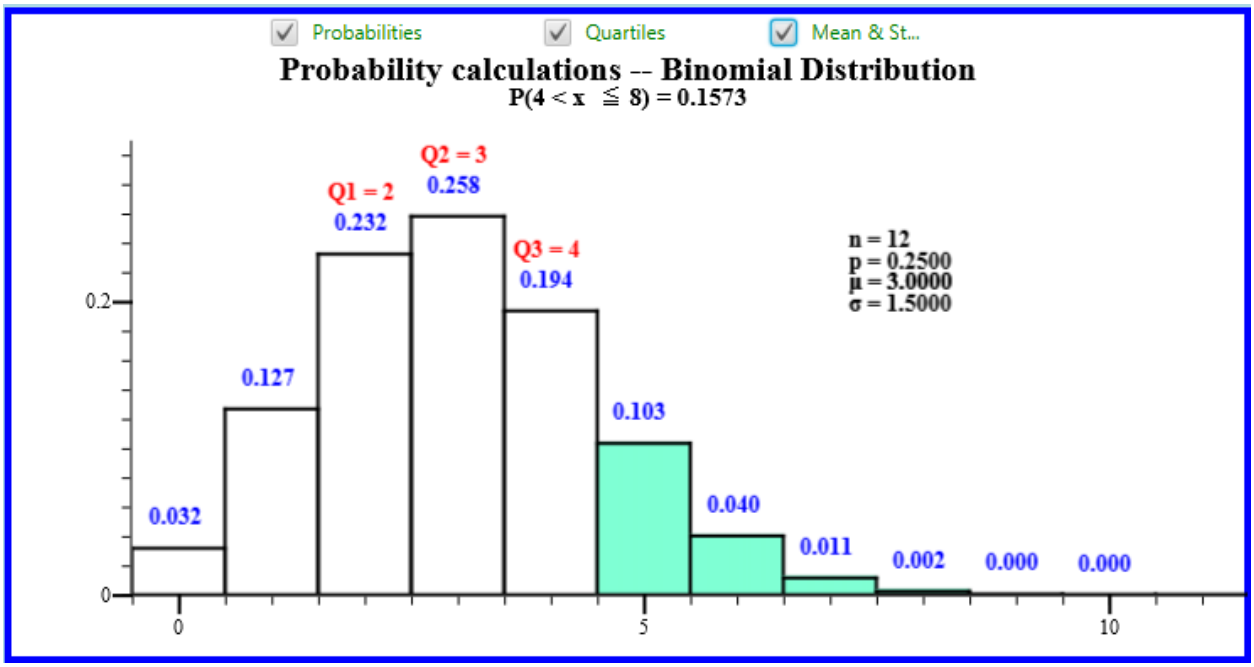
Discrete random variables being different from continuous random variables, I handle them differently. So, for example, suppose you click on "Binomial" in the Probability Distributions panel. This is what you will see:



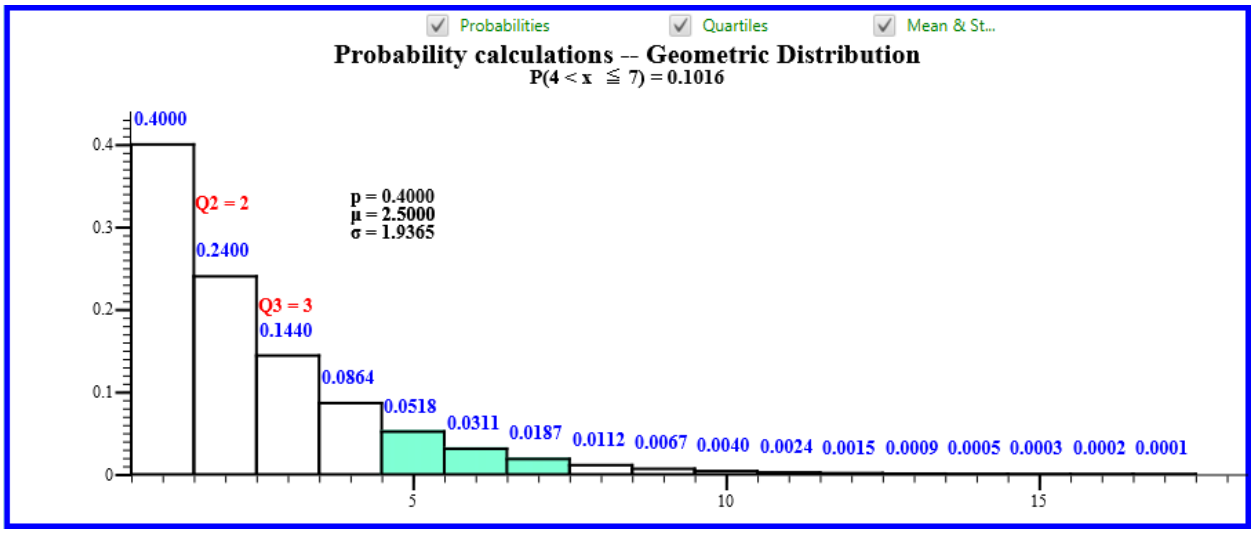
Your task at this point is twofold: first provide an *n*, and a *p* – the parameters of the binomial distribution, and then pick the probability you wish to have calculated. I will use as an example $n = 12$ and $p = .25$. Now suppose you pick this…

$$P(\_\_ < x \leq \_\_)$$

…and supply the values 4, and 8 in the blanks and press Enter (NOT either of the Reset buttons). After your last entry you will – not surprisingly – be presented with the appropriate probability. You will also – after some clicking and dragging -- be presented with the binomial distribution, and assorted information about quartiles, means and standard deviations, and probabilities associated with the random variable:

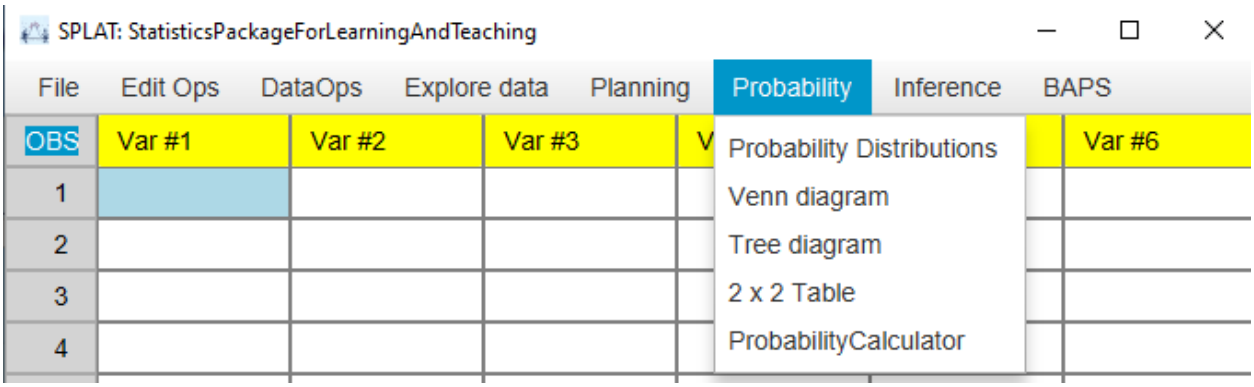


The geometric distribution works in a similar manner…

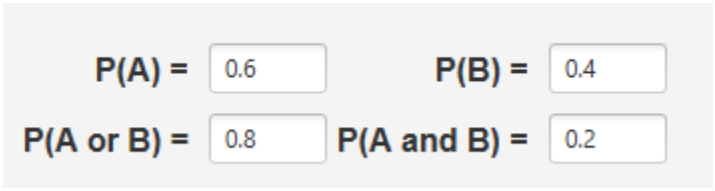


Visual Probabilities

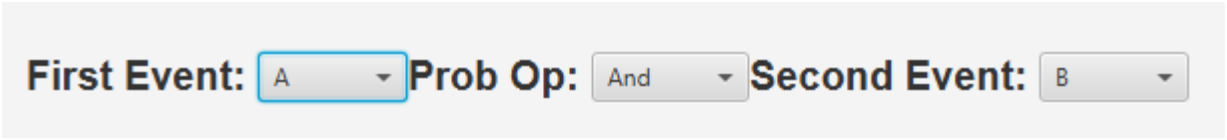
Probability is a perennially difficult topic, and it is frequently the case that a visual representation is helpful when doing probability problems. For elementary probability problems, SPLAT can create Venn diagrams, tree diagrams, and 2 x 2 tables. A straight probability calculation is also available. These options are -- rather cleverly IMHO -- placed in the SPLAT menus under "Probability."



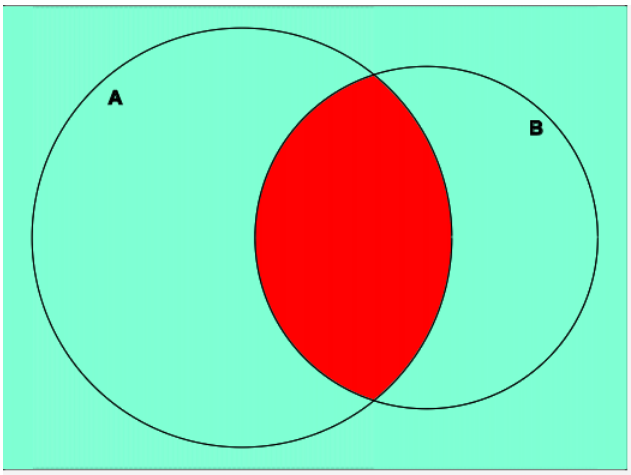
The basic idea in each of these is that when provided probabilities for A, B, and one of A and B and A or B…



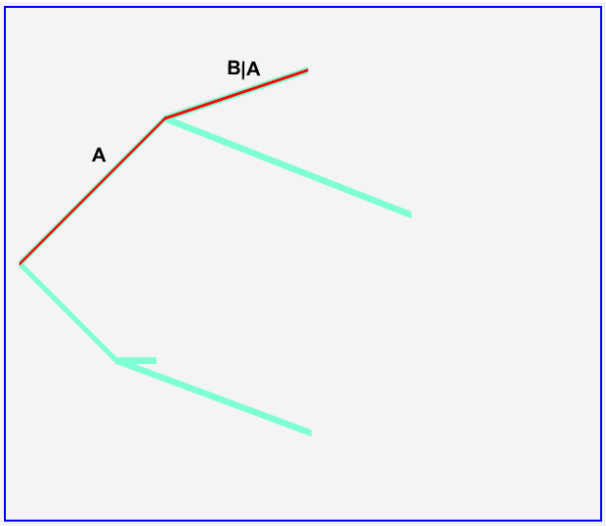
SPLAT can do all the rest. Which probability to display is governed by choices at the top of the screen:



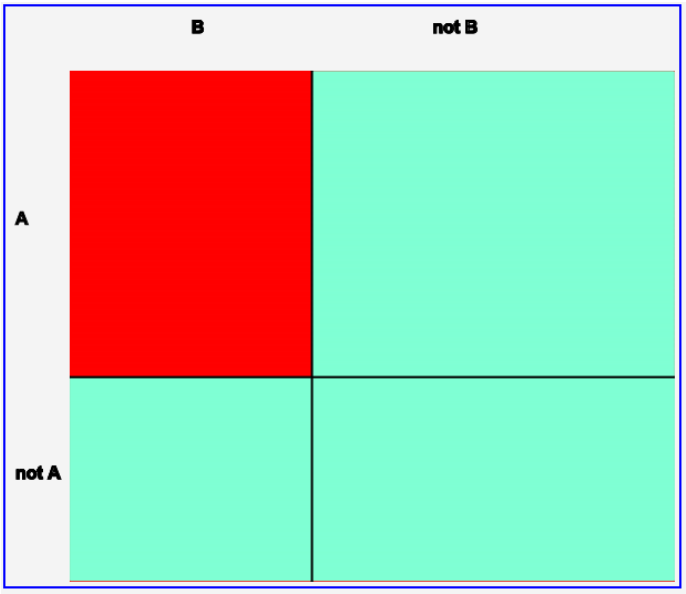
Each of the three visual representations adjusts the "sizes" of the visual components according to the given probabilities. The circles, rectangles, and tree branches are larger when the given probabilities are larger. The Venn, Tree, and Table representations for the probabilities above are shown below. The probability line at the bottom of the panels is there to give some idea of the numerator and denominator that result in the probabilities.



Probability of A and B = $\frac{\text{0.2000}}{\text{1.0000}}$ = 0.2000



Probability of A and B = $\frac{\text{0.2000}}{\text{1.0000}}$ = 0.2000



Probability of A and B = $\frac{\text{0.2000}}{\text{1.0000}}$ = 0.2000

***** The Kitchen Sink!!! *****

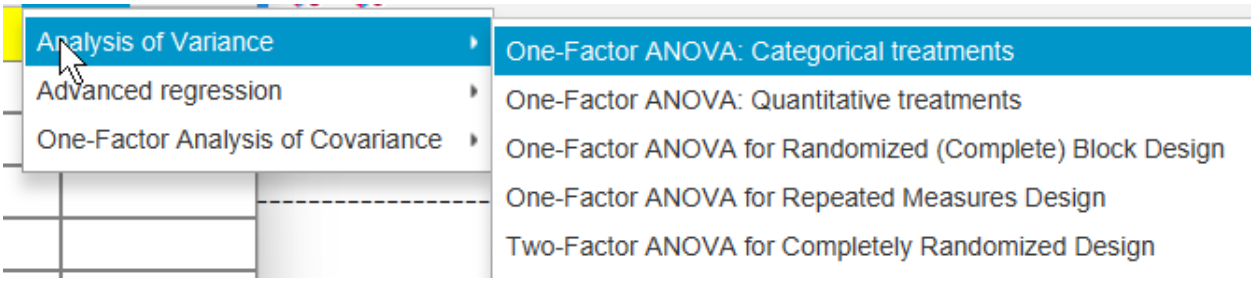
Probability of A = 0.300
Probability of not-A = 0.700
Probability of B = 0.400
Probability of not-B = 0.600
Probability of A and B = 0.200
Probability of A or B = 0.500
Probability of A and not-B = 0.100
Probability of A or not-B = 0.800
Probability of not-A and B = 0.200
Probability of not-A or B = 0.900
Probability of not-A and Not-B = 0.500
Probability of not-A or Not-B = 0.800
Probability of A given B = 0.500
Probability of B given A = 0.667
Probability of not-A given B = 0.500
Probability of not-B given A = 0.333
Probability of not-A given not-B = 0.833
Probability of not-B given not-A = 0.714

Beyond AP Statistics!!

I have programmed into SPLAT some calculations and topics that are not in the Advanced Placement Statistics Course Description. In some cases (e.g., the Anderson-Darling test for normality) I simply wish to suggest to students that there are statistics beyond the AP course. In other cases, the BAPS procedures have been added ad hoc to provide analytic capability for student projects.

Also, I am mindful that for some teachers the academic year calendar extends far beyond the AP exam. For those post-test days when time and inclination permit, SPLAT supports the idea that students could be presented with topics beyond the AP Course. (It could also occur that their projects will take them beyond AP Statistics topics into BAPS territory.) These BAPS topics include the Analysis of Variance, Logistic regression (simple and multiple), ordinary Multiple regression, and a new entry: Analysis of Covariance. For those who might not be familiar with these (and thank your lucky stars for that ignorance), what follows is a short description of what they do. **Note: SPLAT will not in any sense teach any of this stuff – you (and student) should already know about these procedures before clicking away on them!!!)**

Analysis of Variance



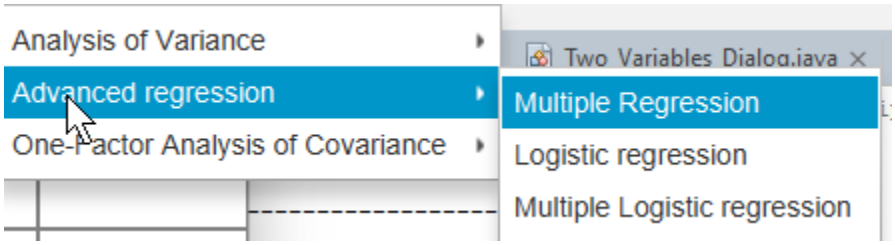
One-way Analysis of Variance (ANOVA) allows one to go beyond the independent means t-test and test a hypothesis of equality of more than two means. If your explanatory variable is quantitative (e.g. dosage levels) the analysis is the same, but the graphs will be slightly different and reflect the quantitative nature of the explanatory variable.

The Randomized (Complete) Block design will analyze data from a Randomized Block Design experiment that students studied in the Planning Studies part of AP Statistics.

The Repeated Measures Design allows one to go beyond the "paired t" to a number of repeated measures past two.

Two-way ANOVA allows you to have more than one (i.e. two!) explanatory variables.

Advanced Regression

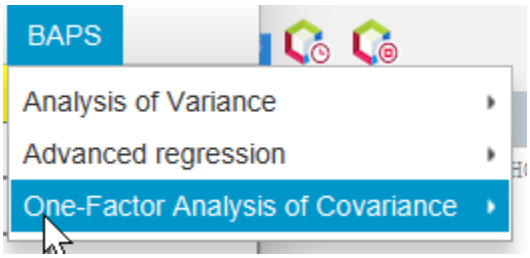


With multiple regression, more than one explanatory variable is allowed.

With logistic regression, the response variable is binary, the explanatory variable quantitative.

With multiple logistic regression, you get both the above.

One Factor Analysis of Covariance



Analysis of covariance is sort of a combination of regression and ANOVA. One could perform an ANCOVA if one had a quantitative rather than categorical blocking variable.

One Way Analysis of Variance (ANOVA)

To perform a One-way Analysis of Variance, execute the clicks: BAPS → Analysis of Variance and take your pick. The single factor ANOVA with categorical treatments is the basic ANOVA everybody learns first.

Before you make chicken soup you need a chicken, and before you do an ANOVA, you need data. Recall that raw data can be organized in SPLAT in two ways as shown at right and below. The organ-ization shown below is familiar to TI-84 users; the data appears in "Lists." It is more common in statistical software to see data in a file in the format shown at right. One column indicates the group or treatment, and one column contains the values of the raw data.

SPLAT: StatisticsPackageForLearningAndTea

File Edit Ops DataOps Explore

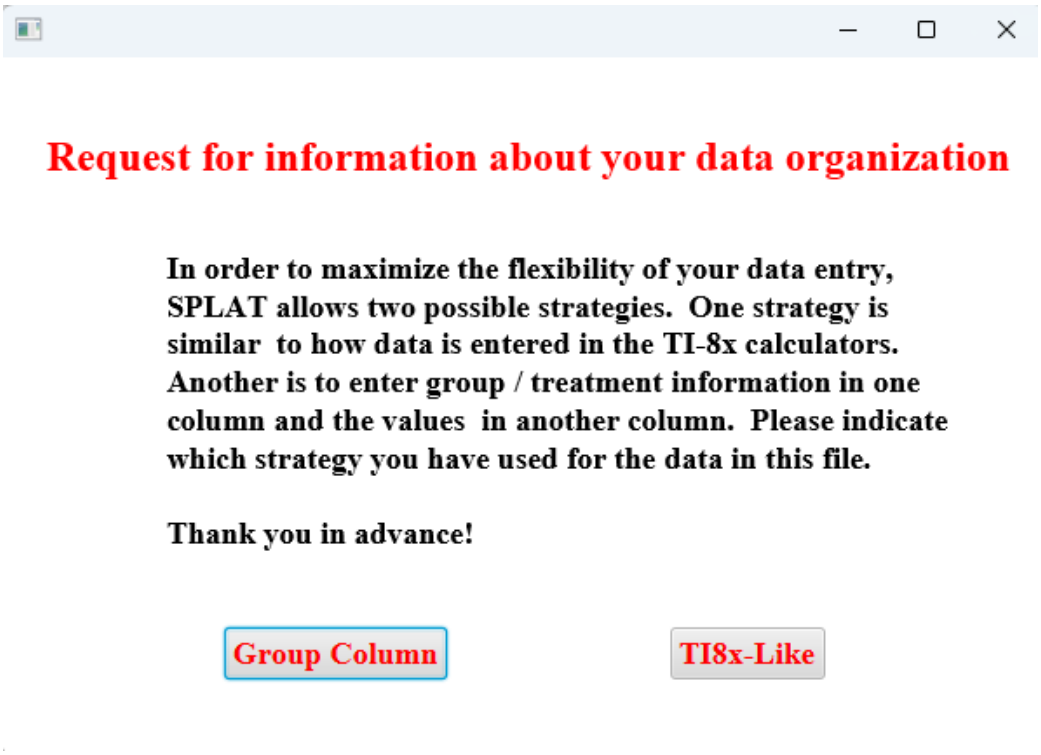
OBS	Base_Var	At_Base
4	Base_Cntrl	74.0
5	Base_Cntrl	78.1
6	Base_Cntrl	88.3
7	Base_Cntrl	87.3
8	Base_Cntrl	75.1
9	Base_Cntrl	80.6
10	Base_Cntrl	78.4
11	Base_Cntrl	77.6
12	Base_Cntrl	88.7
13	Base_Cntrl	81.3
14	Base_Cntrl	78.1
15	Base_Cntrl	70.5
16	Base_Cntrl	77.3

SPLAT: Statistics Package for Learning And Teaching

File Edit DataOps Explore data Planning Probability Inference BAPS

OBS	Base_Cntrl	Base_Be	Base_Fam	Gain_Cntrl	Gain_Bel	Gain_Fam
1	80.7	80.5	83.8	-0.5	1.7	11.4
2	89.4	84.9	83.3	-9.3	0.7	11.0
3	91.8	81.5	86.0	-5.4	-0.1	5.5
4	74.0	82.6	82.5	12.3	-0.7	9.4
5	78.1	79.9	86.7	-2.0	-3.5	13.6
6	88.3	88.7	79.6	-10.2	14.9	-2.9

You care about this because when you pick one-way ANOVA as your procedure DuJour, SPLAT will respond with …



For my One-way (that is, one explanatory variable) ANOVA example I will re-grab the data from the jet lag study above. The data, you will no doubt remember, is in the file, CSV_Circadian, and is stored in separate columns: Aha! The TI8x-Like button is the way to go.

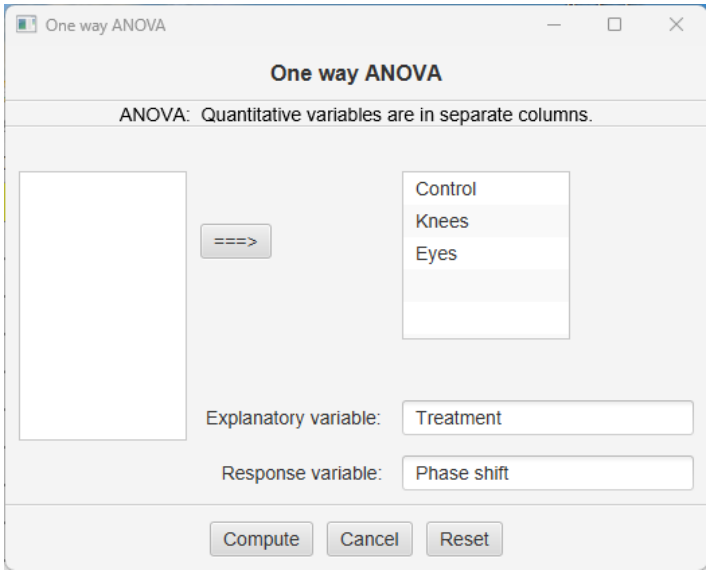
The spreadsheet should look like the following. Notice that the sample sizes are not equal. Equal sample sizes are not particularly a problem with one-way ANOVA, but they are advised.

SPLAT: StatisticsPackageForLearningAndTeaching

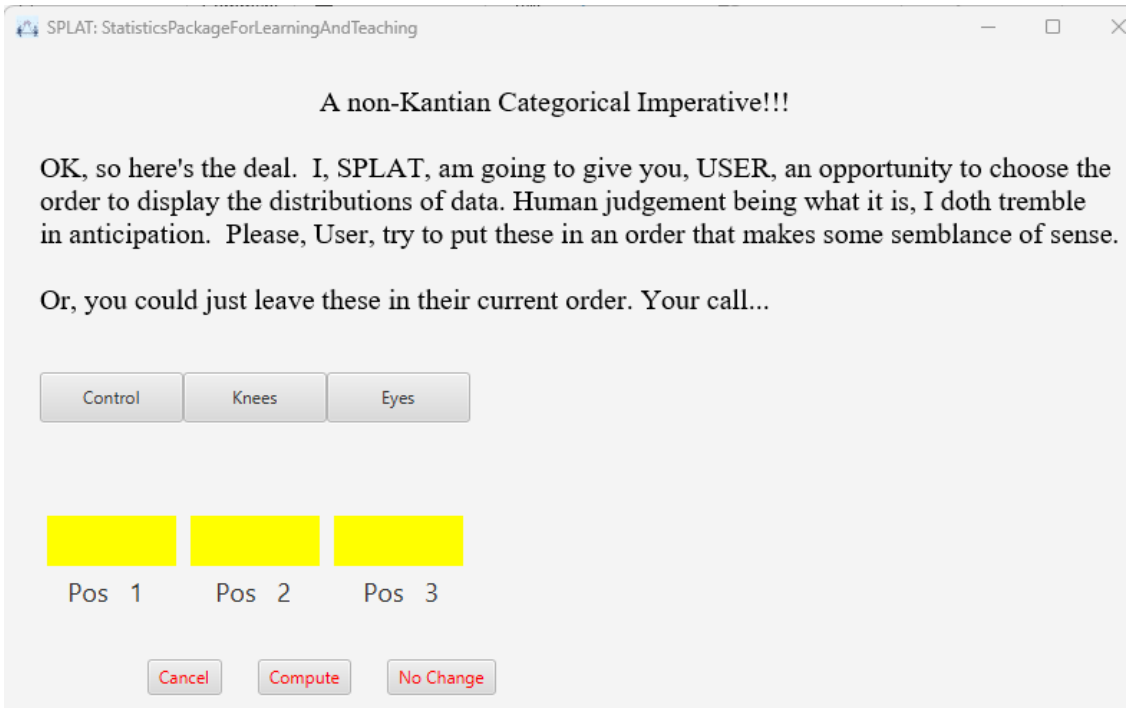
File Edit Ops DataOps Explore data Planning Probability Inference BAPS

OBS	Control	Knees	Eyes	Var #4	Var #5	Var #6
1	0.53	0.73	-0.78			
2	0.36	0.31	-0.86			
3	0.20	0.03	-1.35			
4	-0.37	-0.29	-1.48			

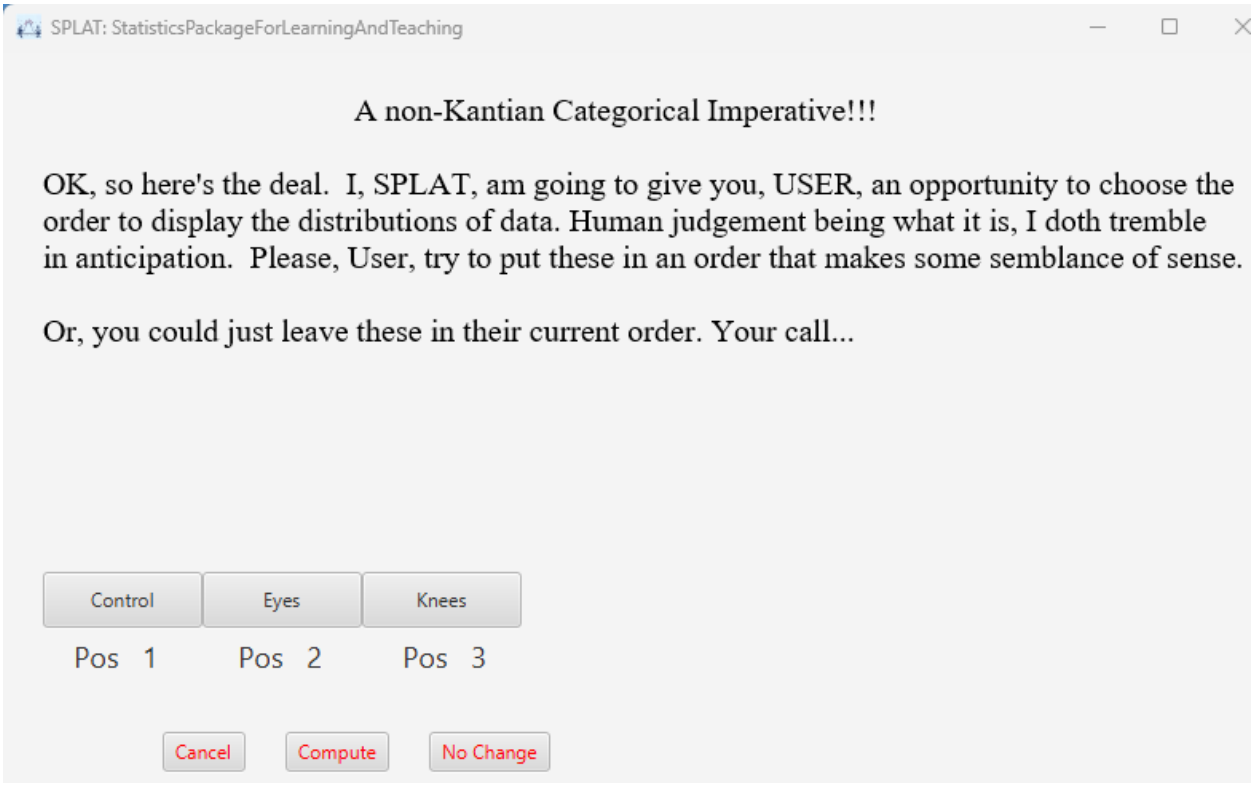
In the menu, click on BAPS→ AnalysisOfVariance→OneFactorANOVA: Categorical treatments.



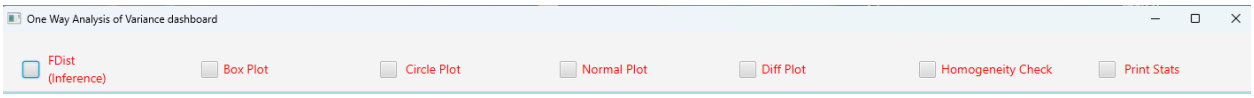
Here the explanatory variable is the "Treatment" with values Control, Knees, and Eyes. The response variable (which, truth be told, I only vaguely understand!) is the phase shift. Click on "Compute" and this will bring us to a choice panel the likes of which we have seen before…



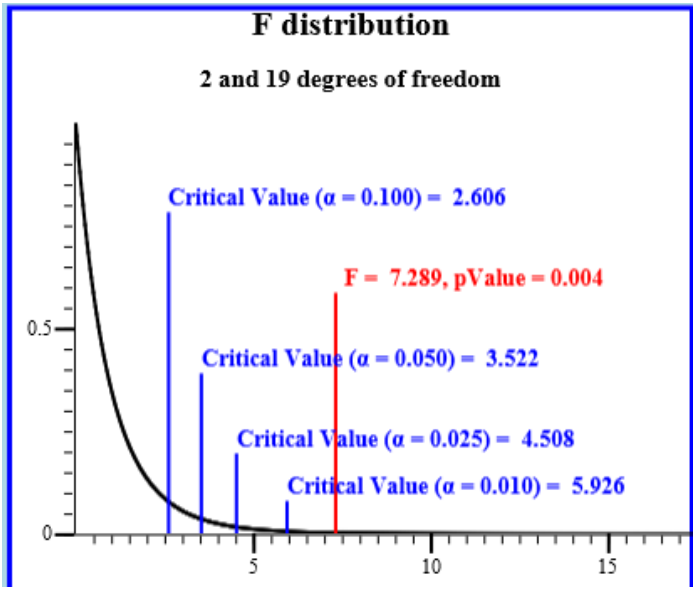
Again, I like alphabetizing…



After Compute-clicking, here are your options:



Holy repetitious look-and-feel, Batman! The usual dashboard precursor to the actual output has appeared! Not much new here – graphic output is provided for checking the assumptions of ANOVA, and the usual-suspect ANOVA statistics are presented. (Remember, you are supposed to already know what these are.) For equal sample sizes SPLAT will give you Tukey's Honestly Significant Difference; for unequal sample sizes you get Tukey-Kramer. Here, in graphic form, is the sampling distribution, F-statistic, and P-value…



Here is all the statistical stuff…

One-way Analysis of Variance

One-way Analysis of Variance: Phase shift vs. Treatment

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F	P-value
Treatments	2	7.22	3.61	7.29	0.0045
Error	19	9.42	0.50		
Total	21	16.64			

Omega Square = 0.364
Cohen's d = 0.756

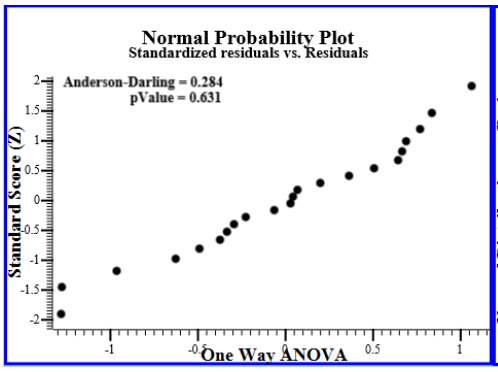
********** Parameter estimates for Levels **********

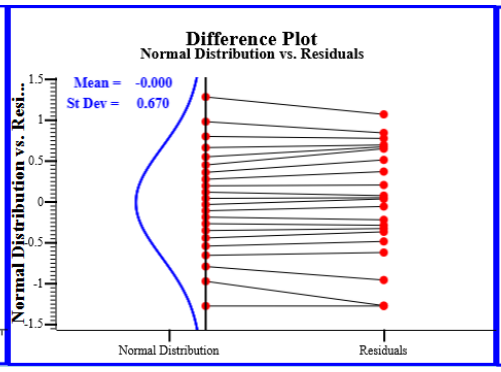
Treatment/ Group	Sample Size	Sample Mean	Sample St Dev	Std Err of mean	Margin of Error	Lower 95PC Bound	Upper 95PC Bound
Control	8	-0.309	0.618	0.233	0.552	-0.861	0.243
Knees	7	-0.336	0.791	0.323	0.790	-1.126	0.454
Eyes	7	-1.551	0.706	0.288	0.706	-2.257	-0.846

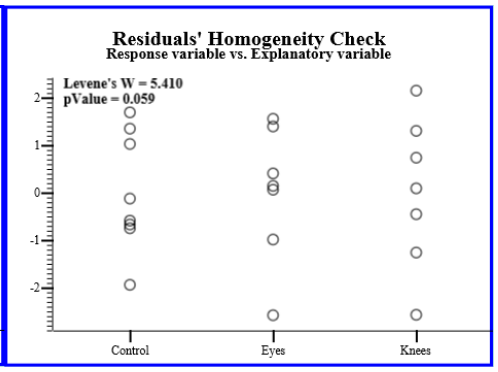
Tukey-Kramer Test Phase shift vs. Treatment

Treatment/ Group	Treatment/ Group	Mean Difference	Critical Q	95PC CI Lower Bound	95PC CI Upper Bound
Control	Knees	0.027	0.926	-0.899	0.953
Control	Eyes	1.243	0.926	0.317	2.168
Knees	Eyes	1.216	0.956	0.260	2.172

Plots for checking assumptions…







Two Way Analysis of Variance:
Factorial, Randomized Block

Two-way analysis of variance extends the concepts introduced in one-way analysis of variance to handle two explanatory variables. Ordinarily one would think that two-way ANOVA is way BAPS, and I would tend to agree. The only reason one might want to go beyond one-way ANOVA is that two-way ANOVA provides a mechanism for analyzing experiments with two categorical explanatory variables, or if analyzing data from an experiment conducted as a Randomized Block design. Some students' projects could conceivably be analyzed with a two-way ANOVA. SPLAT will perform the analysis for a two-variable factorial design, and for the randomized block design both with and without replication of treatments within blocks, and balanced or not.

The tour of two-way ANOVA will use examples from Tamhane, A. C. Statistical Analysis of Designed Experiments: Theory and Applications.

The Data Setup

The SPLAT data setup for all the two-way ANOVA analyses is in this "column format." For the factorial design, two columns are needed for the variables, and one for the response variable. In the data setup for the randomized block design, one column is needed for the treatment, one for the blocking variable, and one for the response variable.

These data are from Example 6.2 in Tamhane, a study of bonding strength of capacitors. The two explanatory variables are the Substrate and Bonding Material. The data file is CSV_Tamhane_Balanced.

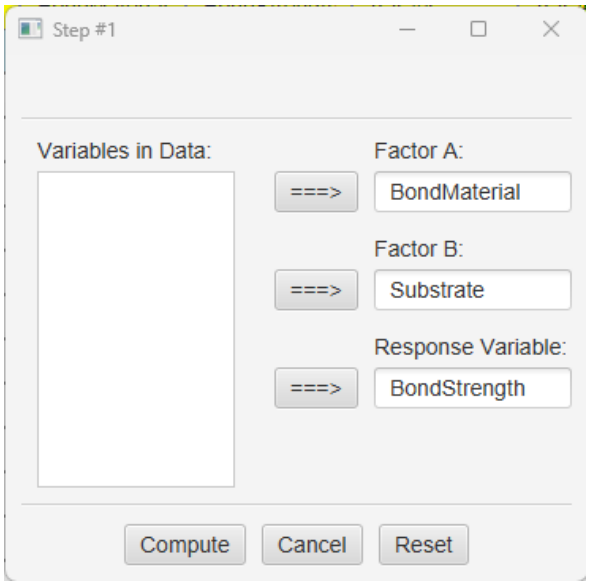
This is a Two Factor ANOVA, Completely Randomized design.

SPLAT: StatisticsPackageForLearningAndTeaching

File Edit Ops DataOps Explore data

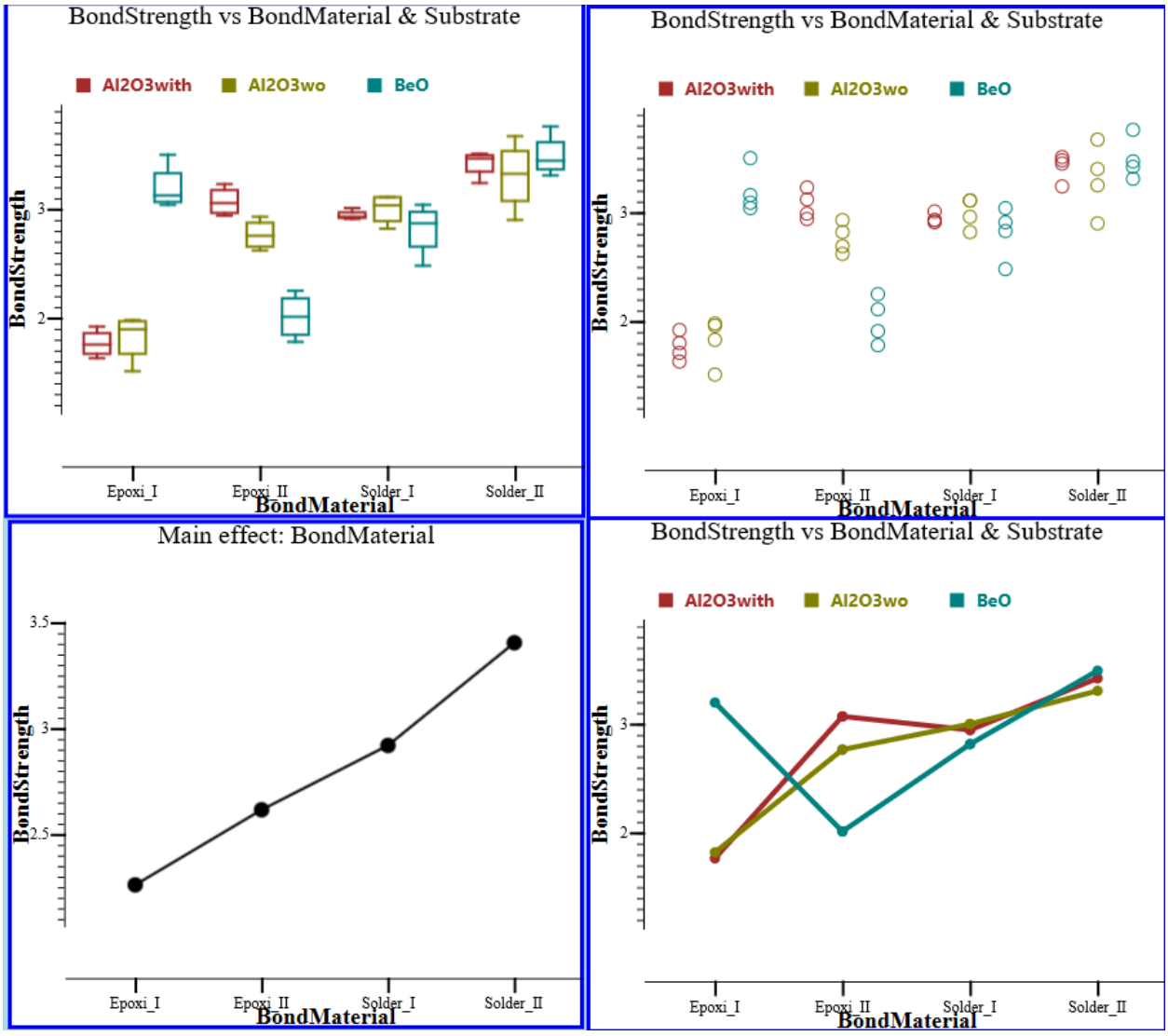
OBS	Substrate	BondMat	BondStre
1	Al2O3wo	Epoxi_I	1.51
2	Al2O3wo	Epoxi_I	1.96
3	Al2O3wo	Epoxi_I	1.83
4	Al2O3wo	Epoxi_I	1.98
5	Al2O3wo	Epoxi_II	2.62
6	Al2O3wo	Epoxi_II	2.82
7	Al2O3wo	Epoxi_II	2.69
8	Al2O3wo	Epoxi_II	2.93
9	Al2O3wo	Solder_I	2.96
10	Al2O3wo	Solder_I	2.82

Click on BAPS→Analysis of Variance
→ Two factor ANOVA for Completely Randomized Design



When you click on Compute, the standard SPLAT dashboard choices will appear:

Your choices are boxplot, circle plot, main effects plots (Only the plot for BondMaterial is shown below) and an interaction plot….



… and, of course you can get the usual two-way ANOVA statistical summary.

Two-way Factorial Analysis of Variance

Source of Variation	df	Sum of Squares	Mean Square	F	P-value
BondMaterial	3	8.461	2.820	80.765	0.0000
Substrate	2	0.195	0.098	2.797	0.0743
BondMat*Substra	6	7.587	1.264	36.213	0.0000
Error	36	1.257	0.035		
Total	47	17.500			

Omega Square for Treatment A = 0.101
Cohen's d for Treatment A = 0.335

Omega Square for Treatment B = 0.769
Cohen's d for Treatment B = 1.823

Omega Square for Interaction = 0.815
Cohen's d for Interaction = 2.098

We will use the same data for the randomized block design, this time treating the Bonding material as if it were assigned as a block in the experiment.

Click on BAPS→Analysis of Variance
→ One factor ANOVA for Randomized (Complete) block design)

Randomized Complete Block ANOVA

Source of Variation	df	Sum of Squares	Mean Square	F	P-value
Substrate	2	0.195	0.098	2.797	0.0743
BondMaterial	3	8.461	2.820	80.765	0.0000
Substra*BondMat	6	7.587	1.264	36.213	0.0000
Error	36	1.257	0.035		
Total	47	17.500			

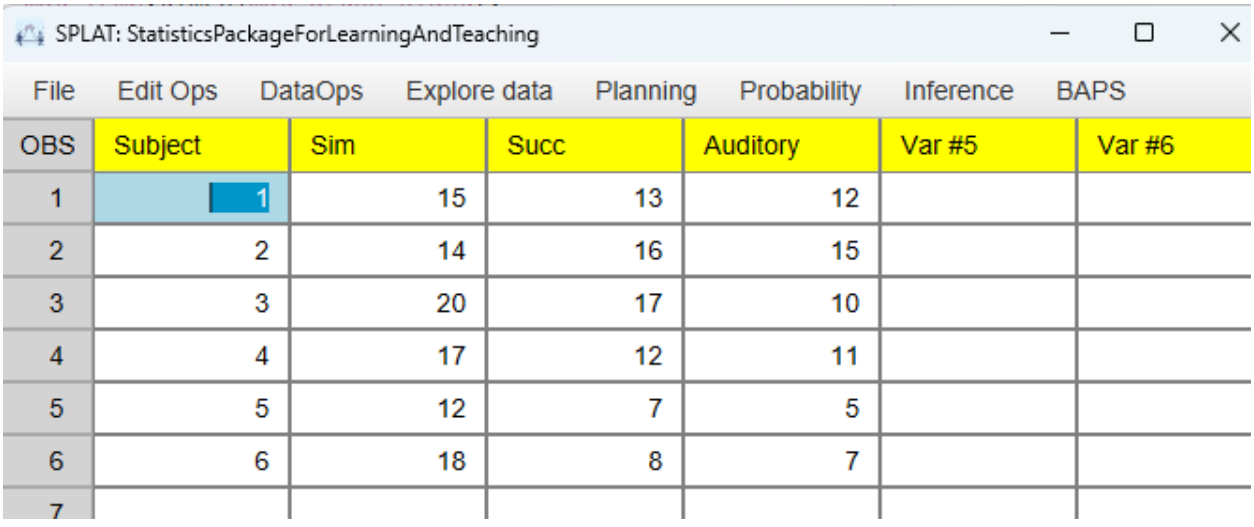
The options and output are, of course, the same; (except for the title); the difference lies in the interpretation of the interaction. With the randomized block design there is not supposed to be any interaction between the explanatory variable and the blocks, and the interaction term is essentially a check on that assumption. There is clearly bow-koo interaction here!

Repeated Measures ANOVA
(Two-way ANOVA cleverly disguised as One-Way ANOVA)
(An extension of the paired/matched *t* test)

The generalization of the independent *t* test to one-way Analysis of Variance (more than two variables) is a step up in terms of the underlying mathematics. The generalization of the paired *t* test to Repeated Measure / Within Groups Analysis of Variance is a leap up in terms of the underlying mathematics. The chief culprit in this leap-making is that the assumption of what is known as "sphericity" is nowhere near as robust as the assumption of normality in the *t* procedures. The check of the sphericity assumption – and how to respond to its failure – is, in my opinion, not something to worry AP Students about unless they are doing some seriously important research. I have included in SPLAT the calculations relevant to the sphericity assumption, but unless a student will be submitting an AP Statistics project in some sort of serious science competition it seems to me that sphericity should not be a concern; interpretation of the data could consist of reporting about the graphs (line plot and boxplots) and the usual output for a two-way ANOVA and be perfectly fine. And for an AP Statistics project, interpretation of just the graphs would be fine.

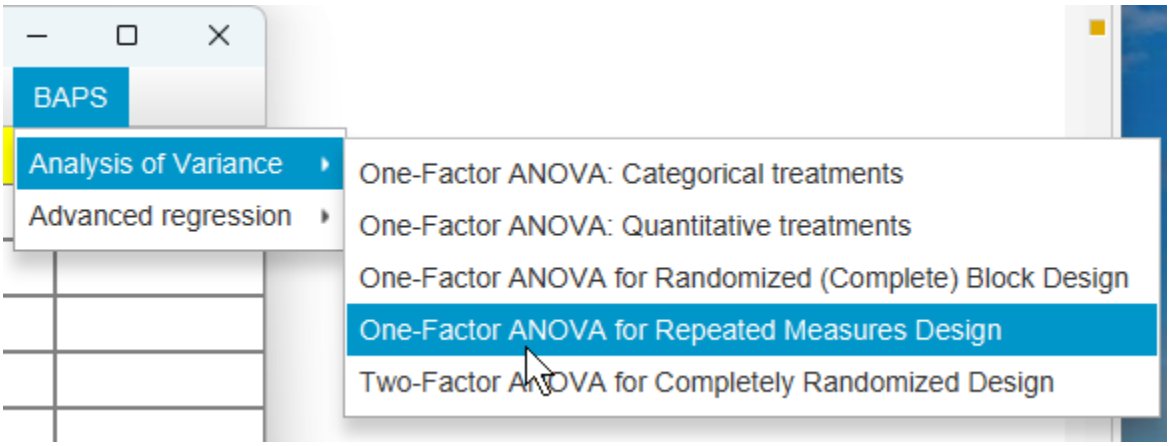
I have not yet uncovered any understandable raw data in the literature. For this example, I will use hypothetical data from Cohen, B. H. (2013). Explaining Psychological Statistics (4th ed.), p513. The hypothetical data consist of responses to three modes of presentation on a computer screen. The response variable is the number of problems correctly solved out of twenty. The file name is CSV_RM_Cohen.

The data layout for repeated measures in SPLAT is shown below. Note that SPLAT demands that the identification of the experimental units / subjects / participants appears in the first column, the repeated measures appear in the next columns, and there are no other columns of data.



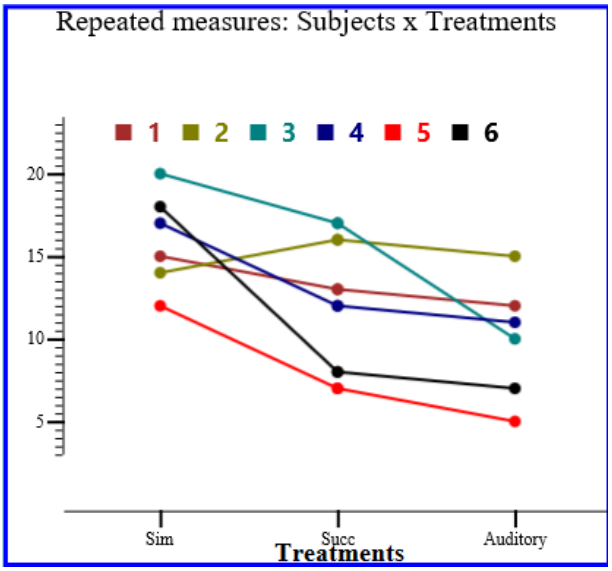
OBS	Subject	Sim	Succ	Auditory	Var #5	Var #6
1	1	15	13	12		
2	2	14	16	15		
3	3	20	17	10		
4	4	17	12	11		
5	5	12	7	5		
6	6	18	8	7		
7						

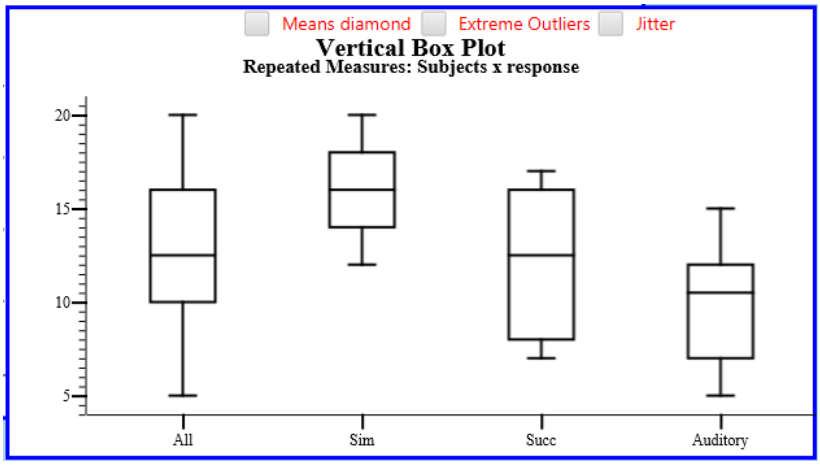
The path for your choices begins at Beyond AP Statistics…



The line plot and box plots are easily interpreted. (Once you have more than a dozen subjects, they are not individually identified in the panel.) For the most part it appears that the responses of the subjects tend to decrease across the treatments in this study. There is no assumption here that the treatments were given through time; there could be crossover in the design. The interpretation of the results would, of course, be different if the subjects were all measured in the same order over time.

Note: For reasons unfathomable to this programmer, you may have to coax the box plot into existence by resizing it. It seems to be shy.





The ANOVA output is a bit more difficult to handle than the graphs, but essentially it follows the usual two-way ANOVA output plus information that looks and acts suspiciously like one-way ANOVA, and can be interpreted as one-way ANOVA output .

Repeated Measures Design

Source of Variation	Sum of Squares	df	Mean Square	F	P-value
Treatments	110.778	2	55.389	8.002	0.0084
Subjects	119.611	5	23.922	3.456	0.0450
Error	69.222	10	6.922		
Total	299.611	17			

Omega Square for Treatments = 0.274
Cohen's f for Treatments = 0.615

Omega Square for Subjects = 0.406
Cohen's f for Subjects = 0.826

********** Parameter estimates for Levels **********

Treatment/ Group	Sample Size	Sample Mean	Sample St Dev	Std Err of mean	Margin of Error	Lower 95PC Bound	Upper 95PC Bound
Sim	6	16.000	2.898	1.296	3.332	12.668	19.332
Succ	6	12.167	4.070	1.820	4.679	7.488	16.846
Auditory	6	10.000	3.578	1.600	4.113	5.887	14.113

********** Tukey-Kramer Post Hoc Tests **********

Treatment/ Group	Treatment/ Group	Mean Difference	Critical Q	95PC CI Lower Bound	95PC CI Upper Bound
Sim	Succ	3.833	4.164	-0.331	7.997
Sim	Auditory	6.000	4.164	1.836	10.164
Succ	Auditory	2.167	4.164	-1.997	6.331

Now, about the interpretation of sphericity…

Sphericity Report

Advanced Repeated Measures Concerns: Sphericity

Mauchly's W	Chi-Square	df	P-value
0.628	1.858	2.000	0.3949

Source of Variation	Sum of Squares	dfNum	dfDen	F	P-value
Unadjusted	110.778	2.000	10.000	8.002	0.0084
Huynh_Feldt	110.778	1.905	9.527	8.002	0.0096
Greenhouse_Geisser	110.778	1.458	7.291	8.002	0.0185
Lower Bound	110.778	1.000	5.000	8.002	0.0367

OK, it may not be THAT bad, but it is close. Should ye choose to Enter Here, start here:

https://statistics.laerd.com/statistical-guides/sphericity-statistical-guide.php

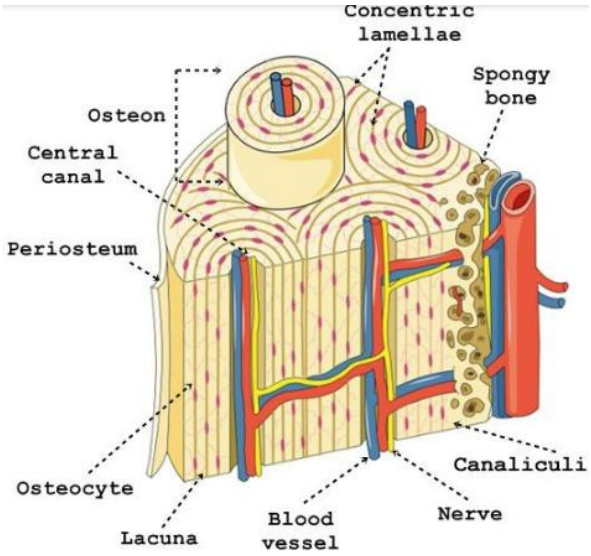
Now, I'm bailing out and heading for the Analysis of Covariance…

Analysis of Covariance (ANCOVA)

Forensic anthropology

In the BONES television series (and the murder mysteries of Kathy Reichs, on which the series is based), one Temperance "Bones" Brennan is a forensic anthropologist / archaeologist who analyses skeletal remains. Biological creatures known as osteons are used to establish the ages at death of the unfortunate and ancient to uncover their mortal identities. One of the measures used is the "osteons per grid," a measure of the prevalence of these in the bones.

An investigation [Botha, D., Lynnerup, N., & Steyn, M. (2020). Inter-population variation of histomorphometric variable used in the estimation of age-at-death. *International Journal of Legal Medicine*, 134:709-719] of the use osteons to establish age in different populations (Danes, South Africans) is the source of the data in this example. Skeletal remains of three groups (Danes, South African Whites, South African Blacks) are the sources of the data. In ANCOVA there are usually two questions to address: are the relationships (osteon density and age in this case) the same among the groups, and if not is the difference only in the intercepts.



The usual ANCOVA is an extension of ANOVA, so means of groups are typically presented. For these data, the issue boils down to the question of equality of ages across groups for each osteon density. Many and varied are the ways that statistical software presents the results of an analysis of covariance! SPLAT provides a bare bones report, just enough to answer your basic question: "Is there sufficient evidence of a difference among the groups?" Estimates of the means will be printed, and the post-hoc Tukey tests will be applied to the results. The bare bones basic ANCOVA results are presented in two tables. The first F-test is about differences in treatment means, the second F-test is about differences in the slopes between each of the treatments. (Recall that homogeneity of slopes is an important assumption in ANCOVA.)

To perform the analysis, click on BAPS → ANOVA → Analysis of Covariance. The treatments are the groups, AVE_OPD – average number of osteons per grid area -- is the Covariate, and Age is the Response. The data are in the file, CSV_OPD_SA.

First, SPLAT provides all that statistical stuff…

Analysis of Covariance

********** Parameter estimates for Levels **********

Treatment/ Group	Sample Size	Sample Mean	Sample St Dev	Std Err of mean	Margin of Error	Lower 95PC Bound	Upper 95PC Bound
Danish	31	44.077	21.238	3.878	7.642	36.435	51.720
SA_Black	98	51.648	16.504	1.676	3.303	48.345	54.951
SA_White	94	67.013	15.612	1.619	3.191	63.822	70.203

Analysis of Covariance

Source of Variation	Sum of Squares	df	Mean Square	F	P-value
Treatments	16991.382	2	8495.691	42.160	0.0000
Error	44130.968	219	201.511		
Total	61122.350	221			

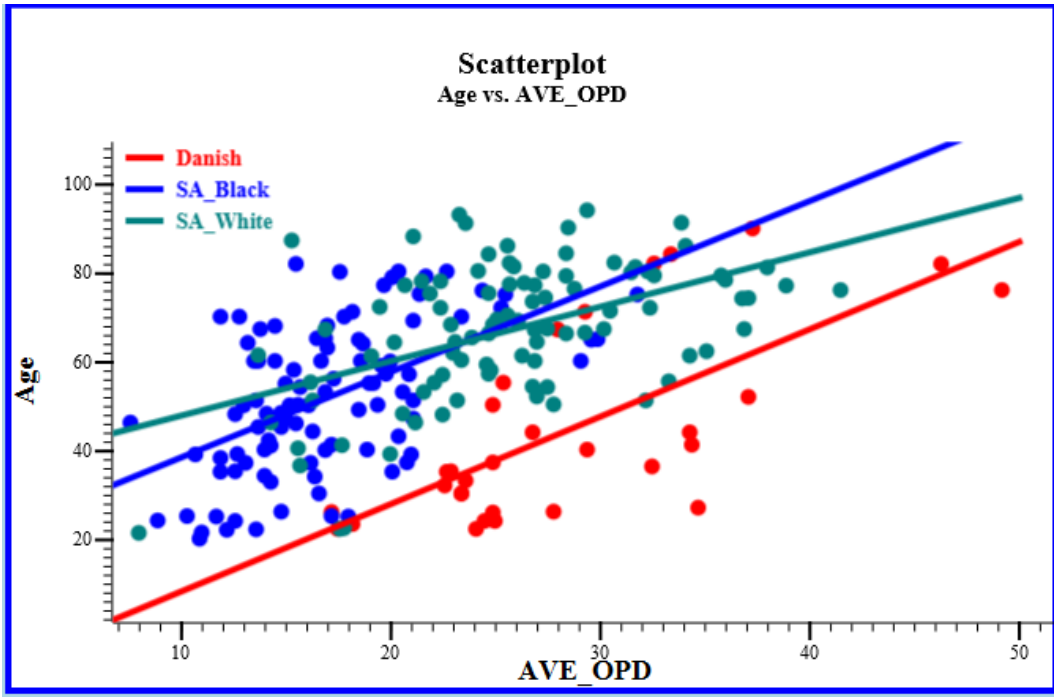
Homogeneity of Slopes

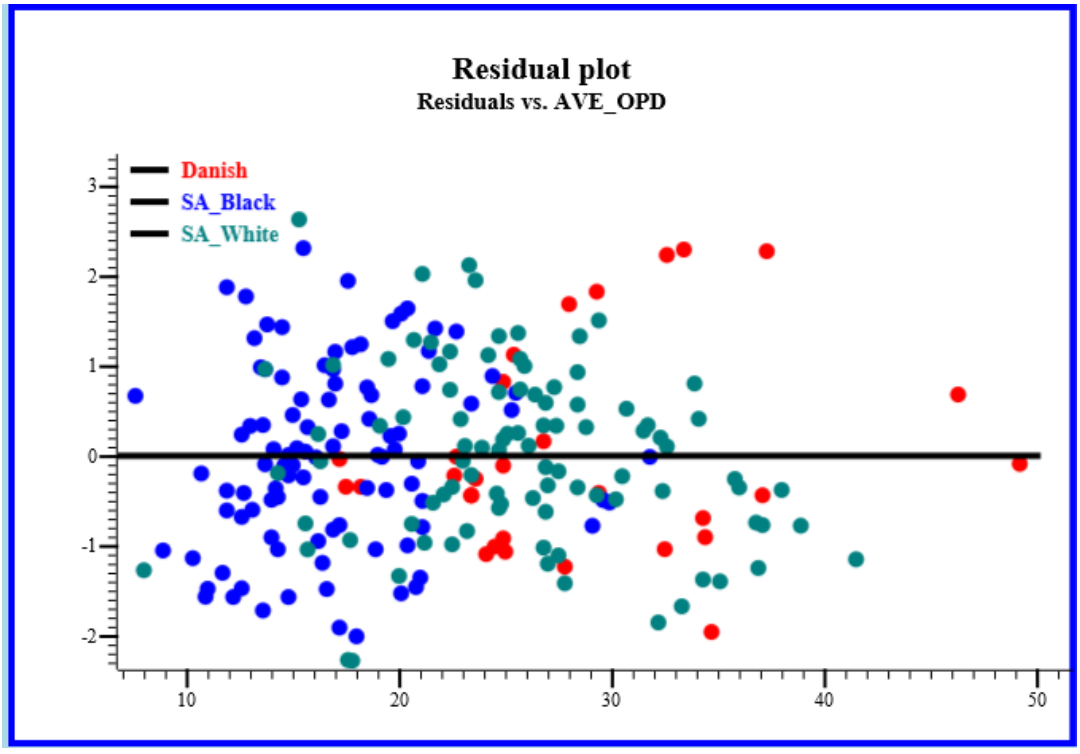
Source of Variation	Sum of Squares	df	Mean Square	F	P-value
Heterogeneity	953.920	2	476.960	2.397	0.0934
Residuals	43177.048	217	198.973		
Within Resids	44130.968	219			

Tukey-Kramer Tests (adjusted means)

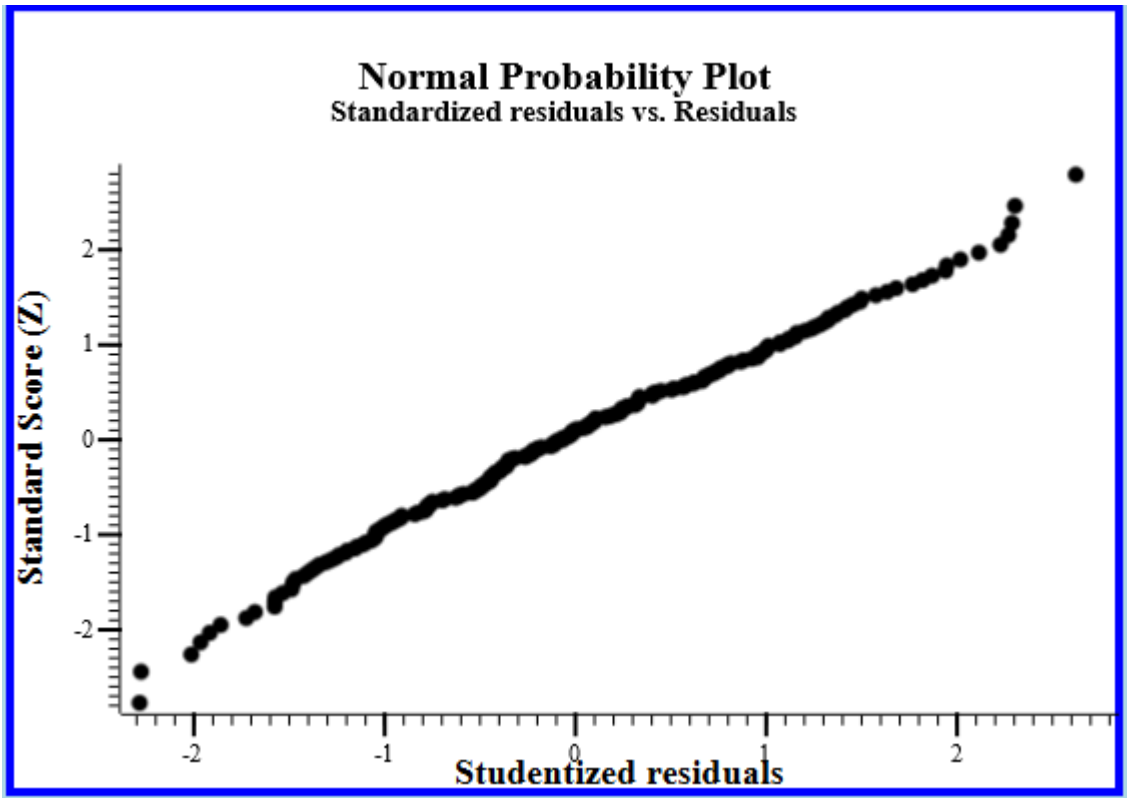
Treatment/ Group	Treatment/ Group	Mean Difference	+/-	95PC CI Lower Bound	95PC CI Upper Bound
Danish	SA_Black	-25.591	8.207	-33.798	-17.384
Danish	SA_White	-26.988	7.009	-33.998	-19.979
SA_Black	SA_White	-1.397	5.935	-7.333	4.538

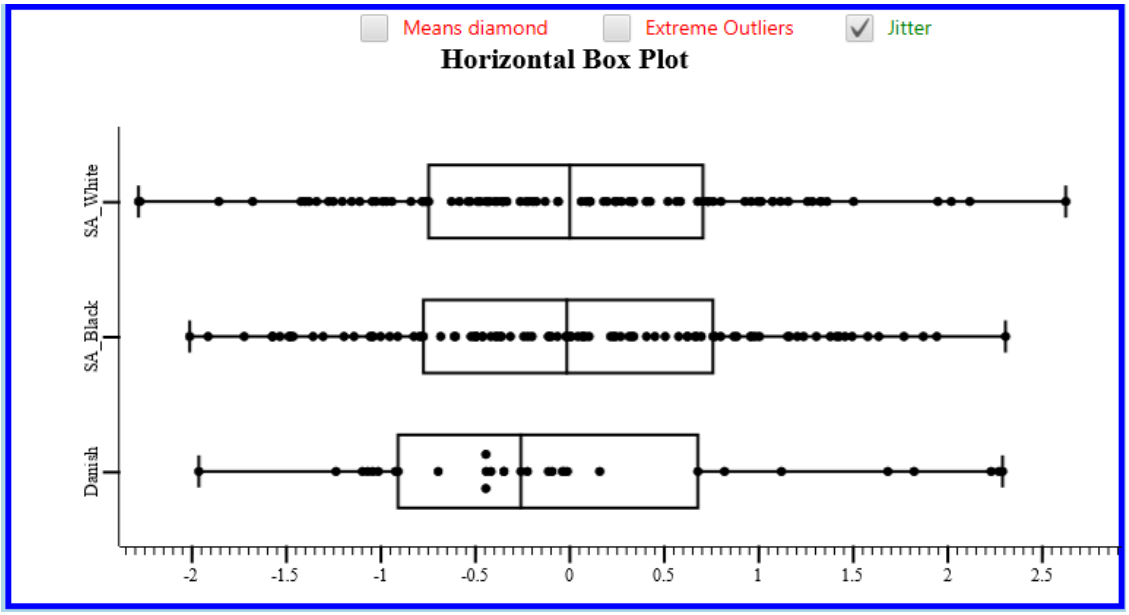
SPLAT also provides the between-treatment regression lines and a residual plot to assess the ANCOVA results:





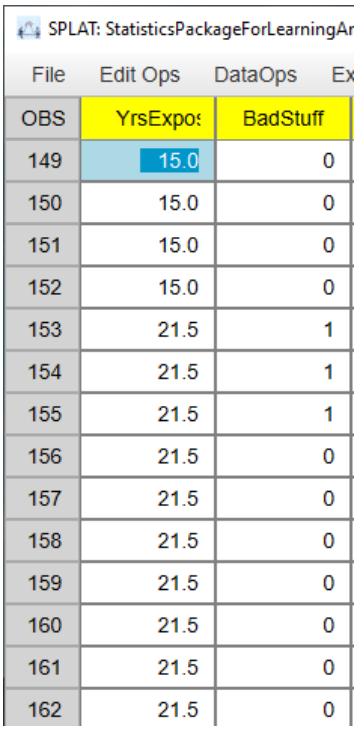
Even more also, graphs are provided for purposes of checking the assumptions of normal and homogeneous residuals:





Simple (Uh-huh, right!) Logistic Regression

For those who might wish to explore another topic after the AP Exam, Logistic regression might fit the bill. Logistic regression is not for the mathematically faint of heart, and I will not try to convince you otherwise. Not only that, since I am a card-carrying member of that faint of heart club, I will only describe the SPLAT output for logistic regression, and not spread around my ignorance of how it is done.



OBS	YrsExpos	BadStuff
149	15.0	0
150	15.0	0
151	15.0	0
152	15.0	0
153	21.5	1
154	21.5	1
155	21.5	1
156	21.5	0
157	21.5	0
158	21.5	0
159	21.5	0
160	21.5	0
161	21.5	0
162	21.5	0

The data for logistic regression consists of a numerical explanatory variable and a binary response variable. The response variable is usually 0 / 1, but SPLAT will accommodate different text values for the response, and assign a 0 / 1 for purposes of calculation. In some textbooks the data are presented in a "grouped" format, but that will not work with SPLAT. The reason I chose to ignore the "grouped" format is that there is a lack of unanimity in the literature about how to "group" data for logistic regression; and I am NOT about to tread where angels fear!!! On the other hand, everyone seems to be happy with – take a deep breath -- "maximum likelihood estimation of the parameters using the Newton-Raphson method with iteratively-reweighted-least-squares." (I would be ecstatic if I had a clue what any of that means!) For that whatever-that-is method to work, raw data are needed.

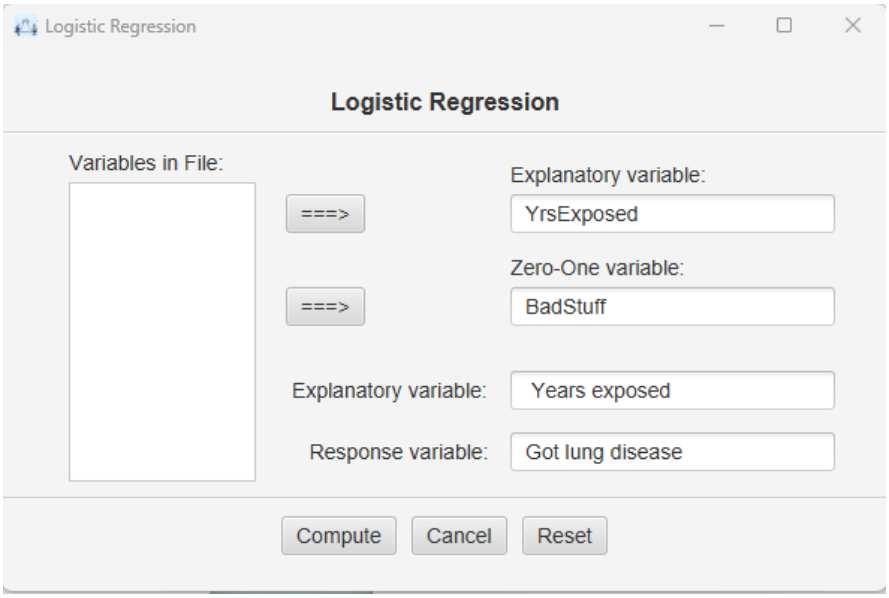
The data file I will use as an example here is CSV_Pneumo.

Clicking on BAPS→Advanced Regression → Logistic regression transports you to the familiar spreadsheet. For these data, YrsExposed is the quantitative explanatory variable; BadStuff is Yes/No, Yes = got lung disease.

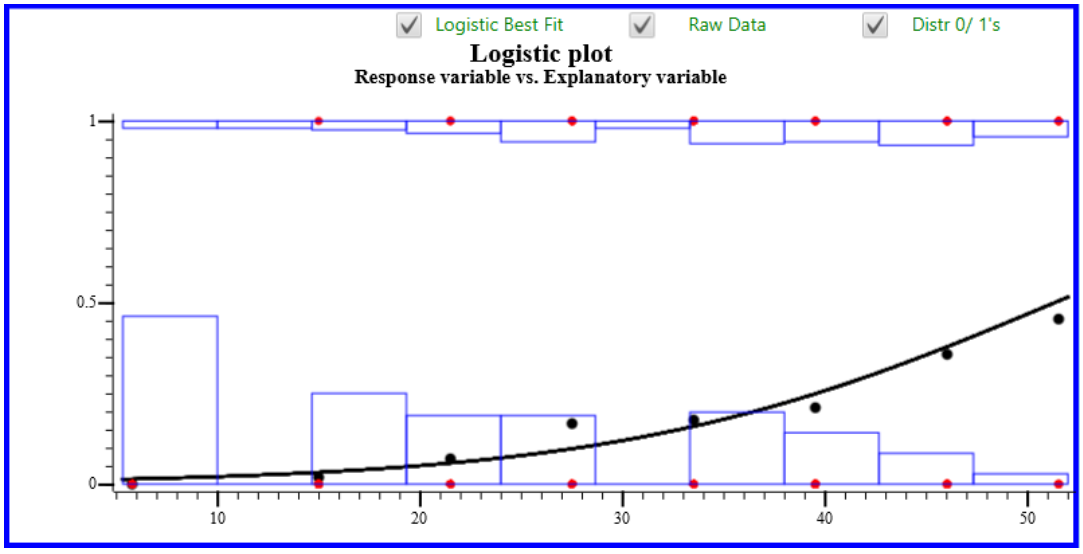
Note: The data used here are hypothetical, and are taken from Table 1.2 in Hosmer, D. W., et al. (2013). Applied Logistic Regression (3rd). John Wiley & Sons. Hoboken, NJ. In "real life" the list of observation values in this panel could be really long!

Another note: If you are running with text values (not 0's and 1's) SPLAT will inform you which value is associated with 0 and which is associated with 1.

Here are your choices:



After making your choices, things can get a bit hairy. I was unable to locate any "standard" visual presentation for logistic regression, so I just took what seemed to be good ideas from various places in the literature. Most graphic presentations of logistic regression data include a scatterplot and the best fit line. Some have relative frequency histograms as shown below.



The red dots are the values of the explanatory variable (the continuous one). Those blue rectangles indicate the relative frequencies of the 0's and 1's (rescaled so that 0. 0 to 0.5 and 1.0 to 0.5 are now proxies for 0.0 to 1.0) in each band of the explanatory variable. Finally, the black dots are the observed proportions of 1's.

After the Logistic plot things get easier. The interpretations of the normal probability plot and the (deviance) residual plot work like usual regression.

Logisic Regression Analysis

********** Logistic Regression Equation **********

$$P(\text{Success}) = \frac{e^{(-4.79648 + 0.09346 \text{ Years exposed})}}{1 + e^{(-4.79648 + 0.09346 \text{ Years exposed})}}$$

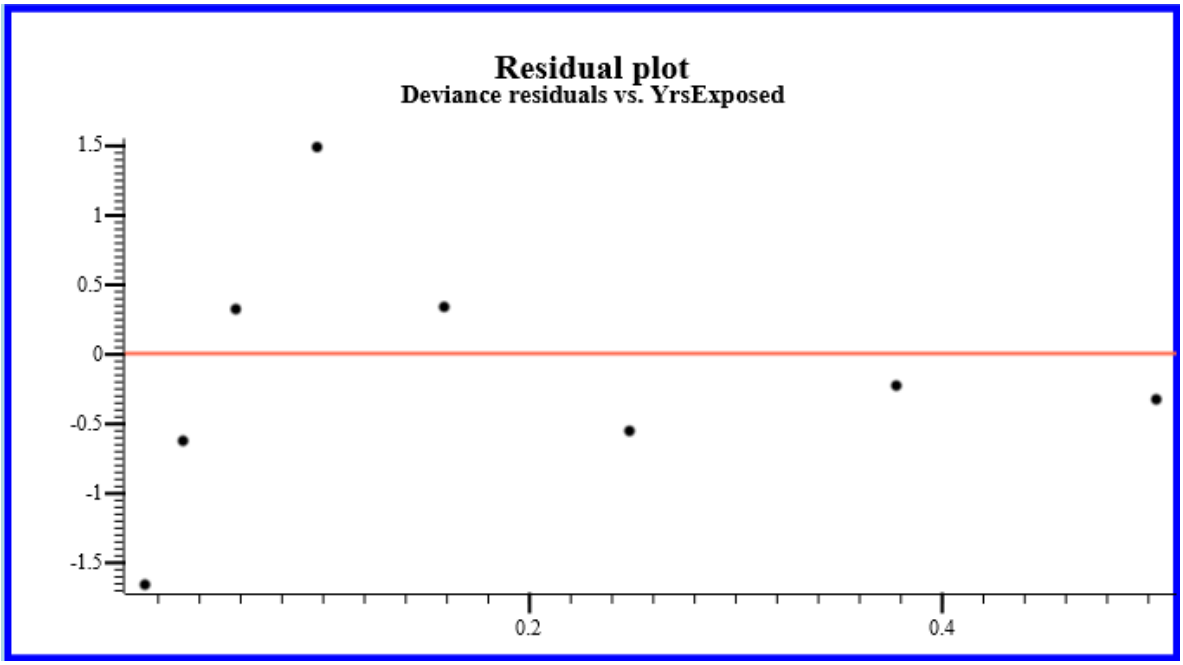
********** Logistic Regression Table **********

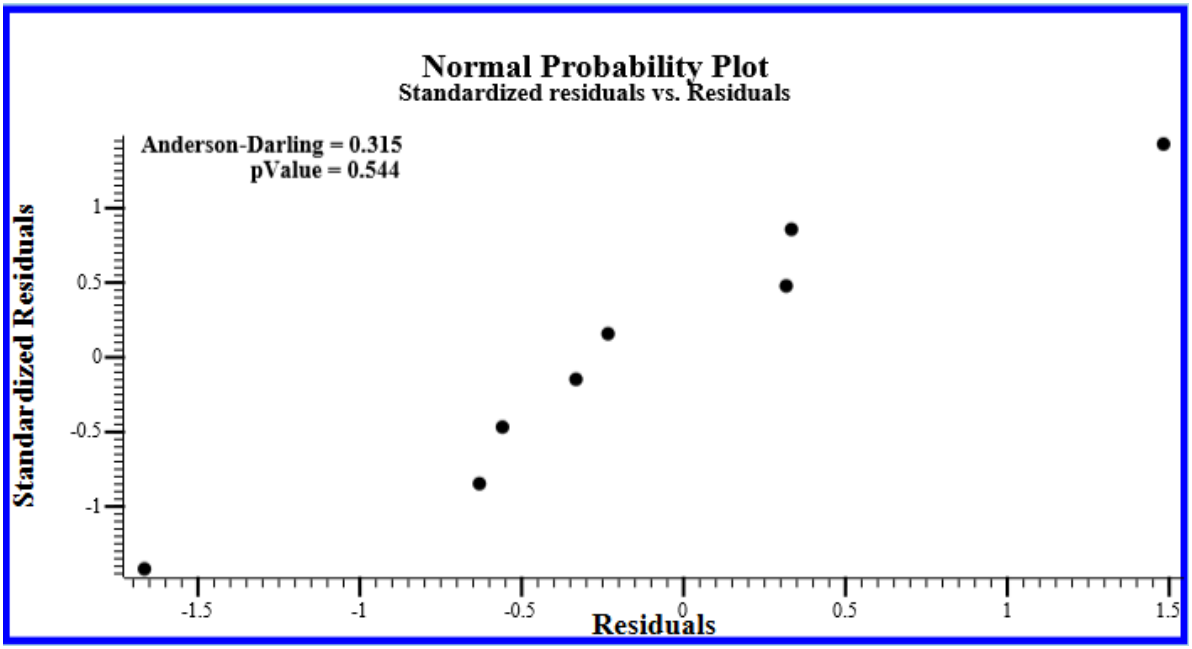
Predictor	Coefficient	StandErr	Z	P-value	Odds Ratio	*** 95% CI *** Lower	Upper
Constant	-4.796	0.569	-8.436				
Years e	0.093	0.015	6.059	0.000	1.10	1.07	1.13

********** Goodness-of-Fit Tests **********

Method	Chi-Square	df	p-value
Pearson	5.029	6	0.540
Deviance	6.051	6	0.418

Log-Likelihood = -109.664
Test for all parameters zero: G = 50.852, df = 6, P-value = 0.000





Logistic Diagnostics

Logistic Regression Diagnostics

Observation	Observed Probability	Estimated Probability	Deviance Residual	Pearson Residual	Stand Pearson Residual
5.800	0.000	0.014	-1.663	-1.180	-1.428
15.000	0.019	0.032	-0.628	-0.578	-0.652
21.500	0.070	0.058	0.320	0.329	0.362
27.500	0.167	0.097	1.485	1.618	1.793
33.500	0.176	0.159	0.336	0.341	0.384
39.500	0.211	0.249	-0.557	-0.547	-0.631
46.000	0.357	0.378	-0.231	-0.230	-0.294
51.500	0.455	0.504	-0.330	-0.329	-0.383

If you MUST take a plunge into logistic regression, be sure to get a good stiff glass of chocolate milk and a good book. I recommend Cannon, A., et al. (2019). STAT2: Modeling with Regression and ANOVA (2nd ed) if you are sufficiently faint of heart. If you want the full mathy I recommend – yet again -- Montgomery, Peck, and Vining (2012). Introduction to Linear Regression Analysis.

Statistics in the Age of Covid-19: The Analysis of Risk

Entering data in a table

The example of logistic regression above hypothesized a quantitative explanatory variable (years working in a coal mine) and a binary response variable (something bad happening). In a context such as this, the logistic regression is an analysis of risk; the estimated probabilities are interpreted as a "risk" of "bad stuff" for a randomly selected person spending that many years working in a coal mine.

Risk analysis in epidemiology usually involves a categorical explanatory variable, complete with a plethora of statistics that are used in observational studies. The different statistics are interpreted differently, and sometimes inappropriately, when mounting an observational study.

In the AP Statistics Course and Exam Description (CED), observational studies are not treated with much respect, which is unfortunate; the lion's share of studies performed in "real life" are observational, not experimental. The CED also makes the utterly false statement that it is not possible to determine causal relationships in an observational study. If that were true, we still would not "know" that smoking causes cancer. (It IS certainly true that establishing a causal relation between variables is more difficult (and expensive) without random assignment to treatments.) If interested in this issue, consult Pearl, J. (2016). Causal Inference in Statistics – A Primer. John Wiley & Sons.

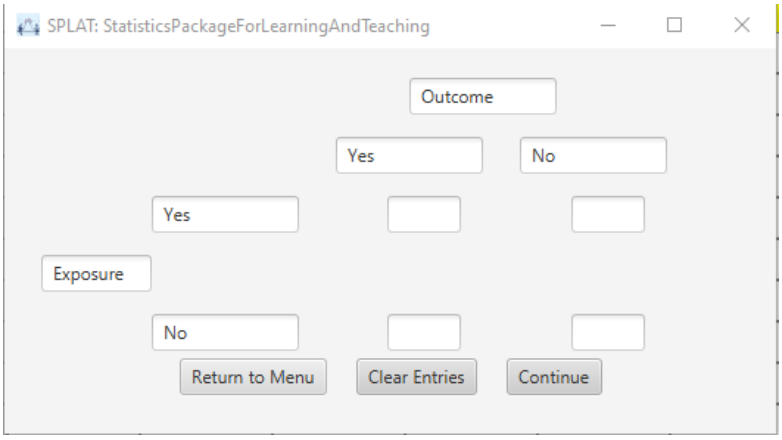
The epidemiological search for causal links between a categorical explanatory variable, commonly referred to as "exposure", and a binary response, commonly referred to as "outcome", is another example of risk analysis. The exposure variable can have many values, but SPLAT only considers binary exposure variables; the visual and statistical presentation of binary results is a great deal easier!

As was the case with logistic regression, I will not attempt to disperse my ignorance. For further study I recommend Webb, P., et. al. (2020). Essential Epidemiology: An Introduction for Students and Health Professionals. Cambridge University Press: New York.

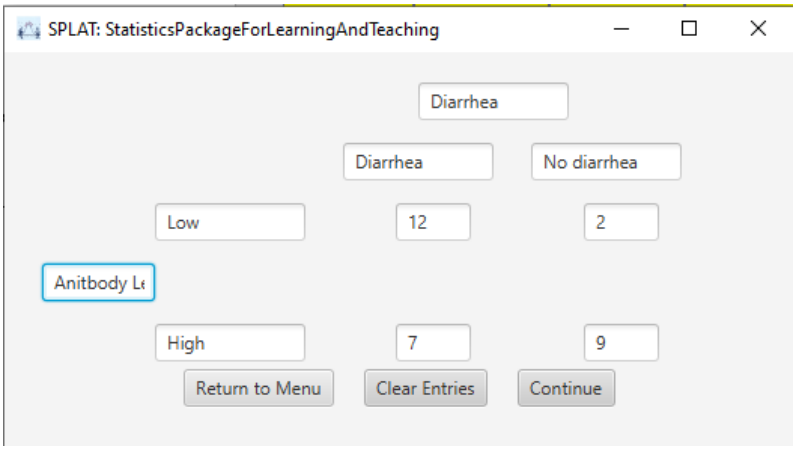
The data I will use as an example of SPLAT's output is from Glass, R., et al. Protection against cholera in breast-fed children by antibiotics in breast milk. *New England Journal of Medicine*, 1983;308;1389-1392. The outcome measure was the occurrence of diarrhea in children in a 10-day follow-up. The children had been exposed to *Vibrio cholera 01*, as determined by Antipolysacchararide antibody titers in the mother's breast milk. A low titer confers elevated risk. (Once again, I have no clue what any of this means – I am just reporting what I read in the article.)

The data in a binary bivariate risk assessment is typically laid out in a 2 x 2 table. To see how this works in SPLAT, do this:

Explore data → Bivariate data → Epidemiology: 2 x 2 categories → Enter 2 x 2 in table.



After writing the information about your variables and changing the blanks to values (you can tab through them forward and backward), here is the data from the study:

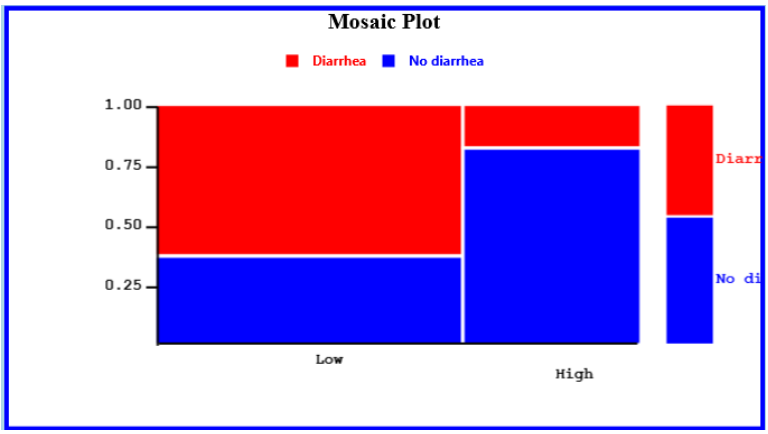


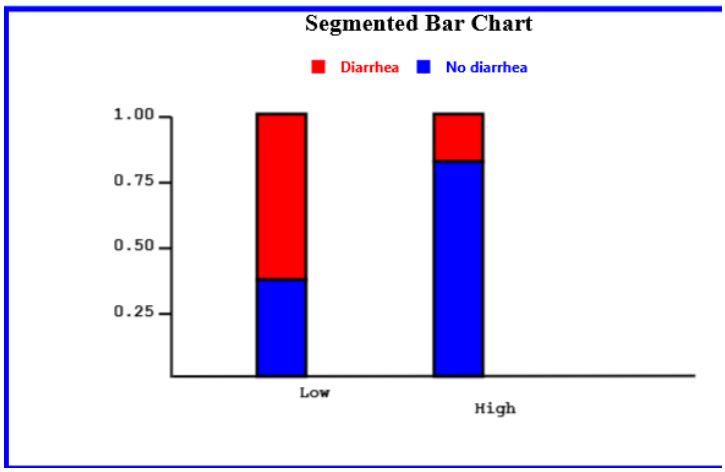
Upon clicking "Continue" some of the usual suspects you saw with Chi square will appear…

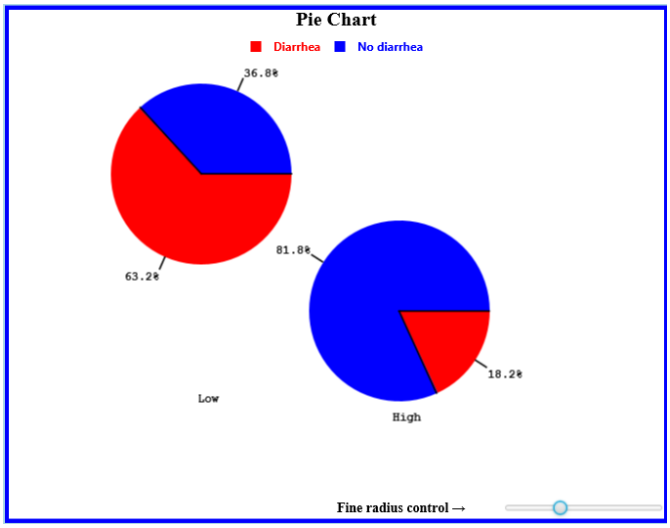
Contingency Table

Diarrhea			
Observed values	12.00	2.00	14.00
Percent of Total	40.00	6.67	46.67
Percent of Row	85.71	14.29	
Percent of Column	63.16	18.18	
No diarrhea			
Observed values	7.00	9.00	16.00
Percent of Total	23.33	30.00	53.33
Percent of Row	43.75	56.25	
Percent of Column	36.84	81.82	
Total	19.00	11.00	30.00
Percent	63.33	36.67	

Cramer's V = 0.434







In addition to the visual presentations above, there are statistics that are presented in a typical epidemiological study. These statistics are presented below without my ignorant comment. The interpretation and appropriateness of the statistics will vary over the "types" of observational studies: Cohort (a.k.a. "Prospective"), Case-control (a.k.a. "Retrospective") and Cross-sectional (a.k.a. "Survey"). Be advised that several definitions of "Prospective" and "Retrospective" have been used in the literature, and it is not necessarily easy to distinguish the two in practice. Here are some good definitions:

A prospective study is one where the units of study are selected at a point in time and data are gathered at that time and into the future.

A retrospective study is one in which the units of study are selected at a point in time and data from the past are gathered.

Notice that the "risk factor" and "subsequent disease" are displayed in what math folks might think of as the axes reversed. The (categorical) "y axis" is the explanatory variable.

********** **Epidemiology** **********

Risk Analysis

Anitbody Level	Diarrhea		Total
	Diarrhea	No diarrhea	
Low	**12**	**2**	**14**
High	**7**	**9**	**16**
Total	**19**	**11**	**30**

	95% Confidence intervals		
Parameter	Lower bound	Estimate	Upper Bound
Risk Difference	0.11521	0.41964	0.72408
Risk Ratio	1.08026	1.95918	3.55322
Log Risk Ratio	0.07720	0.67253	1.26785
Odds Ratio	1.28357	7.71429	46.36311
Log Odds Ratio	0.24964	2.04307	3.83650

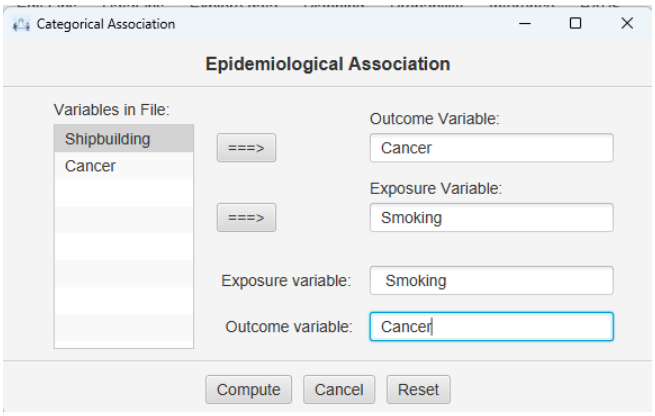
Also notice that there are confidence intervals for the parameters. I debated with myself about whether to put Risk Analysis in the Inference part of SPLAT and decided that the calculations in the Estimates column above would be informative to explorers of data.

Entering data in a file

If your risk data is in a file, the analysis is a bit more complicated, by design. It occured to me that while you may be interested in data of the 2 x 2 categorical persuasion, a variable may actually contain more than two values. Filtering out the two desired values takes a lot of file manipulation, which can be error-prone. To get around that, SPLAT will ask you to choose the values of interest. So, for example, you might be interested in the relationship between minimal and heavy smoking and cancer among shipbuilders. (In Iowa, we talk about nothing else!)

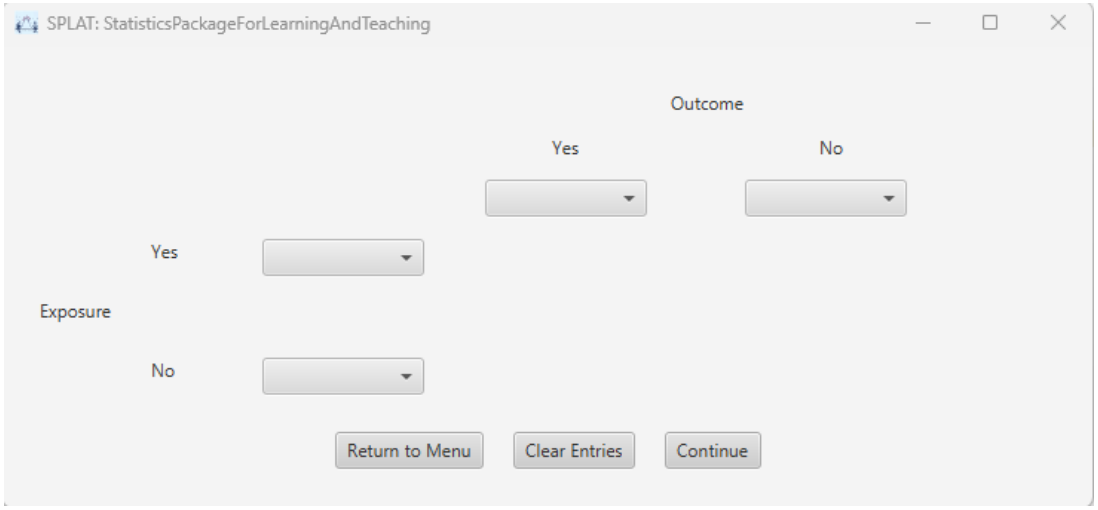
The data used here as an example is from Schlesselman, J. J. (1982). Case-control studies, design, conduct, analysis. Oxford University Press, New York. To make a long story short, J. J. gathered data on shipbuilding workers' smoking history and health, and compared their experience with a comparable group that did not work in the shipbuilding industry. The study's subjects were classified as Minimal, Moderate, or Heavy smokers, and whether they had cancer. The data are in the file, CSV_Schlesselman. The problem (for us) is that we are only interested in the Heavy vs. Minimal smoker. The solution? -- SPLAT will take care of that!

Click the sequence, Explore data → Bivariate data → Epidemiology: 2 x 2 categories
→ The data are in the current file.



Since the values for these variables are categorical, SPLAT will insist that you review the values for data entry errors. Just say "OK" twice (this time) and you should see..

.



For the exposure variable, choose "Heavy" in the "Yes" dropdown box, and "Minimal" in "No." (Recall that we don't care about "Medium" in this example – it was just a nuisance value in the data file.) For the outcome, choose "Yes" for "Yes" and "No" for "No" and click on Continue. SPLAT will process your choices and you will see results just like you did with the Table. FYI, from the SPLAT output it appears that heavy smoking is a serious risk for cancer – but you probably knew that already.

********** **Epidemiology** **********

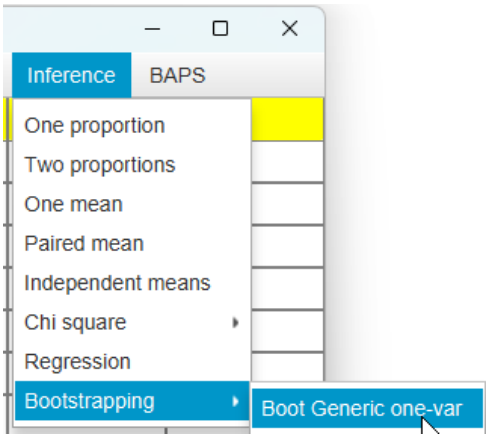
Risk Analysis

Exposure	Exposure		Total
Exposure	Yes	No	
Heavy	**110**	**53**	**163**
Minimal	**61**	**238**	**299**
Total	**171**	**291**	**462**

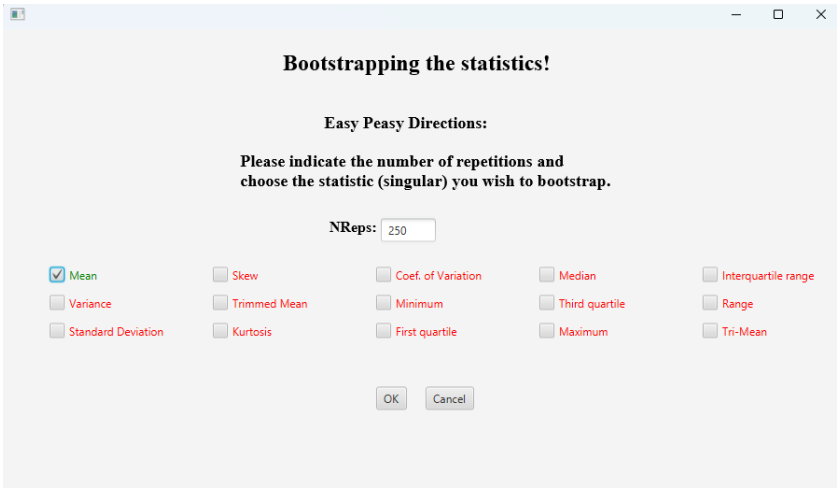
Parameter	95% Confidence intervals Lower bound	Estimate	Upper Bound
Risk Difference	0.38564	0.47083	0.55603
Risk Ratio	2.58143	3.30785	4.23870
Log Risk Ratio	0.94834	1.19630	1.44426
Odds Ratio	5.25778	8.09774	12.47171
Log Odds Ratio	1.65971	2.09159	2.52346

Bootstrapping Univariate Statistics

The bootstrapping in SPLAT provides a mechanism to explore the sampling distributions of various univariate statistics. Bootstrapping is kicked into motion for a particular data set by choosing Bootstrapping in the Inference dropdown menu. Also, one can construct hypothesis tests and confidence intervals with bootstrapping.



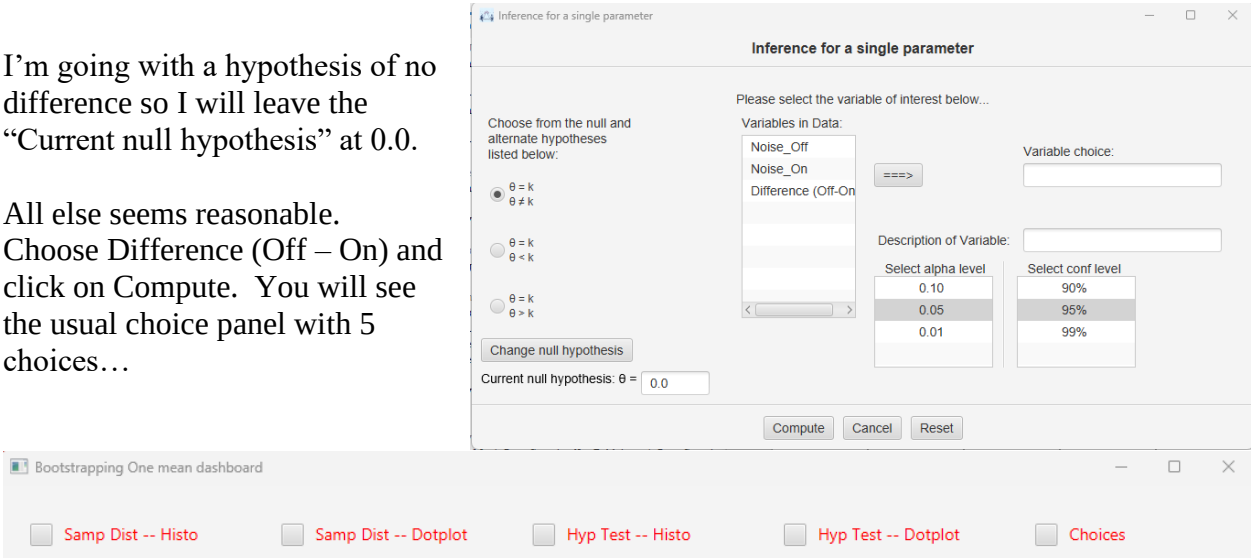
To illustrate SPLAT's bootstrapping approach I will use a modified version of the file previously seen when we discussed the paired-t: CSV_Bootstrap_Chaffinch. (The only difference is that I have already subtracted Off – On to get a difference. Open the file and navigate to "Bootstrapping." The choice panel looks like this:



You can choose any number of repetitions and any of the statistics, but I'm going with the Mean.

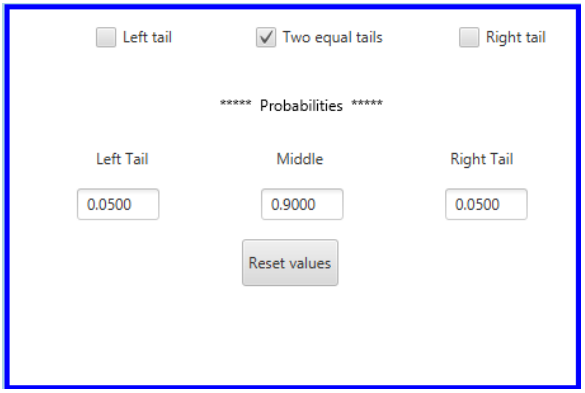
I'm going with a hypothesis of no difference so I will leave the "Current null hypothesis" at 0.0.

All else seems reasonable. Choose Difference (Off – On) and click on Compute. You will see the usual choice panel with 5 choices…

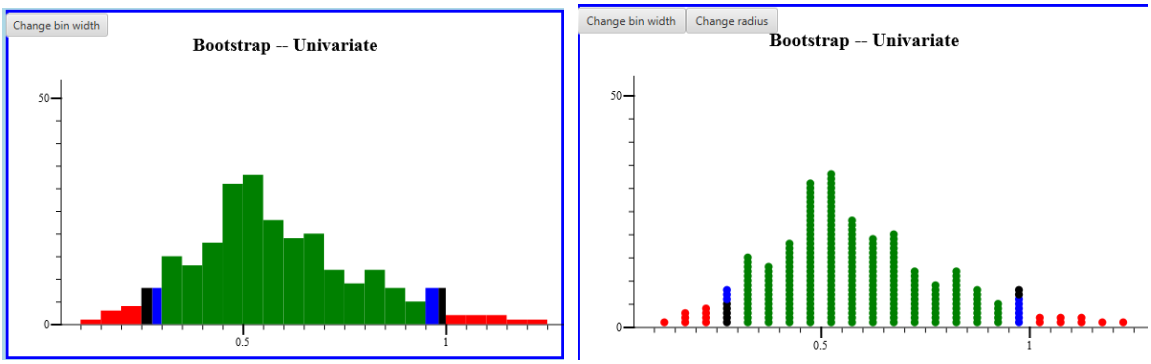


The "Samp Dist" choices present the results of a simulation of taking in this case 250 samples and calculating the sample means. Thus, these are estimates of what the sampling distribution looks like. From these you can get "bootstrapped" confidence intervals. The "Hyp Test" choices are constructed by centering the sampling distributions at the hypothesized mean.

With the "Choices" option the resulting dot plots and histograms can be colored to indicate the middle X% of the distributions. The black and blue parts indicate what bin the transition between in and out of the middle X% occurs and where it occurs in the bin. In these data the transition must have occurred very near a bin limit as it is almost all blue at the left bin that marks the transition point. Generally, the histograms would be used for large samples and dotplots for small samples. On the choices panel, make these choices:

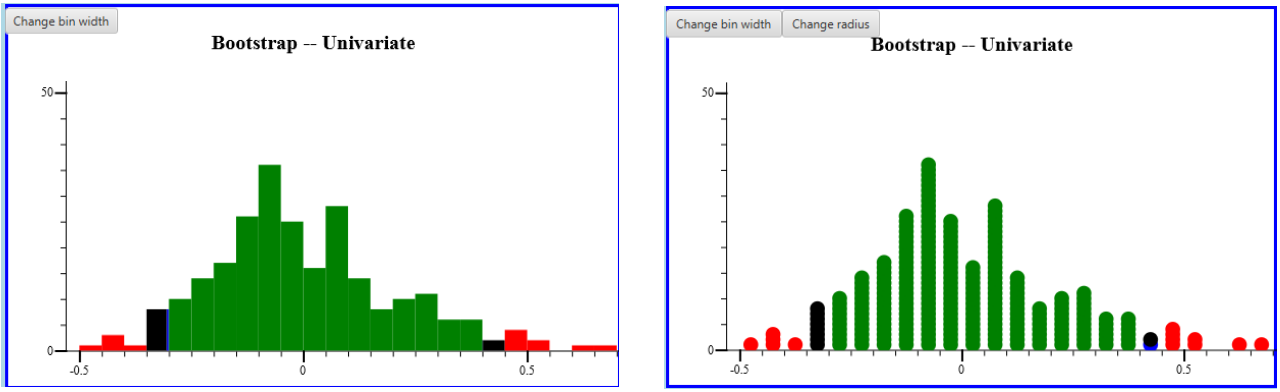


I made some of the usual adjustments to the scales and changed the bin widths using 0.50 and 0.55. Here are the estimated sampling distributions…



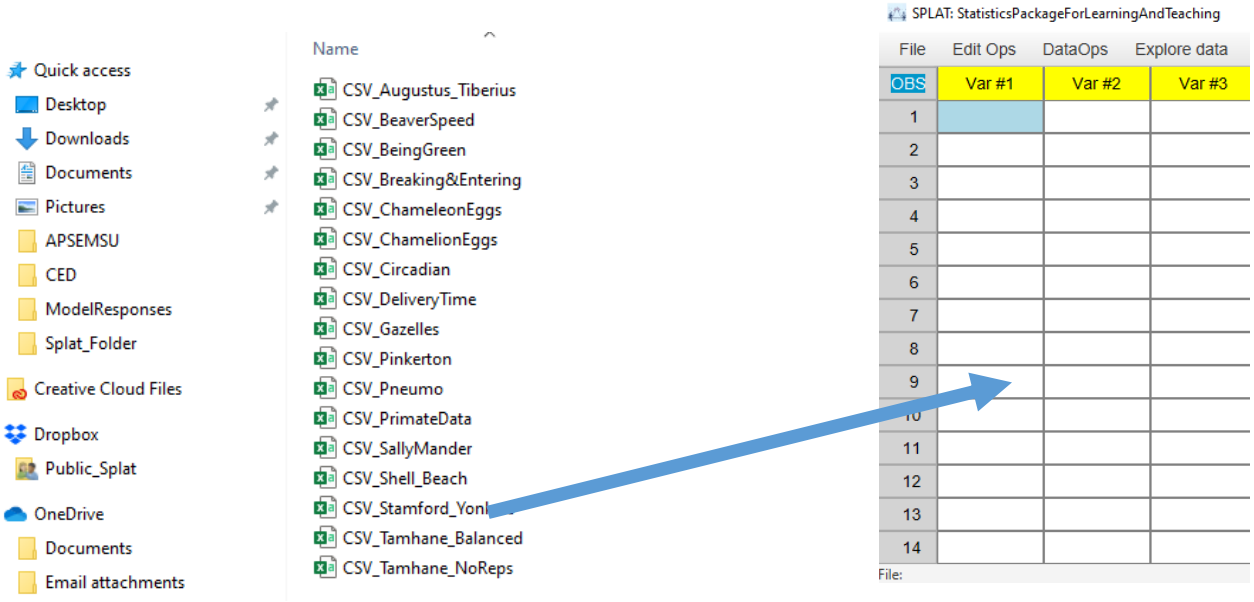
The “bootstrap” confidence intervals can be estimated by locating where the transitions from red to green and green to red occur. The interval where the transition takes place is shown in black and blue to give a sense of where in that interval the transition is made. The language attached to bootstrapping seems unsettled, but the most frequent description of these is “bootstrapped confidence interval.”

In a similar manner the hypothesis can be rejected (or not!) from inspection of the “translated” sampling distributions.

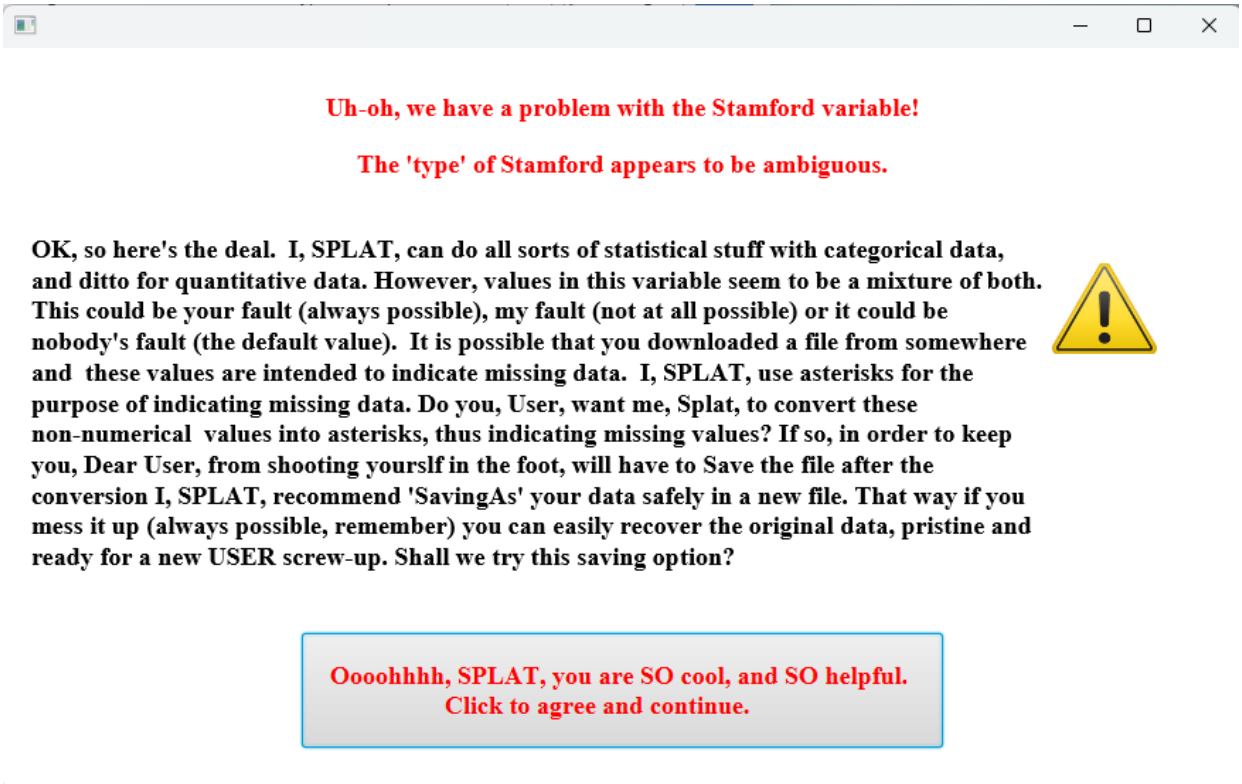


PostScript 1: Missing data

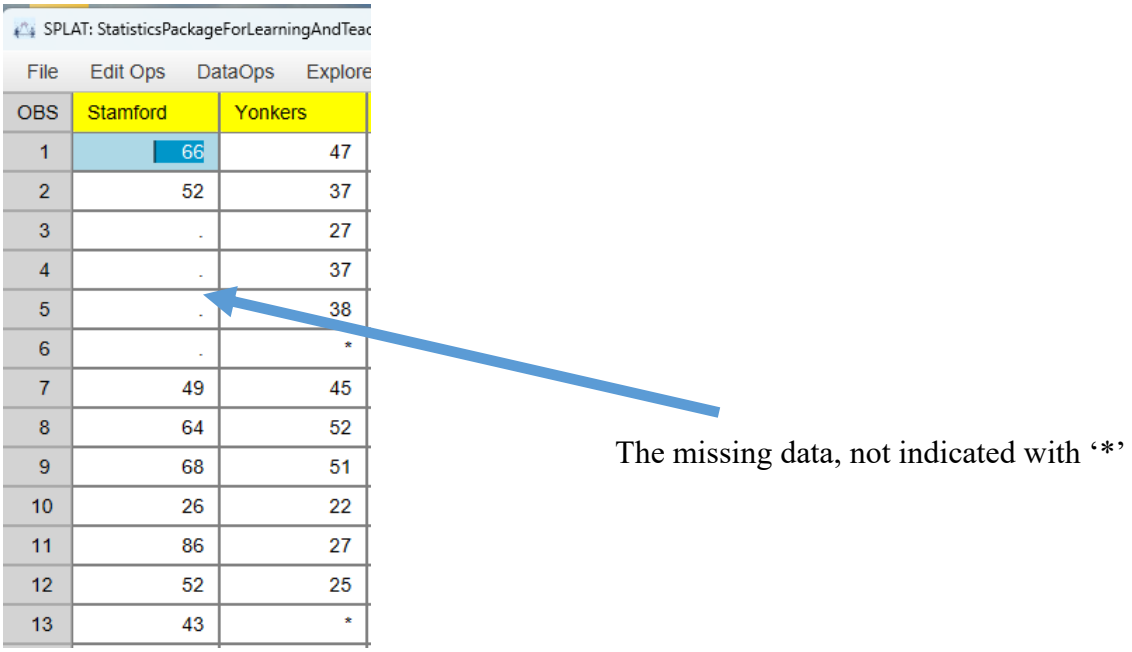
On occasion you might download data from afar (i.e. the web, not the Afar Depression in Ethiopia). As it happens when working with data, different symbols might have been used to indicate missing data. For example, when you downloaded the "CSV_Stamford_Yonkers" file and opened it with SPLAT.



You were be greeted with the following "error" message.



This message is alerting you about a problem with the data in the file. These data are (as I recall) some sorts of weather measures gathered through time in the towns of Stamford and Yonkers. On some days, data were not gathered in one of the towns, and on these days a "." was entered in the file. SPLAT uses an asterisk ("*") to indicate missing data when it occurs in a variable that is allegedly quantitative. SPLAT is offering to change those symbols to SPLAT-compliant asterisks. In general, my counsel would be to make sure you have a backup of the downloaded file just in case, and then tell SPLAT to go ahead and change the symbols. Note that if you have missing non-asterisk data in more than one variable you will see the error message more than once.



OBS	Stamford	Yonkers
1	66	47
2	52	37
3	.	27
4	.	37
5	.	38
6	.	*
7	49	45
8	64	52
9	68	51
10	26	22
11	86	27
12	52	25
13	43	*

PostScript 2: “Bad” files

Only CSV formatted files will work in SPLAT. If you try anything else SPLAT will not be pleased and may even blow up. If SPLAT catches you trying to read a non-CSV file but does not actually swoon and expire on the vine, you will get a gentle and helpful message, laced with customary SPLAT modesty.

A non CSV file has reared its ugly head!!

Scuse me!?!? Do you think I, SPLAT, have no standards!?!?

Not to put too fine a point on it, I am NOT a one trick pony. I, SPLAT, am a one trick STALLION! Just think of me as the statistics program moral equivalent of Bucephalus though sadly, without a programmer of the quality of Alexander the Great. But I digress. The key thing here is that I do not read just any old files; I have very high standards. I read only Comma Separated Value (CSV) files. If you can't handle that, find a less elegant and sophisticated statistics program to do your data analysis.

Yours truly,
SPLAT the Great

Oooohhhh, SPLAT, you are SO cool, and SO helpful.
Click to agree and continue.

In real life if you have downloaded a file from somewhere and you are not familiar with the contents, tell SPLAT not to make the changes. Then, go through the file and see what you are confronted with. If at that point you need to change some values to “missing” (= “*”) you can “Clean the data” one variable at a time – to see how to do this, consult the Postscript at the end of this helpful guide.

PostScript 3: Cleaning Categorical Data with SPLAT…

Cleaning data after entry is not the most fun part of the statistical enterprise, and SPLAT can help a little bit with categorical data. Suppose you have entered some data, and you are checking to make sure the data entry. As you know, data entry may have a somewhat casual relationship with capitalization and spelling.

SPLAT: StatisticsPackageForLearningAn

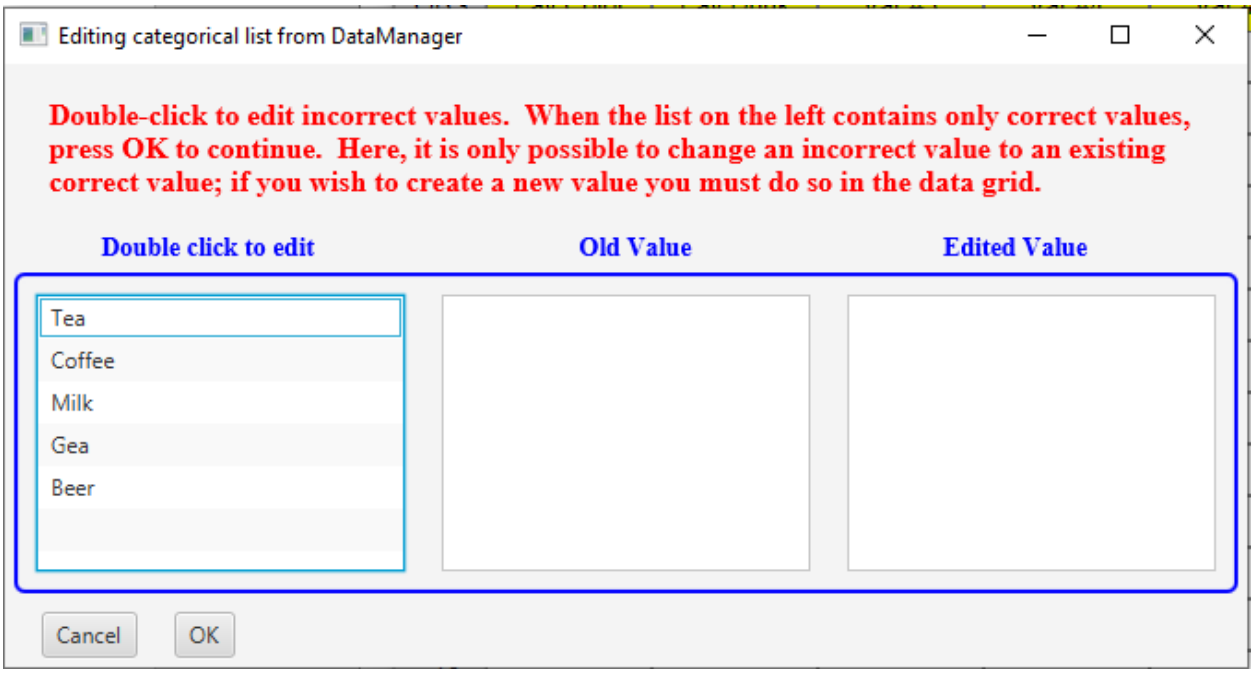
File | Edit Ops | DataOps | Ex

OBS	Fav Color	Fav Drink
1	Green	Tea
2	Blue	Coffee
3	Red	Coffee
4	red	Milk
5	Beige	Tea
6	Bluell	Milk
7	Red	Gea
8	Green	Tea
9	Red	Coffee
10	Blule	Beer
11	Green	Milk

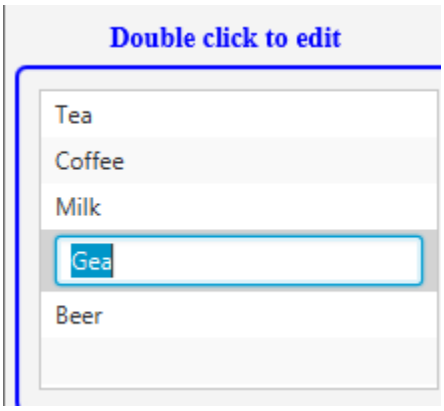
You are shocked, SHOCKED!, to find that your fingers have not been completely faithful to your brain's intentions. It would appear that there is some "creative" spelling and/or capitalization going on, and some variation in both spelling and capitalization. If it is so that "Red" and "red" are the same colors, there exists a bit of a problem; they will be interpreted as different colors by SPLAT.

And what is worse, these are only the first 11 entries out of 500! It would be really nice if all the values of the categorical data could be identified and cleaned in one fell swoop. And, as it happens, SPLAT can do that.

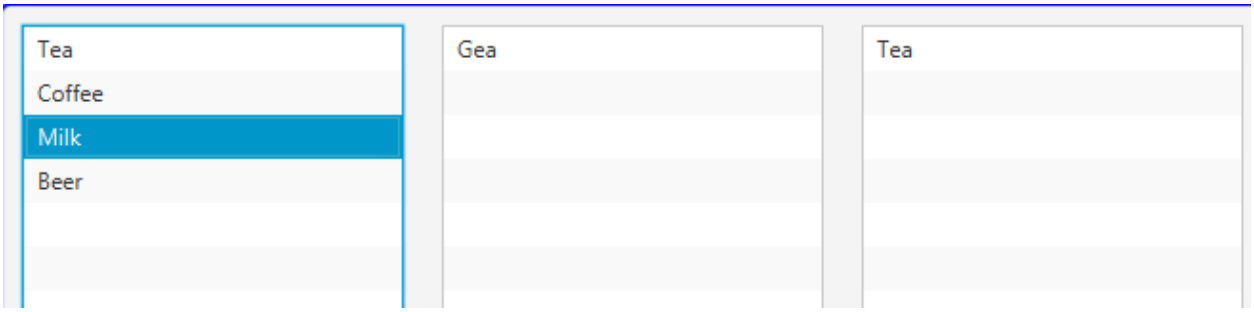
It is possible to initiate the cleaning of a single variable by keying in the sequence, **Edit Ops→Clean Column Data.** For the fav color variable, here is what appears:



What I am trying to say, not very well in this panel, is that in order to correct a bad value, double-click on it. The entry, "Gea," should have been "Tea." OK, double click on the "Gea." Voila!:



Now type in "Tea" in its place and enter.



A few things have happened here. First, bad spelling has been replaced by good spelling. Second, "Gea" has disappeared from the list of values of the categorical variable. And third, you have a list of the corrections you have made. After fixing all the errors and clicking on OK, the replacements indicated will be made for every occurrence of those chosen errors in the data grid.

Note that if the data came from reading in a file, you will need to Save the data to make the changes permanent. (SPLAT will remind you.)

Not only can you initiate a data cleaning of a column, SPLAT will also insist that you check your data if a statistical procedure involves a categorical variable, since spelling is not at the top of the list in many of your student's skill set (but of course IS at the top of the lyst in yore skil sette.)

Ok, end of story! (Or at least end of the SPLAT Guide.)

If you are still reading, you must be a very hardy soul!! I hope this Guide gets you started with SPLAT. Please send any comments, criticisms, and especially bug reports to me at crolsen@fastmail.com. I cannot promise immediate fixes, but certainly soon-fixes. And suggestions for adding procedures are fine also – that, however, may take a little more time…

-- Chris