Univariate & Bivariate Analysis in SAS

Univariate Analysis

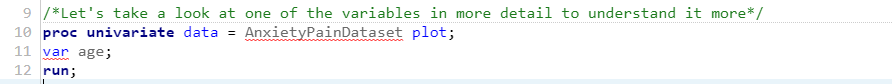
Univariate analysis is essentially just understanding how one variable is distributed. You can find the distribution of a variable, where the outliers are, and how it is split up. To do this in SAS we use the proc univariate command. Refresh the syntax of a proc commands on the SAS basics page if you do not remember.

This is an example of that proc univariate command:

proc univariate data = YOURDATASET plot;  
var VARIABLETOANALYZE SECONDVARTOANALYZE;  
run;

* In the first line, the syntax should be pretty similar to every other proc command except for the word “plot.” Plot allows us to see a histogram of the data
* The second line is where you put what variables you want to analyze all sepearted by a space after the word “var.”
* End with the word “run” and a semicolon

If we are trying to see more information about the age variable for example from our old dataset this is what the code would look like:



And this is what the output would look like. You can see that it has a lot of info about the variable:

Table

Description automatically generated Chart, box and whisker chart

Description automatically generated

Bivariate Analysis

Bivariate analysis is harder that it seems because of all the possible tests that are out there. It is confusing to know what to use when. This is a flowsheet I have always used (I am not sure where I got it so I cannot provide a reference but if someone knows let me know).

Diagram

Description automatically generated

1. I want you to initially ignore the grey squares.
2. The chart first splits at if your response variable is continuous or discrete.
3. If it is continuous, go to the red table above. If there are 2 groups you are comparing, you use a t-test, and if you are comparing 3 groups, you use an ANOVA test. The difference between a normal t-test/ANOVA and a paired t-test is that a paired t-test looks at the same group; that is, if you wanted to compare a group of patient’s scores on a test before and after patient education. If you are comparing 2 independent groups, you do not use a paired test
4. If it is discrete, go to the blue table below and follow the table. Usually you are using chi-square tests here
5. And now to get to the grey boxes. So you only use the grey boxes when the following things are true. If at this point you are considering the grey tests, it may be worth consulting with a statistician
   1. Continuous data is not normally distributed
   2. Discrete data has <5 patients per group
6. And then for correlation – you use this when your explanatory and outcome variables are continuous. You use pearsons correlation the majority of the time.

Now let’s go to the syntax of each of the specific bivariate analyses you are running – t-tests, ANOVA, chi-square, and correlation.

1. T-Tests

Here is a copy of the syntax for a t-test with one variable:

*proc ttest data = YOURDATASET;*

*title "T-Tests";*

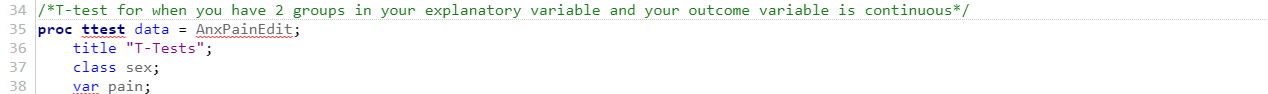
*class EXPLANATORYVARIABLE;*

*var OUTCOMEVARIABLE1 OUTCOMEVARIABLE2;*

*run;*

* *The first line is a standard line with a proc command. Just input the dataset you are going to analyze*
* *The second line just titles the output so you know what you are looking at when you look at the output*
* *The third line is where you list your explanatory variable after the word “class”*
* *The fourth line is where you list your outcome variables after the word “var.” You can list multiple outcome variables but can only list one explanatory variable in the third line.*
* If you are doing a paired analysis here, you do not need the third and fourth line here. So you do not need the class and var statements. Instead, you need a line that says “PAIRED Variable1\*Variable2” where variable 1 is your outcome variable preintervention and variable 2 is your outcome variable postintervention

An example if we wanted to look at differences in gender on pain level in our dataset.



And this is what the output would look like and how we should analyze it:

Chart, line chart

Description automatically generated

A couple of things to note here because it gets to be an overload of information:

* Only look at the 1st, 3rd, and 4th tables
* The first table says that there are 21 people with “0” as their sex who have a mean of 5.67 on the pain scale, and that there are 19 people with “1” as their sex who have a mean of 4.9 on the pain scale.
* The third table is the actual results of the t-test. It is somewhat confusing because there are 2 results here, there is a t-test result using the “pooled” method and a t-test using the “satterhwaite” method. To find out what method to use, go to table 4
* We use the “pooled” method when the 2 groups have equal variances. You don’t really need to understand this, just assume that if in this table the cell with “Pr > F” is <0.05, you use the satterhwaite method, and if it is >0.05, you use the pooled method
* So in our case the Pr>F is >0.05, so we would use the pooled method and the p-value of this t-test would be 0.4035

If I was going to format these results as a table, this is how it would look

|  |  |  |
| --- | --- | --- |
| **Variable** | Average | P-Value |
| Gender |  | 0.4035 |
| Male | 5.67 |  |
| Female | 4.89 |  |

1. ANOVA

Once you understand the above, the rest of the procedures become relatively easy. Let’s do ANOVA next:

*proc ANOVA data = YOURDATASET;*

*title "ANOVA";*

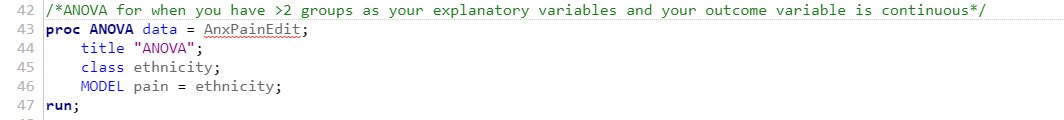
*class EXPLANATORYVARIABLE;*

*MODEL OUTCOMEVARIABLE = EXPLANATORYVARIABLE;*

*run;*

* The first line is the same as the other proc commands
* The second line is just a title line again
* The third line is where you list your explanatory variable after the word “class”
* The fourth line is a little different where you use say the outcome variable = the explanatory variable after the word “model”

This is what it looks like in our dataset if we are looking to investigate the effect of ethnicity on pain score



And this is the output:

Chart, box and whisker chart

Description automatically generated

* Here you can see a box plot of the different pain scores for each ethnicity
* The p-value here would be under “Pr>F” and would be 0.2158 here

1. Chi-Square

Here would be the syntax of the chi-square test

*proc freq data = YOURDATASET;*

*title "Chi-Square";*

*TABLE EXPLANATORYVARIABLE\*OUTCOMEVARIABLE/ CHISQ;*

*run;*

* The first line is the same as most of the proc commands
* The second line is the title statement
* The third line is where you list your explanatory variable\*outcome variable after the word “table.” The unique thing here is you put a slash after the outcome variable and write “CHISQ” to make sure they do the chi-square test. Remember you use chi-square with discrete outcome variables so the outcome variable in our example has changed to SeverePain

A picture containing timeline

Description automatically generated

And here is the output

Table

Description automatically generated

* Look at only the first and second tables
* The first table is just a descriptive. If you look at the first row/first column, it says there are 10 patients with a sex of “0” and no severe pain. This represents 25% of the total patients, 47% of the patients that have a sex of 0, and 45% of the patients with severe pain
* The second table has the result of the chi-square test in the first row

This is how I would make a table out of this data if sex=0 means male:

|  |  |  |  |
| --- | --- | --- | --- |
|  | No severe pain | Severe pain | P-value |
| Gender |  |  | 0.324 |
| Male | 10 (45%) | 11 (61%) |  |
| Female | 12 (55%) | 7 (39%) |  |

I use column percentages here because I think it helps understand the output. If you think about it, a lower percentage of patients with no pain were male, but a higher percentage of patients with severe pain were male. Therefore, there was a trend to severe pain in male patients. However this was not significant because the p-value was >0.324.

1. Pearson’s Correlation

Here is the syntax:

*proc corr data = YOURDATASET;*

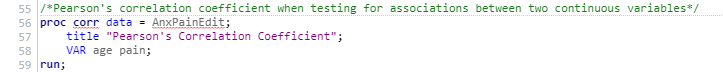
*title "Pearson's Correlation Coefficient";*

*VAR VARIABLE1 VARIABLE2;*

*run;*

* The first 2 lines are similar to the above
* The third line you should list your 2 variables after the word “var” separated by a space

Here is an example in our dataset if we are trying to see if a patient’s age and pain level were correlated



And here is the output:

Table

Description automatically generated

The correlation coefficient between age and pain is -0.05132. And it is not significant because the p-value associated with this statistical test is underneath the first number and is 0.75 which is above 0.05.