A Concise Tutorial for Data Analysis using BlueSky

Introduction and Aims:

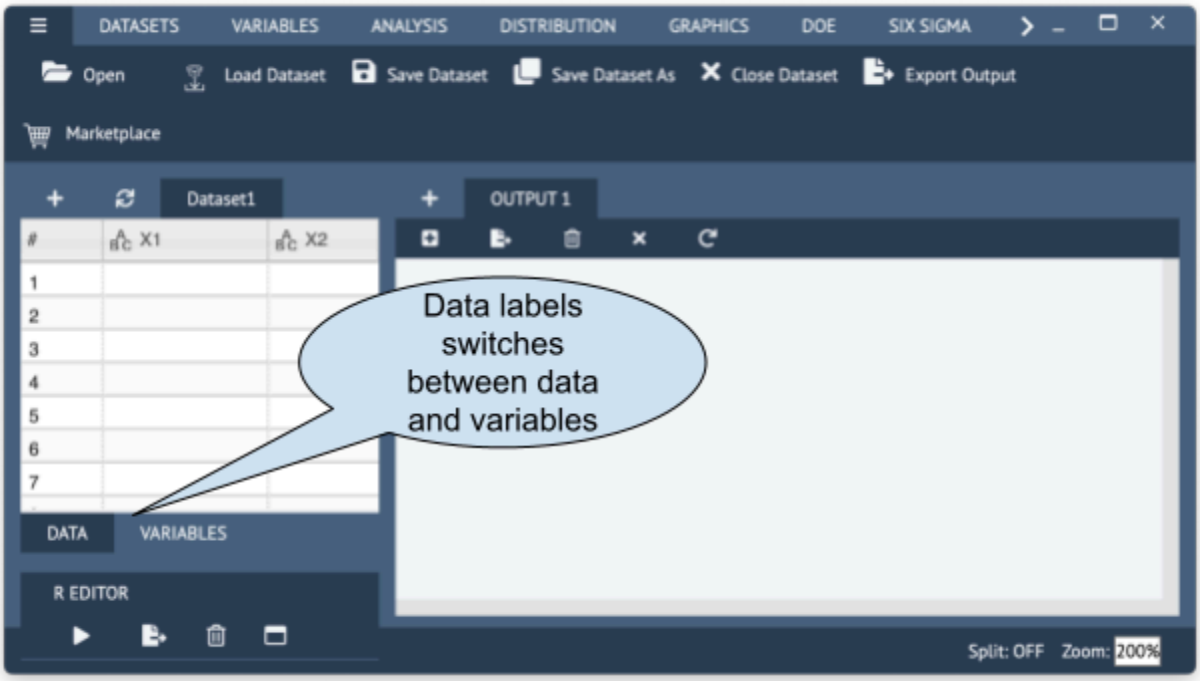
Everyday we deal with numerous data which can be from surveys or experiments. In order to visualize these data and derive a conclusion about our surveys or experiments, it is necessary to analyze data. Data analysis helps us to visualize the data using simple statistics and plotting. For example, we did a survey and asked people how often they drink alcohol. Then, we want to find out which gender is drinking more often. We can simply find this out by using statistical tools like Blue Sky.

Data Input:

Bluesky presents data in two views: data and variable

Data view:

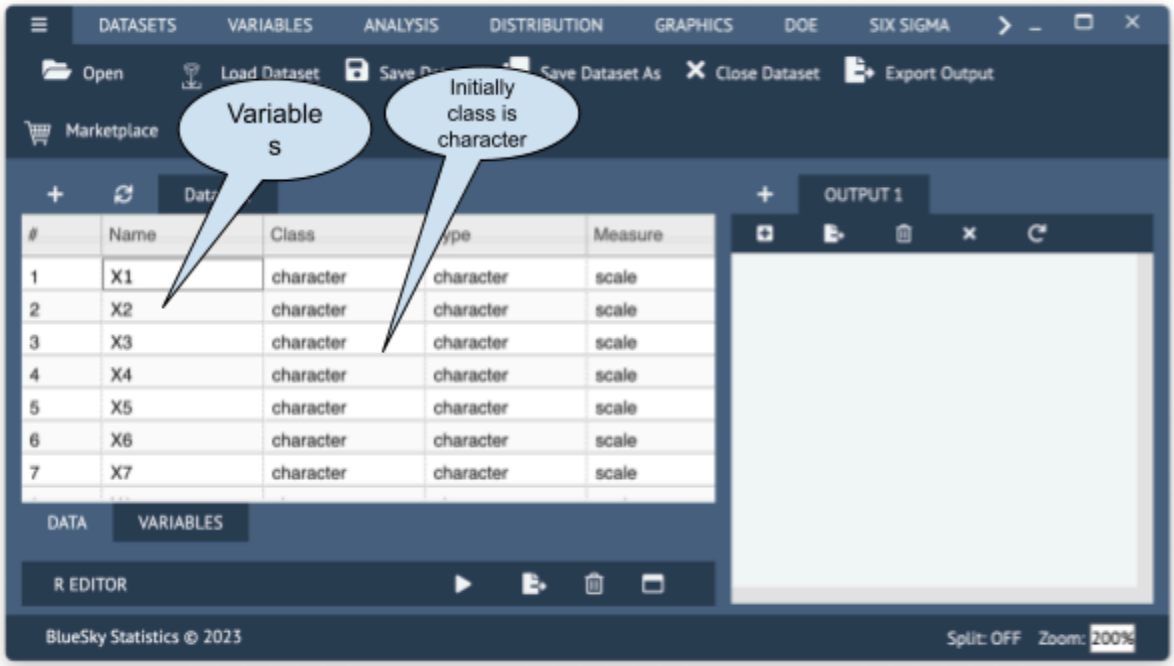
It looks like an excel sheet in which each column is “variable” and each row is “case.” Variables like age, income, education, and case like scores, person.



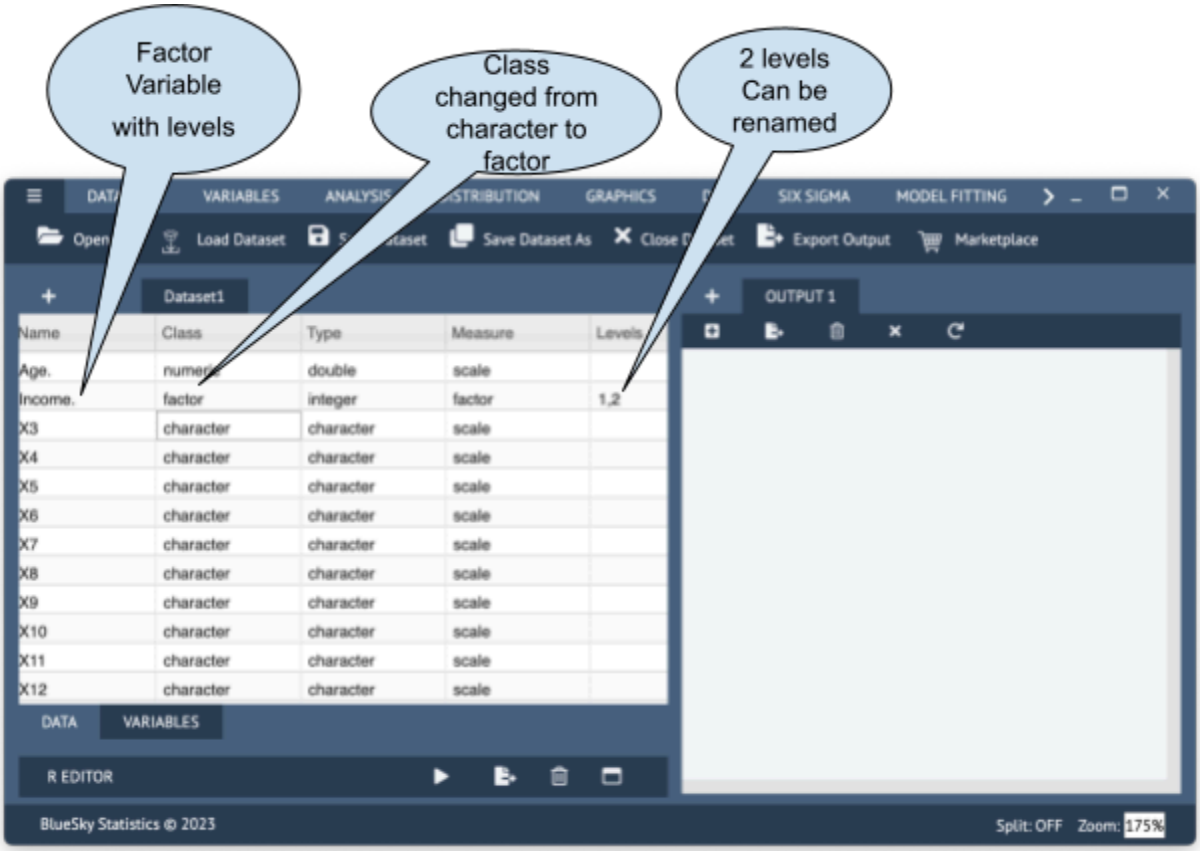
Variable view:

Variable name should not contain any space
Variable class: Numeric, Character, Ordered factor, and Factors
Variable type: Double, Integer, and character
Label: meaningful name for variables used for output.

<u>Measure:</u> always scale, and factor



Variable type changes with the class of variables. For example when we have a numeric variable, its class is double and when it is a factor its class is integer.

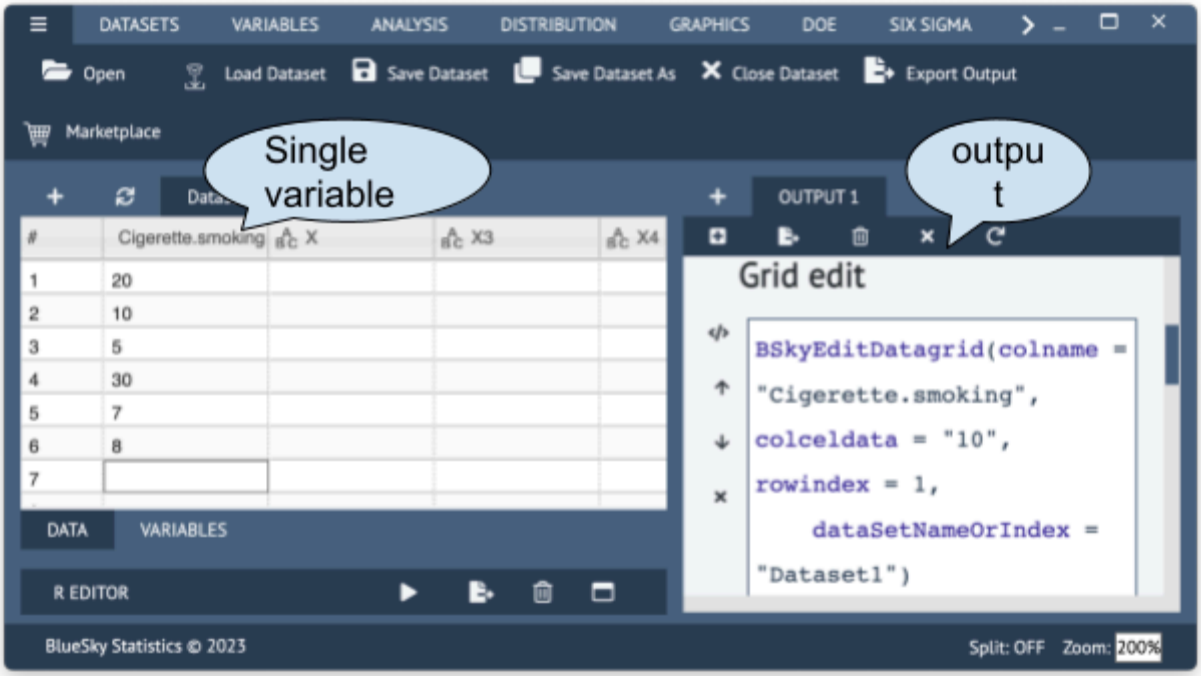


<u>How to enter Data:</u>

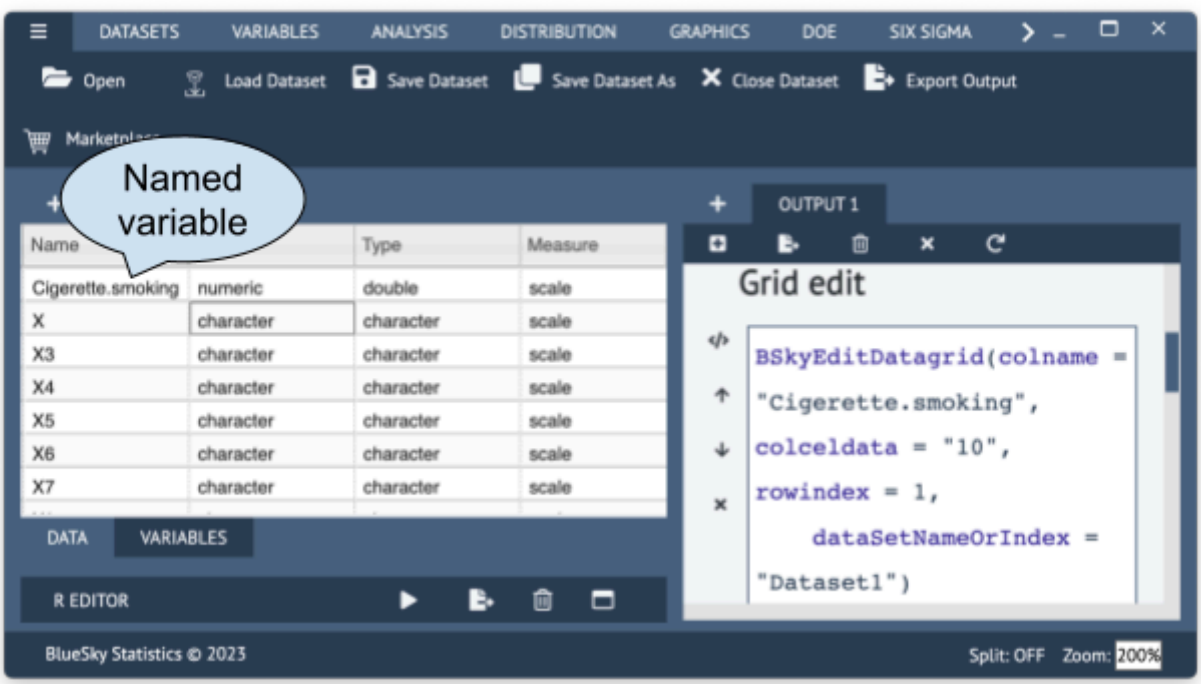
In data view type, enter variables on the column and enter data on the row following each variable. For example if you have data as follow:

Number of people	Number of cigarette smoked per day
1	20
2	10
3	5
4	30
5	7
6	8

Blue sky data entry for this data



We named the variable “cigarette smoking” in the variable section.



Note: Blue Sky cannot do analysis if there are less than 6 data. It also cannot give output if the class is incorrectly assigned while entering the data. So, you have to change the class of variables by yourself, as Blue Sky identifies all the entered data as a character. In the above case I changed my variable class to numeric as my data is numeric.

We can also upload data from excel sheets, R , SPPS, and STATA into Blue sky. For this, you just need to download data in CSV format and open it in Blue Sky. In the above picture you can see "open", there you can click and upload the data.
Additionally, there is an R editor, you can code there then, run the code and you can upload data.

Use: File > Open > Data
OR
R editor > Code > Run

Managing, saving, and exporting Blue Sky output

Saving and opening the output:

Output files contain all your tables, graphs, and plots.
You can save Blue Sky output using:
"Triple bar > Export Output
And open it in future by using:
"Triple bar > Open

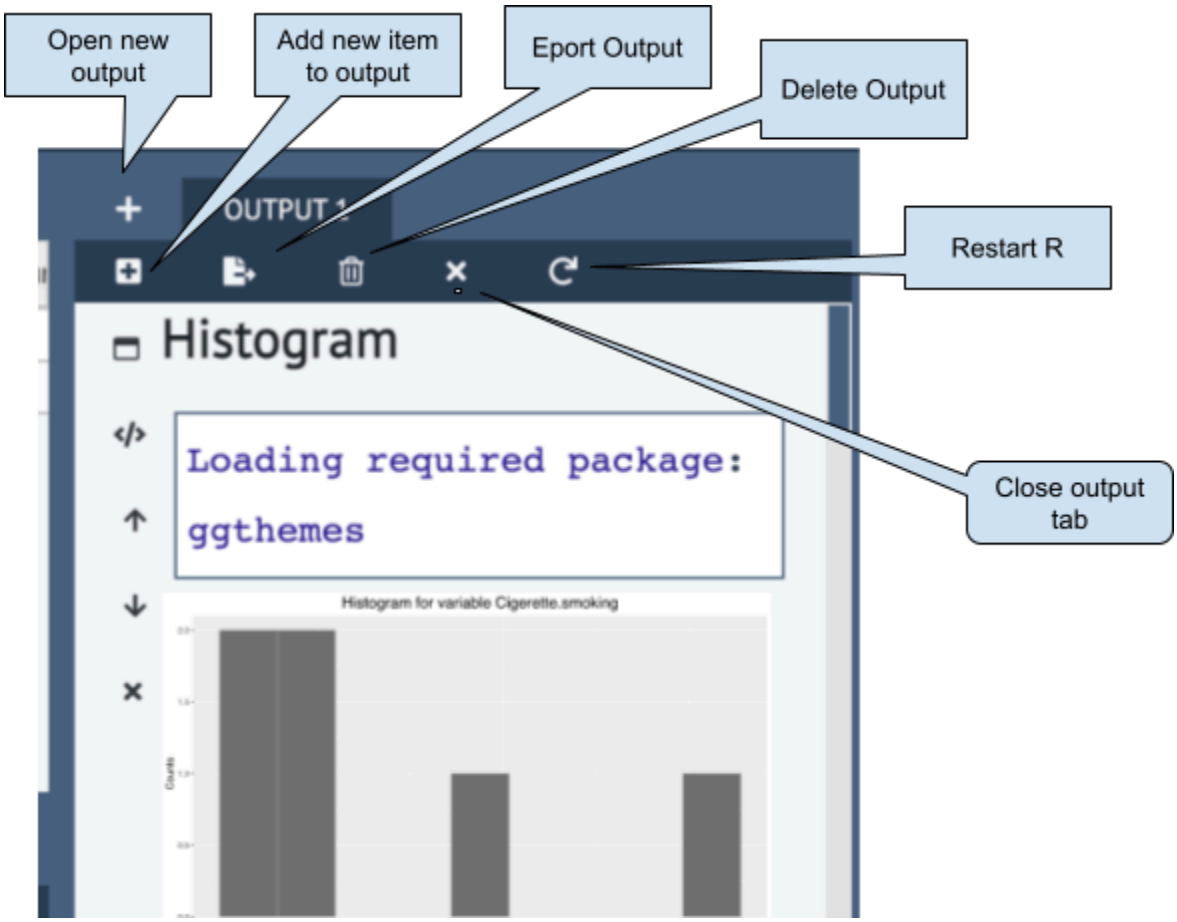


Fig: Output Window with controls annotated

Exporting output

You should export the data in Bmd format in Blue Sky. Bmd means Blue Sky Markdown format or you can share in HTML format as well. You can export by: " Triple Bar > Export Output"

After you have produced several charts, tables and graphics in the Blue Sky output, you might need to incorporate those to the formal documents (like resorts, thesis) To export those outputs to Word or Excel you simply copy those outputs and paste on the designed file; the same works for plots and graphs.

<u>To export plots and graphs in some other program:</u>
You can right click on the plot or graph then, options pop up, copy paste, download as PNG and download as SVG. You can choose any based on your necessity.

<u>To export a table:</u>
Right click on a table and menu pops up with options: copy or copy as texts. You can choose a copy as text and export it in the designated file location.

Bluesky offers several different styles you can apply to its user interface. You can do it by "Triple Dot > Settings > Output". By using this function key, you can export the tables in different formats like APA format, word processor format, and LATEX format. As well as you can change the decimal digit display.

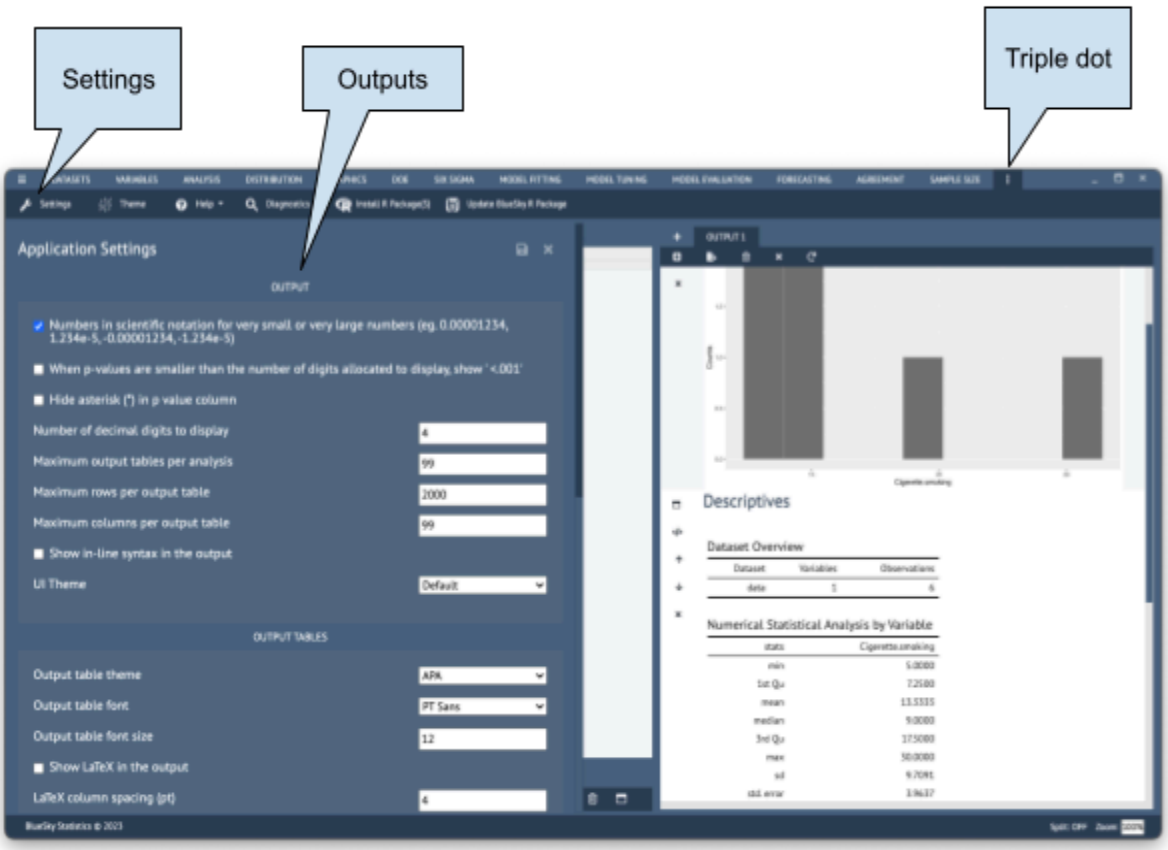


Fig: The output setting Tab

Managing multiple output windows

You can have multiple data sets in a data grid and can open multiple outputs in the output window by simply clicking + sign on the output window. It does not matter which data sets analysis you get in which output window when you are confident, but for simplicity you can open one data set and corresponding to it one output window.

Summary Statistics Analysis and Plots:

We can summarize and analyze data in Blue Sky. It depends on what kind of variables we have. In Blue Sky language “class” of a data. It is numeric, character or a factor.
Basic univariate stats - to calculate and display the data in Blue Sky

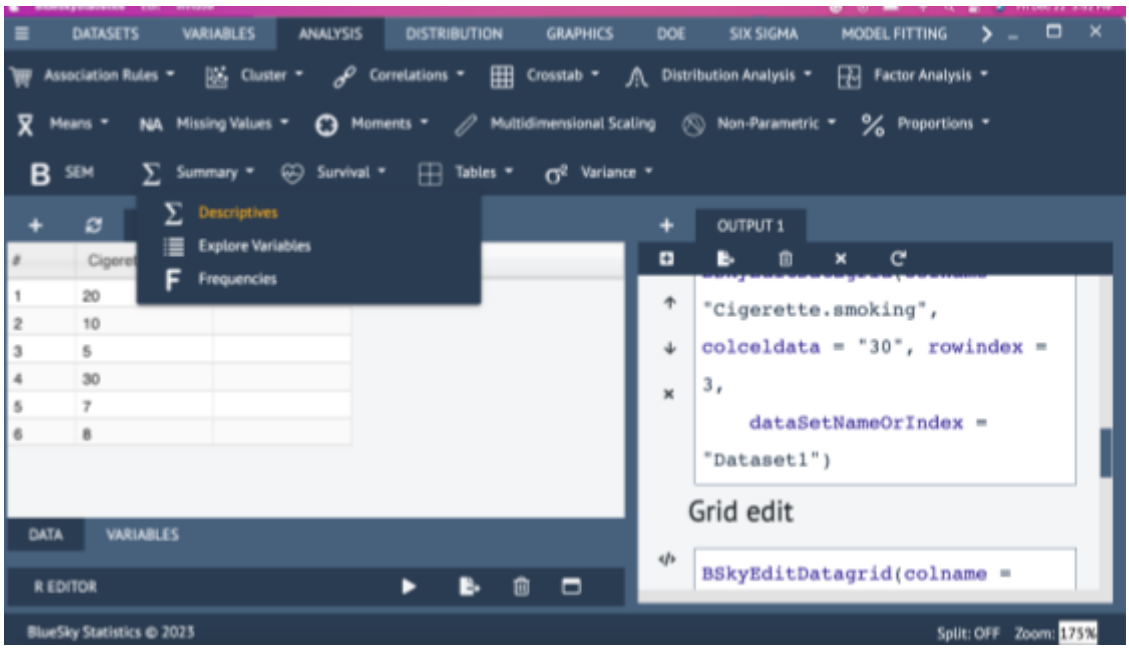
<u>Tutorial 1:</u>

Mean, Sum, Standard Deviation, Variance, Minimum Value, Maximum value, and Range. Summarizing and analyzing data for a single variable can be done in Blue Sky using

descriptive statistics. It includes mean, median, mode, and variability like range, standard deviation, and variance.

For example: We entered data about cigarette smoking in the Blue Sky. We can get a descriptive analysis of it by:

Blue sky > Analysis > summary> Descriptives



Then, push the variable to the 'selected variable box'. Then, run it. You can also choose which of the descriptives you want. You will get the output in the output box on the left hand side.

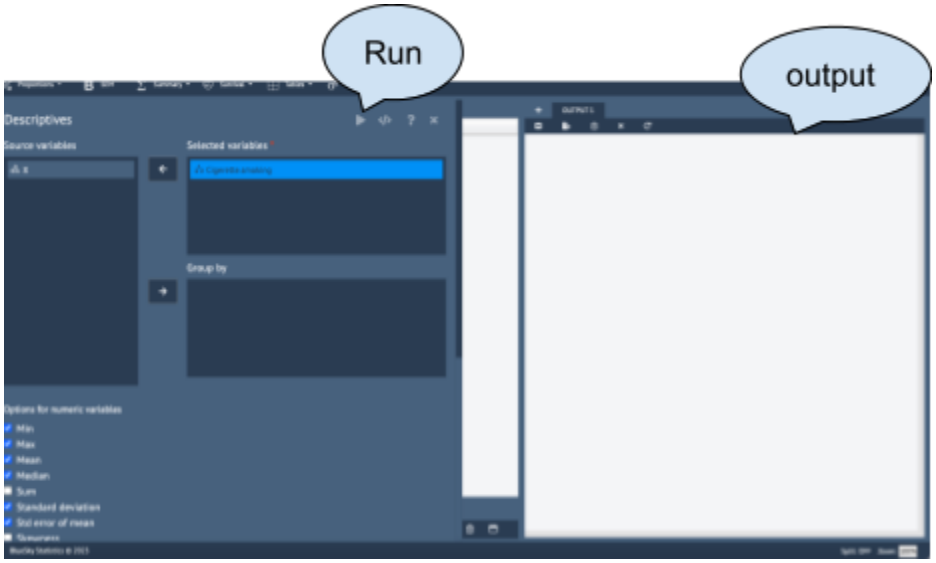


Fig: Variables moved to "selected variable box"

Descriptives

Dataset Overview

Dataset	Variables	Observations
data	1	6

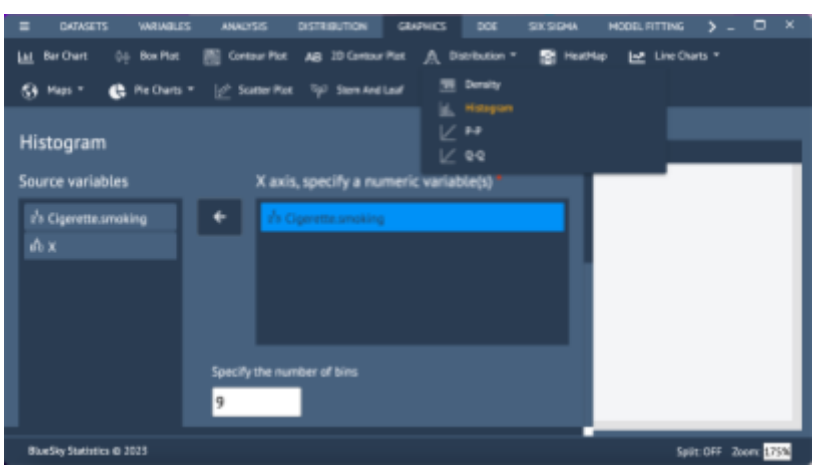
Numerical Statistical Analysis by Variable

stats	Cigerette.smoking
min	5.0000
1st Qu	7.2500
mean	13.3333
median	9.0000
3rd Qu	17.5000
max	30.0000
sd	9.7091
std. error	3.9637
cv	0.7282
var	94.2667
n	6.0000
NAs	0.0000

Flig: Output generated for variable cigarette smoking in Blue Sky

You can also get plots or graphs by:

Graphics > Distributions > Histograms> variable > run



Output:
You can label the output like X axis and Y axis, title, as well as can make it colorful by clicking to the options before running it.

Fig: Histogram

Tutorial 2: Correlation

Two or more variables can be included in a correlation matrix. When generating the correlation matrix, the data editor must be open with appropriate data set before continuing.

Example: Data set

Number of people	Number of cigarette smoked per day	Age of the person
1	20	20
2	10	30
3	5	40
4	30	45
5	7	50
6	8	35

1. Enter this data set in the Blue Sky data grid to do correlation analysis.

2. Then, name the variables. Change the class variables accordingly. In this data set both are numerical variables so change the class from character to numerical.

3. Analysis > Correlation > Pearson Legacy

4. Push the variables from the right box to the left “selected variables” box by clicking the variables.

5. Then, hit the run button. You will get output in the output window.

#	Cigerette.smoking	Age.of.a.person.
1	20	20
2	10	30
3	5	40
4	30	45
5	7	50
6	8	35

Fig: Data set entered in data grid in bluesky

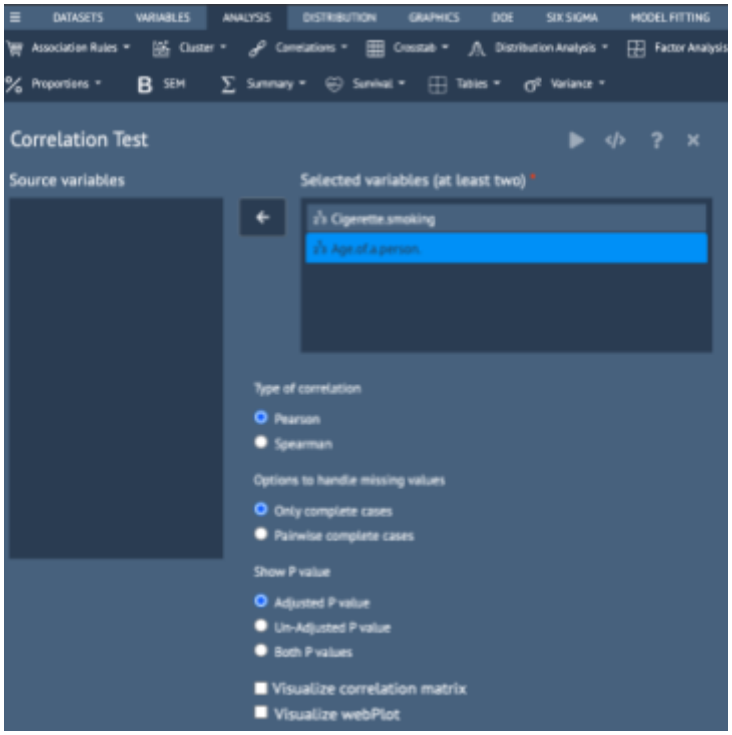


Fig: Doing analysis

		Cigarette.smoking	Age.of.a.person.
Cigarette.smoking	Correlation	1.0000	-0.1017
	Adj-P		0.8480
	n	6.0000	6.0000
Age.of.a.person.	Correlation	-0.1017	1.0000
	Adj-P	0.8480	
	n	6.0000	6.0000

Fig: Output of Correlation analysis showing P value. P value is greater than .05 (0.84) so age and smoking are not correlated.

Statistical Models in Blue Sky

Tutorial 1: Linear Regression

The regression sub-menu under the model fitting on the Statistical menu of the Data Editor provides regression techniques. This tutorial will introduce you to how to perform linear regression in Blue Sky. The output contains goodness of fit statistics and the coefficient of variables.

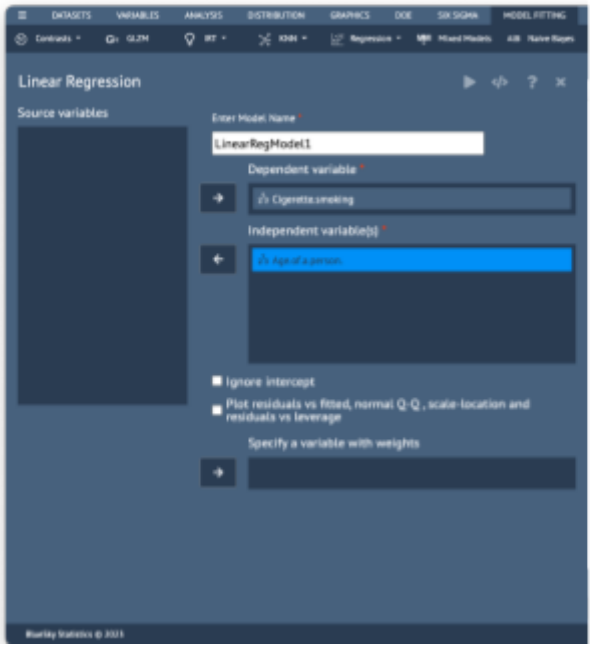
Problem: From the cigarette smoking and age data set, we want to predict cigarette smoking based on the age of the people.

Solutions:

Model Fitting > Regression > Linear basic

From the statistical menu click on the model fitting then on regression. Further, click on linear basics. This will open the linear regression dialog box.

Then you can shift the variables from source variables to the left dependent and independent variable box by clicking in the arrow between boxes. In this data set, the dependent variable is cigarette smoking and the independent variable is age.



Now, you can run it in Blue Sky and you will get the output.

$$\widehat{\text{Cigerette.smoking}} = 16.6857 - 0.0914(\text{Age.of.a.person.})$$

Model: lm(formula = Cigerette.smoking ~ Age.of.a.person., data = Untitled, na.action = na.exclude)

LM Summary

Residual Std. Error	df	R-squared	Adjusted R-squared	F-statistic	numdf	dendf	p-value
10.7988	4	0.0105	-0.2571	0.0418	1	4	0.8480

Residuals

Min	1Q	Median	3Q	Max
-8.0286	-5.3929	-4.5286	2.8714	17.4286

Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)	2.5 %	97.5 %
(Intercept)	16.6857	16.9766	0.9829	0.3813	-30.4488	63.8202
Age.of.a.person.	-0.0914	0.4471	-0.2045	0.8480	-1.3328	1.1500

Anova table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Age.of.a.person.	1	4.8762	4.8762	0.0418	0.8480
Residuals	4	466.4571	116.6143	NA	NA

Sum of squares table

	Values
Sum of squares of regression	4.8762
Sum of squares of residuals	466.4571
Total sum of squares	471.3333

Output Window

ANOVA analysis of Variance:

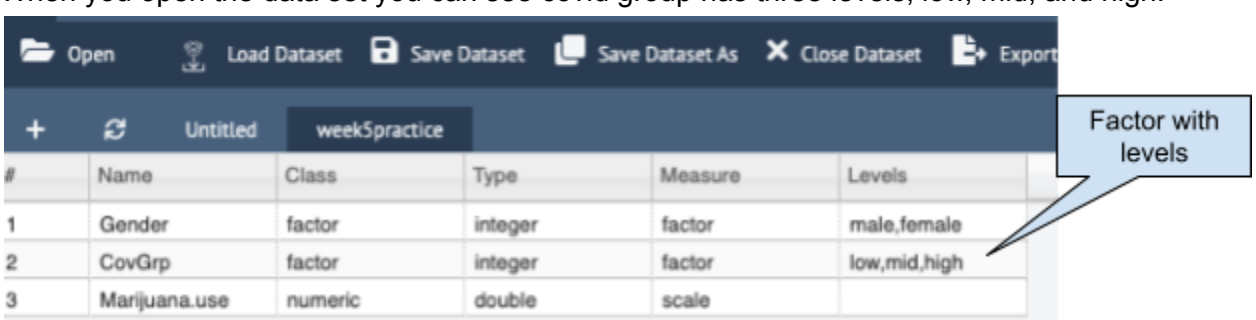
This can be analyzed in Blue sky by going to Analysis then into the means. There pops out ANOVA 1 and 2 ways along with other options. You can click on ANOVA 1 and 2 ways to do an analysis.

Analysis > Means > ANOVA 1 and 2 ways

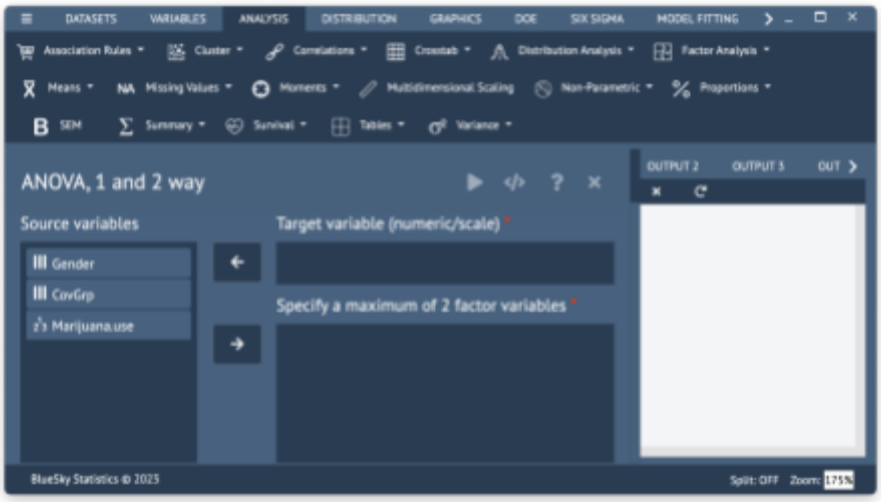
Note: The data set used in this analysis is available on request.

This data is about Marijuana use among different gender groups and covid groups. This data is for Factorial ANOVA analysis as there are 2 factors and each has levels. But I will be doing One Way ANOVA by only taking covid group as an independent variable and Marijuana use as a dependent variable.

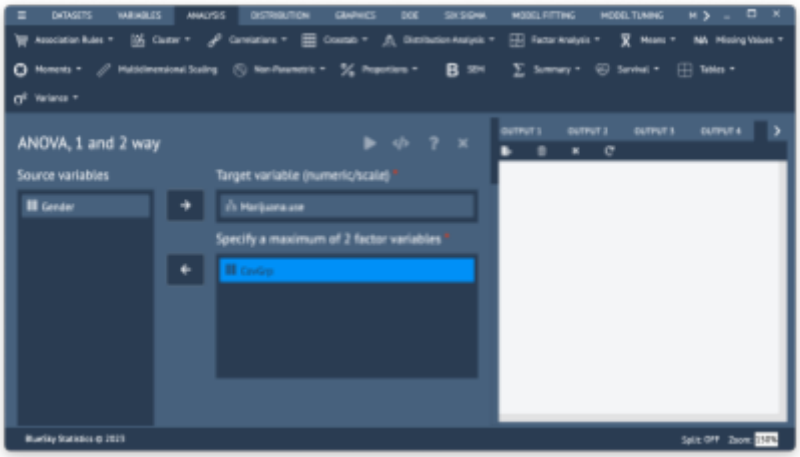
When you open the data set you can see covid group has three levels, low, mid, and high.



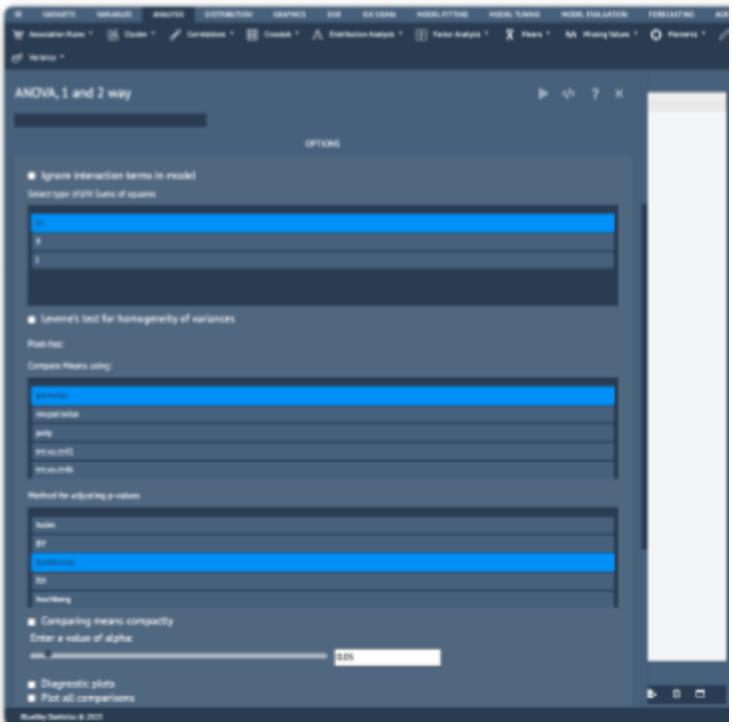
Further, for analysis, go to the analysis and click on it, then the submenu appears. In the submenu you can find means, click on the means then again menu with options pops out. Click on ANOVA 1 and 2 ways. Then you can get the following window.



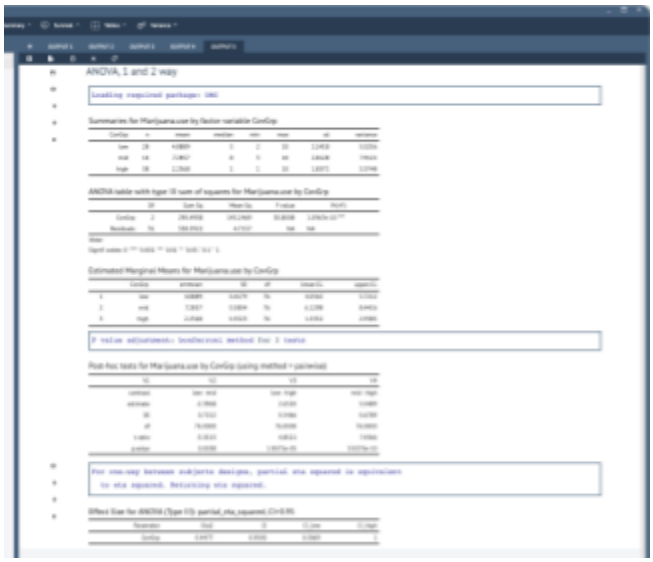
Now, you should choose the variable from source variable and by clicking on each variable you should send it to the target variable box for numerical variable, and factor variable to factor variable box.



Now, in this picture you can see variables are in the left hand side boxes. Further, there are various options which you have to choose before running the analysis.



You will see the above table with options. Most of them are automatically chosen. You must choose the Bonferroni test by yourself.
After choosing options you can run the analysis. Then, you will get the following output.



Output Window

Factorial ANOVA analysis is also done in a similar way. Factorial ANOVA has more than 1 factor variable that is the only difference. Rest of the things are similar.

T Test

For doing T tests in Blue sky, all the steps are similar to ANOVA, only you have to go to the means and then choose T-test. There are two types, T- Test Independent sample and T-Test paired. Other steps are similar. Only difference between ANOVA and T test is when you have more than one factor variable and when a single factor has more than 2 groups then we do ANOVA instead of T test.

Analysis > Means > Independent T- Test/ paired T- Test

When to do T-Test Independent Sample vs Paired T-Test?

T-Test Independent sample you can do when there is study in two different groups of people/samples. Whereas when the study is done in the same group of people/ sample twice, like pre and post survey in a particular sample, then we should do paired T-Test.

Let's see an example from the previous data set we used for ANOVA analysis. In that dataset we have 3 variables, COVID, gender, and marijuana. We used COVID and marijuana use in the ANOVA analysis. We cannot use COVID as an independent variable in the T test because COVID has 3 levels that are low, mid, and high. So, for the T test we will use Gender as an independent variable. Since we have 2 groups of samples male and female we will be doing an independent T test.

You will get this window after you click on the independent T test.
Now, you can shift the variables from the right box to the left box, Gender into factor variables and Marijuana use into select variables.
Then hit the button, run. Before running you can decide what output you want by clicking on options. As well as you can choose which method to use to calculate effect size.

t-test, Independent Samples

Group Statistics

	Gender	N	Mean	Std Deviation	Std Error Mean
Marijuana.use	male	35	6.3429	2.5889	0.4376
	female	44	2.2045	1.4257	0.2146

Independent Samples t-test

		Levene's Test for Equality		t-test Equality of Means						
									confidence interval of the diffs 0.95	confidence interval of the diffs 0.95
		F	Sig.	t	df	Sig.(2-tail)	mean difference	std. error difference	lower	upper
Marijuana.use	Equal variance assumed	21.1763	1.6213e-05 ***	9.0529	77.0000	1.0365e-15 ***	4.1383	0.4571	3.2260	5.0506
	Equal variances not assumed	NA	NA	8.4904	50.0361	2.9114e-11 ***	4.1383	NA	3.1193	5.1573

Note:
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Fig: Output window

As you can see in the output, you cannot see effect size as I didn't choose to display effect size in my output.

You can similarly do the paired t test. Only you have to choose paired t tests in analysis in place of independent t tests. Rest of the things are similar.

Testing for Normality

The most important part of any statistical test is testing the hypothesis with the data that you have for normality. One of the most important checks that you have to do is whether variation in a data is approximately normally distributed or not? The Normal (Gaussian) distribution is famously known as the “bell curve”.

The various statistical tests for normality are:

1. Histogram
2. Kurtosis and skewness
3. Q-Q plot

1. **Histogram:**

It can be used to test normality. Histograms can be drawn with best fitting normal distribution superimposed over it and make a subjective assessment of whether the histogram and normal distribution seems to agree. You can get histogram superimposed with the bell curve by going to graphics then into distribution. In distribution you can see histogram as one of the options. Once you push the variable into the box, you will have the options to choose graphics, labeling axis, and to display a normal curve. You should choose to display a normal curve along with histogram.

Graphics > Distribution > Histogram

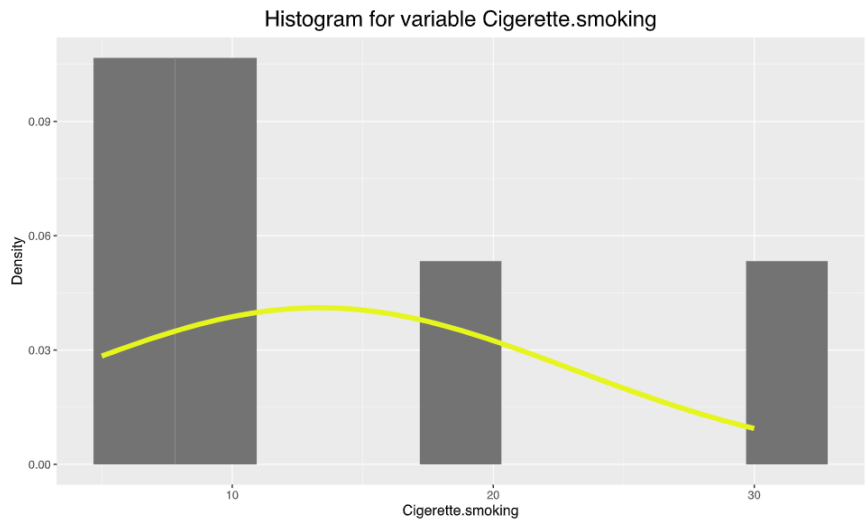


Fig: Histogram superimposed with normal curve(yellow)

Note: The appearance of histogram depends on sample size and number of bins so be careful your sample size is big enough that is more than 100.

2. Kurtosis and Skewness

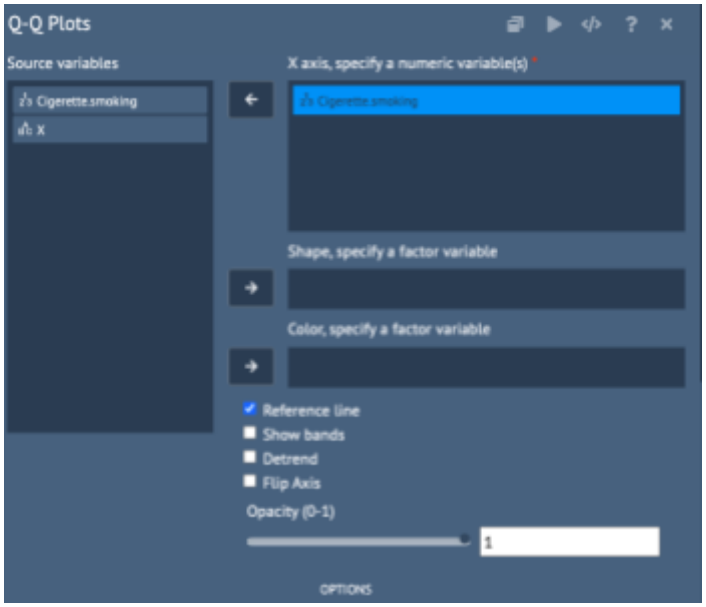
Skew helps to determine the asymmetry in the data. It quantifies to what extent data is skewed or shifted to one side. Positively skewed indicates more low values whereas negatively skewed indicates more high values. It helps in understanding shape and outliers in a data set. Symmetrically distributed means 0 skewness. Kurtosis quantifies the shape of the distribution by providing information about the tails and peakedness of the distribution compared to a normal distribution. Kurtosis > 3 are leptokurtic which have long and thick tails, indicating a larger number of outliers. Kurtosis < 3 is Platykurtic which has a short tail, most data points are concentrated proximity to means. Kurtosis = 3 is Mesokurtic, it is the same as normal distribution.
Note: Small samples from normal distribution may vary considerably in their values of kurtosis and skewness.

3. Q-Q Plot

Best graphical method to determine whether the data is normally distributed or not. The data fits a Normal distribution if the points on a Q-Q plot closely follow a straight line.

Graphics > Distribution > Q-Q plot

Then choose the variable, push it to the box and choose a reference line to display along with a plot from options then run. You will get a Q-Q plot as fig.8.



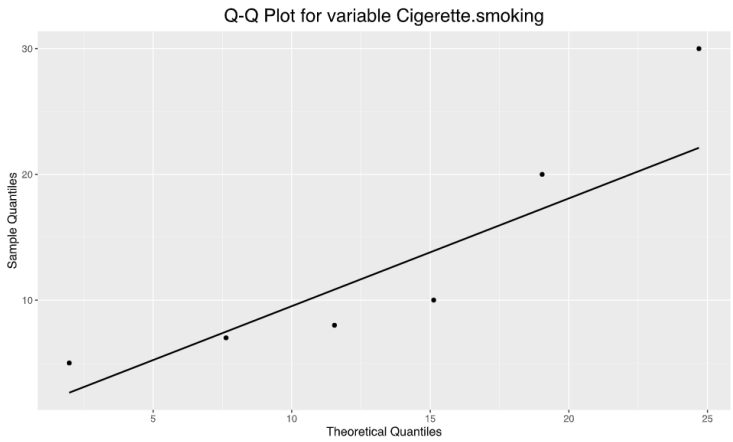


Fig 8: Q-Q plot for variable cigarette smoking with Reference line

This Q-Q plot indicates data is not normally distributed as the data points do not lie close to the reference line.

Reference

Munchen, R.A. (2023). Blue Sky Statistics. Retrieved July 6, 2024, from https://r4stats.com/books/bluesky-statistics-user-guide/