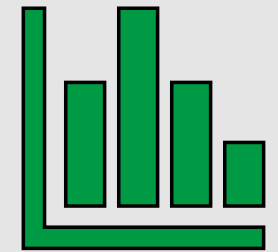




# The Statistical Software Toolkit

BERDC Special Topics Talk 1



**DaCCoTA**  
DAKOTA CANCER COLLABORATIVE  
ON TRANSLATIONAL ACTIVITY

Dr. Mark Williamson  
Biostatistics, Epidemiology,  
and Research Design Core

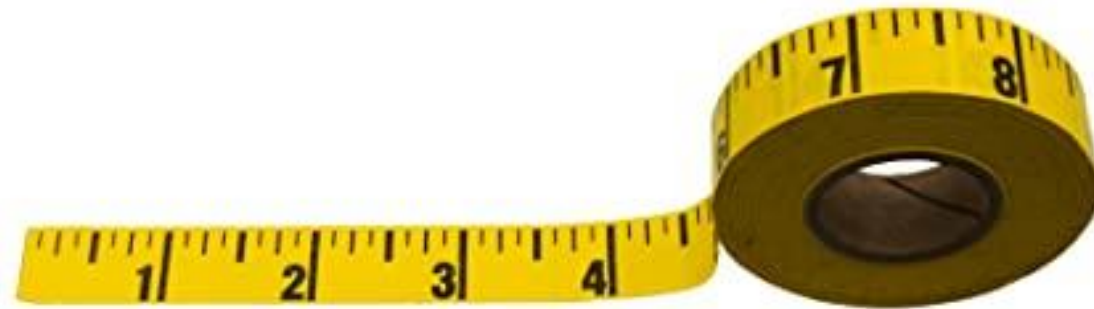
# Introduction

- There are many statistical software options
- Some are better at some things than others
- At the end of the day, they are all tools to aid you in your research
- It is good to know your way around multiple tools
- Here, we'll be counting down the top 5 tools for statistical software



# Inclusion Criteria

- Software is free, has a free version, or has a version that can be accessed through UND for free
- Software I've had at least some experience with





Number 5

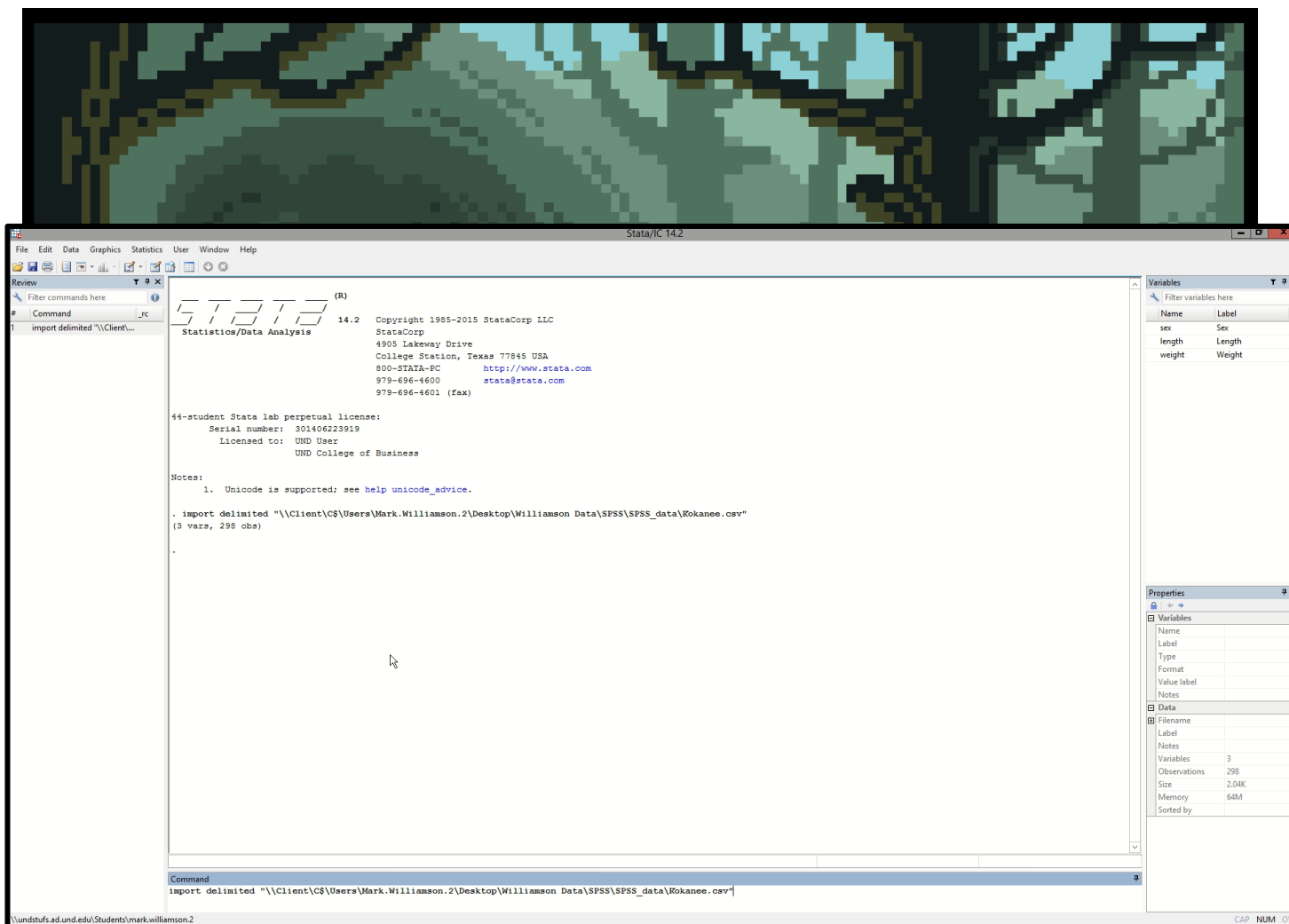
## Overview:

- general-purpose package created by StataCorp
- 1985

Tool: Ratchet wrench



- Access
  - Need Citrix Workspace first
    - <https://und.teamdynamix.com/TDCClient/2048/Portal/KB/ArticleDet?ID=58677>
  - Add Stata-64 to the Citrix App
- Features
  - Command
  - Pull down
  - Allows for user-written programs
- Downside
  - Plug-and-chug problem
  - Coding learning curve



**Suggested Uses: All purpose data**



**Number 4**



## Overview:

- Interactive statistical analysis package created by SPSS Inc. then acquired by IBM
- 1968 (SPSS Inc.); 2009 (IBM)

## Tool: Power Drill







- Access: Citrix App
- Features
  - All pull-down except a few
  - In-package data working
  - Sample Datasets
- Downsides
  - Pain to get data in
  - Plug-and-chug
  - No coding

	agecat	gender	marital	active	bfast	var	var	var	var	var	var	var	var
1	1	0	1	1	3								
2	3	0	1	0	1								
3	4	0	1	0	2								
4	2	1	1	1	2								
5	3	0	1	0	2								
6	4	0	1	0	3								
7	2	1	1	0	1								
8	4	1	0	0	2								
9	2	1	1	1	2								
10	2	1	1	1	1								
11	2	0	1	0	3								
12	1	1	1	0	1								
13	1	1	0	1	1								
14	4	1	1	1	2								
15	4	1	1	1	2								
16	1	0	0	0	1								
17	3	1	1	0	2								
18	2	0	1	0	3								
19	2	0	1	1	1								
20	2	1	0	1	1								
21	3	0	0	1	1								
22	1	1	1	0	3								
23	2	1	1	1	1								
24	4	1	0	1	3								
25	3	0	1	0	2								
26	3	1	1	0	2								
27	1	0	1	1	3								

**Suggested Uses: Very Simple or very complex data**

A large, bright yellow star with a glowing, neon-like outline is centered on a dark blue background. The star has six points and is filled with a solid yellow color. The text "Number 3" is written in a bold, black, sans-serif font across the center of the star.

**Number 3**



## Overview:

- Spreadsheet developed by Microsoft
- 1987

Tool: Hammer





- Access: typically built-in
- Features
  - Interactive spreadsheet
  - Functions
- Downsides
  - Not technically statistical software package
  - Basic

The screenshot shows a Microsoft Excel spreadsheet with the following data:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Islands	SpeciesRi	Saspecies	endemic	percent	Area	altitude	elevation	distanceP	distancev	distances	distancelarge	island							
2	Ecuador	65	57	1	1.5	3487	3500	2397	1	1	1	1								
3	Chiles	36	31	1	2.7	326	3500	1264	36	14	36	36								
4	Las Papas-Cocor	30	25	1	3.3	501	3500	1170	234	13	26	26								
5	Sumapaz	37	32	3	8.1	2031	3500	1060	543	83	80	116								
6	Tolima-Quindio	35	33	9	25.7	989	3500	1900	551	23	25	25								
7	Paramillo	11	10	1	9	25	3500	460	773	45	45	186								
8	Cocuy	21	18	1	4.7	2168	3500	1998	801	14	108	14								
9	Pamplona	11	9	1	9	217	3500	700	950	14	14	14								
10	Cachira	13	12	0	0	143	3250	735	958	5	19	19								
11	Tama	17	14	0	0	46	3000	613	995	29	29	29								
12	Batlalion	13	10	0	0	66	3250	662	1065	55	65	55								
13	Merida	29	26	6	20.6	1798	3500	1502	1167	35	55	204								
14	Perija	4	4	2	50	167	3000	750	1182	75	197	75								
15	Santa Marta	18	16	12	66.6	606	3500	2275	1238	75	75	330								
16	Cende	15	13	0	0	70	3000	552	1380	35	35	35								

**Suggested Uses: Data formatting and exploration**



Number 2

## Overview:

- Multi-purpose package developed by SAS Institute
- 1976

## Tool: Table Saw



- Access: SAS Studio

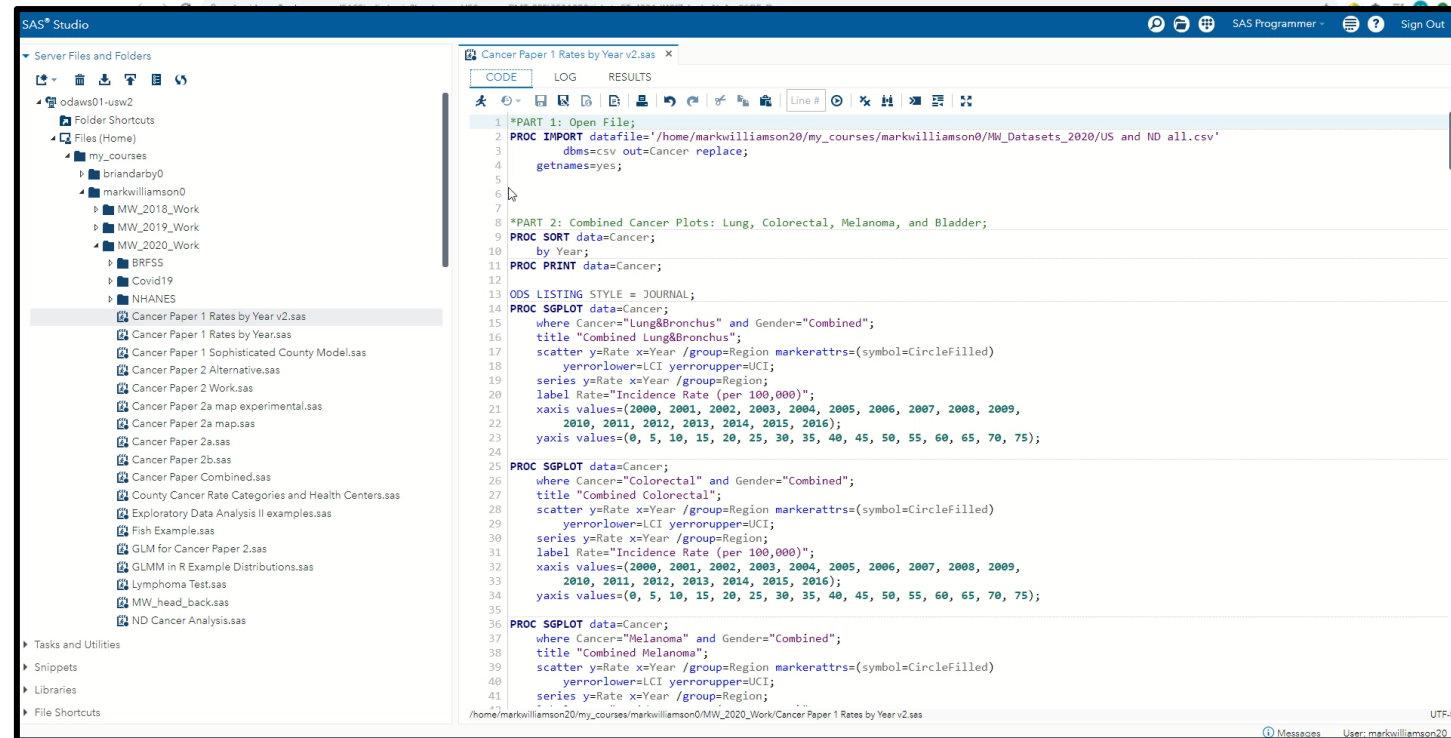
- [https://www.sas.com/en\\_us/software/on-demand-for-academics.html#Features](https://www.sas.com/en_us/software/on-demand-for-academics.html#Features)

- Features

- Consistent structure
- Directories
- Sample data
- Colors
- Anticipatory
- Log
- Transforming data
- Great support

- Downsides

- Not all available in Studio
- Upload limits



```
1 *PART 1: Open File;
2 PROC IMPORT datafile='/home/markwilliamson20/my_courses/markwilliamson0/MW_Datasets_2020/US_and_ND_all.csv'
3   dbms=csv out=Cancer replace;
4   getnames=yes;
5
6
7
8 *PART 2: Combined Cancer Plots: Lung, Colorectal, Melanoma, and Bladder;
9 PROC SORT data=Cancer;
10  by Year;
11 PROC PRINT data=Cancer;
12
13 ODS LISTING STYLE = JOURNAL;
14 PROC SGPLOT data=Cancer;
15   where Cancer="Lung&Bronchus" and Gender="Combined";
16   title "Combined Lung&Bronchus";
17   scatter y=Rate x=Year /group=Region markerattrs=(symbol=CircleFilled)
18     yerrorlower=LCI yerrorupper=UCI;
19   series y=Rate x=Year /group=Region;
20   label Rate="Incidence Rate (per 100,000)";
21   xaxis values=(2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009,
22     2010, 2011, 2012, 2013, 2014, 2015, 2016);
23   yaxis values=(0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75);
24
25 PROC SGPLOT data=Cancer;
26   where Cancer="Colorectal" and Gender="Combined";
27   title "Combined Colorectal";
28   scatter y=Rate x=Year /group=Region markerattrs=(symbol=CircleFilled)
29     yerrorlower=LCI yerrorupper=UCI;
30   series y=Rate x=Year /group=Region;
31   label Rate="Incidence Rate (per 100,000)";
32   xaxis values=(2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009,
33     2010, 2011, 2012, 2013, 2014, 2015, 2016);
34   yaxis values=(0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75);
35
36 PROC SGPLOT data=Cancer;
37   where Cancer="Melanoma" and Gender="Combined";
38   title "Combined Melanoma";
39   scatter y=Rate x=Year /group=Region markerattrs=(symbol=CircleFilled)
40     yerrorlower=LCI yerrorupper=UCI;
41   series y=Rate x=Year /group=Region;
```

**Suggested Uses: All purpose data**

# Honorable Mentions



Minitab®







**Number 1**



## Overview:

- Programming language and software environment developed by the R Core Team
- 1993

Tool: Swiss-army knife made of Swiss-army knives





- Access: Citrix App or direct download (<https://cran.r-project.org/>)
- Features
  - Command line
  - R commander
  - R studio
  - Packages and support
- Downsides
  - Learning curve
  - No standardized notation

A screenshot of the RStudio software interface. The main editor window shows R code for setting up the environment and creating plots. The console window at the bottom displays the R startup message, including copyright information and instructions on how to use R. The interface includes a menu bar, a toolbar, and several panels on the right side for Environment, History, Connections, Files, Plots, Packages, Help, and Viewer.

```
1 #PART 1: Setup
2
3
4 setwd("c:/Users/Mark.williamson.2/Desktop/williamson Presentations/Statistical Modules")
5 library(ggplot2)
6 library(dplyr)
7 library(rgl)
8 #####
9
10 #PART 2: Bubble Plot
11
12 #iris
13 head(iris)
14
15 ggplot(data=iris, aes(x=Sepal.Length, y=Sepal.width, size=Petal.Length))+
16   geom_point(alpha=0.7)
17 ggplot(data=iris, aes(x=Sepal.Length, y=Sepal.width, size=Petal.Length, color=Species))+
18   geom_point(alpha=0.7)
19
20 #insular
21 insular<-read.csv("insular.csv")
22 attach(insular)
23 print(insular)
24
25 ggplot(data=insular, aes(x=Area, y=elevation, size=SpeciesRichness)) +
26   geom_point(alpha=0.7)
```

Copyright (C) 2019 The R Foundation for Statistical Computing  
Platform: x86\_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.

[workspace loaded from ~/.RData]

> |

**Suggested Uses: All purpose data and weird stuff**

# Example

## Kokanee Fish

- 298 observations
- Length and weight of salmon
- Male and female (will not use)

# Summary Handout

- Available online at <https://med.und.edu/daccota/berdc-resources.html>
  - Under Special Topics Talks -> The Statistical Software Toolkit

General Info	STATA	SPSS	Excel	SAS	R
Access	Citrix Workspace -> Stata-64 app	Citrix Workspace -> SPSS Statistics 26 app	(installed on most computers)	SAS website -> On demand for academics -> SAS Studio	<a href="https://cran.r-project.org">https://cran.r-project.org</a> <a href="https://rstudio.com/">https://rstudio.com/</a>
Features	Command line; drop-down menus; user-written programs	Drop-down menus; within-software data spreadsheet	Interactive spreadsheet; functions; add-on functionality	Within-software code sheet; consistent structure; high support	Command line; R commander; R Studio; additional packages
Resources	<p><b>STATA</b></p> <ul style="list-style-type: none"> <li><a href="https://www.stata.com/links/resources-for-learning-stata/">https://www.stata.com/links/resources-for-learning-stata/</a></li> <li><a href="https://data.princeton.edu/stata/">https://data.princeton.edu/stata/</a></li> <li><a href="https://dsc.gmu.edu/files/Advanced_Stata_Skills.pdf">https://dsc.gmu.edu/files/Advanced_Stata_Skills.pdf</a></li> <li><a href="https://www.stata.com/bookstore/statacheatsheets.pdf">https://www.stata.com/bookstore/statacheatsheets.pdf</a></li> </ul> <p><b>SPSS</b></p> <ul style="list-style-type: none"> <li><a href="https://www.spss-tutorials.com/basics/">https://www.spss-tutorials.com/basics/</a></li> <li><a href="http://statistikhowto.com/probability-and-statistics/spss-tutorial-beginners/">statistikhowto.com/probability-and-statistics/spss-tutorial-beginners/</a></li> <li><a href="https://students.shu.ac.uk/life/documents/pdf/analysing_data_using_spss.pdf">https://students.shu.ac.uk/life/documents/pdf/analysing_data_using_spss.pdf</a></li> <li><a href="https://rblissett.com/resources/spss_cheat_sheet/">https://rblissett.com/resources/spss_cheat_sheet/</a></li> </ul> <p><b>XCEL</b></p> <ul style="list-style-type: none"> <li><a href="https://www.excel-easy.com/data-analysis/analysis-toolbak.html">https://www.excel-easy.com/data-analysis/analysis-toolbak.html</a></li> <li><a href="https://www.excel-easy.com/data-analysis/charts.html">https://www.excel-easy.com/data-analysis/charts.html</a></li> <li><a href="https://edu.gcfglobal.org/en/excel2016/functions/1/">https://edu.gcfglobal.org/en/excel2016/functions/1/</a></li> <li><a href="https://www.customguide.com/cheat-sheet/excel-cheat-sheet.pdf">https://www.customguide.com/cheat-sheet/excel-cheat-sheet.pdf</a></li> </ul> <p><b>SAS</b></p> <ul style="list-style-type: none"> <li><a href="https://documentation.sas.com/?cdcId=pgmsascdc&amp;cdcVersion=9.4_3.3&amp;docsetId=pgmsashome&amp;docsetTarget=home.htm&amp;locale=en">https://documentation.sas.com/?cdcId=pgmsascdc&amp;cdcVersion=9.4_3.3&amp;docsetId=pgmsashome&amp;docsetTarget=home.htm&amp;locale=en</a></li> <li><a href="https://www.tutorialspoint.com/sas/index.htm">https://www.tutorialspoint.com/sas/index.htm</a></li> <li><a href="http://listendata.com/p/sas-tutorials.html">listendata.com/p/sas-tutorials.html</a></li> <li><a href="https://sites.ualberta.ca/~ahamann/teaching/enr480/SAS-Cheat.pdf">https://sites.ualberta.ca/~ahamann/teaching/enr480/SAS-Cheat.pdf</a></li> </ul> <p><b>R</b></p> <ul style="list-style-type: none"> <li><a href="http://statmethods.net/r-tutorial/index.html">statmethods.net/r-tutorial/index.html</a></li> <li><a href="https://www.tutorialspoint.com/r/index.htm">https://www.tutorialspoint.com/r/index.htm</a></li> <li><a href="https://davidalpiaz.github.io/appliedstats/applied_statistics.pdf">https://davidalpiaz.github.io/appliedstats/applied_statistics.pdf</a></li> <li><a href="https://rstudio.com/wp-content/uploads/2016/10/r-cheat-sheet-3.pdf">https://rstudio.com/wp-content/uploads/2016/10/r-cheat-sheet-3.pdf</a></li> </ul>				

Examples	STATA	SPSS	Excel	SAS	R
Summary statistics	Data->Describe Data-> Summary Statistics OR summarize <i>num_var</i> .	Analyze-> Descriptive Statistics -> Descriptives	=AVERAGE( <i>num_var</i> ) =MEDIAN( <i>num_var</i> ) =STDEV.S( <i>num_var</i> ) ...	PROC UNIVARIATE; <i>var num_var</i> ;	summary( <i>num_var</i> )
Histogram	Graphics-> Histogram OR histogram <i>num_var</i>	Graphs -> Chart Builder -> Histogram	Insert (Charts)-> Histogram	PROC SGPLLOT; histogram <i>num_var</i> ;	hist( <i>num_var</i> )
Boxplot	Graphics-> Box plot OR graph box <i>num_var</i> , over( <i>cat_var</i> )	Graphs -> Chart Builder -> Boxplot	Insert (Charts)-> Box and Whisker	PROC SGPLLOT; vbox <i>num_var</i> / group= <i>cat_var</i> ;	plot( <i>num_var</i> , <i>cat_var</i> )
Bar plot	Graphics-> Bar Chart OR graph bar (mean) <i>num_var</i> , over( <i>cat_var</i> )	Graphs -> Chart Builder -> Bar	Insert (Charts)-> Column	PROC SGPLLOT; vbar <i>num_var</i> category= <i>cat_var</i> ; treatment= <i>num_mean</i> ;	means <- c( <i>mean_cat1</i> , <i>mean_cat2</i> ) barplot(means)
Scatterplot	Graphics -> Two-way graph OR twoway (scatter <i>num_var1 num_var2</i> )	Graphs -> Chart Builder -> Scatter/Dot	Insert (Charts)-> Scatter	PROC SGPLLOT; Scatter y= <i>num_var1</i> x= <i>num_var2</i> ;	plot( <i>num_var1</i> , <i>num_var2</i> )
T-test	Statistics -> Summaries, tables, and tests -> Classical tests of hypotheses -> t tests OR ttest <i>num_var</i> , by( <i>cat_var</i> )	Analyze -> Compare means-> Independent-Samples T Test	=TTEST( <i>num_var1</i> , <i>num_var2</i> , <i>tails</i> , <i>type</i> )	PROC TTEST; <i>var num_var</i> ; class <i>cat_var</i> ;	t.test( <i>num_var</i> , <i>cat_var</i> )
ANOVA	Statistics -> Linear models and related -> ANOVA/MANOVA -> One-way ANOVA OR odsby <i>num_var cat_var</i>	Analyze -> Compare means-> One-Way ANOVA	Data Analysis (add-on) -> Anova: Single Factor	PROC ANOVA; class <i>cat_var</i> ; model <i>num_var</i> = <i>cat_var</i> ;	anova( <i>num_var</i> , <i>cat_var</i> )
Normal linear regression model	Statistics -> Linear models and related -> Linear regression OR regress <i>num_var1 num_var2</i>	Analyze -> Regression-> Linear	Data Analysis (add-on) -> Regression	PROC REG; model <i>num_var1</i> = <i>num_var2</i> ;	lm( <i>num_var1</i> ~ <i>num_var2</i> )
Logistic regression model	Statistics -> Binary outcomes-> Logistic regression OR logit <i>binary_var num_var</i>	Analyze -> Regression-> Binary Logistic	N/A	PROC LOGISTIC; model <i>event/trial</i> = <i>num_var2</i> ;	glm( <i>binary_var</i> ~ <i>num_var</i> , family=binomial)
Poisson regression model	Statistics -> Count outcomes-> Poisson regression OR poisson <i>count_var num_var</i>	Analyze -> Regression-> Generalized Linear Models	N/A	PROC GLIMMIX; model <i>count_var</i> = <i>num_var</i> ; /dist=Poisson;	glm( <i>count_var</i> ~ <i>num_var</i> , family=Poisson)
Generalized linear mixed model	Statistics -> Multilevel mixed-effects models -> Generalized linear model OR mixed <i>var1 var2</i>    <i>rand_var_ssn</i> , family( <i>distribution</i> ) link( <i>link_function</i> )	Analyze -> Mixed Models-> Generalized Linear	N/A	PROC GLIMMIX; class <i>cat_var</i> ; model <i>num_var1</i> = <i>num_var2</i> <i>cat_var rand_var</i> ; random <i>rand_var</i> ;	Package lme4 lmer( <i>num_var1</i> ~ <i>num_var2</i> + <i>cat_var</i> + (1   <i>rand_var</i> ))