Choose what you use: Teaching different statistical software at the same time.

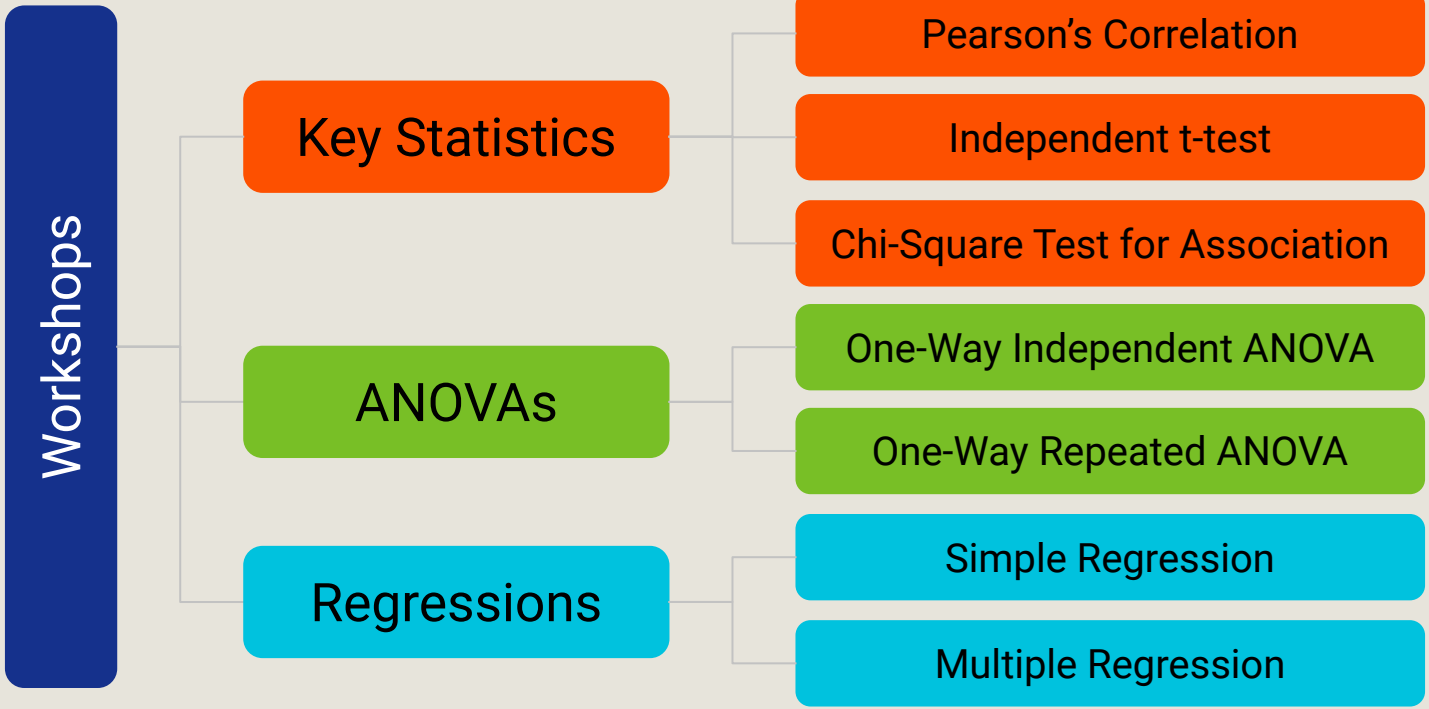
Laura Vinton, Academic Skills Advisor Statistics,

University of York

Context

- Rising number of students using RStudio and STATA compared to SPSS
- Increase resources that we offer
- Comments in appointments such as “I know how to do this in SPSS but my supervisor only knows how to use RStudio.”
- **Focus statistical test workshops on the analysis rather than the software**
- 2 hour workshops to groups of 10-25 online or on campus from any degree level or subject area

Workshops offered



Resources offered

Workshop slides

Cover theory, assumptions, write up and additional information about the statistical tests.

Practical guides

Cover how to run the tests in the software. Including example dataset and code.

One-to-one appointments

For further questions or to discuss their data specifically.

Recordings

Current work in process. Recordings of how to run each test in each software.

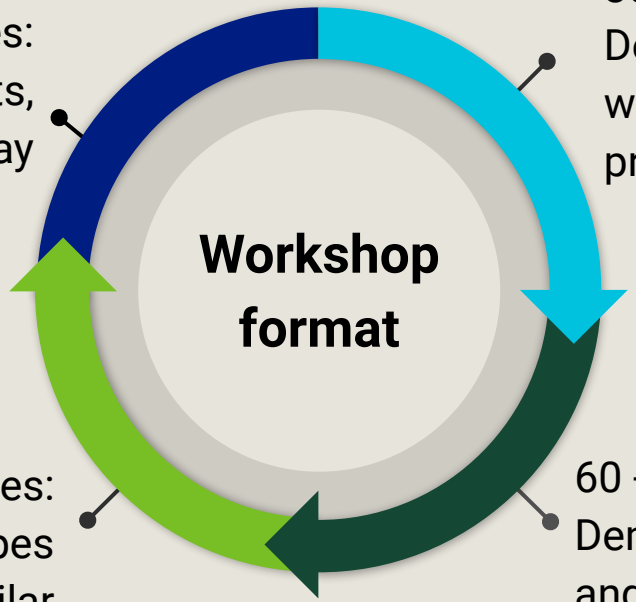
Workshops

Two hours
Offered online or on campus

Workshop format

0 - 30 minutes:
Introduce the tests, assumptions and pathway

30 - 60 minutes:
Demonstrate tests in SPSS, whilst STATA and RStudio prepare



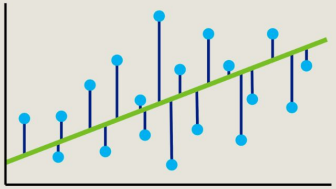
60 - 90 minutes:
Demonstrate tests in STATA and RStudio, whilst the SPSS group practise graphs

90 - 120 minutes:
Writing up results, and other types of tests that are similar

Part 1: Introduction

What can a regression tell us

- Can we reliably predict the values of the outcome variable using the predictor variable?
- Plot a straight line to minimise residual values
  - The difference between the observed value and the predicted



Regression Pathway

Calculate descriptives	Check assumptions	Run regression	Identify significant predictors	Report results
What does the data in your sample look like?	Is your data suitable for a regression?	What does a regression tell us about our data?	Can we dig deeper into the data to look at which of our predictors are significant?	How can we clearly present our results following the appropriate formatting?

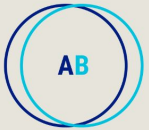
Homoscedasticity

- Variance of residuals should be similar across all of the scores on the continuum.
- Homoscedasticity: similar variance of residuals (errors) across the variable continuum.
- Heteroscedasticity: variance of residuals (errors) differs across the variable continuum.
- Plot fitted values against the standardised residuals

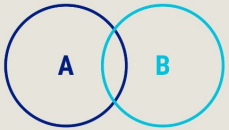
Multicollinearity

- There should be no perfect linear relationship between two or more predictors, else they predict the same thing
- Tolerance statistic should be more than 0.2
- VIF (variance inflation factor) should be less than 5 (low). 5 to 10 is moderate and the variables need to be looked at closely and likely removed. 10 is high and variables should be removed.

High =



Low =

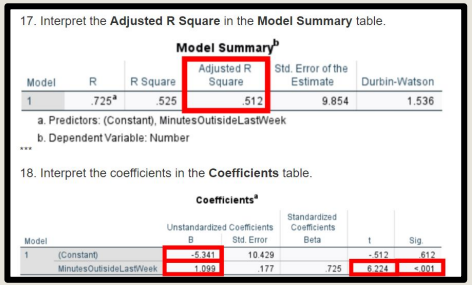


Part 2: First demonstration

Main Demonstration

SPSS Users

- Practical covering importing the data and running the tests
- With our guides if needed



17. Interpret the **Adjusted R Square** in the **Model Summary** table.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.725[a]	.525	.512	9.854	1.536

a. Predictors: (Constant), MinutesOutisideLastWeek

b. Dependent Variable: Number

***

18. Interpret the coefficients in the **Coefficients** table.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-5.341	10.429		-.512	.612
	MinutesOutisideLastWeek	1.099	.177	.725	6.224	<.001

Preparation with GTA

STATA and RStudio Users

- Import data into software
- Install and load any necessary R packages
- Break

Part 3: Second demonstration

Main Demonstration

STATA and RStudio Users

- Practical covering importing the data and running the tests
- Stata .do file and RScript stages line up

Practical with GTA

SPSS Users

- Practice building graphs
- Break

Part 4: Additional tests

Reporting an independent one-way ANOVA

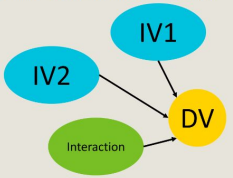
Below is an example of a general format to report result. Please check the correct formatting of results and reporting of statistics for your department and referencing board, (for example APA).

To evaluate if there was a significant difference between students opinion of the wildlife on campus depending on which faculty they belong to an independent one-way ANOVA was performed. The assumption of a normally distributed dependent variable was met for all of the faculties (Arts & Humanities p = .088; Science p = .682; Social Science p = .467). A non-significant Levene's test indicated that the assumption of homogeneity of variance was met (p = .535).

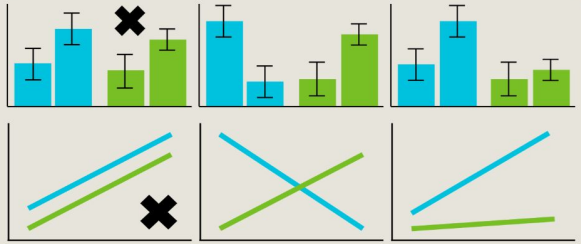
The independent one-way ANOVA indicated that there were significant differences in how much student's from different faculties liked the wildlife on campus (F(2, 81) = 36.83, p<.001). The effect size was large, with a Eta squared of .476. Post-hoc Tukey tests showed that student's from the Science faculty (M = 59.99, SD = 7.06) had significantly better opinions of the wildlife on campus than students from Arts & Humanities (p < .001; M = 44.11, SD = 8.58) and students from Social Sciences (p < .001; M = 43.90, SD = 8.94). There was no significant difference in opinion between Arts & Humanities and Social Sciences students (p = .995).

Two-Way ANOVAs

- Two categorical independent variables and one continuous dependent variable
- Can do three- or more way ANOVAs however you need to consider the theoretical logic to doing this over other statistical tests.
- Can be all independent, all repeated or mixed independent variables.
- For example if you wanted to compare if dogs or cats were faster running 100m both before and after they had been given a treat.

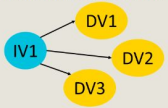


Interaction effects on graphs



MANOVA

- Multivariate Analysis of Variance
- Multiple continuous dependent variables
- For example if you had 3 different football training programs and you wanted to see the difference in players speed, passing, shooting and defending
- Other multivariate tests - such as structural equation modelling - may be used instead



How to manage it

- Stick to timings
- Graduate Teaching Assistant to assist with set up, demonstrations and any questions or technological difficulties
- Select foundation statistics for demonstrations
- Line up steps of the analysis in STATA and RStudio

.do file and RScript

```
////// Pearson's Correlation //////

//Calculate descriptive statistics for your variables
sum(number)
sum(distancewalked)

//Plot the data on a scatter plot to check for linearity
scatter distancewalked number

//Plot the data on box plot to look at the distribution and check for outliers
graph box number
graph box distancewalked

//Check that the dependent variable is normally distributed for both groups of you indpendent
variable by plotting the variables on a histogram and running a Shapiro-Wilk test.
histogram number
histogram distancewalked
swilk number
swilk distancewalked

//Run the Pearson's correlation
pwcorr number distancewalked, sig

//If you would like you could produce a scatter graph to visualise your results.
graph twoway (lfitci number distancewalked) (scatter number distancewalked)
```

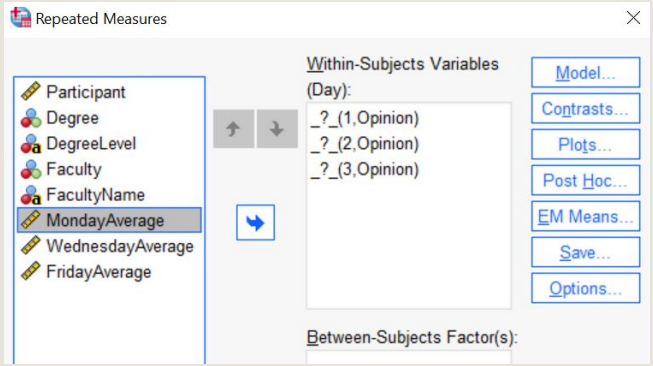
```
### Pearson's Correlation

### If you don't have the following packages installed use install.packages("") to install the
package first, then library() to load the packages.

# Load packages
library(ggplot2)
library(tidyverse)
library(psych)

# Calculate the mean and standard deviation of the variables
describe(data[, c('Number', 'DistanceWalked')], fast = TRUE)

# Plot the data on a scatter plot
ggplot(data, aes(x = DistanceWalked, y = Number)) +
  geom_point()

# Plot the first variable on a box plot
ggplot(data, aes(y = Number)) +
  geom_boxplot()
# Plot the second variable on a box plot
ggplot(data, aes(y = DistanceWalked)) +
  geom_boxplot()

# Create histograms for the variables
ggplot(data = data) +
  geom_histogram(aes(x = Number))
ggplot(data = data) +
  geom_histogram(aes(x = DistanceWalked))

# Run two Shapiro-Wilk tests to check normal distribution
shapiro.test(data$Number)
shapiro.test(data$DistanceWalked)

# Run the Pearson's correlation
cor(data$Number, data$DistanceWalked, method = "pearson")

# Plot the data on a scatter plot
ggplot(data, aes(x = DistanceWalked, y = Number)) +
  geom_point() +
  geom_smooth(method = "lm") +
  theme_classic() +
  labs(x = "Distance walked on campus (meters)", y = "Number of animals")
```

Benefits for students

- Can join in with the practical elements or watch the demonstration and use the complementing guides to try the practical after
- Can watch demonstrations in different software
- May have previously been taught in multiple software, or be being encouraged to a new software by supervisors
- Can compare the different software and make their own decision on which to use

Different methods to run tests

- Coding allowing for easier replication
- Menu-based only include a more concise range of options for an analysis
- Between variations of tests (e.g. independent to repeated) SPSS often requires a different set up, whereas STATA and RStudio often it is the same function with changes to the arguments




```
//Run the Repeated One-Way ANOVA
anova opinion participant day, repeated(day)
```

```
# Repeated Measures One-Way ANOVA
test <- anova_test(data = dataLong, dv = Opinion,
                   wid = Participant, within = Day)
```

Different options for assumptions

- Different software have different methods to check assumptions
- Some check certain assumptions automatically, some you can run additional functions to check and others you can't check

Mauchly's Test of Sphericity[a]

Measure: Opinion

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[b] Greenhouse-Geisser	Epsilon[b] Huynh-Feldt	Epsilon[b] Lower-bound
Day	.997	.272	2	.873	.997	1.000	.500

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. Design: Intercept
Within Subjects Design: Day

b. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

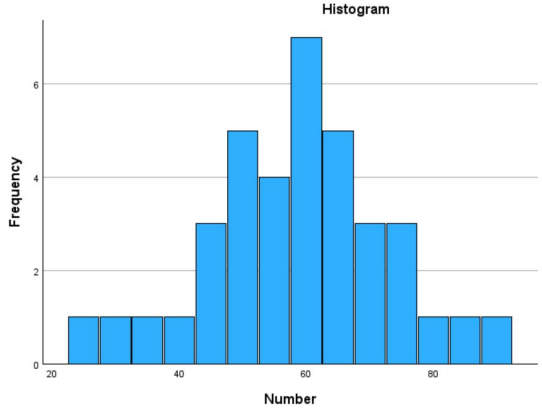
```
> # Check Mauchly's Test for Sphericity
> test$`Mauchly's Test for Sphericity`
  Effect     W     p p<.05
1    Day 0.997 0.873
```

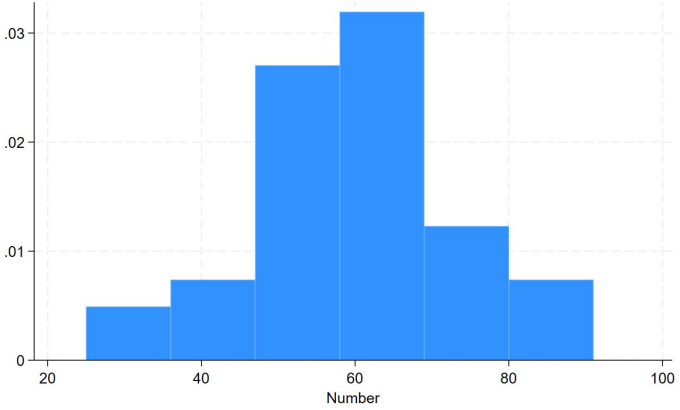
```
Huynh-Feldt epsilon        =  1.0212
*Huynh-Feldt epsilon reset to  1.0000
Greenhouse-Geisser epsilon =  0.9967
Box's conservative epsilon =  0.5000
```

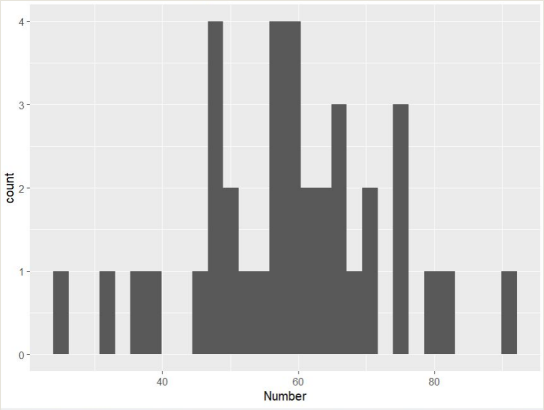
Source	df	F	Prob > F Regular	H-F	G-G	Box
day	2	24.89	0.0000	0.0000	0.0000	0.0000
Residual	166					

Different outputs

- SPSS generally gives you more information, not all of which is relevant
- STATA and RStudio sometimes you have to request certain outputs
- Default outputs look different between the softwares



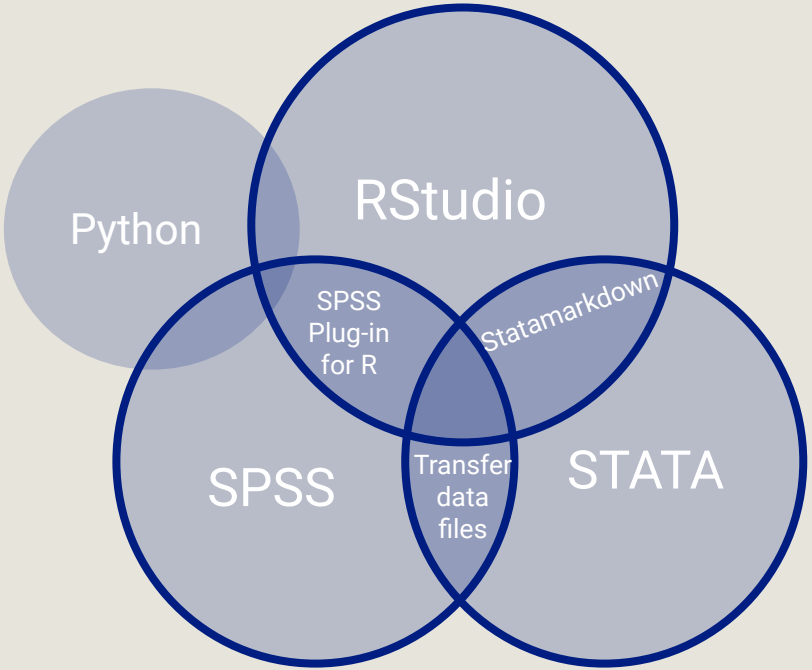




Different Access

	SPSS	STATA	RStudio
Number of file types accepted	7	14	Many packages for even niche file types
Data format	Wide	Wide or long	Wide or long
Number of data sets	SPSS can have multiple open in different editor windows where one is set as active.	STATA can only have multiple open in multiple STATA windows.	RStudio can have multiple open in the same window.
Software format	Menu-based	Menu or coding	Coding
Requires paid licence	Yes	Yes	No
Screen reader accessibility	Can work with JAWS however found to be inaccessible.	Can work with JAWS however reads content as plain text.	Can work with screen readers. BrailleR, tactileR, and sonifyR.

Increasingly integrated software



Pitfalls as a teacher

- Control + D in STATA runs the code, whereas in RStudio it deletes the code
- Control + Enter in RStudio runs the code, whereas in STATA it deletes the code
- Preparation is key
- Time management can be tricky
- At the mercy of the software working smoothly
- Students do come when using other software - requires GTA to look up resources
- Students coming to sessions unfamiliar with the basics of the software

Student feedback

- “Thank you very much, the staff was very good, **Helpful, and easy to follow.**”
- “**Good introduction to the topic and practical element to practice** with a toy dataset for all the tests we went through.”
- “Explanations of the models were clear with great attention to detail. She also answered all of my questions very well and made the models far easier to explain. The second half of the workshop was just as good and **I learnt a lot about using Stata** that I will certainly take forward into my projects.”
- “Thank you for running the session, it was **excellent**!”
- “Having the workshop be for the different stats programs all at the same time meant that **some parts of the session was not as relevant for others**. The technical difficulty that they had was unlucky and affected the progress we made to address some additional learning about the **other versions of multiple regression (forward, stepwise, etc.), which would have been super useful**.”
- “**Pace** - this was the first workshop of its kind and we went over + rushed towards the end (very understandable). **Might take several tries to find the balance between the initial slides and practical aspect.** It might be that this would suit a 2.5 hour session”

Conclusion

- Opportunities to expand teaching offer and resources
- Allow students to decide which software they prefer and to focus on the statistical analysis rather than the software
- However workshops need to be well prepared for and supported with additional resources in case of timing or software issues