INTERNSHIP PRESENTATION

INTERNSHIP PRESENTATION

Biometry Hub Internship

Alec McCallum

24/07/2020

ABOUT ME

WEEK 1

- Intro to R
- Intro to Experimental Design
- Talk: Russel (Functions and Programming)
- Talk: Pete (Tidyverse)
- Talk: Sam (Rmarkdown)

INTRO TO R

- Rstudio console
- R Basics
- Variables, vectors, data frames
- More advanced functions
- Graphics

INTRO TO R

Key Takeaways

- Always have good record keeping and data management, keep original files, back up your files
- Google everything you dont know
- Always be very accurate when typing, capitals and punctuation matter

INTRO TO EXPERIMENTAL DESIGN

- Completely Randomised Design
- Randomised Complete Block Design
- Latin Square Design
- Factorial RCBD

INTRO TO EXPERIMENTAL DESIGN

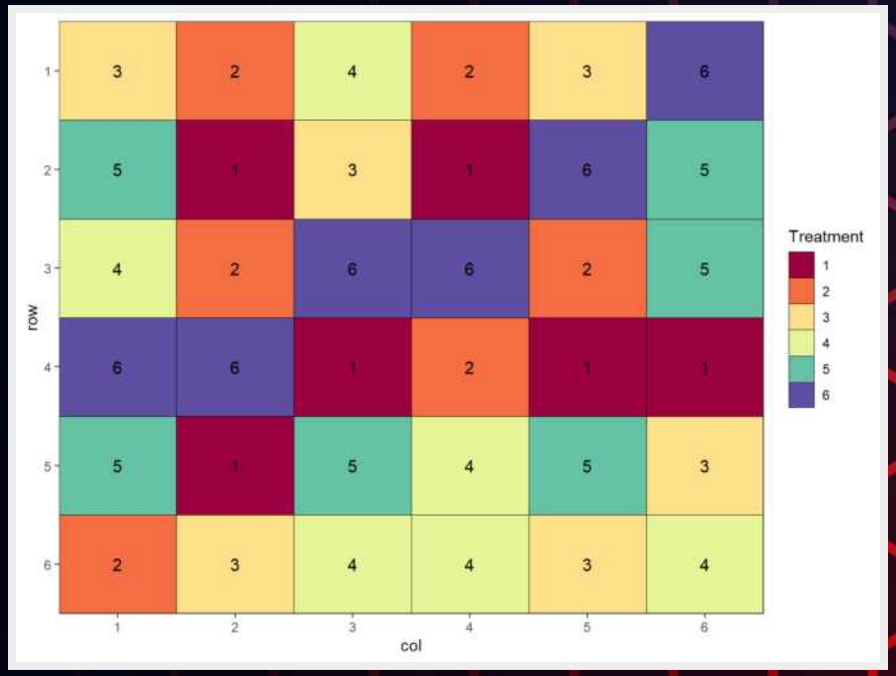
Completely Randomised Design

```
#aim- Yield response to N fertiliser
#obs- 36 plots
#arr- 6 rows x 6 col
#trt- 6
#rep- 6
#des- CRD
#blk- NA

trt<-c(1:6)
rep<-6

outdesign<-design.crd(trt,rep)
des.out<-des.info(design.obj = outdesign,
                  nrows=6,ncols=6)
```

INTRO TO EXPERIMENTAL DESIGN

Completely Randomised Design

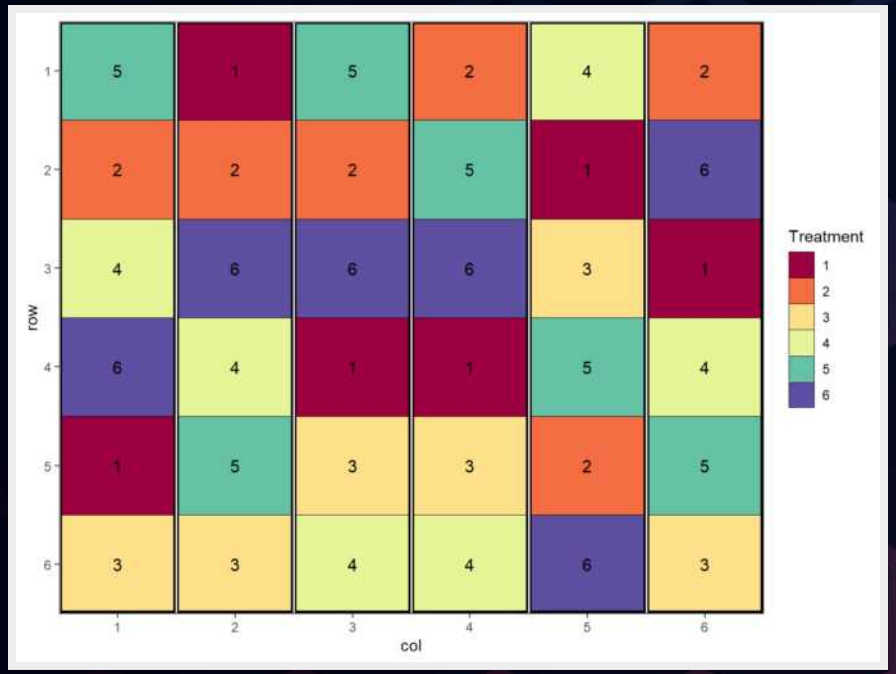
Source of Variation	df
trt	5
Residual	30
Total	35



INTRO TO EXPERIMENTAL DESIGN

Randomised Complete Block Design

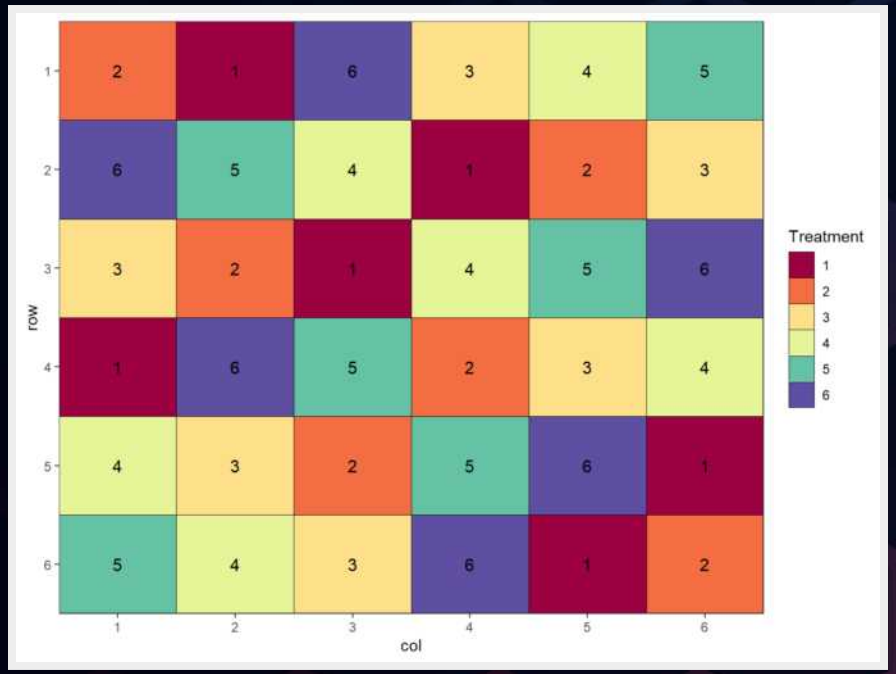
Source of Variation	df
Block stratum	5
trt	5
Residual	25
Total	35



INTRO TO EXPERIMENTAL DESIGN

Latin Square Design

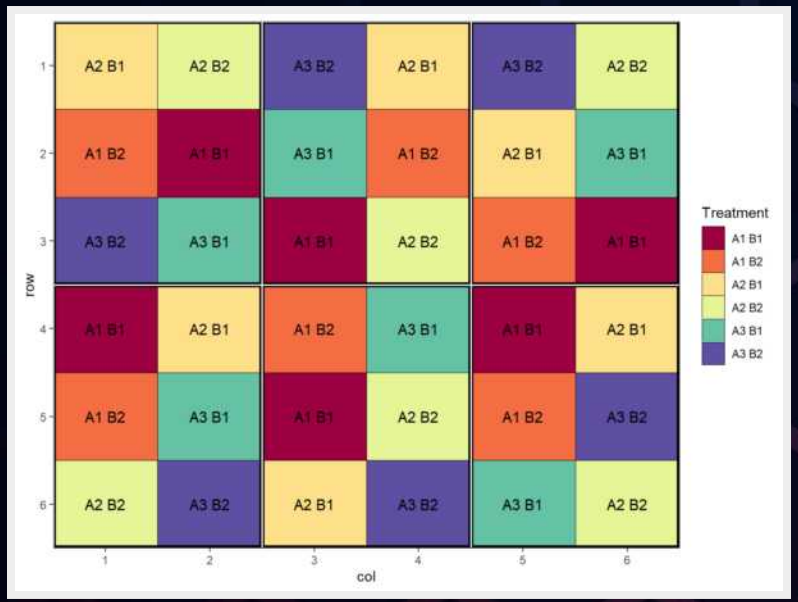
Source of Variation	df
Row	5
Column	5
trt	5
Residual	20
Total	35



INTRO TO EXPERIMENTAL DESIGN

Factorial RCBD

Source of Variation	df
Block stratum	5
A	2
B	1
AB	2
Residual	25
Total	35



INTRO TO EXPERIMENTAL DESIGN

Key Takeaways

- For every design write down: Aim, Observations, Arrangement, Treatments, Replicates, Design, Blocking Arrangement. It helps with creating the right design.
- A bad design can mean the data cannot be analysed accurately
- Look at residual degrees of freedom when choosing a design. Residual df should be greater than 12. Increasing the complexity of the design reduces the residual df.
- Keep in mind the constraints of the trial to get the most appropriate design, e.g. spatial trend in a field

TALK: RUSSEL (FUNCTIONS AND PROGRAMMING)

- How to write functions in R
- How to manage complexity
- Euler Problems

EULER PROBLEM

```
# The four adjacent digits in the 1000-digit number that have
# the greatest product are 9 x 9 x 8 x 9 = 5832.

# Find the thirteen adjacent digits in the 1000-digit number that
# have the greatest product.
# What is the value of this product?

options(scipen = 999)

value<-list()
value[[1]]<-
        c("731671765313306249192251196744265747423553491949349698352031

value[[2]]<-
        c("622298934233803081353362766142828064444866452387493035890729

value[[3]]<-
```

```
[1] 23514624000
```

EULER PROBLEM

```
# A Pythagorean triplet is a set of three natural numbers,
# a < b < c, for which, a^2 + b^2 = c^2
# For example, 3^2 + 4^2 = 9 + 16 = 25 = 5^2.
# There exists exactly one Pythagorean triplet for
# which a + b + c = 1000.
# Find the product abc.

# total and range of values
total<-1000
range<-c(1:total)

# function to find triplets
triplet<- function(A,B) {
    C<-sqrt(A^2+B^2)
    if((C%%1)==0){
    return(c(A,B,C))
    }
```

```
[1] 200 375 425
[1] 31875000
```

```
[1] TRUE TRUE TRUE
```

TALK: PETE (TIDYVERSE)

- Useful for data management and cleaning to get the data into a table with the proper layout needed for analysis
- Easier to use and more versatile than Excel functions
- Don’t try to memorise them all, just know where to find them
- e.g. pipe, gather, spread, separate, filter, arrange, select, group by, mutate, etc.

TALK: SAM (RMARKDOWN)

Compared to Word:

- More fiddly and less user friendly when starting out
- Much easier to get consistent formatting throughout the document
- Easier to include plots and tables

MORE KEY TAKEAWAYS

- Google Everything!
- Break complicated problems into simple, easy steps
- Don’t need to memorise every function, just know that there is a function for almost everything and use Google to find it
- Important to write notes and comments in the code so you know what you did 3 months later and someone else can figure out what you did

WEEK 2

- Workbook 10: Genstat -> R
- Meeting: Hotdesk
- Talk: Pete (ggplot)
- Stats PD @ Waite
- Talk: Wendy (Exact Permutation Tests)
- Talk: Mario (Bioinformatics)
- Talk: Mexiuan (Honours)
- Demonstration: Pete (Drones & Machine Learning)
- Talk: Sam (CV & Website)
- Website

WORKBOOK 10: GENSTAT -> R

- I found R is a bit easier to use than Genstat, there are way more resources online for R
- Took some time to get the right code

WORKBOOK 10: GENSTAT -> R

```
            Df Sum Sq Mean Sq F value Pr(>F)
diet         5   4613   922.6     4.3 0.0023 **
Residuals   54  11586   214.6
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

R ANOVA 1

Analysis of variance

Variate: Gain

Source of variation	d.f.	s.s.	m.s.	v.r.	F pr.
Diet	5	4612.9	922.6	4.30	0.002
Residual	54	11586.0	214.6		
Total	59	16198.9			

Genstat ANOVA 1

WORKBOOK 10: GENSTAT -> R

	Gain	groups
Beef High	100.0	a
Pork High	99.5	a
Cereal High	85.9	b
Cereal Low	83.9	b
Beef Low	79.2	b
Pork Low	78.7	b

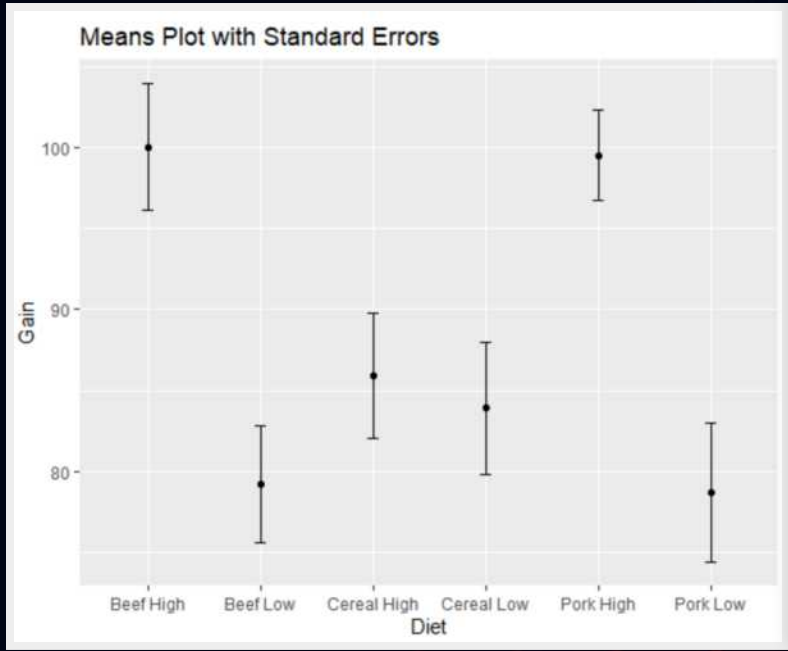
R LSD

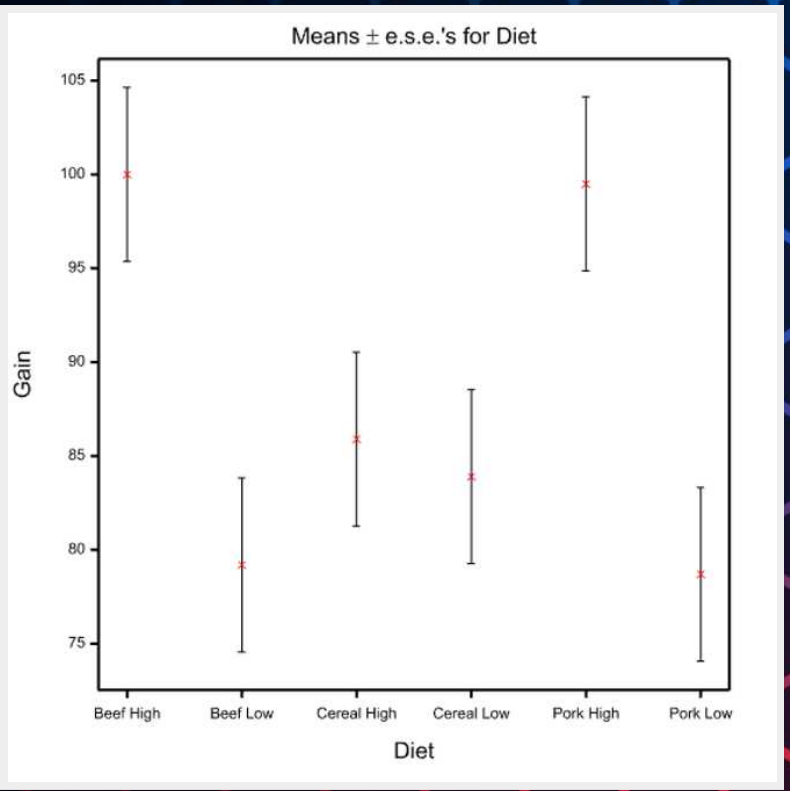
Fisher's unprotected least significant difference test
Diet

	Mean	
Pork Low	78.70	a
Beef Low	79.20	a
Cereal Low	83.90	a
Cereal High	85.90	a
Pork High	99.50	b
Beef High	100.00	b

Genstat LSD

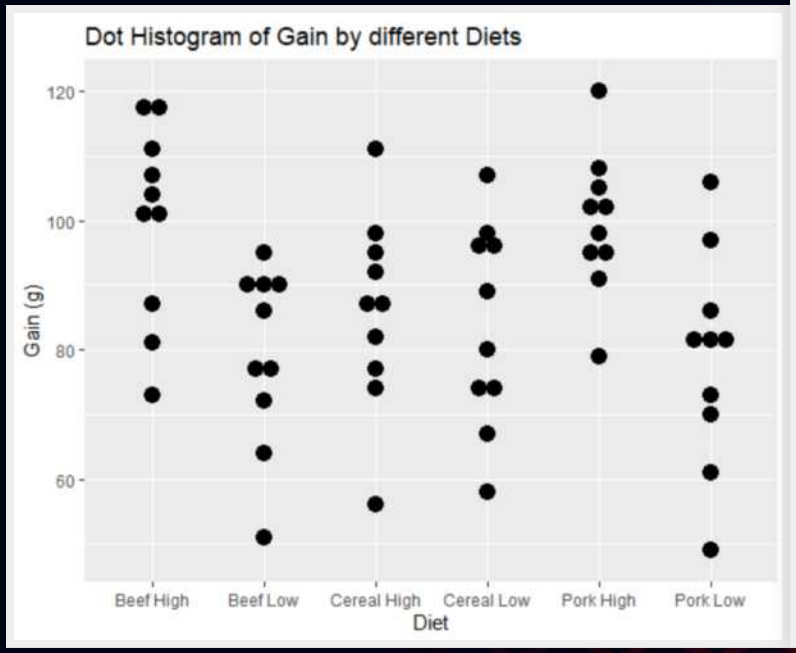
WORKBOOK 10: GENSTAT -> R

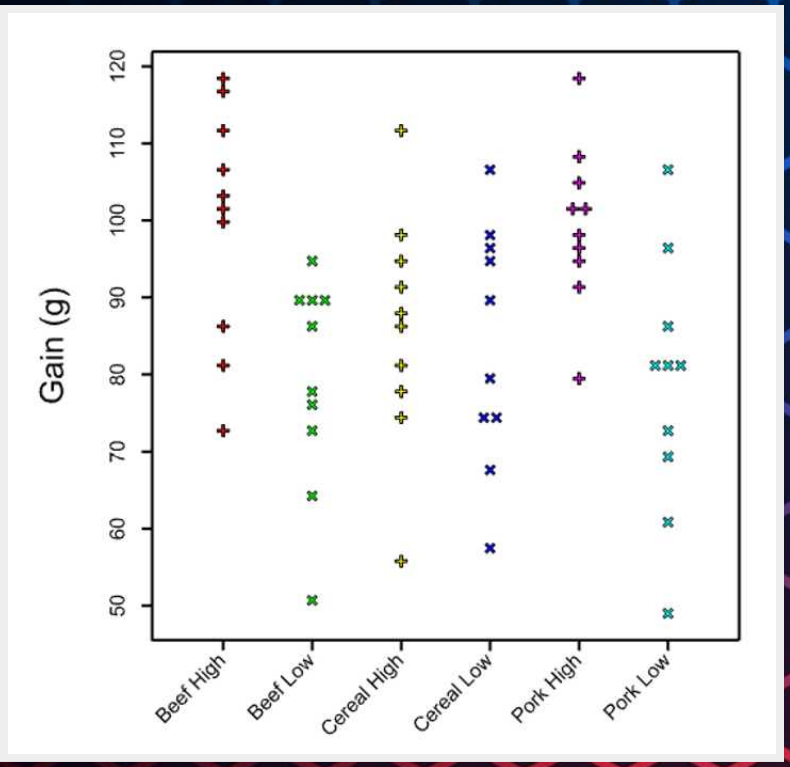




R Means Plot vs Genstat Means Plot of weight gain in rats with different diets

WORKBOOK 10: GENSTAT -> R





R Dotplot vs Genstat Dotplot of weight gain in rats with different diets

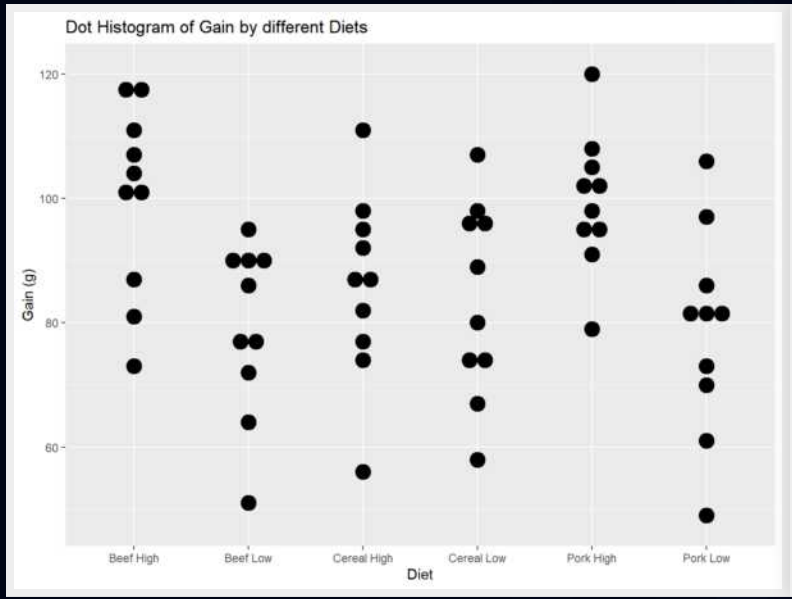
MEETING: HOTDESK

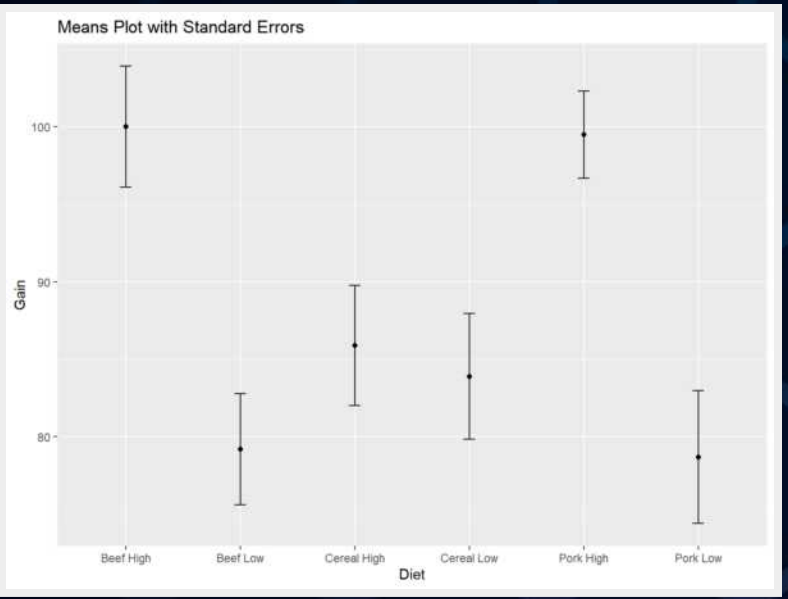
- Have a clear idea of what the experiment is, the goal and the limitations
- Consult with a statistician to make sure the design is right, they ask questions about things you haven’t thought about

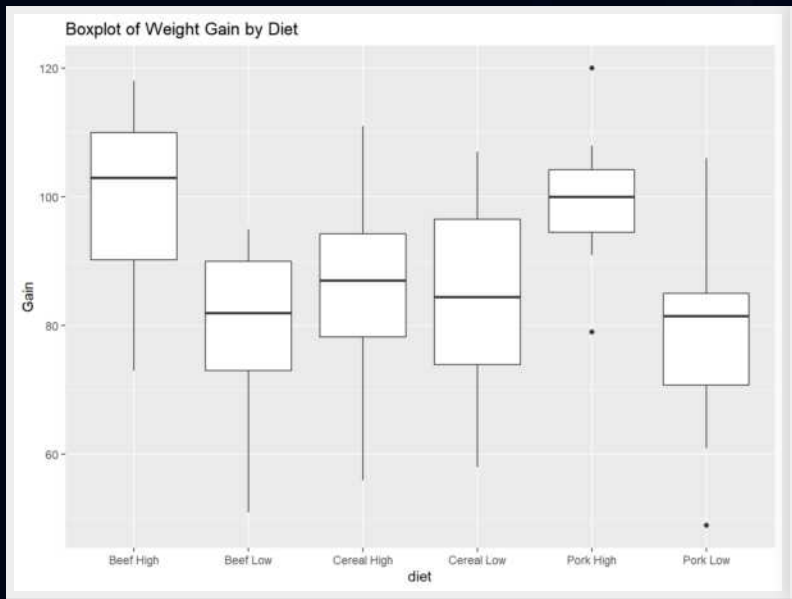
TALK: PETE (GGPLOT)

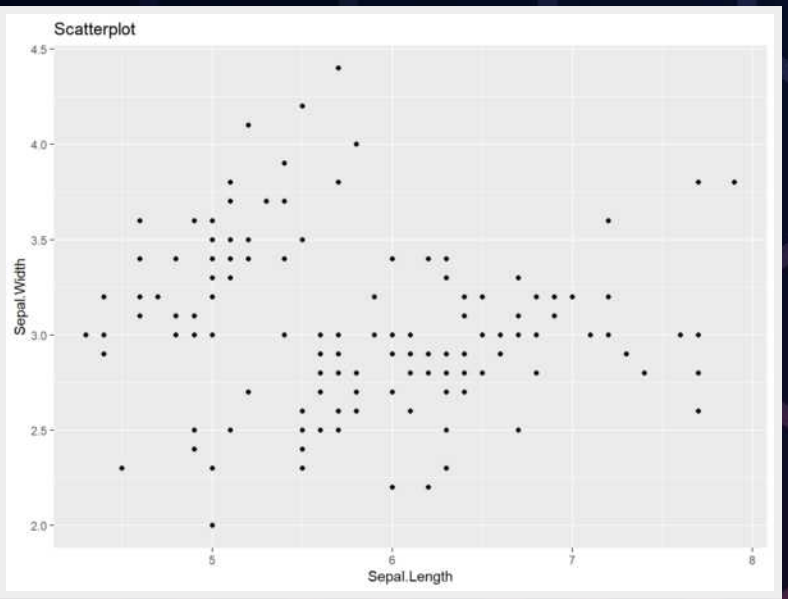
- hard to use at first
- more versatile than excel but a bit less user friendly
- easier to import into documents with rmarkdown

TALK: PETE (GGPLOT)









STATS PD @ WAITE

Takeaways

- Find all sources of errors
- Understand the whole structure
- Use simplified examples for better understanding
- Analyse in different ways and compare outputs for differences or consistency
- Consult an expert

TALK: WENDY (EXACT PERMUTATION TESTS)

- Take every permutation of the responses and assign them to experimental units
- Calculate a test statistic for each permutation
- Create a distribution from these permutations
- Take all the permutations that have a test statistic equal to or more extreme than the observations and determine how likely the observations are to occur
- Makes no assumptions of the underlying distribution of observations
- Can use any test statistic

TALK: MARIO (BIOINFORMATICS)

- Exists as a link between biology and statistics therefore need to have a good understanding of the biology and the statistics behind the experiments
- Determine which genes have a statistically higher or lower gene expression in a treatment compared to a control
- Function of these genes and any genes related to this gene also need to be determined
- This information directly helps the biologist by telling them which genes to study further

TALK: MEIXUAN (HONOURS)

- Find the statistical model that best describes canola seedling emergence
- Time management

DEMONSTRATION: PETE (DRONES + MACHINE LEARNING)

- Using drones and machine learning to count and map Faba bean seedling emergence
- Same concept can apply to different crops
- Real world applications for monitoring crops objectively, not relying on observations in one corner of the paddock

TALK: SAM (CV AND WEBSITE)

- Sell yourself
- Keep online profiles consistent and up-to-date

www.alecmccallum.netlify.com

KEY TAKEAWAYS

- ggplot is way more versatile than excel
- For future projects: Have a clear idea of what the project is, its limitations and constraints and how big the project will be. That will make it easier to design.
- There is a huge variety in areas of statistics with real-world applications

WEEK 3

- Talk: Beata (Genetic Association Analysis)
- Talk: Paul (Personal Experiences)
- Meeting: Olena (Honours)

TALK: BEATA (GENETIC ASSOCIATION ANALYSIS)

- Determining which markers/SNPs have a significant effect on a trait
- Helps researchers know what genes to do more research on
- Helps breeders in marker-assisted and genomic selection
- Related well to the Plant Breeding course

TALK: PAUL (PERSONAL EXPERIENCES)

Mistakes from researchers

- Not designing the experiment properly
- Trying to make the data analysis fit their preconceived idea
- Getting help on the analysis just before the due date because the experiment didn’t work

MEETING: OLENA (HONOURS)

- Doing Honours will make me more prepared for a job and more competitive in the job market
- The agricultural industry wants and needs people to be trained in data management and analytics
- Continually learn and develop new skills

REFLECTION

Skills

- Coding in R
- Engaging with speakers
- Problem Solving
- Self-motivation

THANKS FOR LISTENING :)

Here is my puppy, Fred.