Chapter 1

Mark Farrell

09/10/2018

Chapter 1 starts on t-tests then simple linear models for the wheat dataset from agridat.

Part 1 - T-tests and simple ANOVA

First task is to sort data and get it ready for analysis:

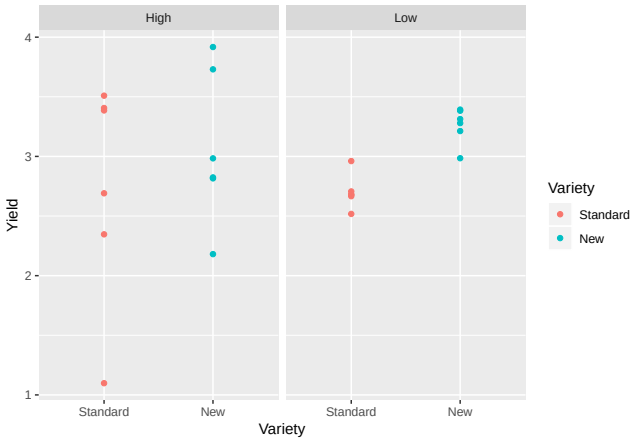
```
setwd("~/DATASCHOOL/r-learning/stats-terry")
wheat <- read_csv("data/working/wheat yield.csv")
```

```
## Parsed with column specification:
## cols(
##   ExpID = col_integer(),
##   PlotID = col_integer(),
##   Variation = col_character(),
##   Variety = col_character(),
##   Yield = col_double()
## )
```

```
as_factor(wheat$Variety)
wheat$Variety <- factor(wheat$Variety, levels = c("Standard", "New"))
wheat_H <- filter(wheat, Variation == "High")
wheat_L <- filter(wheat, Variation == "Low")
```

A simple plot lets us see the data

```
ggplot(wheat, aes(Variety, Yield, colour=Variety)) +
  geom_point() +
  facet_wrap(~Variation)
```



Now data is all sorted and we've had a peek at what it looks like, simple t-tests run easily.

```
t.test(Yield~Variety, data = wheat_H, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  Yield by Variety
## t = -0.7282, df = 10, p-value = 0.4832
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.3635528  0.6918159
## sample estimates:
## mean in group Standard      mean in group New
##               2.739381               3.075249
```

```
t.test(Yield~Variety, data = wheat_L, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  Yield by Variety
## t = -6.5726, df = 10, p-value = 6.291e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.7485274 -0.3695078
## sample estimates:
## mean in group Standard      mean in group New
##               2.701900               3.260917
```

As can be seen, there's no significant effect of Variety in the high variance half of the dataset, but in the low variance half, $P << 0.05$

Given that t-tests are a bit pasé, and also ultimately inappropriate for a dataset like this (noting trhere are really two factors), we'll explore linear modelling.

```
lm1 <- lm(Yield ~ Variety, data = wheat_L)
anova(lm1)
```

```
## Analysis of Variance Table
##
## Response: Yield
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Variety    1 0.93750  0.9375  43.199 6.291e-05 ***
## Residuals 10 0.21702  0.0217
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here we see that for the low variance half of the dataset, there is a significant effect of variety. To look closer at the fitted model, we'll run `summary(lm1)` to give info on the residuals and coefficients, goodness of fit, etc. and `emmeans(lm1, ~Variety)` to provide output of estimated marginal means of model.

```
summary(lm1)
```

```
##
## Call:
## lm(formula = Yield ~ Variety, data = wheat_L)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.275981 -0.038703 -0.008592  0.069784  0.258975
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.70190    0.06014  44.926 7.18e-13 ***
## VarietyNew   0.55902    0.08505   6.573 6.29e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1473 on 10 degrees of freedom
## Multiple R-squared:  0.812,  Adjusted R-squared:  0.7932
## F-statistic:  43.2 on 1 and 10 DF,  p-value: 6.291e-05
```

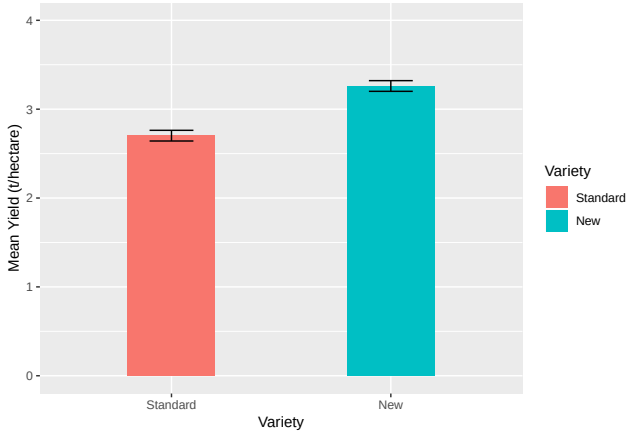
```
emmeans(lm1, ~Variety)
```

```
##  Variety    emmean         SE df lower.CL upper.CL
##  Standard 2.701900 0.06014152 10 2.567896 2.835903
##  New      3.260917 0.06014152 10 3.126914 3.394921
##
## Confidence level used: 0.95
```

The final task is to plot the output of the model. This took a bit of wrangling, but a passable ggplot output solution is presented below:

```
lm1.results<-summary(emmeans(lm1,~Variety))

ggplot(lm1.results, aes(Variety, emmean, fill=Variety)) +
  geom_bar(stat="identity", width=.4) +
  geom_errorbar(aes(ymin = emmean-SE, ymax = emmean+SE), width=.2) +
                  ylim(0,4) +
                  labs(y = 'Mean Yield (t/hectare)')
```



Going though the structure of this ggplot command, a few things are happening:

1. First we make a new data frame from the summary outputs of `emmeans`
2. This is then plotted with Variety on the x-axis and emmean on the y-axis, which is the output of the model. Fill is purely for prettiness
3. `stat="identity"` tells R to use the y values for the height of the bars
4. Error bars are calculated from the emmean +/- SE within the `lm1.results` data frame
5. Y axis limits are imposed
6. The proper y-axis label is then added. Note this was very particular and didn't like double quotes. There's lots of info on getting sub/superscript, symbols, etc. on axis titles available. Best searched on a case-by-case basis.

==========

Part 2 - Multi-factor ANOVA

```
setwd("~/DATASCHOOL/r-learning/stats-terry")
wheat2 <- read_csv("data/working/wheat yield PLUS.csv")
```

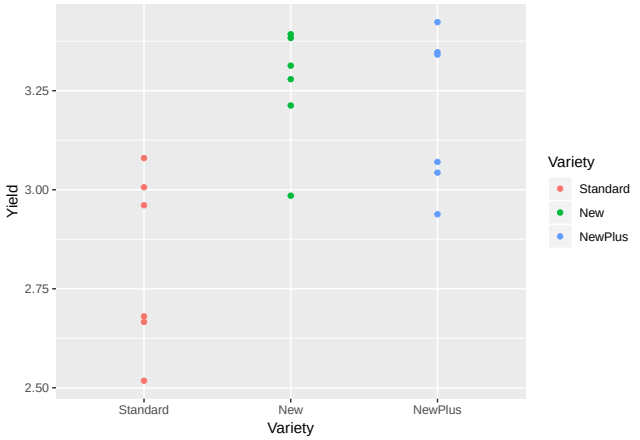
```
## Parsed with column specification:
## cols(
##   X1 = col_integer(),
##   PlotID = col_integer(),
##   Variety = col_character(),
##   Yield = col_double()
## )
```

```
str(wheat2)
```

Then, we force "Standard" to be plotted first, with the other varieties following, and do the plot

```
wheat2$Variety <- factor(wheat2$Variety, levels = c("Standard", "New", "NewPlus"))

ggplot(wheat2, aes(Variety, Yield, colour=Variety)) +
  geom_point()
```



Next, a simple ANOVA to test for significance of difference in Yield between the three Varieties

```
lm2 <- lm(Yield ~ Variety, data = wheat2)
anova(lm2)
```

```
## Analysis of Variance Table
##
## Response: Yield
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## Variety    2 0.68203 0.34101  8.9513 0.002764 **
## Residuals 15 0.57145 0.03810
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(lm2)
```

```
##
## Call:
## lm(formula = Yield ~ Variety, data = wheat2)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.30081 -0.14751  0.03522  0.14616  0.26121
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.81857    0.07968  35.372 7.26e-16 ***
## VarietyNew      0.44235    0.11269   3.925  0.00135 **
## VarietyNewPlus  0.37529    0.11269   3.330  0.00457 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1952 on 15 degrees of freedom
## Multiple R-squared:  0.5441, Adjusted R-squared:  0.4833
## F-statistic: 8.951 on 2 and 15 DF,  p-value: 0.002764
```

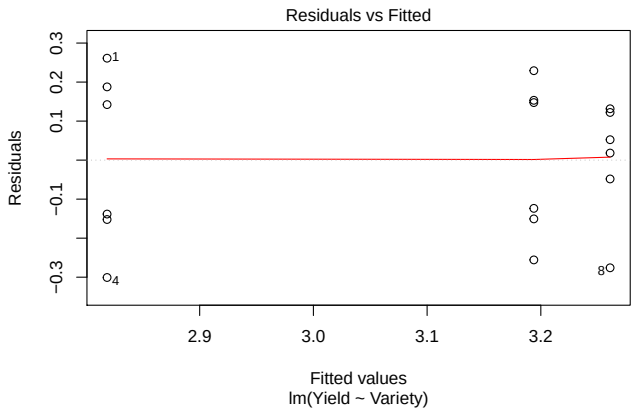
```
emmeans(lm2, pairwise~Variety)
```

```
## $emmeans
##  Variety    emmean         SE df lower.CL upper.CL
##  Standard 2.818566 0.07968335 15 2.648725 2.988407
##  New      3.260917 0.07968335 15 3.091076 3.430758
##  NewPlus  3.193854 0.07968335 15 3.024013 3.363695
##
## Confidence level used: 0.95
##
## $contrasts
##  contrast              estimate        SE df t.ratio p.value
##  Standard - New     -0.44235095 0.1126893 15  -3.925  0.0036
##  Standard - NewPlus -0.37528787 0.1126893 15  -3.330  0.0120
##  New - NewPlus       0.06706308 0.1126893 15   0.595  0.8248
##
## P value adjustment: tukey method for comparing a family of 3 estimates
```

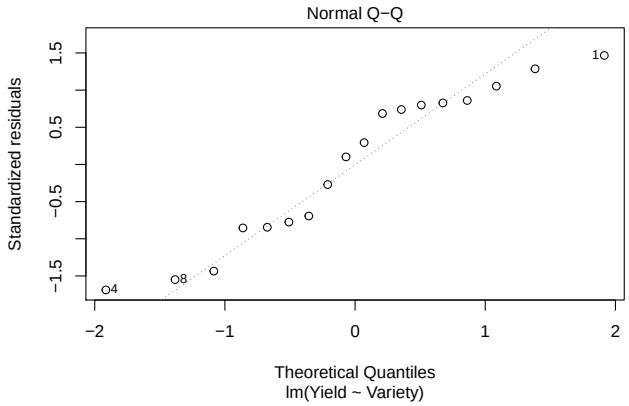
Here, because we have three factors, we ask `emmeans` to conduct Tukey's pairwise comparisons with the `pairwise` function in front of the `~Variety`. We can see that both New and NewPlus differ from Standard, but not from each other.

Next is a quick eyeball of the residuals:

```
plot(lm2,which=1)
```




```
plot(lm2,which=2)
```



As all looks good, it's on to the final masterpiece, code slightly modified from above:

```
lm2.results<-summary(emmeans(lm2,~Variety))

ggplot(lm2.results, aes(Variety, emmean, fill=Variety)) +
  geom_bar(stat="identity", width=.4) +
  geom_errorbar(aes(ymin = emmean-SE, ymax = emmean+SE), width=.2) +
  ylim(0,4) +
  labs(y = 'Mean Yield (t/hectare)')
```