ST517-HW1

Ben Tankus

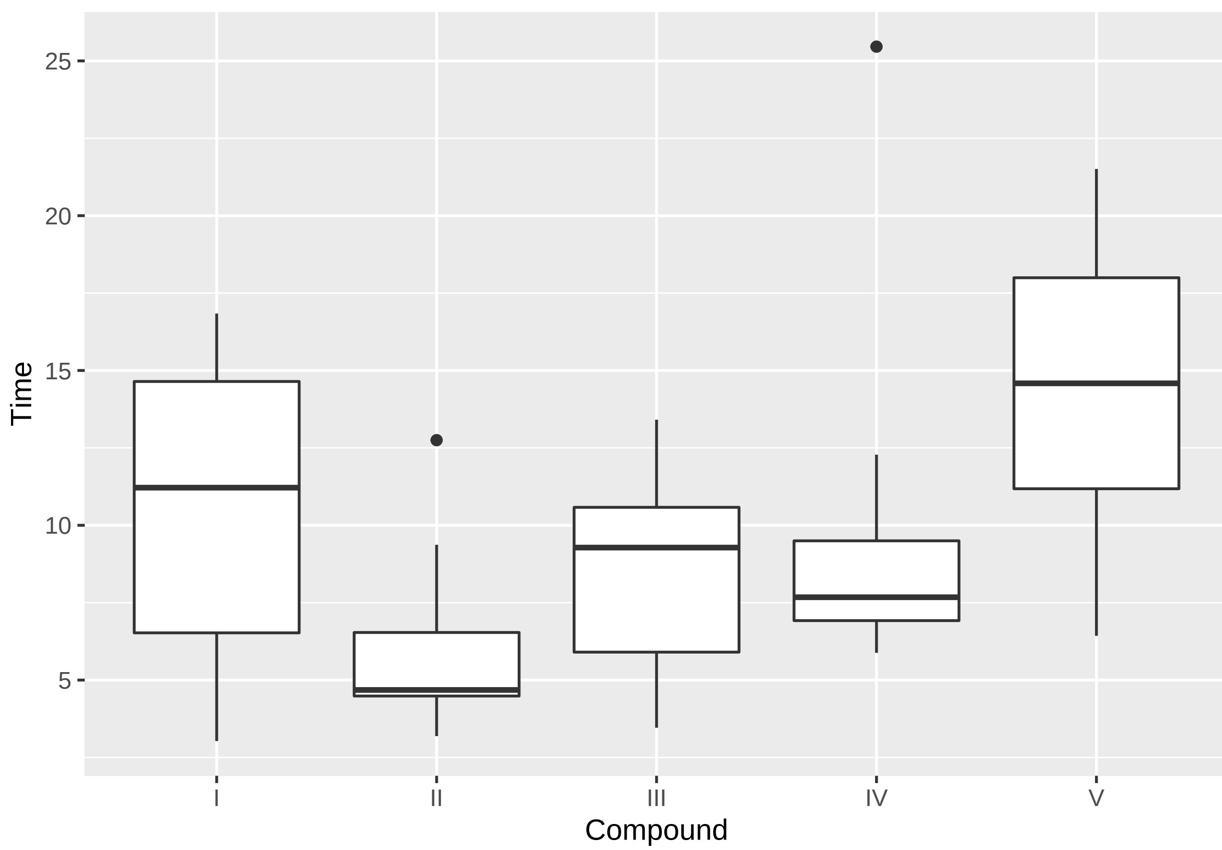
1. (3 points) (Adapted from Exercise 21, Section 6.7, The Statistical Sleuth, 2nd Ed.) The dataset bearings.csv contains failure times (measured in millions of cycles) of engine bearings made from five different compounds.

(a) (1 point) Read in the data. What type of data object is bearings? With ggplot2, create side-by-side boxplots of the failure times by compound.

```
bearing <- read.csv(file = 'bearings.csv')
print(paste("data is stored in a", typeof(bearing)))
```

```
## [1] "data is stored in a list"
```

```
qplot(Compound, Time , data = bearing, geom = 'boxplot')
```



```r
#length(bearing$Time[bearing$Compound == 'III'])
```

(b) (2 points) Determine the pairs of engine ball bearing compounds for which there is a significant difference in mean failure times. Present your findings in a short statistical report (4 sentences).

$$H_0 : \mu_I = \mu_{II}, \mu_I = \mu_{III}, \mu_I = \mu_{IV}, \mu_I = \mu_V ... for\ all\ pairs$$

$$H_A : \mu_I \neq \mu_{II}, \mu_I \neq \mu_{III}, \mu_I \neq \mu_{IV}, \mu_I \neq \mu_V ... for\ any\ pairs$$

```r
# Equal var, false. paired = false, independant = true, outliers = true, equal N = True
# METHOD - Use Tukey Kramer adjustment as we are doing all pairwise comparisons
fit <- aov(Time ~ Compound, data = bearing)
summary(fit)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## Compound     4  401.3  100.32    5.02 0.00197 **
## Residuals   45  899.2   19.98
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
TukeyHSD(fit)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Time ~ Compound, data = bearing)
##
## $Compound
##              diff         lwr       upr     p adj
## II-I   -4.6430001 -10.3234907  1.037490 0.1567083
## III-I  -2.0570003  -7.7374909  3.623490 0.8406451
## IV-I   -0.8950002  -6.5754908  4.785490 0.9914058
## V-I     4.0129998  -1.6674907  9.693490 0.2791467
## III-II  2.5859998  -3.0944908  8.266490 0.6964484
## IV-II   3.7479999  -1.9324907  9.428490 0.3453895
## V-II    8.6559999   2.9755094 14.336491 0.0007521
## IV-III  1.1620001  -4.5184905  6.842491 0.9771836
## V-III   6.0700001   0.3895095 11.750491 0.0308937
## V-IV    4.9080000  -0.7724905 10.588491 0.1195580
```

Significant V - II & V - III

There were two comparisons that reject the null hypothesis of equal means, V - II (p-val = 0.0008), and V - III (p-val = 0.031). Looking at the Box Plot, one would expect more pairs with statistical significance (V-IV, IV-II), but the high variance of V and the outlier of IV make comparisons less likely to be rejected. ANOVA procedures are not robust to outliers, and a high variance makes a sample much less likely to reject the null hypothesis. Due to the manufacturing quality nature of this testing and the strength of the outlier in IV, I reccomend removing the outlier in IV and retesting that sample to improve accuracy of the test.

2. (3 points) Using the data from the last weeks lab and homework, case0501 in the Sleuth3, answer the question "Which diets differ in their mean lifetime?"

```
dietData <- case0501
# METHOD - Use Tukey Kramer adjustment as we are doing all pairwise comparisons
#head(dietData)
fitDiets <- aov(Lifetime ~ Diet,data = dietData)

diets <- TukeyHSD(fitDiets)

diets
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Lifetime ~ Diet, data = dietData)
##
## $Diet
##                    diff        lwr         upr     p adj
## N/R40-N/N85  12.4254386   8.885436  15.9654413 0.0000000
## N/R50-N/N85   9.6059550   6.202170  13.0097399 0.0000000
## NP-N/N85     -5.2891873  -9.017748  -1.5606269 0.0008380
## R/R50-N/N85  10.1944862   6.593417  13.7955556 0.0000000
## lopro-N/N85   6.9944862   3.393417  10.5955556 0.0000008
## N/R50-N/R40  -2.8194836  -6.175736   0.5367684 0.1564608
## NP-N/R40    -17.7146259 -21.399845 -14.0294069 0.0000000
## R/R50-N/R40  -2.2309524  -5.787127   1.3252222 0.4684413
## lopro-N/R40  -5.4309524  -8.987127  -1.8747778 0.0002306
## NP-N/R50    -14.8951423 -18.449713 -11.3405719 0.0000000
## R/R50-N/R50   0.5885312  -2.832070   4.0091319 0.9963976
## lopro-N/R50  -2.6114688  -6.032070   0.8091319 0.2460200
## R/R50-NP     15.4836735  11.739756  19.2275913 0.0000000
## lopro-NP     12.2836735   8.539756  16.0275913 0.0000000
## lopro-R/R50  -3.2000000  -6.816968   0.4169683 0.1167873
```

Diets that Differ:

N/R40-N/N85 12.4254386 8.885436 15.9654413 0.0000000
N/R50-N/N85 9.6059550 6.202170 13.0097399 0.0000000
NP-N/N85 -5.2891873 -9.017748 -1.5606269 0.0008380
R/R50-N/N85 10.1944862 6.593417 13.7955556 0.0000000

lopro-N/N85 6.9944862 3.393417 10.5955556 0.0000008
NP-N/R40 -17.7146259 -21.399845 -14.0294069 0.0000000
lopro-N/R40 -5.4309524 -8.987127 -1.8747778 0.0002306
NP-N/R50 -14.8951423 -18.449713 -11.3405719 0.0000000
R/R50-NP 15.4836735 11.739756 19.2275913 0.0000000
lopro-NP 12.2836735 8.539756 16.0275913 0.0000000

3. (2 points) A soda company is developing a new soda. They are tying to determine how much sugar to put in it to give it the best taste. In order to evaluate this, they have made samples with ten different sugar levels. Each level of sugar is a assigned to a random sample of seven people, and each person rates the soda on a scale from 0 to 10. The company would like to make inference on the difference between mean ratings between each pair of sugar level.

(a) How many pairwise comparisons are there?

There are 10c2 comparisons, or *45 combinations*

(b) Name 2 procedures that the company could use to control the familywise Type I error rate on the differences of means? Explain.

I reccomend using either the *TukeyHSD*, or *Bonferoni Adjustment.*

4. (2 points) A consumer research group names seven types of department stores. They take a random sample of six department stores for each type and record their yearly sales. They wish to find significant pairwise differences in mean yearly sales for seven types of stores.

(a) How many pairwise comparisons are there?

There are 7c2 comparisons, or *21 combinations*

(b) The group wishes to control the familywise Type I error rate at 1 Percent using Bonferroni methods. What should be the Type I error rate of each pairwise comparison?

The type I error rate should be:

$$(1 - \alpha/21)$$

, or

$$(1 - 0.01/21)$$

Which Equals *0.9995238*