Case Study 5

Kamryn, Minh, Alexis, Addison

A consumer advocacy group wants to examine the variability of tire durability for 175/80R13 size tires. The group randomly selects five tire types (which are brand/model combinations like Goodyear/Arrive) of the given size and six tires of each type are taken at random from warehouses. The tires are then placed (in rando order) on a machine that tests tread durability in thousands of miles.

```r
#read in data
tires <- read_csv('~/MCS-243/tires.csv')
```

```
## Rows: 30 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (2): Brand, Miles
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
tires
```

```
## # A tibble: 30 x 2
##    Brand Miles
##    <dbl> <dbl>
##  1     1    55
##  2     1    56
##  3     1    59
##  4     1    55
##  5     1    60
##  6     1    57
##  7     2    39
##  8     2    42
##  9     2    43
## 10     2    41
## # i 20 more rows
```

```r
#convert variables to factor as needed
tires <- tires %>%
  mutate(Brand = as_factor(Brand))
tires
```

```
## # A tibble: 30 x 2
##    Brand Miles
##    <fct> <dbl>
##  1 1        55
##  2 1        56
```

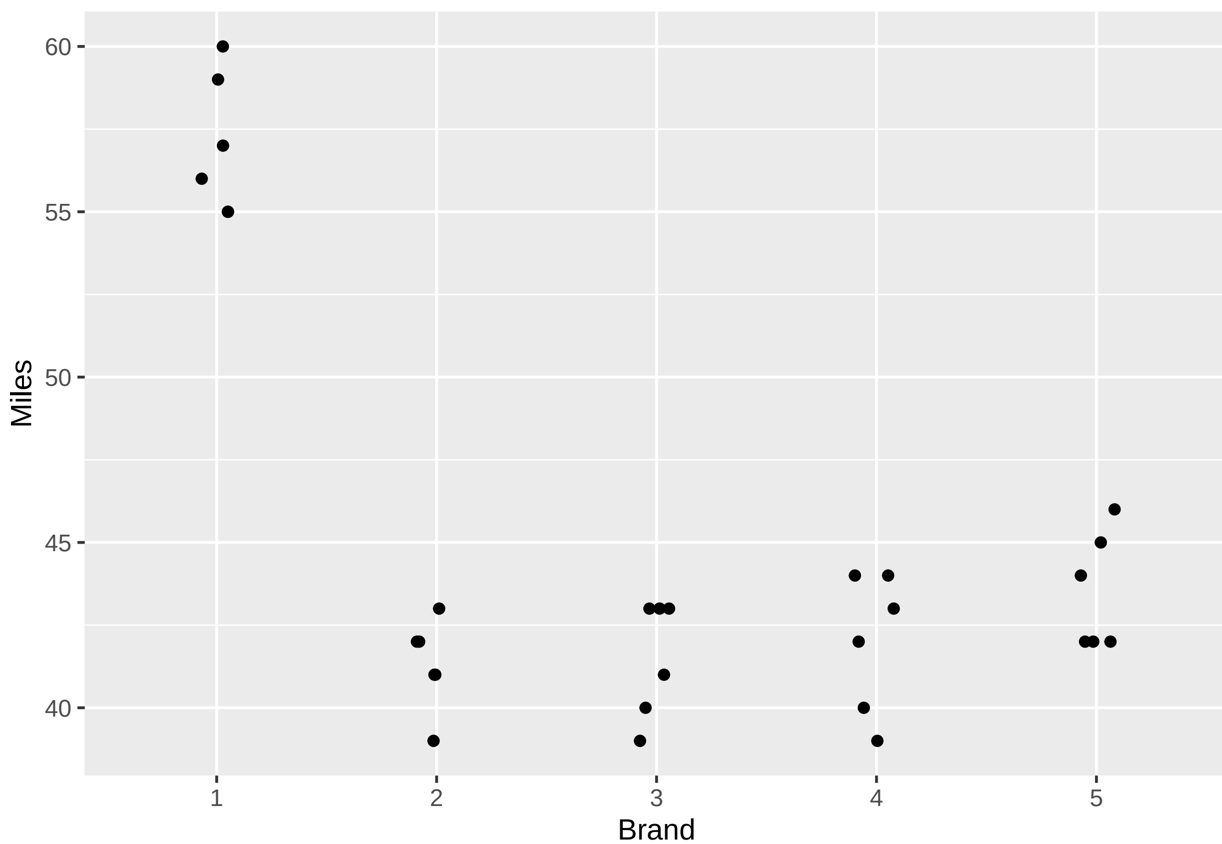
```
##   3 1         59
##   4 1         55
##   5 1         60
##   6 1         57
##   7 2         39
##   8 2         42
##   9 2         43
## 10 2         41
## # i 20 more rows
```

1. [2pts] Explain why tire type should be considered a random effect.

Answer: Tire type should be considered a random effect because the different tire types were randomly selected, not intentionally chosen, and results will be applied to all tire types.

2. [2pts] Construct an appropriate plot to visualize the data. Based on this plot, explain if the variability between brands is a major contributor to the variability in durability.

```
tires %>% ggplot(aes(x = Brand, y = Miles))+
  geom_point(position = position_jitter(width=0.1, height=0.0))
```



Answer: It appears that the variability between brands is a major contributor to the variability in durability because the average durability differs greatly between brands. Brands 2, 3, 4, and 5 have similar distributions centering around 42,000 to 44,000 miles, comparatively, the distribution for brand 1 has a higher average durability centering around 57,000 miles.

3. The model for our analysis is $Miles_{ij} = \mu + \alpha_i + \epsilon_{ij}$.

a. [1pt] What assumptions are being made about μ?

Answer: We are assuming that μ is the average durability of all tires, which is a fixed effect.

b. [1pt] What assumptions are being made about α_i?

Answer: α_i is the random brand effects which is assumed to be normally distributed with a mean of 0 and a variance of σ^2_{brand}. It is also assumed to be independent from ϵ_{ij}.

c. [1pt] What assumptions are being made about ϵ_{ij}?

Answer: ϵ_{ij} is the random error which is assumed to be normally distributed with a mean of 0 and a variance of σ^2. We are assuming that ϵ_{ij} and α_i are independent of one another.

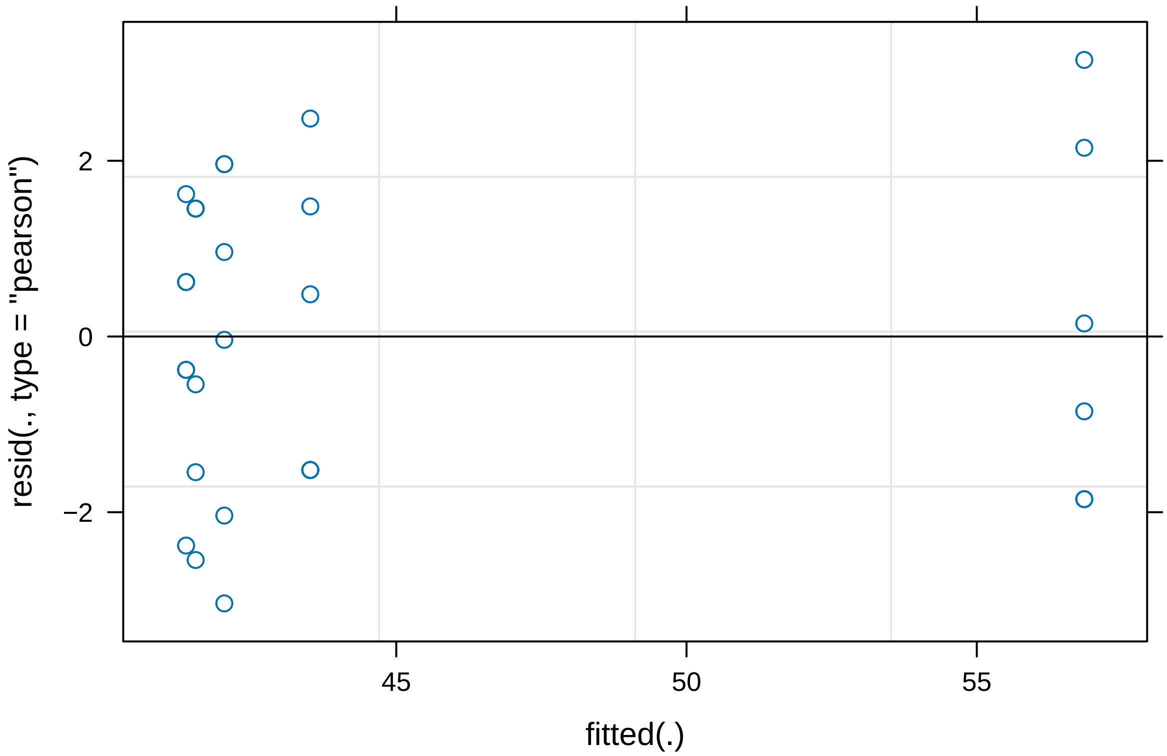
4. [3pts] First let's check whether the three "core" assumptions are met for this model. Fit the appropriate model and then evaluate whether the three assumptions are met.

```r
#fit model
tires_reml <- lmer(Miles ~ (1|Brand), data = tires, REML = T)
```

Answer: Independence. After accounting for the effect of tire type, independence is reasonably satisfied. The tires are not collected from groups or clusters, the eu and the ou are the same (tire), and the response for each tire is only measured once.

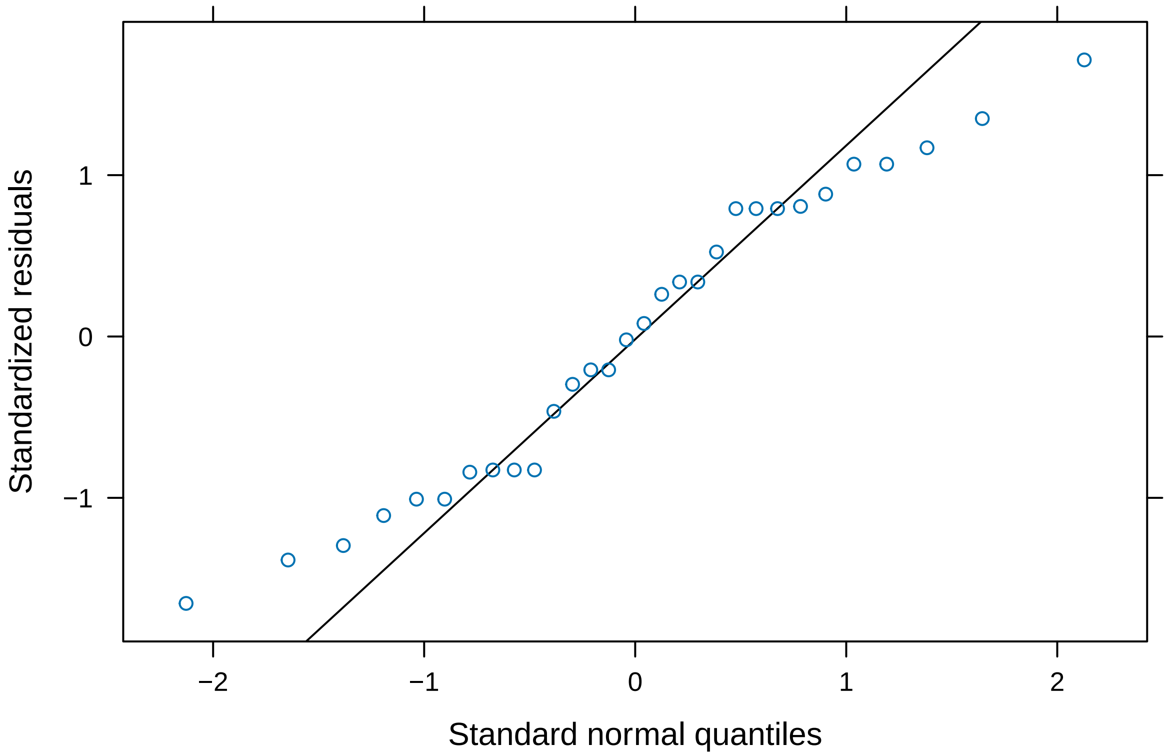
Answer: Equal Variance. The equal variance assumption is reasonably satisfied as the length of the smallest band in the residual plot is greater than half of the size of the largest band in the residual plot.

```r
plot(tires_reml)
```



Answer: Normality. The normality assumption is reasonably satisfied as in the QQ plot, the observations roughly follow the line representing the ideal relationship. There are some deviations in the tails, but they are not too severe considering our sample size of 30.

```
qqmath(tires_reml)
```



5. [4pts] Conduct the appropriate test to test whether between brands variability is a significant contributor to the variability in durability.

```
ranova(tires_reml)
```

```
## ANOVA-like table for random-effects: Single term deletions
##
## Model:
## Miles ~ (1 | Brand)
##              npar  logLik    AIC    LRT Df Pr(>Chisq)
## <none>          3 -69.256 144.51
## (1 | Brand)     2 -96.461 196.92 54.411  1  1.626e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer: Hypotheses. $H_0 : \sigma^2_\alpha = 0$ vs $H_a : \sigma^2_\alpha > 0$

Answer: Conclusion. We reject the null hypothesis and conclude that there is very strong evidence that there is significant variability between tire types (χ^2 : 54.411, df: 1, p-value: <0.0001)

6. [3pts] What is a bigger contributor to durability variability-between brand variability or within brand variability? Explain.

```r
summary(tires_reml)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: Miles ~ (1 | Brand)
##    Data: tires
##
## REML criterion at convergence: 138.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.6542 -0.8273  0.0300  0.7926  1.7142
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Brand    (Intercept) 44.668   6.683
##  Residual              3.373   1.837
## Number of obs: 30, groups:  Brand, 5
##
## Fixed effects:
##             Estimate Std. Error     df t value Pr(>|t|)
## (Intercept)   45.067      3.008  4.000   14.98 0.000116 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer: Between brand variability has a larger effect on durability ($/sigma_A^2 = 44.668$) vs in-between brand variability effect ($sigma^2 = 3.373$).

7. [2pts] Suppose, instead, that the five brands were intentionally chosen and it was of interest to compare the five brands. How would this change your analysis? You are NOT being asked to redo the analysis but briefly explain what would change.

Answer: If the brands were intentionally chosen, tire type would be a fixed effect and the experiment would instead be a CRD. A one-way ANOVA would be appropriate to analyze the data.

Milk is tested after pasteurization to ensure the process was effective. In this study, it was of interest to estimate the variability of test results between laboratories and to examine if interlaboratory difference depend on bacteria concentration. Five contract laboratories are selected at random. Four levels of contamination are chosen at random by choosing four samples of milk from a collection of samples at various stages of spoilage. A batch of fresh milk was split into 40 units. Each unit was randomly assigned to the 20 combinations of laboratory and contamination. Each unit is then contaminated with 5 ml from its selected sample, marked with a numeric code, and sent to the selected laboratory. The laboratories count the bacteria in each sample by serial diluation plate counts without knowing that they received four pairs rather than eight separate samples. The response is colony forming units per μL.

```r
#read in data
milk_p <- read_csv('milk_p.csv')
```

```
## Rows: 40 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (3): Lab, Sample, Response
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
milk_p
```

```
## # A tibble: 40 x 3
##      Lab Sample Response
##    <dbl>  <dbl>    <dbl>
##  1     1      1     2200
##  2     1      1     2200
##  3     1      2     3000
##  4     1      2     2900
##  5     1      3      210
##  6     1      3      200
##  7     1      4      270
##  8     1      4      260
##  9     2      1     2600
## 10     2      1     2500
## # i 30 more rows
```

```r
#convert to factor as needed
milk_p <- milk_p %>%
  mutate(Lab = as_factor(Lab),
         Sample = as_factor(Sample))
milk_p
```

```
## # A tibble: 40 x 3
##    Lab   Sample Response
##    <fct> <fct>     <dbl>
##  1 1     1          2200
##  2 1     1          2200
##  3 1     2          3000
##  4 1     2          2900
##  5 1     3           210
##  6 1     3           200
##  7 1     4           270
##  8 1     4           260
##  9 2     1          2600
## 10 2     1          2500
## # i 30 more rows
```

8. [2pts] The two factors in this study are laboratory and sample. Are these two factors nested or crossed? Explain.

Answer: These factors are crossed because each unit was assigned to the 20 combinations of lab and sample. Since there is a unit for each combination of the factor levels, the factors are crossed.

9. [4pts] Let's address the questions of interest. Note that in practice you should first look at your data and also check assumptions. If you do check the assumptions you'll note that they are a bit suspect but alas let's continue on. The first question of interest is whether interlaboratory differences depend on bacteria concentration. Test the appropriate variance component to answer this question.

```r
milkp_reml <- lmer(Response ~ (1|Lab) + (1|Sample) + (1|Lab:Sample), data = milk_p)
ranova(milkp_reml)
```

```
## boundary (singular) fit: see help('isSingular')
```

```
## ANOVA-like table for random-effects: Single term deletions
##
## Model:
## Response ~ (1 | Lab) + (1 | Sample) + (1 | Lab:Sample)
##                   npar  logLik    AIC     LRT Df Pr(>Chisq)
## <none>               5 -306.83 623.66
## (1 | Lab)            4 -307.64 623.27  1.6101  1  0.2044756
## (1 | Sample)         4 -318.33 644.66 22.9947  1  1.624e-06 ***
## (1 | Lab:Sample)     4 -313.22 634.45 12.7858  1  0.0003493 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer: Hypotheses. $H_0 : \sigma^2_{AB} = 0$ vs $H_a : \sigma^2_{AB} > 0$

Answer: Conclusion. We reject the null hypothesis and conclude that there is moderately strong evidence to conclude that the interlaboratory variability depends on bacteria concentration (chi^2 : 12.786, df: 1, p-value: 0.0004)

10. [2pts] We can estimate the variability of test results among laboratories using $\hat{\sigma}^2_{lab} + \hat{\sigma}^2_{L \times S}$. Obtain this value and using this value, calculate what percentage of variability of readings is due to variability of test results among laboratories.

```r
summary(milkp_reml)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: Response ~ (1 | Lab) + (1 | Sample) + (1 | Lab:Sample)
##    Data: milk_p
##
## REML criterion at convergence: 613.7
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.32058 -0.16427  0.02271  0.09838  3.00904
##
## Random effects:
##  Groups     Name        Variance Std.Dev.
##  Lab:Sample (Intercept)  266029   515.8
##  Lab        (Intercept)  129667   360.1
##  Sample     (Intercept) 2404245  1550.6
##  Residual                101743   319.0
## Number of obs: 40, groups:  Lab:Sample, 20; Lab, 5; Sample, 4
##
## Fixed effects:
##             Estimate Std. Error       df t value Pr(>|t|)
## (Intercept) 1655.750    801.773    3.246   2.065    0.124
```

```r
(129667 + 266029) / (129667 + 266029 + 2404245 + 101743)
```

```
## [1] 0.1363677
```

Answer: Approximately 13.64% of the variability of readings is due to variability of test results among laboratories.