# One Way Analysis of Variance

Blake Pappas

12/3/2021

## Example: NFL Weights

The file `NLF_weights.csv` contains weights of NFL players from several teams.

a. Is there evidence that the average weight differs across teams? State your hypotheses, test statistic, p-value, and conclusion.

```
NFL <- read.csv("NFL_weights.csv")

lm_NFL <- lm(weight ~ team, data = NFL)
anova(lm_NFL)
```

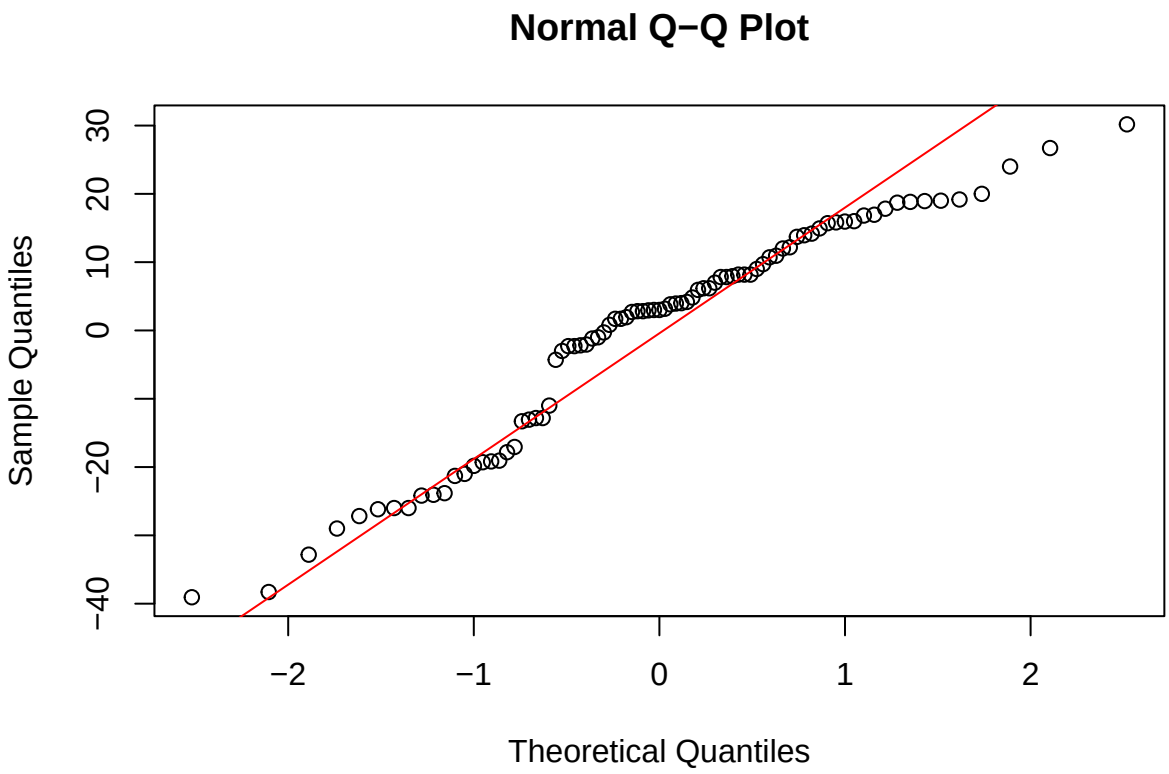
```
## Analysis of Variance Table
##
## Response: weight
##           Df  Sum Sq Mean Sq F value Pr(>F)
## team       4  1713.8  428.44   1.575  0.189
## Residuals 80 21761.4  272.02
```

**Answer: See above for the ANOVA between the `weight` and `team` variables. My null hypothesis was H0 : μ1 = μ2 = μ3 = μ4 = μ5. The alternative hypothesis was HA : Not all population means are equal. My test statistic was 1.575, the p-value was 0.189, and conclusion was to fail to reject the null hypothesis. There is insufficient evidence to conclude that the group means differ.**

b. Create a normal quantile plot of the residuals. Does it seem reasonable to assume that the residuals are normal?

```
NFL_residuals <- residuals(lm_NFL)

qqnorm(NFL_residuals)
qqline(NFL_residuals, col = 'red')
```

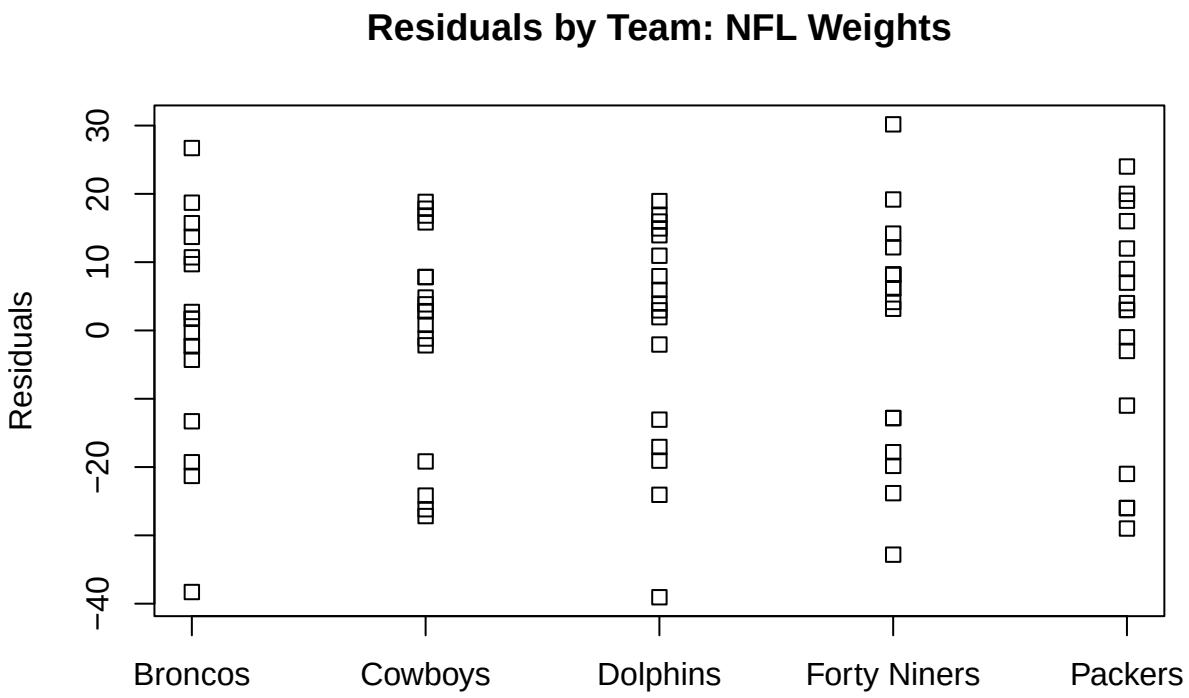



**Answer: See above for the plot of the normal quantile plot of the residuals. Based on the plot, the modeling assumption of normality appears to be reasonable. The pattern of the plot is very close to the trend line.**

c. Create a plot of residuals by group. Does it seem reasonable to assume that the groups have equal variances?

```
NFL_residuals <- residuals(lm_NFL)

stripchart(NFL_residuals ~ NFL$team, vertical = TRUE,
           main = 'Residuals by Team: NFL Weights',
           ylab = 'Residuals')
```




**Answer: See above for the plot of the `team` variable vs residuals. Based on this plot, the modeling assumption of equal variances appears to be reasonable, as there is relatively little variance between the residuals.**

## Example: Red40

An experiment was conducted in which three groups of laboratory mice were given dosages of the dye Red40 (low, medium, or high) and one control group received no treatment. The file `Red40_dosage.csv` contains the age at death (in weeks) of all mice in the study.

a. Does the data provide evidence that the mean age at death differs significantly across dosages? State the hypotheses, test statistic, the p-value, decision, and your conclusion in the context of the problem. Use $\alpha = 0.05$.

```
Red40 <- read.csv("Red40_dosage.csv")

lm_Red40 <- lm(age_at_death_weeks ~ dosage_red40, data = Red40)
anova(lm_Red40)
```

```
## Analysis of Variance Table
##
## Response: age_at_death_weeks
##              Df Sum Sq Mean Sq F value  Pr(>F)
## dosage_red40  3   4052 1350.65  3.5496 0.02447 *
```

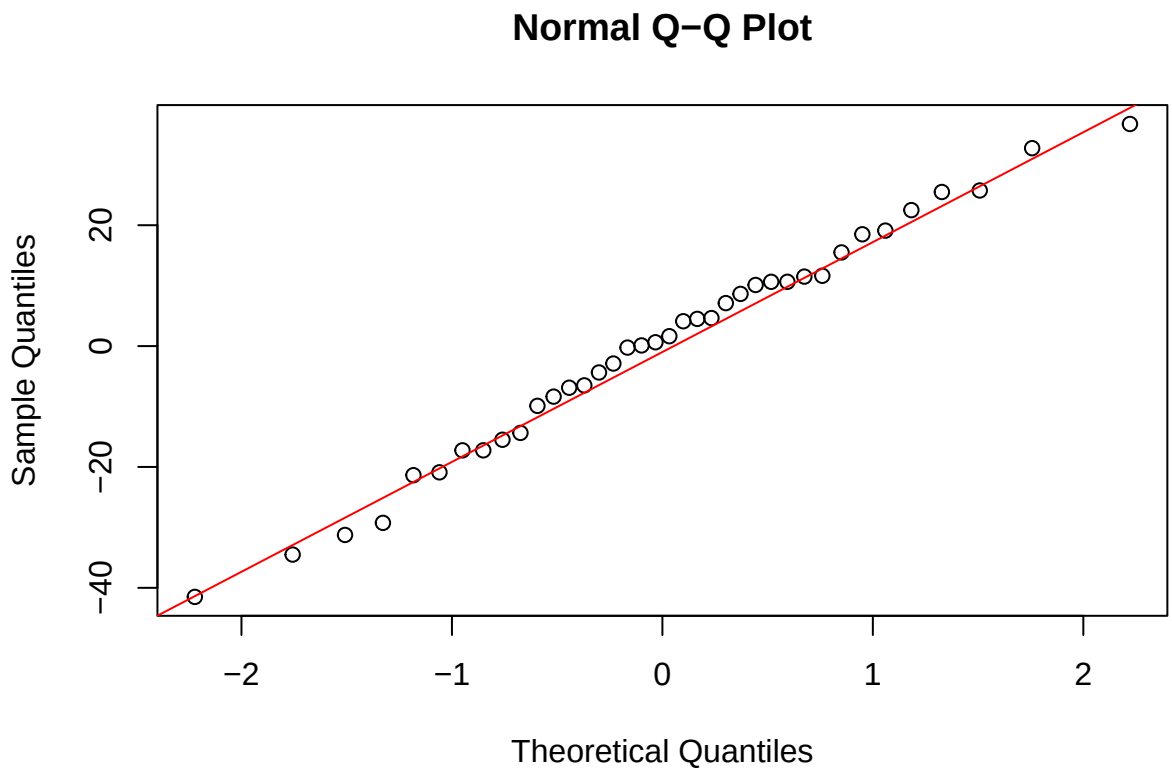
```
## Residuals    34  12937  380.51
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Answer: See above for the ANOVA between the `age_at_death_weeks` and `dosage`. My null hypothesis was H0 : μ1 = μ2 = μ3 = μ4. The alternative hypothesis was HA : Not all population means are equal. My test statistic was 3.5496, the p-value was 0.02447, and conclusion was to reject the null hypothesis. There is sufficient evidence to conclude that the mean age at death differs significantly across dosages.**

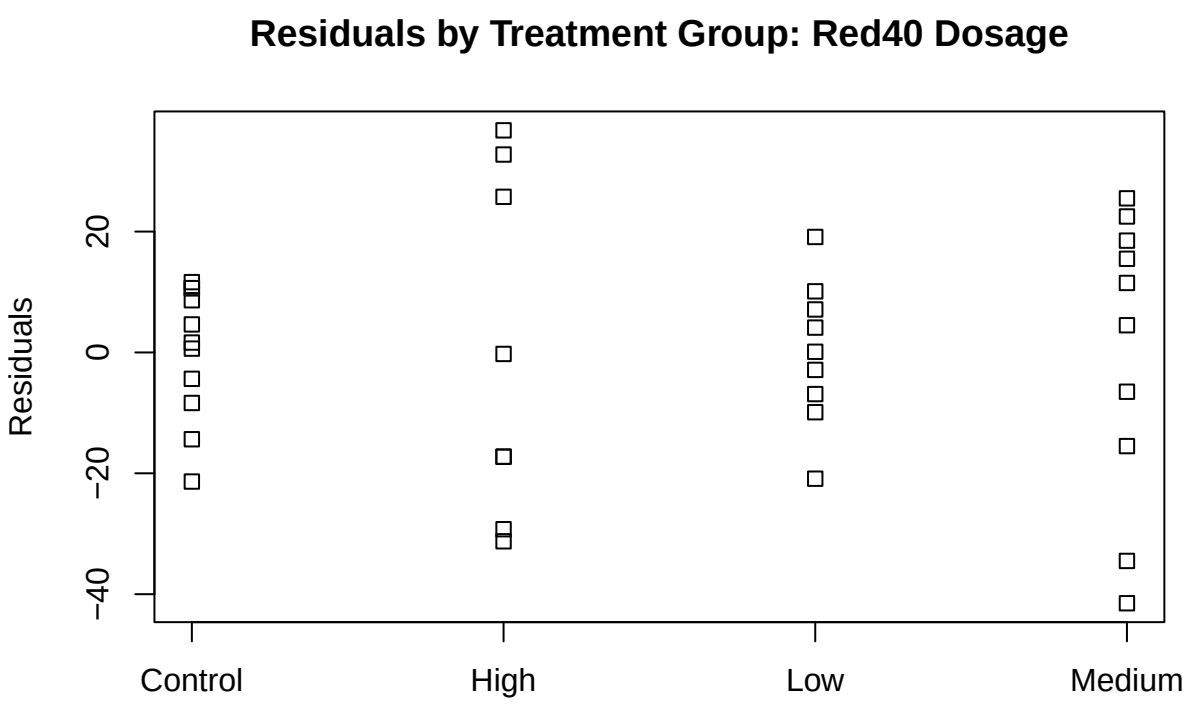
b. Make a normal quantile plot and plot of residuals by group. Do the assumptions of normality and equal variances seem reasonable?

```r
Red40_residuals <- residuals(lm_Red40)

qqnorm(Red40_residuals)
qqline(Red40_residuals, col = 'red')
```




```r
stripchart(Red40_residuals ~ Red40$dosage_red40, vertical = TRUE,
           main = 'Residuals by Treatment Group: Red40 Dosage',
           ylab = 'Residuals')
```




**Answer: See above for the normal quantile plot of the residuals, as well as the plot of the x variable vs residuals. Based on the first plot, the modeling assumption of normality appears to be reasonable, as the pattern of the plot is very close to the trend line. However, based on the second plot, the assumption of equal variances does not seem to be reasonable.**

c. Which treatment groups have statistically significant differences? Use Tukey's method with $\alpha_E = 0.05$.

```
anova(lm_Red40)
```

```
## Analysis of Variance Table
##
## Response: age_at_death_weeks
##              Df Sum Sq Mean Sq F value  Pr(>F)
## dosage_red40  3   4052 1350.65  3.5496 0.02447 *
## Residuals    34  12937  380.51
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Red40_pairwise_comparisons <- TukeyHSD(aov(lm_Red40), conf.level = 0.95)
Red40_pairwise_comparisons
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = lm_Red40)
```

```
##
## $dosage_red40
##                       diff       lwr       upr     p adj
## High-Control   -26.113636 -50.59372 -1.633551 0.0328848
## Low-Control    -21.474747 -45.15437  2.204879 0.0868094
## Medium-Control -19.863636 -42.88286  3.155592 0.1109953
## Low-High         4.638889 -20.96086 30.238636 0.9609302
## Medium-High      6.250000 -18.74014 31.240141 0.9056532
## Medium-Low       1.611111 -22.59544 25.817667 0.9978960
```

**Answer: The High-Control treatment groups have statistically significant differences.**

## Example: Mushroooms

Revisit the mushroom data in `mushrooms.csv`. For this question, consider the variables `edible` (e = edible, p = poisonous) and `population`, which describes how abundant the species is (abundant = a, clustered = c, numerous = n, scattered = s, several = v, solitary = y).

a. Make a contingency table with `population` as the row variable and `edible` as the column variable. Do certain population types appear to be more likely to be poisonous?

```r
mushrooms <- read.csv("mushrooms.csv")

mushrooms.table <- table(mushrooms$population, mushrooms$edible)
rownames(mushrooms.table) <- c("abundant", "custered", "numerous", "scattered", "several", "solitary")
colnames(mushrooms.table) <- c("edible", "poisonous")
mushrooms.table
```

```
##
##             edible poisonous
##   abundant      22         0
##   custered      11         3
##   numerous      21         0
##   scattered     35        20
##   several       69       148
##   solitary      61        35
```

**Answer: See above for the contingency table, comparing the `population` and `edible` variables. Looking at the table, the "several" population appears more likely to be poisonous.**

b. Perform a chi-square test using $\alpha = 0.01$ to assess whether the data provide strong evidence of an association. Report the hypotheses, test statistic, p-value, and conclusion.

```r
chisq.test(mushrooms.table)
```

```
##
##  Pearson's Chi-squared test
##
## data:  mushrooms.table
## X-squared = 87.148, df = 5, p-value < 2.2e-16
```

**Answer: See above for the chi-square test. My null hypothesis was H0 : variables are independent. The alternative hypothesis was HA : variables are not independent. My test statistic was 87.148, the p-value was 2.2e-16, and conclusion was to reject the null hypothesis. There is some association between the two variables.**

c. What is the expected number of mushrooms that are edible and scattered? What is the expected number of mushrooms that are poisonous and abundant?

```
mushrooms_results <- chisq.test(mushrooms.table)

mushrooms_results$expected
```

```
##
##                edible  poisonous
##   abundant   11.336471  10.663529
##   custered    7.214118   6.785882
##   numerous   10.821176  10.178824
##   scattered  28.341176  26.658824
##   several   111.818824 105.181176
##   solitary   49.468235  46.531765
```

**Answer: The expected number of mushrooms that are edible and scattered is 28.341176. The expected number of mushrooms that are poisonous and abundant is 10.663529.**

## Example: Wine

The data in the file `wines_big.csv` give ratings scraped from the web of a large number of wines. The variables for each wine include an expert's subjective rating on a scale of 0 to 100, the price of the wine, the variety, and its country of origin. Use these data to answer the following questions.

a. Make a contingency table in which the country of origin is the row variable and the variety is the column variable.

```
wine <- read.csv("wines_big.csv")

wine.table <- table(wine$country, wine$variety)
wine.table
```

```
##
##             Cabernet Sauvignon Chardonnay Merlot Riesling Sauvignon Blanc
##   Argentina                 96         67      9        1              17
##   Chile                    154         91     55        3             139
##   France                     9        527     21      134             166
##   Italy                     26         46     30        9              10
##   US                      1426       1312    474      326             410
```

**Answer: See above for the contingency table, comparing the `country` and `variety` variables.**

b. Make a proportion table that displays, for each country, the proportion of wines of each variety (row proportions).

```
wine.prop.table <- prop.table(wine.table, margin = 1)
wine.prop.table
```

```
##
##             Cabernet Sauvignon  Chardonnay      Merlot    Riesling
##   Argentina        0.505263158 0.352631579 0.047368421 0.005263158
##   Chile            0.348416290 0.205882353 0.124434389 0.006787330
##   France           0.010501750 0.614935823 0.024504084 0.156359393
##   Italy            0.214876033 0.380165289 0.247933884 0.074380165
##   US               0.361195542 0.332320162 0.120060790 0.082573455
##
##             Sauvignon Blanc
##   Argentina     0.089473684
##   Chile         0.314479638
##   France        0.193698950
##   Italy         0.082644628
##   US            0.103850051
```

**Answer: See above for the proportion table, comparing the `country` and `variety` variables.**

c. Perform a hypothesis test using $\alpha = 0.01$ to see if the data provide evidence that the type of wine produced will vary significantly by country. Report the test statistic, p-value, and conclusion.

```
chisq.test(wine.table)
```

```
##
##  Pearson's Chi-squared test
##
## data:  wine.table
## X-squared = 852.2, df = 16, p-value < 2.2e-16
```

**Answer: See above for the chi-square test. My null hypothesis was H0 : variables are independent. The alternative hypothesis was HA : variables are not independent. My test statistic was 852.2, the p-value was 2.2e-16, and conclusion was to reject the null hypothesis. There is some association between the two variables, which indicates the type of wine produced will indeed vary significantly by country.**

## Example: Insurance

The insurance data in the file `insurance.csv` contain several variables measured on insured individuals who are clients of a particular provider. Use these data to answer the following questions about associations between variables.

a. Is there a statistically significant association between whether a client is a smoker and which geographic region they come from? Use the variables `smoker` and `region`. Use $\alpha = 0.05$. Report the hypotheses, test statistic, p-value, and conclusion.

```r
# Note: The `smoker` field was converted into a binary variable. Values of "yes" were replaced with 1 a

insurance <- read.csv("insurance.csv")

lm_insurance <- lm(smoker_bin ~ region, data = insurance)
anova(lm_insurance)
```

```
## Analysis of Variance Table
##
## Response: smoker_bin
##           Df Sum Sq Mean Sq F value Pr(>F)
## region     3 0.4559 0.15196   0.958 0.4212
## Residuals 43 6.8207 0.15862
```

**Answer: See above for the ANOVA between the `smoker` and `region` variables. My null hypothesis was H0 : μ1 = μ2 = μ3 = μ4. The alternative hypothesis was HA : Not all population means are equal. My test statistic was 0.958, the p-value was 0.4212, and conclusion was to fail to reject the null hypothesis. There is insufficient evidence to conclude that there is a statistically significant association between whether a client is a smoker and which geographic region they come from.**

b. Is there a statistically significant difference in the mean claim amount across the four geographic regions? Use the variables `expenses` and `region`. Use $\alpha = 0.05$. Report the hypotheses, test statistic, p-value, and conclusion.

```r
lm_insurance <- lm(expenses ~ region, data = insurance)
anova(lm_insurance)
```

```
## Analysis of Variance Table
##
## Response: expenses
##           Df     Sum Sq   Mean Sq F value  Pr(>F)
## region     3  844257624 281419208  2.2749 0.09343 .
## Residuals 43 5319341237 123705610
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Answer: See above for the ANOVA between the `expenses` and `region` variables. My null hypothesis was H0 : μ1 = μ2 = μ3 = μ4. The alternative hypothesis was HA : Not all population means are equal. My test statistic was 2.2749, the p-value was 0.09343, and conclusion was to fail to reject the null hypothesis. There is insufficient evidence to conclude that there is a statistically significant difference in the mean claim amount across the four geographic regions.**

c. Is there evidence that the mean age differs between smokers and non-smokers? Use the variables `age` and `smoker`. Use $\alpha = 0.05$. Report the hypotheses, test statistic, p-value, and conclusion.

```r
lm_insurance <- lm(age ~ smoker, data = insurance)
anova(lm_insurance)
```

```
## Analysis of Variance Table
##
```

```
## Response: age
##           Df Sum Sq Mean Sq F value Pr(>F)
## smoker     1   84.9  84.871  0.5899 0.4465
## Residuals 45 6474.4 143.875
```

**Answer: See above for the ANOVA between the `age` and `smoker` variables. My null hypothesis was H0 : μ1 = μ2. The alternative hypothesis was HA : Not all population means are equal. My test statistic was 0.5899, the p-value was 0.4465, and conclusion was to fail to reject the null hypothesis. There is insufficient evidence to conclude that the mean age differs between smokers and non-smokers.**