Stats 101B HW 1

Anna Piskun

4/8/2020

Island Exploration

Your goal is to determine if energy drinks raise the Islanders' pulse. You must collect data on at least 10 subjects, 5 males and 5 females following two designs:

1. Collect data so that you can use the paired t-test to compare the mean pulse rate before and after the sports drink. Analyze the data and state your conclusion. Your analysis should include appropriate statistical graphics, the statement of the null and alternative hypotheses, and a conclusion based on the outcome of your data. You should also discuss sources of variation. Which variables did you hold constant?

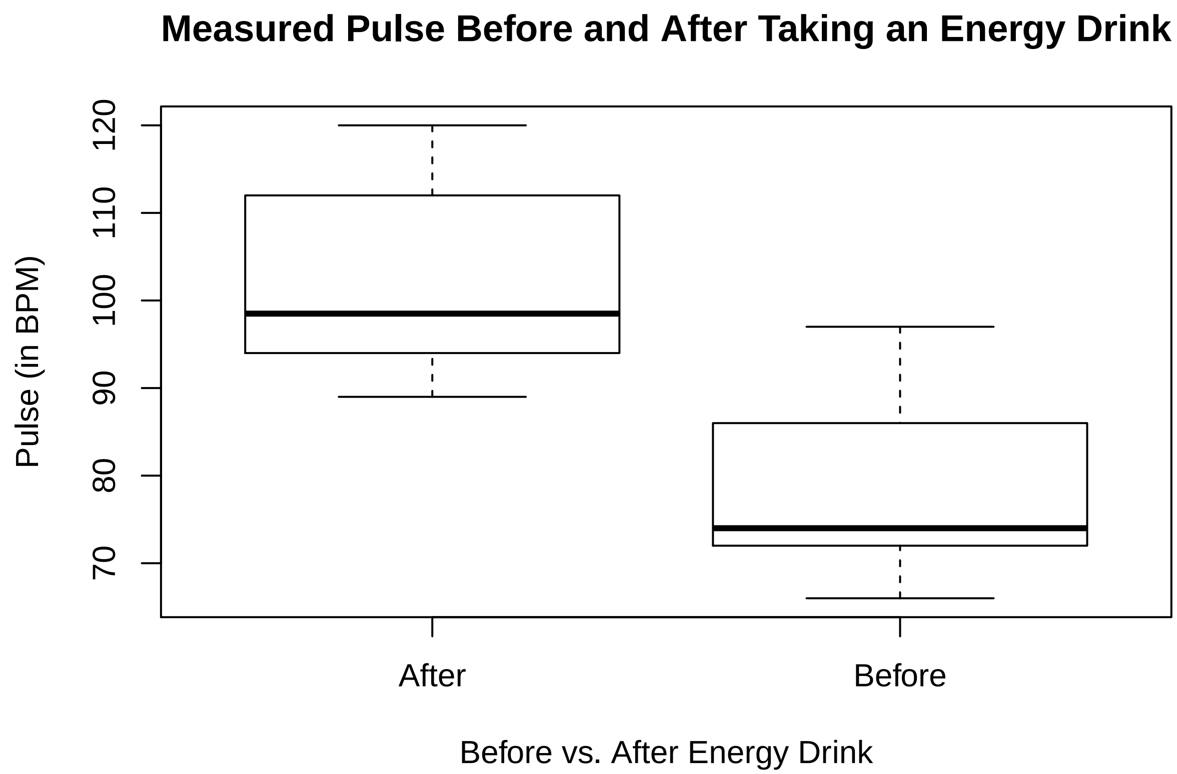
H0 : µ2 - µ1 = 0

HA : µ2 - µ1 does not = 0

```r
mydata <- read.csv("hw1.2.csv")
t.test(Pulse~Time, data = mydata, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  Pulse by Time
## t = 22.855, df = 9, p-value = 2.791e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  21.44427 26.15573
## sample estimates:
## mean of the differences
##                    23.8
```

```r
boxplot(Pulse~Time, data = mydata, xlab = "Before vs. After Energy Drink", ylab = "Pulse (in BPM)", mai
```



Looking at the output of our paired t test, since the p-value is very small, we reject the null hypothesis. Thus, we can conclude that there was a difference in mean pulse rate before and after drinking 250 mL of an energy drink. Likewise, since the mean of the differences is 23.8 we can conclude that the energy drink raised test subjects' pulse (because in order to get a positive mean of the differences this indicates that the mean pulse after was significantly higher than the mean pulse before). There are many potential sources of variation such as age, weight, BMI, fitness level, pre-existing health conditions, and more. People with higher fitness levels may have naturally higher pulses while others with pre-existing health conditions such as diabetes may be affected more significantly by a sugary energy drink. These are just some examples of sources of variation that can affect the outcome of our experiment. While I did not follow a particular sampling scheme, I chose to keep the village variable constant - meaning all 10 subjects were from the same village called Maconado.

2. Using the data from (1), carry out an analysis by applying the two-sample t-test between males and females. Why do you get different results from those you got in (1)?

```r
t.test(Pulse ~ Gender, data = subset(mydata, Time == "Before"))
```

```
##
##  Welch Two Sample t-test
##
## data:  Pulse by Gender
## t = 0.62167, df = 5.7372, p-value = 0.558
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.9206  19.9206
## sample estimates:
## mean in group female   mean in group male
##                   80                   76
```

Looking at the output of our t-test, since the p-value is greater than 0.05 it is not statistically significant, and we fail to reject the null hypothesis. Thus we can conclude that there is no difference in mean pulse between males and females. These results are different than in 1 because here we are testing if gender affects pulse vs. in question 1 we are testing if an energy drink affects pulse which is a completely different experimental goal.

3. How might you have changed your data collection if the goal was instead to "determine the effect of energy drinks on the Islanders' pulse"? How would you analyze these data?

If the goal was to specifically determine the effect of energy drinks on the Islanders' pulse I would have improved on the experimental design by controlling for the sources of variation listed in the answer to question 1. This means I would control for variables such as age, weight, fitness level, and pre-existing conditions. I would analyze this data by again using a paired t test to compare the pulses of the test subjects before taking the energy drink and after taking the enrgy drink.

Textbook Questions: Warm-up Questions

2.9) A computer program has produced the following output for a hypothesis-testing problem:

Difference in sample means: 2.35

Degrees of freedom: 18

Standard error of the difference in sample means: ?

Test statistic: t0 = 2.01

P-value: 0.0298

a) What is the missing value for the standard error?

```
t = (xbar - mu)/SE(xbar)

SE(xbar) = (xbar - mu) / t

Standard Error = (2.35 - 0) / 2.01 = 1.17
```

b) is this a two-sided or a one-sided test?

```
pt(-2.01, 18)
```

```
## [1] 0.02983103
```

If this were a two-sided test we would have to multiply the output of pt() by two. However, since the output is the same as the one provided above we can conclude that this is a one-sided test.

c) If alpha = 0.05, what are your conclusions?

```
Since the pvalue is less than 0.05 (0.0298), we reject the null hypothesis.
```

d) Find a 90% two-sided CI on the difference in means.

```
[0.3211456, 4.378854]
```

```
t <- qt(.95, 18)
me <- t*(1.17)

lb <- 2.35 - me
lb
```

```
## [1] 0.3211456
```

```
ub <- 2.35 + me
ub
```

```
## [1] 4.378854
```

2.15) Consider the computer output shown below

a) Can the null hypothesis be rejected at the 0.05 level? Why?

   Yes, the null hypothesis can be rejected because the p-value is 0.001 which is less than 0.05. This means that there is a difference betwen µ1 and µ2.

b) is this a one-sided or a two-sided test?

   This is a two-sided test as made evident by the output which specifies that the t-test is testing the null hypothesis of there being no difference in means vs. there being a difference in means.

c) If the hypotheses had been H0 : mu1 - mu2 = 2 versus H1: mu1 - mu2   2 would you reject the null hypothesis at the 0.05 level?

   We would still reject the null hypothesis because -2.33 is even farther away from 2 than it is from 0, and since the test for 0 was statisticially signficant it will also be significant for 2.

d) if the hypothesis had been H0 : µ1 - µ2 = 2 versus H1: µ1-µ2 < 2 would you reject the null hypothesis at the 0.05 level? can you answer this questions without doing any additional calculations? Why?

Generally speaking, a two-sided hypothesis test is more conservative than a one-sided test, so if the null hypothesis was rejected for the two-sided test we will naturally also reject for a one-sided test.

e) Use the output and the t table to find a 95% upper confidence bound on the difference in means.

se = sp * sqrt(1/n1 + 1/n2)

Upper Confidence Bound = -1.199036

```r
SE <- 2.1277 * sqrt(1/20 + 1/20)
qt(.95, 38)
```

```
## [1] 1.685954
```

```r
UB <- -2.33341 + 1.685954*SE

#95% upper confidence bound on the difference in means
UB
```

```
## [1] -1.199036
```

f) what is the p-value if the hypotheses are H0 : µ1 - µ2 = 2 versus H1: µ1-µ2   2?

Because we know we are more likely to reject using the above hypotheses rather than those given in the output, the pvalue will be even smaller than the one given in the output. As such, the p-value will be very small, ending up being approximately zero.

3.7) The tensile strength of Portland cement is being studied. Four different mixing techniques can be used economically. A completely randomized experiment was conducted and the following data were collected:

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
setwd("~/Desktop")
data <- read.csv("hw1.csv")
df <- dplyr::select(data, Technique, Tensile.Strength)
```

a) Test the hypothesis that mixing techniques affect the strength of the cement. Use alpha = 0.05.

```
H0: T1 = T2 = T3 = T4

HA: Ti does not = 0 for at least one i (meaning at least one Ti is different)
```

```
m1 <- aov(Tensile.Strength~factor(Technique), data = df)
summary(m1)
```

```
##                   Df Sum Sq Mean Sq F value   Pr(>F)    
## factor(Technique)  3 489740  163247   12.73 0.000489 ***
## Residuals         12 153908   12826                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the pvalue is 0.000489 which is less than 0.05 (and the F value of 12.73 is sufficiently large), we reject the null hypothesis that the average strength of cement is the same across all four mixing techniques. Therefore, we can conclude that mixing techniques does affect the strength of the cement, but we don't know which specific ones are different.

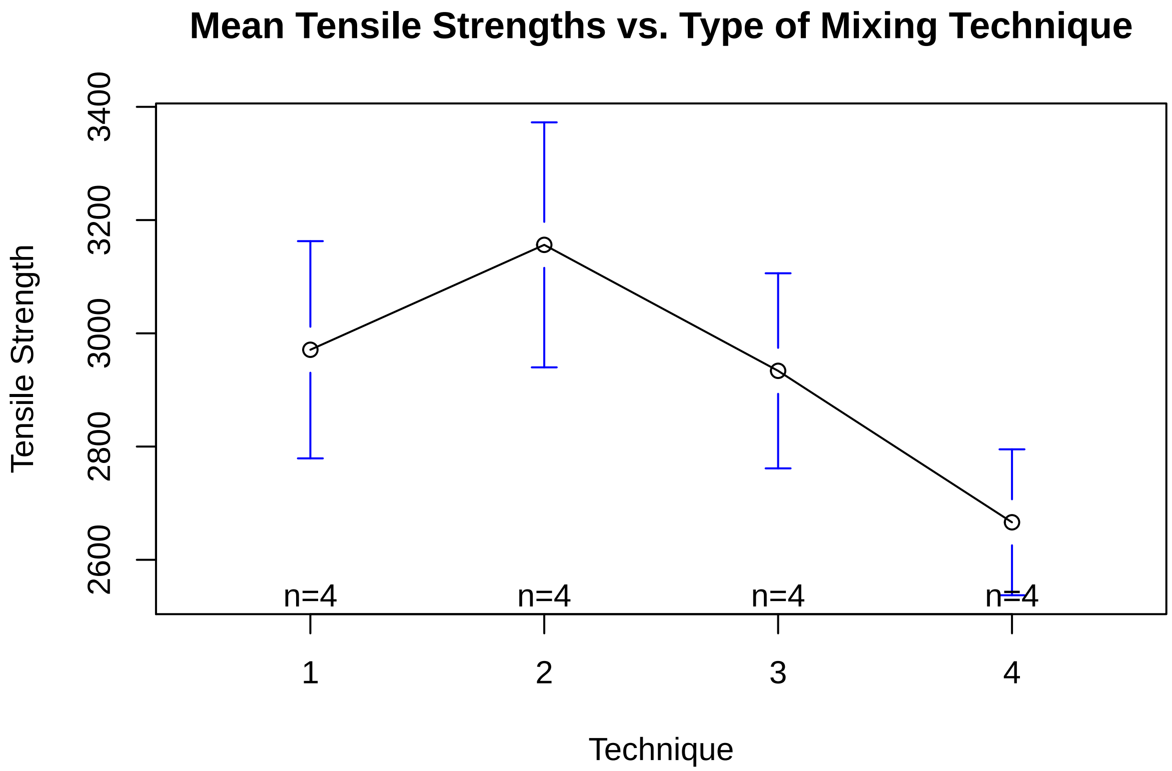
b) Construct a graphical display as described in Section 3.5.3 to compare the mean tensile strengths for the four mixing techniques. What are your conclusions?

```
library(gplots)
```

```
## 
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
## 
##     lowess
```

```
plotmeans(Tensile.Strength~Technique, data = df,
          xlab = "Technique",
          ylab = "Tensile Strength",
          main = "Mean Tensile Strengths vs. Type of Mixing Technique")
```



Looking at the plot of the mean tensile strengths for the four different mixing techniques we see that generally there is a difference between them. For example, mixing technique number 2 results in higher mean cement tensile strengths, while mixing technique number 4 results in lower mean cement tensile strengths. Mixing techniques 1 and 3 report having approximately the same mean tensile strength. There is sufficient graphical evidence to conclude that the tensile strength varies from one type of mixing technique to another (except for in the case of type 1 and 3) with technique 2 being the strongest and technique 4 being the weakest.

c) Use the Fisher LSD method with alpha = 0.05 to make comparisons between pairs of means.

a = 4

N = 16

MS = 12826

t0.025,16-4 = 2.17881

```
#Fisher LSD method
LSD <- 2.17881*sqrt(((2*12826)/4))
LSD
```

```
## [1] 174.4817
```

Y1 - Y2 = 2971 - 3156.25 = -185.25 = absolute value(-185.25) > 174.4817

Y1 - Y3 = 2971 - 2933.75 = 37.25 < 174.4817

Y1 - Y4 = 2971 - 2666.25 = 304.75 > 174.4817

Y2 - Y3 = 3156.25 - 2933.75 = 222.5 > 174.4817

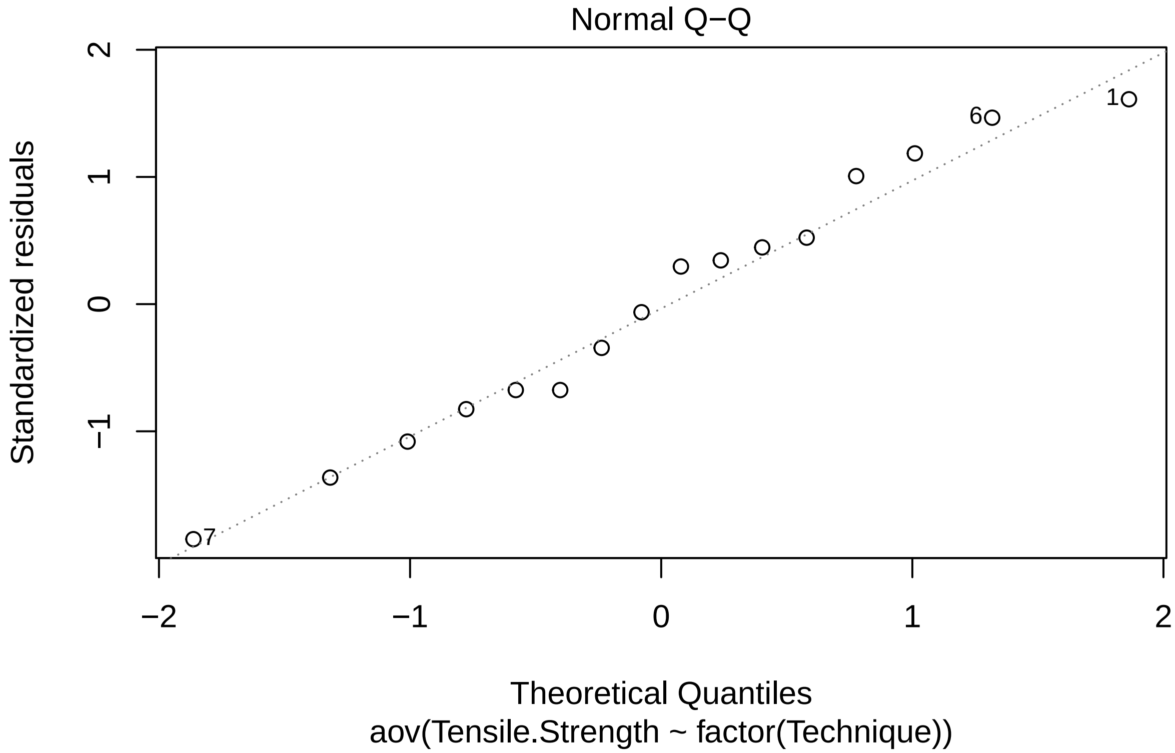
Y2 - Y4 = 3156.25 - 2666.25 = 490 > 174.4817

Y3 - Y4 = 2933.75 - 2666.25 = 267.5 > 174.4817

From this analysis, we see that there are significant differences between all pairs of treatment means except 1 and 3 (which supports our findings from our graph in part b). This implies that mixing techniques 1 and 3 produced approximately the same cement tensile strength and that all other mixing techniques produced different tensile strengths.

d) Construct a normal probability plot of the residuals. What conclusion would you draw about the validity of the normality assumption?

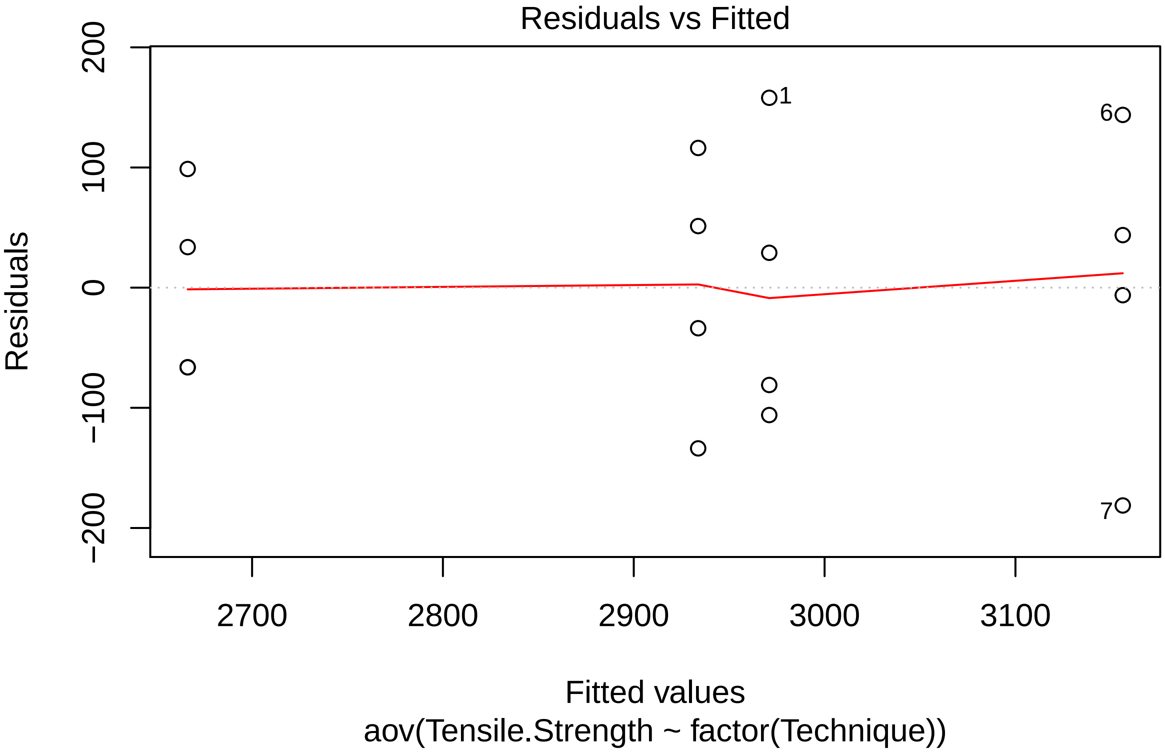
```r
plot(m1, which = 2)
```



While there is some slight deviation in the normal qq plot, the points still follow a relatively straight line indicating that the residuals satisfied the normality assumption.

e) Plot the residuals versus the predicted tensile strength. Comment on the plot.

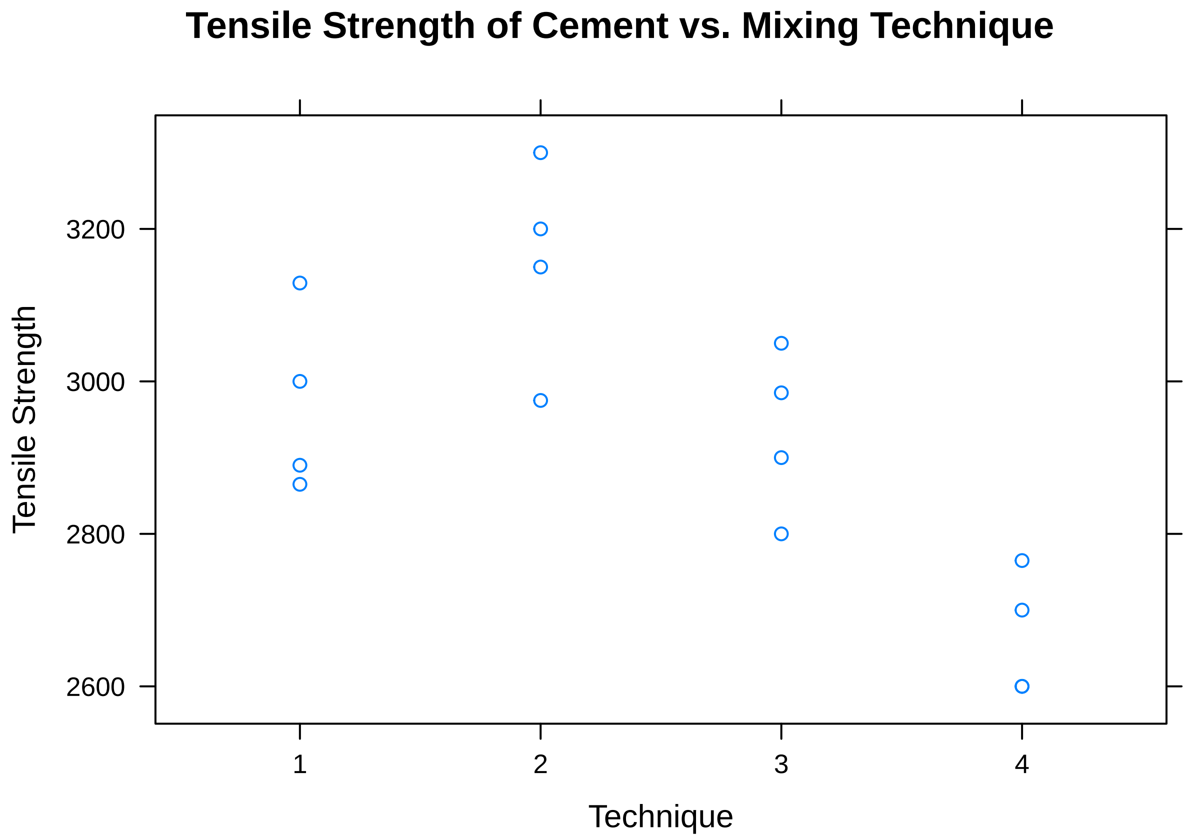
```r
plot(m1, which = 1)
```



There is a slight fan-shape pattern in the plot, indicating non-constant variance in the errors.

f) Prepare a scatter plot of the results to aid the interpretation of the results of this experiment.

```
library(lattice)
xyplot(Tensile.Strength~factor(Technique), data = df,
       xlab = "Technique",
       ylab = "Tensile Strength",
       main = "Tensile Strength of Cement vs. Mixing Technique")
```



```r
by(df$Tensile.Strength, factor(df$Technique), function(x) mean(x))
```

```
## factor(df$Technique): 1
## [1] 2971
## ------------------------------------------------------------
## factor(df$Technique): 2
## [1] 3156.25
## ------------------------------------------------------------
## factor(df$Technique): 3
## [1] 2933.75
## ------------------------------------------------------------
## factor(df$Technique): 4
## [1] 2666.25
```