Appendix

Jadon Fowler, STA 570 Section 1

Homework 7, 2023/04/04

```
library(ggplot2)
library(dplyr)
library(mosaic)
library(Lock5Data)
library(tidyr)
library(coin)
```

1)

b)

```
# calculate variance from sample stddev and sample size
V <- function(s, n) { s^2/n }

# calculate the test statistic under unequal variance conditions
test.statistic <- function(xbar1, s1, n1, xbar2, s2, n2) {
  (xbar1 - xbar2) / sqrt(V(s1,n1) + V(s2,n2))
}

# Satterthwaite's Approximation
# this is used to find the degrees of freedom for a two-sample t-test
degrees.of.freedom <- function(s1, n1, s2, n2) {
  V1 <- V(s1,n1)
  V2 <- V(s2,n2)
  (V1 + V2)^2 / (V1^2/(n1-1) + V2^2/(n2-1))
}

# in proximity to a fracking well
n1 = 21
xbar1 = 19.2
s1 = 30

# sites in the same region with no fracking wells
n2 = 13
xbar2 = 1.1
s2 = 6.3

t.delta <- test.statistic(xbar1, s1, n1, xbar2, s2, n2)
dof <- degrees.of.freedom(s1, n1, s2, n2)
cat(sprintf("test statistic = %f\n", t.delta))
```

```
## test statistic = 2.671308
```

```
cat(sprintf("degrees of freedom = %f\n", dof))
```

```
## degrees of freedom = 22.758541
```

c)

```
p.value <- pt(t.delta, df=dof, ncp=0)
cat(sprintf("p-value = %f\n", p.value))
```

```
## p-value = 0.993145
```

2)

b)

```
# this is s^2
pooled.var <- function(s1, n1, s2, n2) {
  (1/(n1+n2-2)) * (s1^2*(n1-1) + s2^2*(n2-1))
}

test.statistic.pooled.var <- function(xbar1, s1, n1, xbar2, s2, n2) {
  (xbar1 - xbar2) / sqrt(pooled.var(s1, n1, s2, n2)) * sqrt(1/n1 + 1/n2)
}

# money the male candidates raised
n1 = 30
xbar1 = 350000
s1 = 61900

# money the female candidates raised
n2 = 30
xbar2 = 245000
s2 = 52100

t.delta <- test.statistic.pooled.var(xbar1, s1, n1, xbar2, s2, n2)
dof <- n1 + n2 - 2
cat(sprintf("test statistic = %f\n", t.delta))
```

```
## test statistic = 0.473882
```

```
cat(sprintf("degrees of freedom = %f\n", dof))
```

```
## degrees of freedom = 58.000000
```

c)

```
p.value <- pt(t.delta, df=dof, ncp=0)
cat(sprintf("p-value = %f\n", p.value))
```

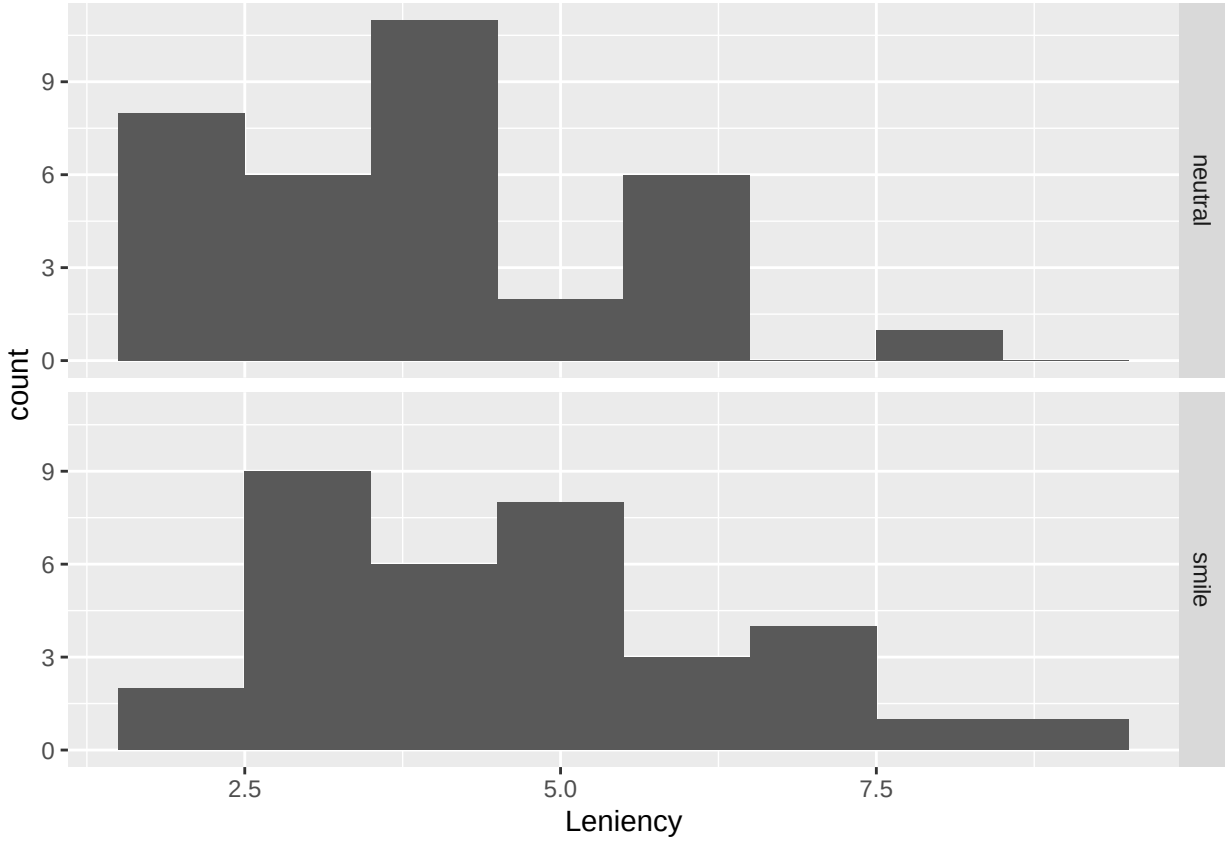
```
## p-value = 0.681318
```

3)

a)

```
data("Smiles",package = "Lock5Data")

ggplot(Smiles, aes(x=Leniency)) +
  #geom_dotplot(binwidth=.1) +
  geom_histogram(binwidth=1) +
  facet_grid(Group ~ .)
```



b)

```
smiles.summary <- Smiles %>%
  group_by(Group) %>%
  summarize(x.bar = mean(Leniency), stddev = sd(Leniency), n=length(Leniency))
smiles.summary
```

```
## # A tibble: 2 x 4
##   Group   x.bar stddev     n
##   <fct>   <dbl>  <dbl> <int>
## 1 neutral  4.12   1.52    34
## 2 smile    4.91   1.68    34
```

c)

```r
#mosaic::t.test(Leniency ~ Group, data=Smiles, var.equal=FALSE, conf.level=0.95)

t.delta <- test.statistic.pooled.var(
  smiles.summary$x.bar[1],
  smiles.summary$stddev[1],
  smiles.summary$n[1],
  smiles.summary$x.bar[2],
  smiles.summary$stddev[2],
  smiles.summary$n[2])
dof <- smiles.summary$n[1] + smiles.summary$n[2] - 2
critical.t <- qt(0.975, dof)
range <- critical.t * sqrt((smiles.summary$stddev[1]^2/smiles.summary$n[1]) +
                  (smiles.summary$stddev[2]^2/smiles.summary$n[2]))
x.diff <- smiles.summary$x.bar[1] - smiles.summary$x.bar[2]

p.value <- pt(t.delta, df=dof, ncp=0)
cat(sprintf("test statistic = %f\n", t.delta))
```

```
## test statistic = -0.120090
```

```r
cat(sprintf("degrees of freedom = %f\n", dof))
```

```
## degrees of freedom = 66.000000
```

```r
cat(sprintf("p-value = %f\n", p.value))
```

```
## p-value = 0.452388
```

```r
cat(sprintf("95 percent CI: (%f, %f)\n", x.diff-range, x.diff+range))
```

```
## 95 percent CI: (-1.570741, -0.017494)
```

d)

```r
Smiles %>% group_by(Group) %>%
  summarise(xbar=mean(Leniency)) %>%
  summarise(d = diff(xbar))
```

```
## # A tibble: 1 x 1
##       d
##   <dbl>
## 1 0.794
```

```r
observed.d <- 0.7941176

PermutationDist <- mosaic::do(1000) * {
  Smiles %>%
  mutate( ShuffledGroup = mosaic::shuffle(Group) ) %>%
  group_by( ShuffledGroup )  %>%
```

```
  summarise(xbar=mean(Leniency)) %>%
  summarise(d.star = diff(xbar))
}

#ggplot(PermutationDist, aes(x=d.star)) +
#  geom_histogram(binwidth=.2) +
#  ggtitle('Permutation dist. of d* assuming H0 is true') +
#  xlab('d*') +
#  geom_vline(xintercept = c(-observed.d, observed.d), lwd=1.5, col='red')

PermutationDist %>%
  mutate( MoreExtreme = ifelse( abs(d.star) >= observed.d, 1, 0)) %>%
  summarise( p.value = mean(MoreExtreme))
```

```
##   p.value
## 1   0.067
```

```
p.value <- 0.0514

BootDist <- mosaic::do(1000)*{
    Smiles %>%
    group_by(Group) %>%
    mosaic::resample() %>%
    summarise( xbar.i = mean(Leniency) ) %>%
    summarise( d.star = diff(xbar.i) )
}
CI <- quantile( BootDist$d.star, probs=c(0.025, 0.975) )

cat(sprintf("p-value = %f\n", p.value))
```

```
## p-value = 0.051400
```

```
cat("95% CI: ", CI)
```

```
## 95% CI:  0.08559244 1.526765
```

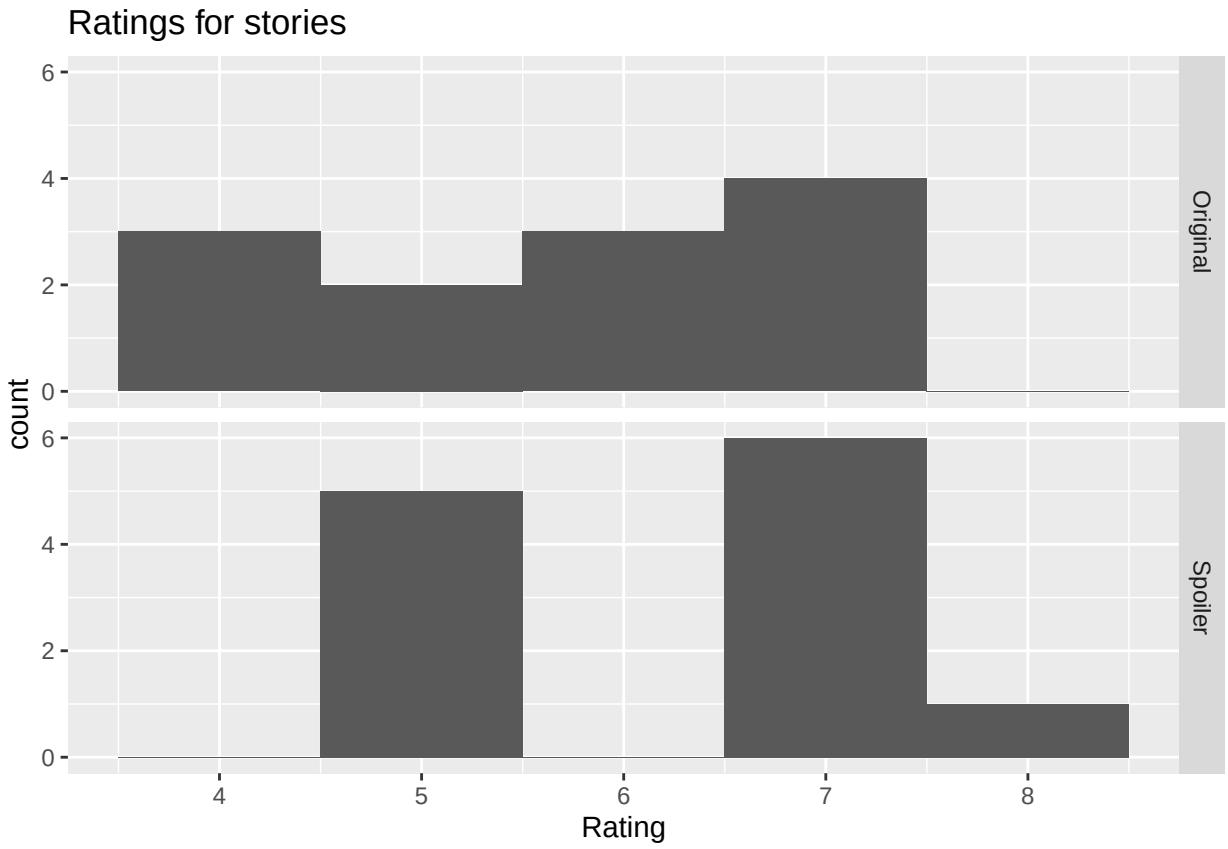
4)

```
data("StorySpoilers",package = "Lock5Data")

StorySpoilers.Long <- StorySpoilers %>%
 gather('Type', 'Rating', Spoiler, Original) %>%
 mutate( Story = factor(Story), # make Story and Type into
 Type = factor(Type) ) %>% # categorical variables
 arrange(Story)
```

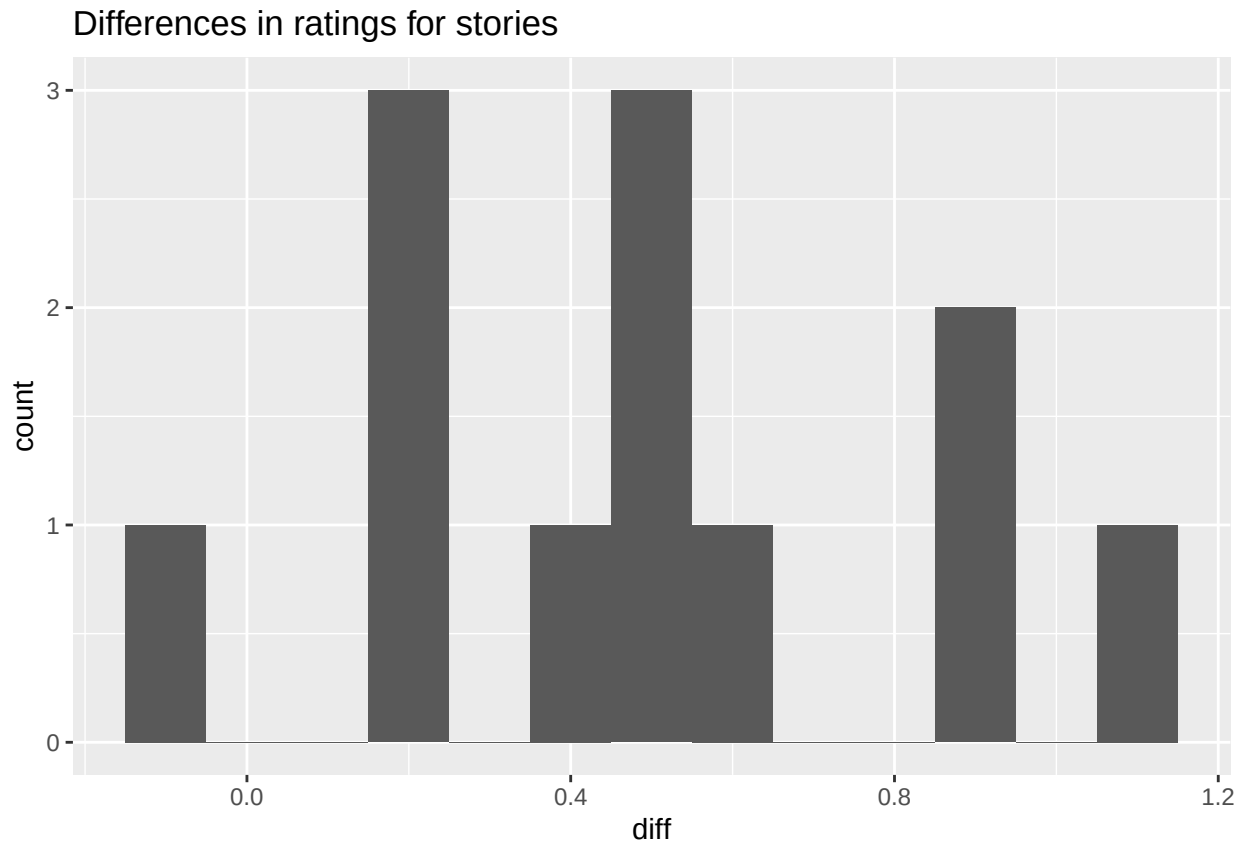
b)

```
ggplot(StorySpoilers.Long, aes(x=Rating)) +
  #geom_dotplot(binwidth=.1) +
  geom_histogram(binwidth=1) +
  facet_grid(Type ~ .) +
  ggtitle("Ratings for stories")
```



c)

```
StorySpoilers %>%
  reframe(diff = Spoiler - Original) %>%
  ggplot(aes(x=diff)) +
  geom_histogram(binwidth=0.1) +
  ggtitle("Differences in ratings for stories")
```



d)

```
spoilers.summary <- StorySpoilers.Long %>%
  group_by(Type) %>%
  summarize(x.bar = mean(Rating), stddev = sd(Rating), n=length(Rating))

t.delta <- test.statistic.pooled.var(
  spoilers.summary$x.bar[1],
  spoilers.summary$stddev[1],
  spoilers.summary$n[1],
  spoilers.summary$x.bar[2],
  spoilers.summary$stddev[2],
  spoilers.summary$n[2])
dof <- spoilers.summary$n[1] + spoilers.summary$n[2] - 2
critical.t <- qt(0.975, dof)
range <- critical.t * sqrt((spoilers.summary$stddev[1]^2/spoilers.summary$n[1]) +
                   (spoilers.summary$stddev[2]^2/spoilers.summary$n[2]))
x.diff <- spoilers.summary$x.bar[1] - spoilers.summary$x.bar[2]

p.value <- pt(t.delta, df=dof, ncp=0)
cat(sprintf("test statistic = %f\n", t.delta))
```

```
## test statistic = -0.162084
```

```
cat(sprintf("degrees of freedom = %f\n", dof))
```

```
## degrees of freedom = 22.000000
```

```
cat(sprintf("p-value = %f\n", p.value))
```

```
## p-value = 0.436360
```

```
cat(sprintf("95 percent CI: (%f, %f)\n", x.diff-range, x.diff+range))
```

```
## 95 percent CI: (-1.540152, 0.556819)
```

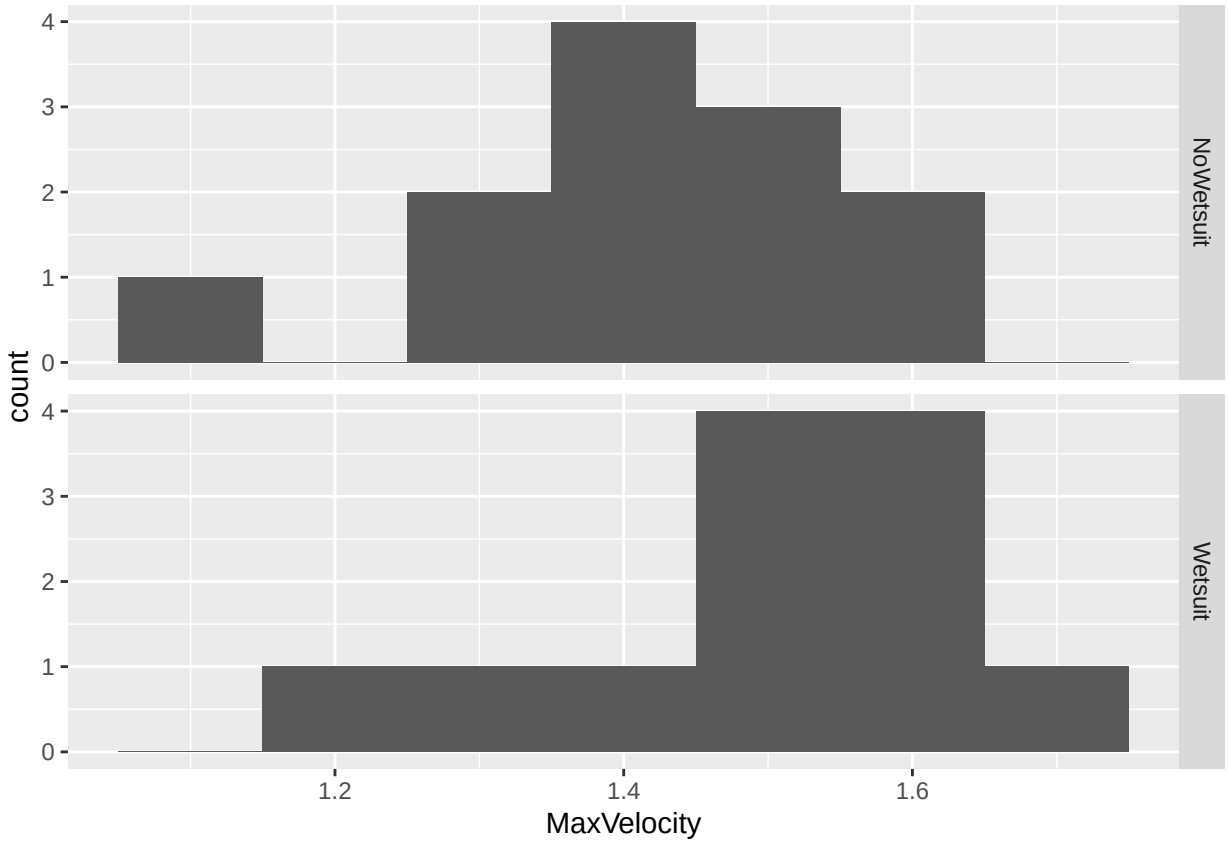
f) The t-test shows there isn't a significant difference between spoilers and not spoiling.

5)

```
data('Wetsuits', package='Lock5Data')
Wetsuits.Long <- Wetsuits %>%
  mutate(Participant = factor(1:12)) %>%
  gather('Suit', 'MaxVelocity', Wetsuit,NoWetsuit) %>%
  arrange(Participant, Suit) %>%
  mutate(Suit = factor(Suit))
```

b)

```
ggplot(Wetsuits.Long) +
  geom_histogram(aes(x=MaxVelocity), binwidth=.1) +
  facet_grid(Suit ~ .)
```



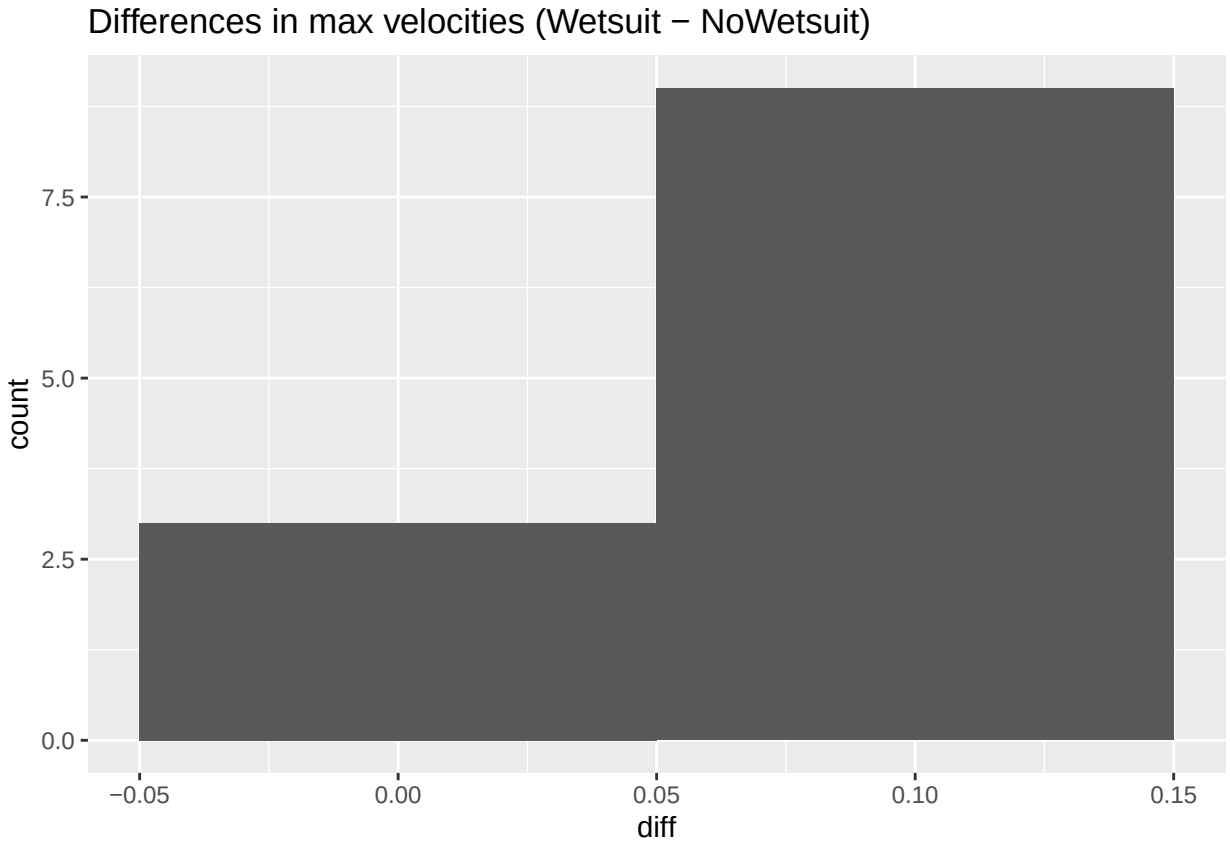
c)

```
mosaic::t.test(MaxVelocity ~ Suit, data=Wetsuits.Long, var.equal=FALSE, conf.level=0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  MaxVelocity by Suit
## t = -1.3688, df = 21.974, p-value = 0.1849
## alternative hypothesis: true difference in means between group NoWetsuit and group Wetsuit is not eq
## 95 percent confidence interval:
##  -0.19492937  0.03992937
## sample estimates:
## mean in group NoWetsuit   mean in group Wetsuit
##                1.429167                1.506667
```

d)

```
Wetsuits %>%
  reframe(diff = Wetsuit - NoWetsuit) %>%
  ggplot(aes(x=diff)) +
  geom_histogram(binwidth=0.1) +
  ggtitle("Differences in max velocities (Wetsuit - NoWetsuit)")
```



e)

```
mosaic::t.test(Wetsuits$Wetsuit, Wetsuits$NoWetsuit, var.equal=FALSE, conf.level=0.95, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  Wetsuits$Wetsuit and Wetsuits$NoWetsuit
## t = 12.318, df = 11, p-value = 8.885e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.06365244 0.09134756
## sample estimates:
## mean of the differences
##                  0.0775
```