Homework #7

Derek Sonderegger

```
library(ggplot2)
library(dplyr)
library(tidyr)
```

1.

In the 2011 article "Methane contamination of drinking water accompanying gas-well drilling and hydraulic fracturing" in the Proceedings of the National Academy of Sciences, $n_1 = 21$ sites in proximity to a fracking well had a mean methane level of $\bar{x}_1 = 19.2$ mg CH_4 L^{-1} with a sample standard deviation $s_1 = 30.3$. The $n_2 = 13$ sites in the same region with no fracking wells within 1 kilometer had mean methane levels of $\bar{x}_2 = 1.1$ mg CH_4 L^{-1} and standard deviation $s_2 = 6.3$. Perform a one-sided, two-sample t-test with unpooled variance at an $\alpha = 0.05$ level to investigate if the presence of fracking wells increases the methane level in drinking-water wells in this region. *Notice that because I don't give you the data, you can only analyze the data using the asymptotic method and plugging in the give quantities into the formulas presented.*

(a) State an appropriate null and alternative hypothesis. (Be sure to use correct notation!)

$$\begin{aligned} H_0 &: \mu_{diff} = 0 \\ H_a &: \mu_{diff} > 0 \end{aligned}$$

*where $\mu_{diff} = \mu_{gas} - \mu_{control}$.

(b) Calculate an appropriate test statistic (making sure to denote the appropriate degrees of freedom, if necessary).

$$t_{???} = \frac{(\bar{x}_{gas} - \bar{x}_{control}) - 0}{\sqrt{\frac{s^2_{gas}}{n_{gas}} + \frac{s^2_{control}}{n_{control}}}} = \frac{(19.2 - 1.1) - 0}{\sqrt{\frac{30.3^2}{21} + \frac{6.3^2}{13}}} = 2.6466$$

$$V_{gas} = \frac{30.3^2}{21} = 43.719$$

$$V_{control} = \frac{6.3^2}{13} = 3.053$$

$$df = \frac{(V_{gas} + V_{control})^2}{\frac{V^2_{gas}}{n_{gas}-1} + \frac{V^2_{control}}{n_{control}-1}} = \frac{(43.719 + 3.053)^2}{\frac{43.719^2}{20} + \frac{3.053^2}{12}} = 22.71$$

(c) Calculate an appropriate p-value.

$$\text{p.value} = P(T_{22.71} > 2.6466) =$$

```
1-pt(2.6466, df=22.71)
```

```
## [1] 0.007254139
```

(d) At an significance level of $\alpha = 0.05$, do you reject or fail to reject the null hypothesis?

Because the p-value is smaller than $\alpha = 0.05$ we will reject the null hypothesis.

(e) Restate your conclusion in terms of the problem. *We have statistically significant to conclude that wells near these fracking sites have higher levels of methane gas than wells that are not near fracking wells.*

2.

All persons running for public office must report the amount of money raised and spent during their campaign. Political scientists contend that it is more difficult for female candidates to raise money. Suppose that we randomly sample 30 male and 30 female candidates for state legislature and observe the male candidates raised, on average, $\bar{y} = \$350,000$ with a standard deviation of $s_y = \$61,900$ and the females raised on average $\bar{x} = \$245,000$ with a standard deviation of $s_x = \$52,100$. Perform a one-sided, two-sample t-test with pooled variance to test if female candidates generally raise and spend less in their campaigns that male candidates. *Notice that because I don't give you the data, you can only analyze the data using the asymptotic method and plugging in the givenxs quantities into the formulas presented.*

(a) State an appropriate null and alternative hypothesis. (Be sure to use correct notation!)

$$\begin{aligned} H_0 &: \mu_y - \mu_x = 0 \\ H_a &: \mu_y - \mu_x > 0 \end{aligned}$$

(b) Calculate an appropriate test statistic (making sure to denote the appropriate degrees of freedom, if necessary).

$$s_{pooled} = \sqrt{\frac{(n_x - 1)\, s_x^2 + (n_y - 1)\, s_y^2}{n_x + n_y - 2}} = \sqrt{\frac{29 * 52,100^2 + 29 * 61,900^2}{58}} = 57,210.23$$

$$t = \frac{(\bar{y} - \bar{x}) - 0}{s_{pooled}\sqrt{\frac{1}{n_y} + \frac{1}{n_x}}} = \frac{350,000 - 245,000}{57,210.23\sqrt{\frac{1}{30} + \frac{1}{30}}} = 7.11$$

where t has $n_y + n_x - 2 = 58$ degrees of freedom.

(c) Calculate an appropriate p-value.

$$\text{p.value} = P\left(T_{58} > 7.11\right) =$$

```
1-pt(7.11, df=58)
```

```
## [1] 9.576877e-10
```

(d) At an significance level of $\alpha = 0.05$, do you reject or fail to reject the null hypothesis? *Reject H_0.*

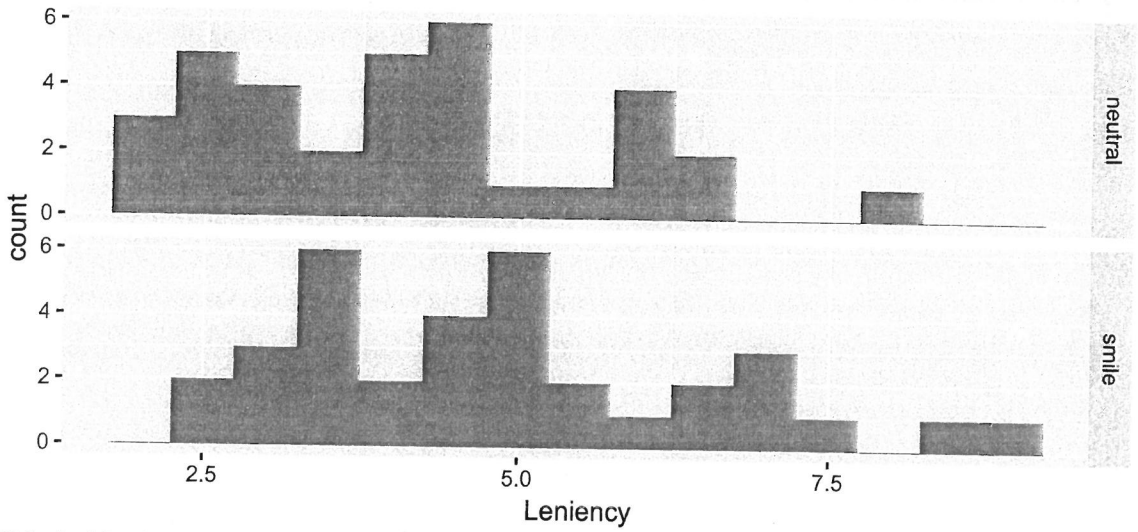
(e) Restate your conclusion in terms of the problem. *There is statistically significant evidence ($t = 7.11$, $df = 58$, $p = 9.56E-10$) to reject the null hypothesis that female candidates rase as much money as male candidates and conclude that female candidates raise less money.*

3.

In the `Lock5Data` package, the dataset `Smiles` gives data "...from a study examining the effect of a smile on the leniency of disciplinary action for wrongdoers. Participants in the experiment took on the role of members of a college disciplinary panel judging students accused of cheating. For each suspect, along with a description of the offense, a picture was provided with either a smile or neutral facial expression. A leniency score was calculated based on the disciplinary decisions made by the participants."

(a) Graph the leniency score for the smiling and non-smiling groups. Comment on if you can visually detect any difference in leniency score.

```
data(Smiles, package='Lock5Data')
ggplot(Smiles, aes(x=Leniency)) +
  geom_histogram(binwidth=.5) +
  facet_grid( Group ~ .)
```



It looks like the smile group has a slightly higher leniency score than the neutral faced group. The spread of the two distributions appears similar.

(b) Calculate the mean and standard deviation of the leniencies for each group. Does it seem reasonable that the standard deviation of each group is the same?

```
Smiles %>% group_by(Group) %>%
  summarise(xbar=mean(Leniency), s=sd(Leniency))
```

```
## # A tibble: 2 × 3
##     Group     xbar        s
##    <fctr>    <dbl>    <dbl>
## 1 neutral 4.117647 1.522850
## 2   smile 4.911765 1.680866
```

So we see that the mean of the smile group is about 0.8 *larger than the neutral group. Could that be just due to sampling variability?*

The two standard deviations are pretty close, and only differ by 0.10%*, so it probably is ok to assume equal variances in the two groups. (Next chapter we'll see how to test for that!) Finally we'll calculate the observed difference in means between the two groups.*

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

```
## # A tibble: 1 × 1
##           d
##       <dbl>
## 1 0.7941176
```

(c) Do a two-sided two-sample t-test using pooled variance using the asymptotic method. Report the test statistic, p-value, and a 95% CI.

```
mosaic::t.test(Leniency ~ Group, data=Smiles, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  Leniency by Group
## t = -2.0415, df = 66, p-value = 0.0452
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.57074128 -0.01749402
## sample estimates:
## mean in group neutral   mean in group smile
##              4.117647              4.911765
```

We see that these data provide moderate evidence suggesting that disciplinary boards are more lenient in cases where the profile picture is smiling (t=-2.04, df=66, p=0.0452). A 95% confidence interval is $(-1.57, -0.0174)$.

(d) Do a two-side two-sample t-test using resampling methods. Report the p-value and a 95% CI.

```
BootDist <- mosaic::do(10000)*{ Smiles %>%
    group_by(Group) %>%
    mosaic::resample() %>%
    summarise(xbar = mean(Leniency)) %>%
    summarise(d.star = diff(xbar))
}
PermDist <- mosaic::do(10000)*{ Smiles %>%
    mutate( Group = mosaic::shuffle(Group) ) %>%
    group_by(Group) %>%
    summarise(xbar = mean(Leniency)) %>%
    summarise(d.star = diff(xbar))
}

CI <- quantile(BootDist$d.star, probs=c(0.025, 0.975))
p.value <- 2 * mean( PermDist$d.star >= 0.7941176)
CI
```

```
##       2.5%      97.5%
## 0.05259285 1.54687747
```

```
p.value
```

```
## [1] 0.0508
```

So we see that a 95% CI for the difference in leniency (Smile - neutral) is $(0.053, 1.55)$ *and we fail reject the null hypothesis of no leniency at a* $\alpha = 0.05$ *level.*

(e) What do you conclude at an $\alpha = 0.05$ level? Do you feel we should have used a more stringent α level?

Because the p.value is quite close to the α *level for both the asymptotic and permuation test, and that we actually have the a difference in inference (the asymptotoic method would reject vs the permutation*

would fail-to-reject) this shouldn't be considered particularly strong evidence. Furthermore, the observed difference of 0.794 *should be considered on the scale of leniency.*

Given these results, it is reasonable to think that perhaps the student-to-student variability is large relative to the effect of smiling and we might consider how to do a similar experiment but using a paired design.

As far as the chosen α level is concerned, we fall back to the idea that extraordinary claims require extraordinary evidence. But because it isn't particular surprising that a smiling picture would elicit more leniency, we shouldn't feel too bad about our α choice, and perhaps we ought to have chosen a less stringent value of $\alpha = 0.10$. This choice seems particularly appropriate because the result of a rejection would be to recommend that a review board NOT be shown pictures of the student and you have to wonder why anyone ever thought it was a good idea to show photos.

4.

In the `Lock5Data` package, the dataset `StorySpoilers` is data from an experiment where the researchers are testing if a "spoiler" at the beginning of a short story negatively affects the enjoyment of the story. A set of $n = 12$ stories were selected and a spoiler introduction was created. Each version of each story was read by at least 30 people and rated. Reported are the average ratings for the spoiler and non-spoiler versions. The following code creates the "long" version of the data.

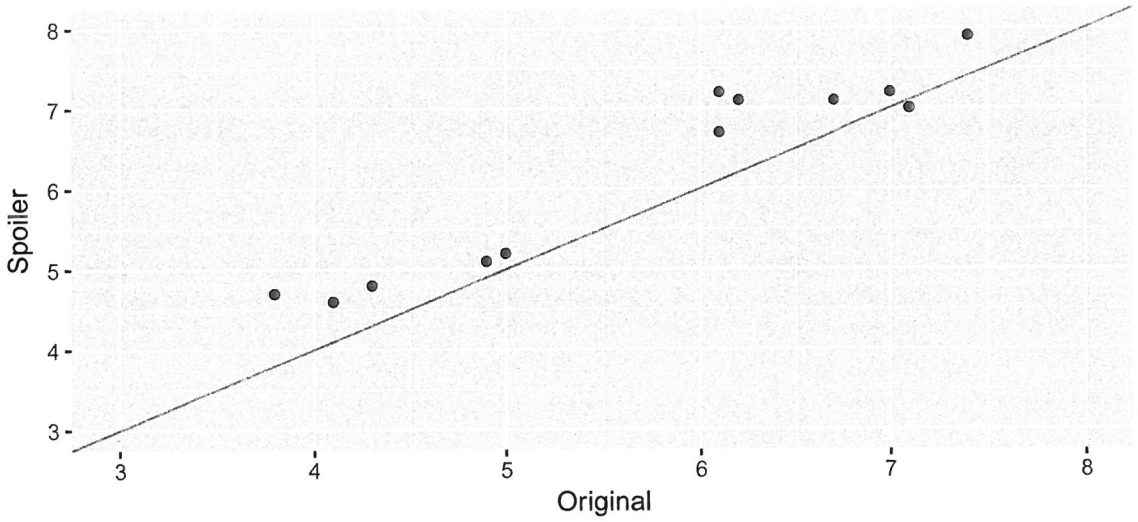
```
data(StorySpoilers, package='Lock5Data')
StorySpoilers.Long <- StorySpoilers %>%
  gather('Type', 'Rating', Spoiler, Original) %>%
  arrange(Story) %>%
  mutate( Story = factor(Story),
          Type  = factor(Type))
```

(a) Based on the description, a 1-sided test is appropriate. Explain why.

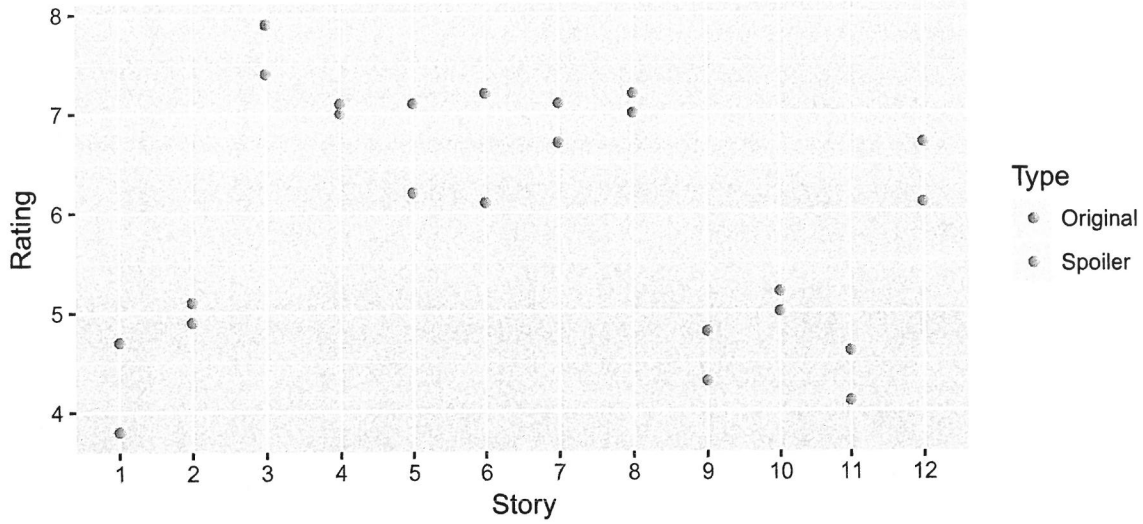
For this scenario, the idea is that a "Spoiler" spoiles the story and people won't enjoy it. So in this case, we only are interested in seeing if the enjoyment is decreased.

(b) Graph the ratings for the original stories and the modified spoiler version. Comment on if you detect any difference in ratings between the two. *I will show three different ways I came up with for showing the differences.*

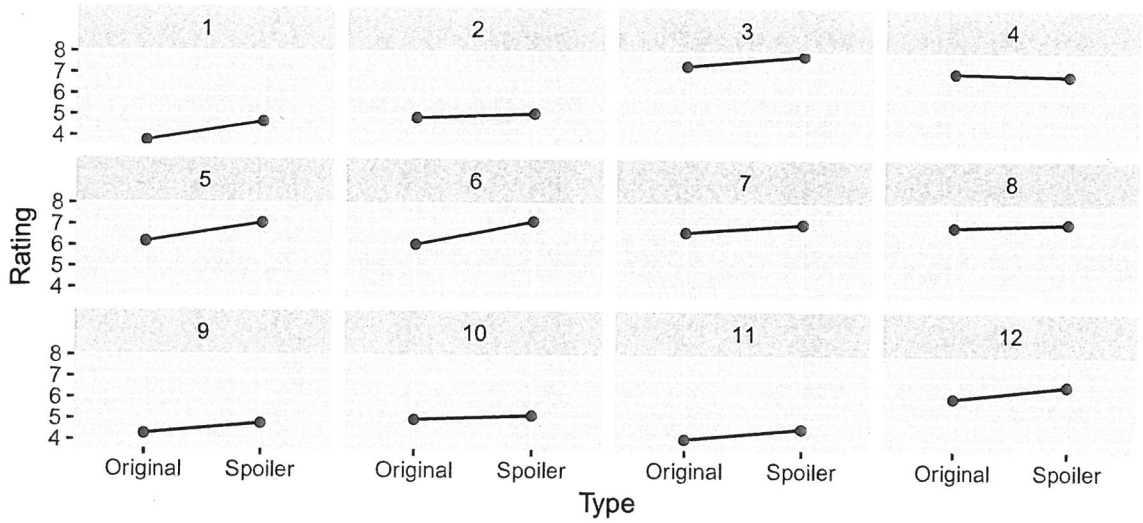
```
ggplot(StorySpoilers, aes(x=Original, y=Spoiler)) +
  geom_point() +
  geom_abline(slope=1, intercept=0, color='dark grey') +
  coord_cartesian(xlim=c(3,8), ylim=c(3,8))
```




```
ggplot(StorySpoilers.Long, aes(y=Rating, x=Story, color=Type)) +
  geom_point()
```




```
ggplot(StorySpoilers.Long, aes(y=Rating)) +
  geom_point(aes(x=Type)) +
  facet_wrap(~Story) +
  geom_line( aes(y=Rating, x=as.numeric(Type)) )
```



For 11 out of 12 stories, we see that the spoiler version has a higher rating! Sometimes by quite alot! This is exactly the opposite result that we expected. So in some ways, we could stop right here because we this data certainly doesn't support the alternative hypothesis that the ratings should decrease with the spoiler.

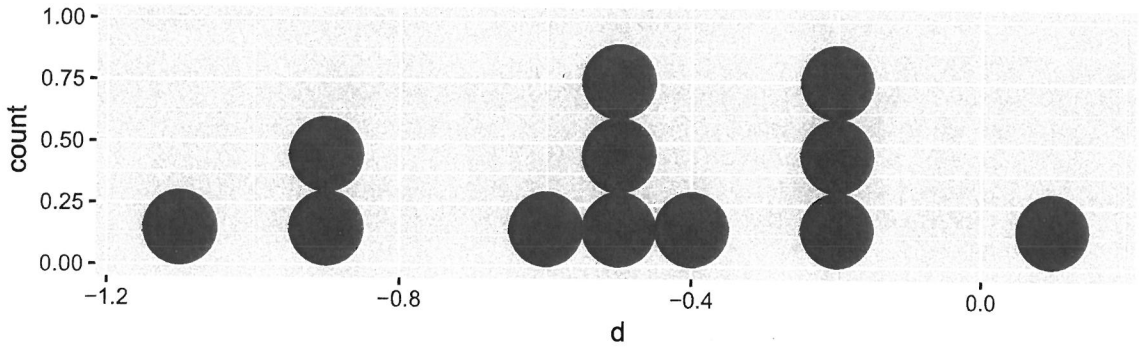
(c) Graph the difference in ratings for each story. Comment on if the distribution of the differences seems to suggest that a spoiler lowers the rating.

```
diff.data <- StorySpoilers.Long %>%
   group_by(Story) %>%
   arrange( Story, desc(Type) ) %>%      # order is Spoiler then Original
   summarise( d = diff(Rating))          # d = Original - Spoiler

diff.data %>% summarise(dbar = mean(d))    # observed mean difference
```

```
## # A tibble: 1 × 1
##         dbar
##        <dbl>
## 1 -0.4916667
```

```
ggplot(diff.data, aes(x=d)) +
  geom_dotplot( binwidth=.1)
```



This makes it even more clear that most of the differences are negative so we have Spoiler - Original > 0 and thus it it appears the additional spoiler at the beginning did not decrease the enjoyment, but rather it increased the enjoyment!

(d) Do a paired t-test using the asymptotic method. Also calculate a 95% confidence interval.

```
t.test( diff.data$d, mu=0, alternative='less')

##
##  One Sample t-test
##
## data:  diff.data$d
## t = -4.8997, df = 11, p-value = 0.0002359
## alternative hypothesis: true mean is less than 0
## 95 percent confidence interval:
##        -Inf -0.3114557
## sample estimates:
##  mean of x
## -0.4916667
```

We find insufficient evidence to conclude that the spoiler actually spoils the enjoyment. Rather we actually found evidence to suggest that the spoiler increases the listeners enjoyment. Because this is so contrary to what we expectd at the outset of the experiment, the follow up analysis where we consider the not equal alternative will be performed at a much lower α level due to the extraordinarily nature of the claim.

```
t.test( diff.data$d, mu=0 )

##
##  One Sample t-test
##
## data:  diff.data$d
## t = -4.8997, df = 11, p-value = 0.0004719
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.7125281 -0.2708052
## sample estimates:
##  mean of x
## -0.4916667
```

The post-hoc test where we allow a two-sided alternative is highly significant (t=4.90, df=11, p=0.0005) is highly significant even for extremely small signifcance threshold such as $\alpha = 0.001$.

For the post-hoc test, the resulting two-sided confidence interval for the increase in enjoyment from adding the spoiler is (0.27, 0.71) points.

(e) Do a paired t-test using the permutation method. Also calculate a 95% confidence interval using bootstrapping.

```
PermDist <- mosaic::do(10000)*{ StorySpoilers.Long %>%
    group_by(Story) %>%
    mutate( Group = mosaic::shuffle(Type) ) %>%
    arrange( desc(Type) ) %>%          # order is Original then Spoiler
    summarise( d = diff(Rating))  %>%  # d = Spoiler - Original
    summarise(dbar.star = mean(d)) }
BootDist <- mosaic::do(10000)*{ StorySpoilers.Long %>%
    group_by(Story) %>%
    arrange( desc(Type) ) %>%         # order is Original then Spoiler
    summarise( d = diff(Rating)) %>%  # d = Spoiler - Original
    mosaic::resample() %>%
    summarise(dbar.star = mean(d)) }
```

```
p.value <- mean( PermDist$dbar.star >= -0.491667)
p.value
```

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

As expected the p-value for the 1-sided test was nearly 1. The follow up test where we consider the two sided case has a p-value of and 95% confidence interval of:

```
CI <- quantile(BootDist$dbar.star, probs=c(0.025, 0.975))
p.value <- 2 * mean( PermDist$dbar.star <= -0.491667)
CI
```

```
##       2.5%      97.5%
## -0.6833333 -0.3083333
```

```
p.value
```

```
## [1] 0
```

This p-value is extradinarily small and the best we can say is that the p-value is less than 1 / 10,000.

(f) Based on your results in parts (c) and (d), what do you conclude?

In this case we are flabergasted to learn that our initial suspicion that a spoiler would decrease enjoyment was not supported but instead we found the opposite effect and that effect is so strong that even at an extremely low α the effect is still present.

Because this result is so surprising, we should attempt to replicate this study with more stories as it is possible that our selection of the stories might have made a difference here.