HW2

Solutions

The code below just loads some packages and makes it so that enough digits are printed that you won't get confused by rounding errors.

```
library(dplyr) # functions like summarize
```

```
## Warning: package 'dplyr' was built under R version 3.5.2
```

```
library(ggplot2) # for making plots
```

```
## Warning: package 'ggplot2' was built under R version 3.5.2
```

```
library(mosaic) # convenient interface to t.test function
```

```
## Warning: package 'mosaic' was built under R version 3.5.2
```

```
## Warning: package 'ggformula' was built under R version 3.5.2
```

```
options("pillar.sigfig" = 10) # print 10 significant digits in summarize output
```

Problem 1: Adapted from Sleuth3 2.20

Researchers assigned 7 volunteer males to a special fish oil diet. They recorded each subject's diastolic blood pressure at baseline and again after four weeks on the diet. The researcher's interest was in the reduction diastolic blood pressure (mm of mercury) between baseline and 4 weeks later (a positive reduction is a good thing). The R code below reads in the data set. The variable `BP` records each subject's reduction in blood pressure.

```
fish_oil <- read.csv("http://www.evanlray.com/data/sleuth3/ex0112_oil_diets.csv") %>%
  filter(Diet == "FishOil")
```

(a) Define the relevant parameter(s) the researchers wanted to estimate.

The parameter of interest is the average reduction in diastolic blood pressure after 4 weeks on a fish oil diet, in a population of people similar to the 7 volunteers in this study.

If you want, you can be less specific about the target population here as long as you address the scope of your conclusions when you draw conclusions for the hypothesis test and interpret the confidence interval.

(b) Calculate the sample mean and standard deviation of the blood pressure reduction.

```
fish_oil %>%
  summarize(
    mean_bp = mean(BP),
    sd_bp = sd(BP)
  )
```

```
##    mean_bp  sd_bp
## 1 6.571429 5.8554
```

(c) Compute the standard error for the sample mean. What are the degrees of freedom?

Confirming that there are 7 rows in our data set:

```
dim(fish_oil)
```

```
## [1] 7 2
```

```
5.8554 / sqrt(7)
```

```
## [1] 2.213133
```

The degrees of freedom is 7 - 1 = 6.

(d) Conduct a relevant hypothesis test.

i. Define the null and alternative hypotheses for the test

H_0: $\mu = 0$. On average, the fish oil diet has no impact on a person's blood pressure, in the population people like those who were volunteers for this study.

H_A: $\mu \neq 0$. On average, there is some impact of the fish oil diet on a person's blood pressure, in the population people like those who were volunteers for this study.

If you went into the study believing that the fish oil diet would lead to a reduction in blood pressure, you might instead specify a one-sided alternative:

H_A: $\mu > 0$. On average, the fish oil diet leads to a positive reduction in a person's blood pressure, in the population people like those who were volunteers for this study.

Note: You need to either make a statement about scope of conclusions as part of defining the parameter here, or as part of drawing conclusions in part v.

ii. Find the value of the t statistic for this test. What is the interpretation of this statistic?

```
6.571429 / (5.8554 / sqrt(7))
```

```
## [1] 2.969288
```

The sample mean reduction in blood pressure after 4 weeks on the fish oil diet is about 3 standard errors larger than the difference of 0 specified in the null hypothesis.

iii. Find the p-value for the test using the `t.test` function in R. Don't forget to specify the value of `mu` from the null hypothesis. (See the corrected handout on the course website from Mon, Sep 9 – the original handout had incorrect code.)

```
t.test(~ BP, mu = 0, data = fish_oil)
```

```
##
##  One Sample t-test
##
## data:  BP
## t = 2.9693, df = 6, p-value = 0.02498
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   1.156086 11.986771
## sample estimates:
## mean of x
##  6.571429
```

It turns it it's actually ok if you didn't specify `mu = 0` in your code since that's the default.

The p-value for a two-sided test is 0.02498. If we had specified a one-sided test in part i, the p-value would instead be 0.01249, which is half of the p-value above. It is important that your p-value match the form of your alternative hypothesis.

iv. What is the interpretation of the p-value? (I'm not looking for a conclusion about strength of evidence yet, just a statement of what the p-value is in the context of this example.)

If you used a two-sided hypothesis test:

If on average the fish oil diet had no impact on a person's blood pressure, the probability of observing a t statistic at least as large (in abolute value) as 2.9693 is 0.02498.

If you used a one-sided hypothesis test:

If on average the fish oil diet had no impact on a person's blood pressure, the probability of observing a t statistic at least as large as 2.9693 is 0.02498.

v. What is the conclusion of the test? Please state this in terms of strength of evidence about the null hypothesis.

The data provide a moderate amount of evidence against the null hypothesis that fish oil has no effect on a person's blood pressure, in the population of people similar to those who volunteered to participate in the study.

(e) Find a relevant confidence interval.

i. Find a confidence interval from the formula, using output from the `qt` function as needed. Confirm that your interval matches the interval from the output of `t.test` in part (d) iii, up to rounding error.

```
qt(0.975, df = 6)
```

```
## [1] 2.446912
```

```
6.571429 - 2.447 * (5.8554 / sqrt(7))
```

```
## [1] 1.155892
```

```
6.571429 + 2.447 * (5.8554 / sqrt(7))
```

```
## [1] 11.98697
```

The confidence interval calculated here is [1.156, 11.987], which matches the results from part (d) iii up to rounding error.

ii. Interpret your interval in context, including a statement of what the phrase "95% confident" means.

We are 95% confident that the interval [1.156, 11.987] contains the mean change in a person's blood pressure after 4 weeks on a fish oil diet, among the population of people similar to those who participated in this study. If we were to take many different samples from this population and calculate a similar 95% confidence interval based on each of those samples, approximately 95% of those confidence intervals would contain the mean change in a person's blood pressure after 4 weeks on a fish oil diet, among the population of people similar to those who participated in this study.

iii. Is the confidence interval from part ii guaranteed to contain the sample mean?

Yes: the interval is obtained by subtracting a margin of error from the sample mean and adding a margin of error to the sample mean, so the sample mean is guaranteed to be contained in the confidence interval.

iv. Is the confidence interval from part ii guaranteed to contain the population mean?

No. For 95% of samples, a confidence interval calculated based on that sample will contain the population mean. However, for 5% of samples, a confidence interval calculated based on that sample will not contain the population mean. For a given sample, there is no way to know whether or not the confidence interval we have calculated contains the population parameter.

Problem 2: Adapted from Sleuth3 2.23

Ther National Highway System Designation Act was signed into law in the United States on November 29, 1995. Among other things, the act abolished the federal mandate of 55-mile-per-hour maximum speed limits on roads in the United States and permitted states to establish their own limits. Of the 50 states (plus the District of Columbia), 32 increased their speed limits either at the beginning of 1996 or sometime during 1996.

The R code below reads in data with the percentage changes in interstate highway traffic fatalities from 1995 to 1996 (the variable is called `PctChange` in the data frame). Among the states where the speed limit increased, what evidence is there that the average percent change in fatalities was different from 0?

Conduct a full analysis, including:

- an appropriate plot with informative axis labels,
- summary statistics that would be meaningful to someone who had not taken a statistics class (i.e., don't report the t statistic),
- a confidence interval, and
- a hypothesis test.

Interpret all of your results in context. Explain how to interpret the p-value for the test and the conclusions that can be drawn from it as though to someone who had not taken a statistics class. Similarly, explain how to interpret your confidence interval. You do not need to calculate the p-value or the confidence interval by hand; you can use output from the `t.test` function. What conclusions can be drawn about whether this policy change was a good idea?

```
fatalities <- read.csv("http://www.evanlray.com/data/sleuth3/ex0223_highway_safety.csv") %>%
  filter(SpeedLimit == "Inc")
```

```
dim(fatalities)
```

```
## [1] 32  5
```

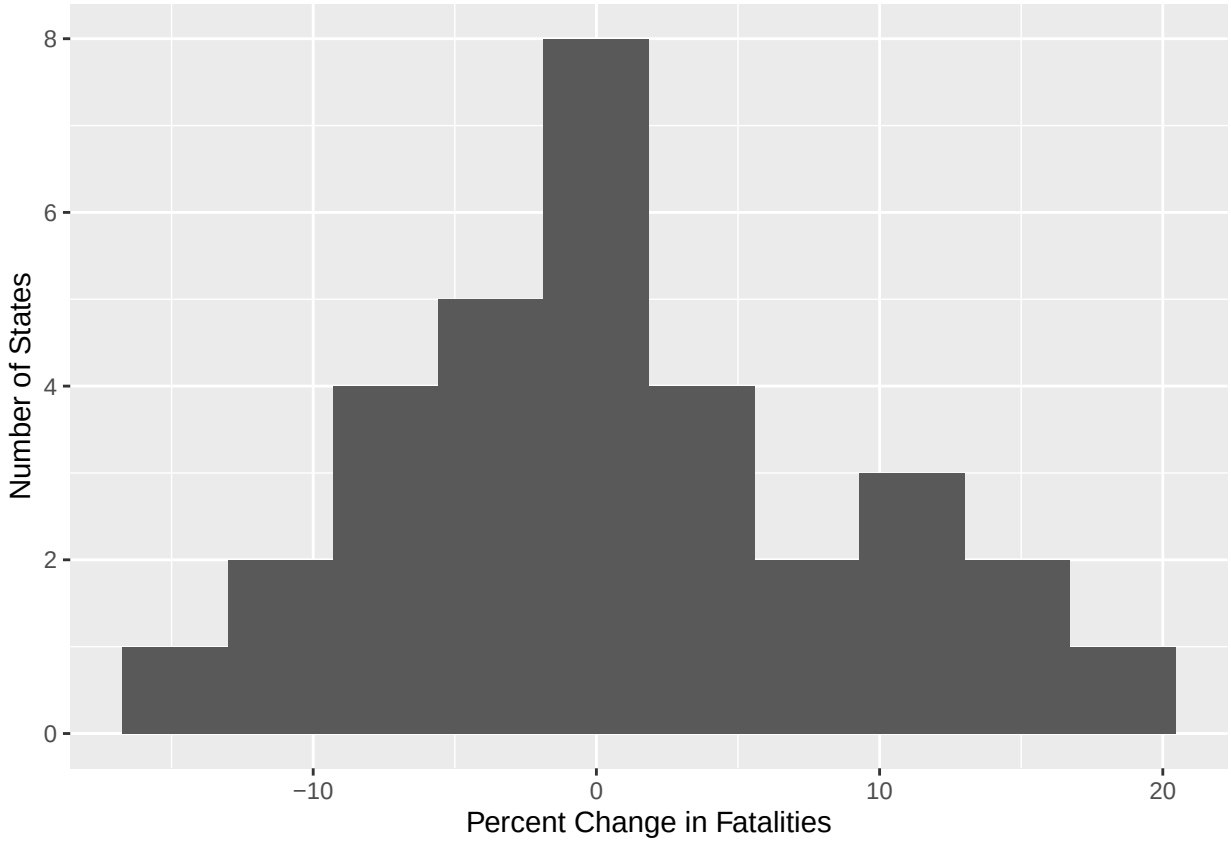
```
fatalities %>%
  summarize(
    mean_pct_change = mean(PctChange),
    min_pct_change = min(PctChange),
    max_pct_change = max(PctChange)
  )
```

```
##   mean_pct_change min_pct_change max_pct_change
## 1         0.49375         -15.88          17.56
```

```
t.test( ~ PctChange, data = fatalities)
```

```
##
##  One Sample t-test
##
## data:  PctChange
## t = 0.34503, df = 31, p-value = 0.7324
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -2.424853  3.412353
## sample estimates:
## mean of x
##   0.49375
```

```
ggplot(data = fatalities, mapping = aes(x = PctChange)) +
  geom_histogram(bins = 10) +
  xlab("Percent Change in Fatalities") +
  ylab("Number of States")
```



First, we should note that this was an observational study, so it is difficult to use this data set to make any definitive claims about a causal link between speed limits and changes in the numbers of fatalities due to traffic accidents. Additionally, it would be difficult to frame the states in our data set as a random sample from any population. However, these data could tell us what we might see if a state made an increase to its speed limit under circumstances similar to the states in this sample.

Among the 32 states that increased the speed limit, there was a roughly even split in terms of states that had reduced frequency of fatalities due to traffic accidents accidents and states that had increased frequency of fatalities. Across these states, the mean percent change in fatalities was about 0.5%, with a 95% confidence interval of approximately -2.4% to 3.4%. We can think of this interval as a range of plausible values for the average percent change in fatalities due to traffic accidents that we might see if a state similar to those in

this sample increased its speed limit; in 95% of samples, a similar interval would contain the average percent change in fatalities.

A hypothesis test of the claim that the speed limit increases had no association with changes in the number fatal accidents had a p-value of 0.73. This means that if in fact the speed limit increases were not associated with any change in the frequency of fatal accidents, there would be a 73% chance of seeing an average difference of at least 0.5% (or -0.5%). Our data are therefore consistent with that hypothesis of no effect; the data do not offer any evidence to rule out the possibility that the speed limit increase had no association with a change in the frequency of fatal accidents.

Overall, the data suggest that if the speed limit increase had any effect on the frequency of fatal accidents, it was small in percentage terms.