Problem Set 6

Daniel Shapiro

10/20/2022

Question 1 Background:

Download the votes.csv dataset from the course website. These data describe the number of votes obtained by the Democratic and Republican candidates in each presidential election from 1932 to 2008.

```
votes <- read.csv("votes.csv")
```

1a) Delete observations for Alaska, DC, and Hawaii (since residents of these states did not have the vote throughout this entire time period). Convert the data from wide to long format, with four variables (state, party, election, and vote).

```
# I'm going to do this in kind of an annoying way.

votes <- votes %>%
  filter(state != "Hawaii") %>%
  filter(state != "Alaska") %>%
  filter(state != "District of Columbia") %>%
  pivot_longer(!state, names_to = "election", values_to = "vote") %>%
  separate(election, c("party", "election"))
```

1b) Transform the data to a state-election year dataset, with four variables (election, state, pcDem, turnout).

```
# Had to suppress this really annoying message

options(dplyr.summarise.inform = FALSE)

newvotes <- votes %>%
  group_by(election, state) %>%
  summarize(turnout = sum(vote))

demvotes <- votes %>%
  filter(party == "d")

votemerge <- merge(newvotes, demvotes, by = c("election", "state")) %>%
  select(-party) %>%
  mutate(pcDem = vote/turnout*100) %>%
```

```
  select(-vote)

votemerge <- votemerge[,c(1, 2, 4, 3)]
```

1c) Plot democratic vote share over time for Pennsylvania, South Carolina, and West Virginia using a `ggplot()` line graph with clear and appropriately labeled legends and axes.

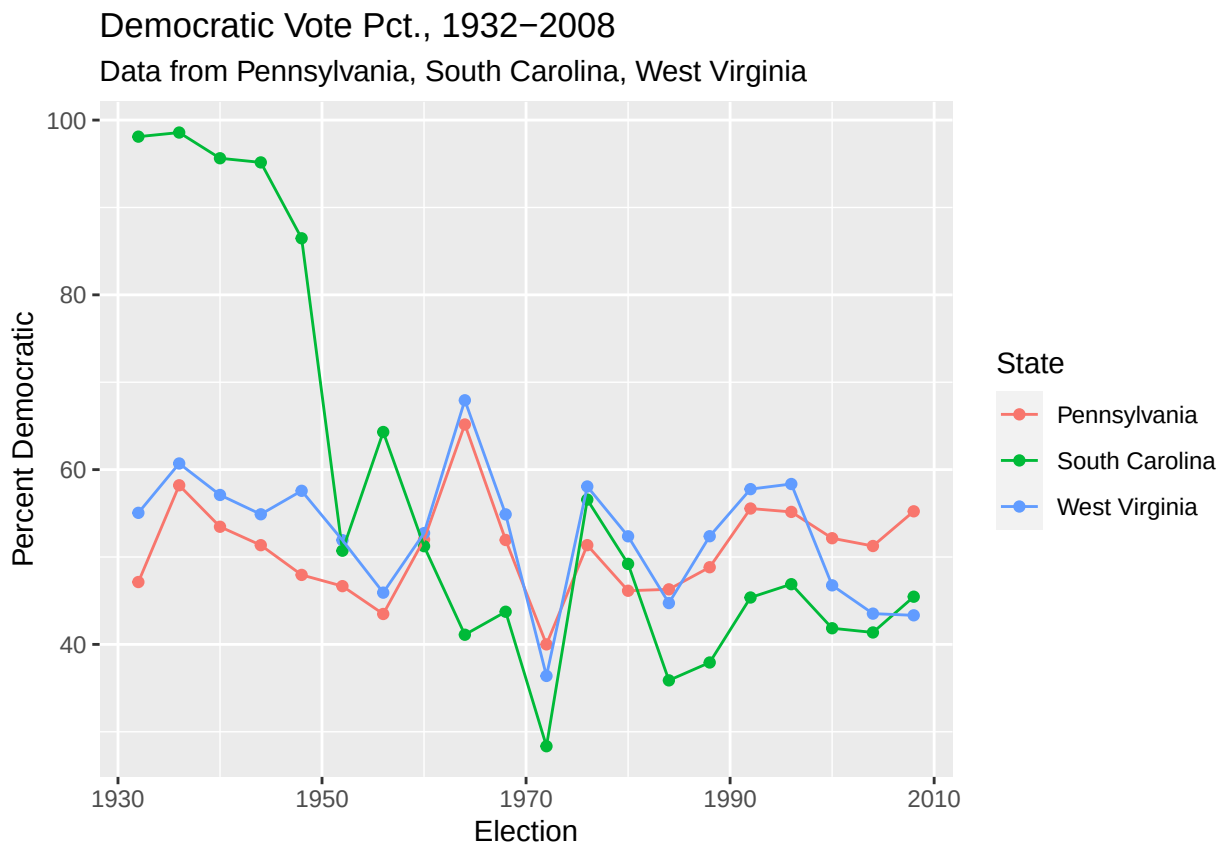
```
# West Virginia is misspelled...edit out with recode() below

cdata <- votemerge %>%
  filter(state %in% c("Pennsylvania", "South Carolina", "West Vriginia"))

cdata$state <- recode(cdata$state, `West Vriginia` = "West Virginia")

cdata$election <- as.numeric(cdata$election)

ggplot(cdata, aes(x = election, y = pcDem)) +
  geom_point() +
  geom_line() +
  aes(color = state) +
  labs(x = "Election",
       y = "Percent Democratic",
       color = "State",
       title = "Democratic Vote Pct., 1932-2008",
       subtitle = "Data from Pennsylvania, South Carolina, West Virginia")
```

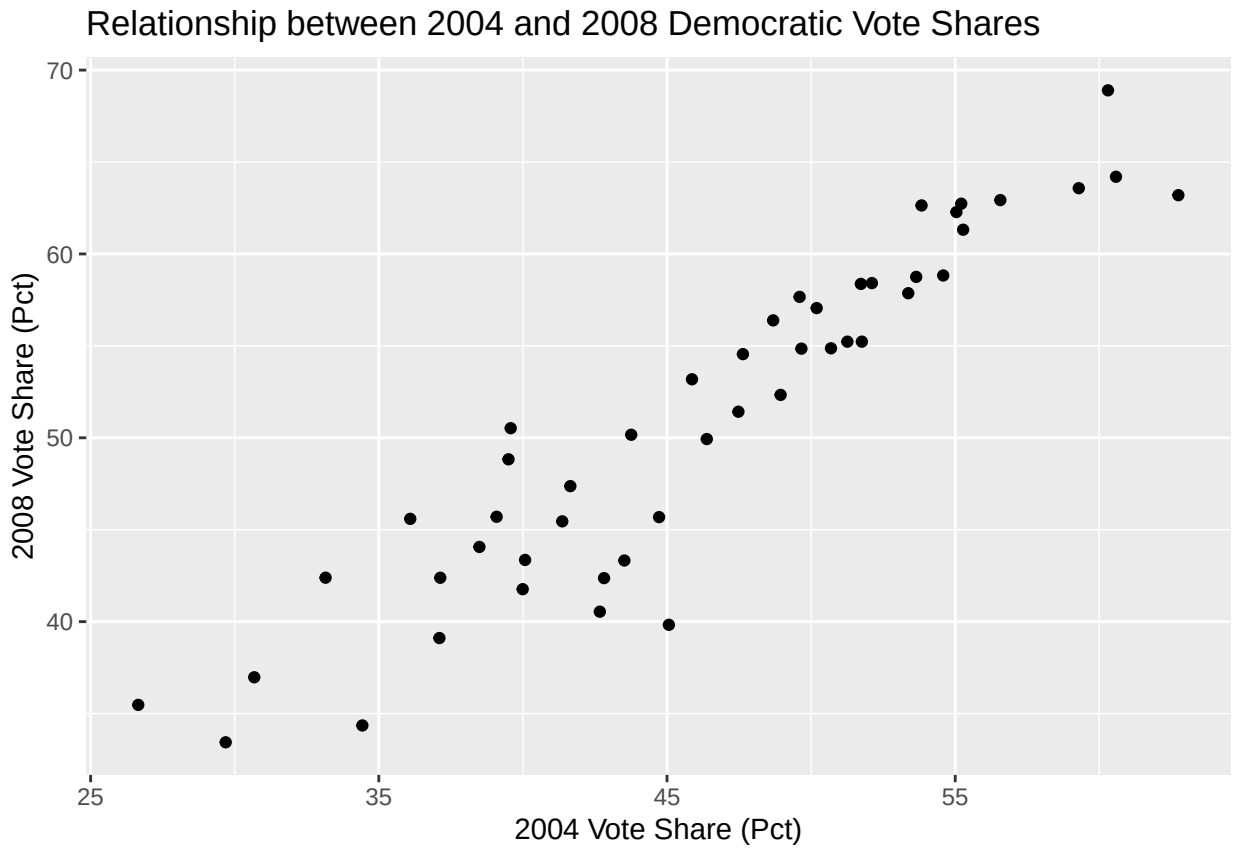


1d) Use `ggplot()` to make a scatterplot showing the relationship between 2004 and 2008 democratic vote share.

```
ddata <- votemerge %>%
  filter(election %in% c(2004, 2008))
```

```
graphdata <- ddata %>%
  select(-turnout) %>%
  pivot_wider(names_from = election, values_from = pcDem)
```

```
ggplot(graphdata, aes(x = `2004`, y = `2008`)) +
  geom_point() +
  labs(x = "2004 Vote Share (Pct)",
       y = "2008 Vote Share (Pct)",
       title = "Relationship between 2004 and 2008 Democratic Vote Shares")
```



1e) Use a paired t-test to test whether the difference between the 2004 and 2008 democratic vote share was significantly different across states. Interpret your results.

```
# First, I split my data into two -- 2004 and 2008 and select pcDem.

e2004 <- ddata %>%
  filter(election == 2004) %>%
  select(pcDem)

e2008 <- ddata %>%
  filter(election == 2008) %>%
  select(pcDem)

t.test(e2004[,1], e2008[,1], paired = TRUE, alternative = "two.sided")
```

```
##
##  Paired t-test
##
## data:  e2004[, 1] and e2008[, 1]
## t = -10.415, df = 47, p-value = 8.493e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.861352 -3.963520
## sample estimates:
```

```
## mean of the differences
##               -4.912436
```

The p-value of the test is 8.493^{-14}, which is lower than the significance level $\alpha = 0.05$. Thus, we can reject the null hypothesis and say that the average Democratic vote share in 2008 is significantly different than the average Democratic vote share in 2004, with a p-value at 8.493^{-14}.

Question 2 Background:

Download the olken.csv dataset from the course website, which comes from an experiment described in Benjamin A. Olken. 2007. "Monitoring Corruption: Evidence from a Field Experiment in Indonesia.'' Journal of Political Economy. 115(2): 200-249.

Missing data from the dataset were omitted. The experiment sought to test the effect of efforts to encourage communities to monitor road building projects in Indonesian villages on reducing corruption. The main treatment variable is whether or not residents in a particular village were invited to participate at accountability meetings in which project officials account for how they spent project funds. The main dependent variable is a measure of the difference between what the villages claimed they spent on road construction and an independent estimate of what the villages actually spent. The variables in the dataset are:

- *pct_missing*: Percent expenditures missing
- *treat_invite*: Treatment assignment
- *mosques*: Mosques per 1,000

```
olken <- read.csv("olken.csv")
```

2a) Estimate the population average μ of the *pct_missing* variable with the sample mean. Report the standard error and a 95% confidence interval for your estimate as well.

```
missing_mean <- mean(olken$pct_missing)
std_err <- sd(olken$pct_missing) / sqrt(length(olken$pct_missing))
confidence = quantile(olken$pct_missing, c(.025, .975))

missing_mean
```

```
## [1] 0.2363279
```

```
std_err
```

```
## [1] 0.01585687
```

```
confidence
```

```
##       2.5%      97.5%
## -0.3349367  1.0007106
```

2b) Conduct a two-sided t-test with the null hypothesis that $\mu_0 = 0$ with $\alpha = 0.05$. Report your test statistic and p-value. What is the interpretation of this p-value? Can you reject the null hypothesis?

```
t.test(olken$pct_missing)
```

```
##
##  One Sample t-test
##
## data:  olken$pct_missing
## t = 14.904, df = 471, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.2051690 0.2674869
## sample estimates:
## mean of x
## 0.2363279
```

Here, the test statistic is 14.904, and the p-value is $2.2 * 10^{-16}$. The p-value is way lower than 0.05, so the null hypothesis can be rejected at the 5% level of significance.

2c) Repeat b) with the null hypothesis $\mu_0 = 0.25$. Hint: check the documentation for `t.test` using `?t.test` to figure out how to do this.

```
t.test(olken$pct_missing, mu = 0.25)
```

```
##
##  One Sample t-test
##
## data:  olken$pct_missing
## t = -0.86222, df = 471, p-value = 0.389
## alternative hypothesis: true mean is not equal to 0.25
## 95 percent confidence interval:
##  0.2051690 0.2674869
## sample estimates:
## mean of x
## 0.2363279
```

Here, the test statistic is -0.86222, and the p-value is 0.389. This time, the p-value is way higher than 0.05, meaning that we cannot reject this null hypothesis at the 5% level of significance.

2d) Calculate the t-statistic of the null hypothesis in parts b) and c) analytically. Confirm you get the same results. Explain what the test statistic means.

To get our test statistic, we can use the following formula: $\frac{\mu_a - \mu_0}{\hat{SE}(\mu_a)}$. In the first case, $\mu_0 = 0$, so we can simply say $\frac{\mu_a}{\hat{SE}(\mu_a)}$, or our missing_mean value over the std_err value from our previous question.

```
tvalue <- missing_mean/std_err
tvalue
```

```
## [1] 14.90382
```

This is the same result as the t-value we got in 2b).

For 2c), we can get our test statistic using the same formula, but instead, it turns into $\frac{\mu_a - .25}{\hat{SE}(\mu_a)}$. Here it is:

```
tvaluec <- (missing_mean-.25)/std_err
tvaluec
```

```
## [1] -0.8622165
```

Again, this confirms the value that we got in 2c).

The test statistic is any function of the data that dictates whether the null should be rejected or not. In this case, we are using the t-value, given that we're doing t-tests.

2e) Let Y_t denote *pct_missing* for villages that received the treatment and Y_c denote *pct_missing* for villages that did not receive treatment. Assume that Y_t and Y_c are independent and have unequal variances. Conduct a two-sided t-test for the equality of means of Y_t and Y_c where the null hypothesis is $H_0 : \mu_{Y_t} = \mu_{Y_c}$. Report your test statistic and p-value, as well as a 95% confidence interval. Can you reject the null hypothesis at the $\alpha = 0.05$ level? Give a brief substantive explanation of your result (i.e. what can you say about the effect of the treatment on corruption?).

```
# First, I need to split these into two vectors: one that was treated and one
# that was not.

treated <- olken %>%
  filter(treat_invite == 1)

untreated <- olken %>%
  filter(treat_invite == 0)

# Now, I can insert these two vectors into a t-test.

t.test(treated$pct_missing, untreated$pct_missing)
```

```
##
##  Welch Two Sample t-test
##
## data:  treated$pct_missing and untreated$pct_missing
## t = -0.75376, df = 333.66, p-value = 0.4515
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.09006088  0.04016183
## sample estimates:
## mean of x mean of y
## 0.2278176 0.2527671
```

Here, we used Welch's Two Sample t-test, which can also basically be described as an unequal variances t-test. The test statistic is the t-value, at -0.75376, and the p-value is 0.4515. The lower bound of the 95% confidence interval is -0.09, and the upper bound is 0.04. The p-value is quite high – 0.4515, so we cannot reject the null hypothesis at the 5% level of significance.

Putting this into more applied language, basically, the null hypothesis is that we should observe *no* difference in the amount of money "missing" between villages that have been treated (i.e. where villagers have been invited to participate in accountability meetings) and villages that were untreated (where villagers were not invited to said meetings). If we cannot reject the null hypothesis at the 5% level of significance, it means that we cannot argue (at this level of significance) that inviting villagers to participate in accountability meetings had a significant impact on corrupt activity. This does not mean we can *confirm* the null hypothesis; simply that we cannot reject it.

2f) Suppose you were actually running this experiment and your research assistant comes to you and says "I think we can get better results if we ran a one-sided test instead." Do you think this is a good idea? Why or why not? What assumptions would you be making?

I would say that it's most likely not a good idea. As Jane said in class, using a one-sided test "comes off as shady." There are a number of assumptions that you're making, most notably that the consequences of not recording an effect in the direction that is getting skipped are negligible. In this case, this could potentially look like the research assistant saying that he/she is only interested in seeing whether the treated villages experienced less corruption than the untreated villages, and that he/she does not care at all about any potential data in which the untreated villages experienced less corruption than the treated villages.

This does not look like a great idea. The research question does not say that this is what the researchers are looking for; rather, it just says that the experiment tests "the effect of efforts." This "effect," as listed in the research question, is value-neutral. If I would even consider running a one-sided test instead of a normal two-sided one, I'd want to have a well-written explanation as to why exactly we only care about effects in one direction, and not the other. Given that I don't see a particularly good rationale why this would be the case in this instance, I don't think I'd be likely to accept my research assistant's proposition.

2g) Repeat e) using a one-sided test where the alternative hypothesis is $H_1 : \mu_{Y_t} < \mu_{Y_c}$.