Problem Set 1: Answer Key

Quantitative Political Methodology (U25 363)

Due: February 6, 2018

Instructions

- *Please show your work if possible. You may lose points by simply writing in the answer. If the problem requires you to execute commands in* `R`*, please include the code you used to get your answers. Please also include the* `.R` *file that contains your code. If you have plots, attach them as well within your written document. Make sure you label clearly which question the codes correspond to. If you are not sure if work needs to be shown for a particular problem, please ask me.*
- *Your homework should be submitted electronically on the course GitHub page.*
- *This problem set is due before the beginning of class on Wednesday February 27, 2019. No late assignments will be accepted.*
- *Total available points for this homework is 90.*

Question 1 (8 points)

Go to the GSS Web site, sda.berkeley.edu/GSS/ and click on "GSS – with 'No Weight' as the default (SDA 4.0)". By entering LDCGAP, find a summary of responses to the question, "Turning to international differences, do you agree or disagree: Present economic differences between rich and poor countries are too large."

The screen shot below is taken from the GSS website.

LDCGAP DIFFERENCES BETWEEN RICH AND POOR TOO LARGE

Description of the Variable

1266. Turning to international differences, do you agree or disagree: a. Present economic differences between rich and poor countries are too large.

Percent	N	Value	Label
20.7	237	1	STRONGLY AGREE
44.6	511	2	AGREE
27.4	314	3	NEITHER AGREE NOR DISAGREE
6.0	69	4	DISAGREE
1.2	14	5	STRONGLY DISAGREE
	58,327	0	IAP
	78	8	CANT CHOOSE
	49	9	NA
100.0	59,599		Total

Properties

Data type:	numeric
Missing-data codes:	0,8,9
Mean:	2.22
Std Dev:	.88
Record/column:	1/5838

(a) (2 points) *What is the mean response?*

The mean response is 2.22.

(b) (2 points) *What was the most common response?*

The most common response is 2 ("Agree"), with about 45% (N=511) of all respondents.

(c) (4 points) *Is you answer in (b) a descriptive statistic or an inferential statistic? Explain.*

The mode (most common response) is a descriptive statistic because it summarizes the observed data, and by definition is a function of the data.

Question 2 (8 points)

In the 2016 presidential election, an exit poll sampled 1,941 of the 2,808,605 people who voted in Missouri. The poll stated that 57.23% of respondents reported voting for the Republican candidate, Donald Trump. Of all 2,808,605 voters, 56.8% voted for Trump.

(a) (2 points) *For this exit poll, what was the population?*

The population for the exit poll was all individuals who voted in the state of Missouri (N=2.8 million).

(b) (2 points) *For this exit poll, what was the sample?*

The sample for the exit poll was all individuals who participated in the exit poll (N=1,941).

(c) (2 points) *Identify a statistic.*

See definition on page 5 of AF: "A parameter is a numerical summary of the population. A statistic is a numerical summary of the sample data". The percentage of respondents that reported that they would vote for the Republican candidate is a statistic, it describes a numerical summary of the sample.

(d) (2 points) *Identify a parameter.*

The percentage of Missouri voters that voted for President Trump.

Question 3 (24 points)

Which scale of measurement (nominal, ordinal, or interval) is most appropriate for:

(a) (2 points) *Educational attainment (less than high school, some high school, high school, some college, college degree, graduate or professional degree)*

Ordinal

(b) (2 points) *Race (White, Black or African American, Asian, . . .)*

Nominal

(c) (2 points) *Letter grades*

Ordinal

(d) (2 points) *Statewide murder rate (number of murders per 1000 population)*

Interval

(e) (2 points) *Distance (in miles) commuted to work*

Interval

(f) (2 points) *Hair color (Blond, Brunette, Red, Black)*

Nominal

(g) (2 points) *Number of people you have known who volunteered for the Obama campaign*

Interval

(h) (2 points) *Partisan affiliation (Republican, Democrat, Green,....)*

Nominal

(i) (2 points) *Zip code*

Nominal

(j) (2 points) *Religious affiliation (Catholic, Protestant, Jewish, Muslim, Buddhist, other)*

Nominal

(k) (2 points) *Government spending on environment (up, same, down)*

Ordinal

(l) (2 points) *GPA (4.00, 3.00, 2.00, etc)*

Interval

Question 4 (14 points)

Suppose you are comparing the age of 20 men and 20 women. For the men, you get the following responses (in years): 56, 60, 50, 26, 45, 35, 41, 43, 34, 42, 37, 39, 33, 28, 52, 48, 27, 20, 44, 32. For the women, you get the following responses (in years): 47, 49, 20, 46, 43, 44, 45, 60, 57, 28, 55, 27, 25, 50, 52, 48, 23, 42, 33, 59.

(a) (12 points) *In R, make a box plot for both men and women. Make sure the mean and interquartile range are clearly marked for each box plot. (Make sure you attach a page with the code and figure to the back of this homework.)*

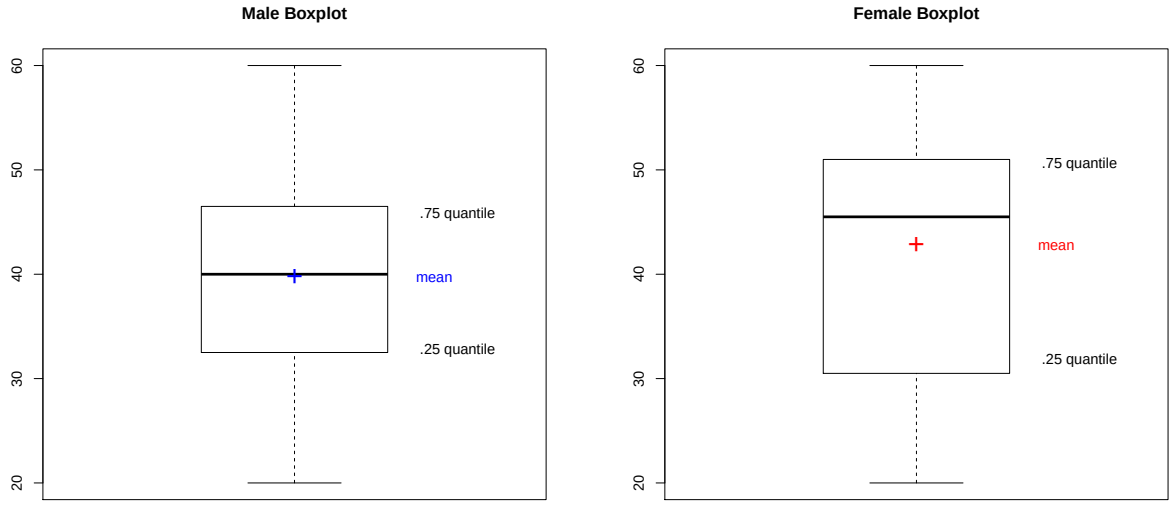
```
# set seed so that anyone else can replicate my graph
set.seed(23409)
# create a random sample of men and women
m <- sample(20:60, 20)
f <- sample(20:60, 20)

# open up .pdf to save male boxplot
pdf("maleBoxplot.pdf")
boxplot(m, main = "Male Boxplot")
points(mean(m), col = "blue", pch = "+", cex = 2)
text(labels = "mean", x = 1.3, y = mean(m), col = "blue")
text(labels = c(".25 quantile", ".75 quantile"),
     x = 1.35, y = quantile(m, c(.25, .75)))
#  25%   75%
# 32.75 45.75
dev.off()

# open up .pdf to save
pdf("femaleBoxplot.pdf")
boxplot(f, main = "Female Boxplot")
points(mean(f), col = "red", pch = "+", cex = 2)
text(labels = "mean", x = 1.3, y = mean(f), col = "red")
text(labels = c(".25 quantile", ".75 quantile"),
     x = 1.35, y = quantile(f, c(.25, .75)))
#  25%   75%
# 31.75 50.50
dev.off()
```

The left panel of Figure 1 displays the boxplot for men, and the right panel displays the boxplot for women. For each plot, in addition to the inter-quartile range, I've also clearly labeled the mean (which differs substantially in the female sample from the median).

Figure 1: Boxplots by gender with mean and interquartile range labeled.



(b) (2 points) *Is this likely to be a random sample of the American female population? Why or why not?*

No, it is not representative. No children or elderly.

Question 5 (24 points)

Identify each variable as discrete or continuous. If you think that more than one answer might be correct, justify your answer.

(a) (2 points) *Attitudes toward legalization of marijuana (favor, neutral, oppose)*

Discrete

(b) (2 points) *Number of political parties in a country*

Discrete

(c) (2 points) *Religious affiliation (Catholic, Jewish, Protestant, Muslim, ...)*

Discrete

(d) (2 points) *Choice of candidate a person will vote for*

Discrete

(e) (2 points) *Distance (in miles) commute to work*

Continuous

(f) (2 points) *Years of school completed (0, 1, 2, ...)*

Discrete

(g) (2 points) *Number of people you have known who volunteered for the Clinton campaign*

Discrete

(h) (2 points) *Partisan affiliation (Republican, Democrat, Green,....)*

Discrete

(i) (2 points) *Attitude towards the health care reform (favor, oppose, neutral)*

Discrete

(j) (2 points) *Political ideology (very liberal, somewhat liberal, moderate, somewhere moderate, somewhat conservative, conservative)*

Discrete

(k) (2 points) *Government spending on environment (up, same, down)*

Discrete

(l) (2 points) *GPA (4.00, 3.00, 2.00, etc)*

Continuous

Question 6 (EXTRA CREDIT! 10 points)

Assume that the probability of an American high school graduate participating in sports is 0.45. Further assume that of those who participate in sports, about 34% play an instrument. Again, assume that the probability of a high school graduate attending college is 0.69 (which is the case for 2015 high school graduates according to Bureau of Labor Statistics). Estimate the following probabilities (show your work):

(a) (2 point) *that a randomly chosen American high school graduate does not participate in sports.*

$$\begin{aligned} \text{P(not participate in sports)} &= 1 - \text{P(participate in sports)} \\ &= 1 - 0.45 \\ &= 0.55 \end{aligned}$$

(b) (4 points) *that a randomly chosen American high school graduate both participates in sports and plays an instrument.*

$$\begin{aligned} \mathrm{P}(\text{sports} \cap \text{instrument}) &= \mathrm{P}(\text{sports}) * \mathrm{P}(\text{instrument} \mid \text{sports}) \\ &= 0.45 * 0.34 \\ &= 0.153 \end{aligned}$$

(c) (4 points) *that a randomly chosen American high school graduate participate in sports and attends to college.*

$$\begin{aligned} \mathrm{P}(\text{sports} \cap \text{college}) &= \mathrm{P}(\text{sports}) * \mathrm{P}(\text{college}) \\ &= 0.45 * 0.69 \\ &= 0.3105 \end{aligned}$$

Question 7 (12 points)

Find an article that uses a statistic in the news and attach it to your homework. Please comment on the following: What kind of underlying data is used to calculate the statistic? Is it discrete or continuous? Is it nominal, ordinal, or interval? Is the statistic inferential or descriptive?

The answer is dependent on students' responses.