Problem Set 2

Quantitative Political Methodology (U25 363) Due: February 27, 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 100.

Question 1 (5 points)

You would like to find the proportion of bills passed by Congress that were vetoed by the President in the last congressional session. After checking congressional records, you see that for the population of all 40 bills passed, 2 were vetoed. Does it make sense to construct a confidence interval using these data to answer your question? Explain.

No, because confidence intervals are used to estimate a range of numbers in which an unknown parameter is believed to be found. In this case, the exact population and the necessary statistics to construct a proportion are known.

Question 2 (25 points)

The distribution of family size in a particular tribal society is skewed to the right, with μ = 5.2 and σ = 3. Those values are unknown to an anthropologist, who samples families to estimate mean family size. For a random sample of 36 families, she gets a mean of 4.6 and a standard deviation of 3.2.

1. Identify the population distribution. State its mean and standard deviation. Is the data skewed?

The population distribution consists of the distribution of mean family size for all the families in a particular tribal society. It is skewed to the right with a mean of 5.2 members and a standard deviation of 3 members.

1. Identify the sample data distribution. State its mean and standard deviation. Is the data skewed?

The sample data distribution consists of the distribution of the mean family size for the 36 families that the anthropologist samples. Its mean is 4.6 members and its standard deviation is 3.2 members.

1. Identify the sampling distribution of y ̄. State its mean and standard error and explain what it describes.

The sampling distribution of y bar is the theoretical distribution of the mean of many random samples within the entire tribal society. Additionally, this is a sampling with a large sample size n (above 30). Thus, the sampling

distribution of the sample mean y ̄ is approximately normal.

1. Find the probability that her sample mean falls within 0.5 of the population mean. – p 94 A + F. I worked with Andrew. Work attached.

68.26 %

1. Suppose she takes a random sample of size 100. Find the probability that the sample mean falls within 0.5 of the true mean, and compare the answer to that in (d).

90.4406%

1. Refer to (e). If the sample were truly random, would you be surprised if the anthropologist obtained y ̄ = 4. Why?

Yes, because this is so many standard deviations away from the population mean that for a normal distribution that would not make sense.

Question 3 (10 points)

The GSS asks respondents to rate their political views on a seven-point scale, where 1= extremely liberal, 4=moderate, and 7=extremely conservative. A researcher analyzing data from 2011 has the following data

Variable N Mean St. Dev SE Mean Polviews 1294 4.23 1.39 0.0387

1. Show how to construct a 95% confidence interval from the information provided.

CI = 4.23 + 1.96 (1.39/sq root of 1294) = [4.1546, 4.3057]

1.96 is the z score, which is obtained by calculating (1-.95)/2=.025 and using a z-table to find the 1.96 z score.

(b)  Interpret the confidence interval you found in (a).

The margin of error for the confidence interval is very narrow, suggesting that there is a high probability (19 times out of 20) that the sample mean would fall within the small range of [4.1546, 4.3057].

1. Would the confidence interval be wider or narrower (i) if you constructed a 90% confidence interval, (ii) if you found the 95% confidence interval only for those who called themselves strong Democrats on political party identification (PARTYID), for whom the mean was 3.50 with standard deviation 1.61?

The confidence interval would be narrower for a 90% confidence interval since there are fewer samples included in the measurement. The confidence interval for the CI of strong Democrats would depend on the standard deviation and sample size. If the sample size were similar, then since the standard deviation is larger, the confidence interval has to be larger to account for the larger spread of data.

Question 4 (5 points)

For a normal distribution with μ = 50 and σ2 = 36, find the probability that an observation falls (Hint: type help(Normal) in R):

(a) At or below the value 57.75 (b) At or above the value of 50.45

A:

above: pnorm(57.75, mean=50, sd=6) = .90176

B:

pnorm(50.45, mean=50, sd=6, lower.tail = FALSE) = 0.4701074

(c) Between the values of 52.4 and 59.4

x<-pnorm(52.4, mean=50, sd=6)

y<-pnorm(59.4, mean=50, sd=6)

y-x = 28.59819%

Question 5 (5 points)

R has a number of functions that make it simple to simulate from a variety of distributions.

One thing to note is that when sampling you want to set a seed in R. Setting the seed allows you to replicate your results. It doesn’t matter what it is set to. So, for the purposes of this question, type:

set.seed(12345)

Suppose that salaries follow a normal distribution with mean 40000 and standard deviation 15000. We can sample from this distribution using the rnorm() command. Type the followng into R to generate a sample with 10000 observations:

salaries <- rnorm(n=10000,mean=40000,sd=15000)

Plot the distribution. Add a title to this plot. Save this plot as a .pdf file.

plot(density(salaries),

main="Distribution of Salaries",

xlab="salaries", ylab="Probability Density")pdf(file = "PS2PDF")

dev.off

**Question 6 (10 points)**

Plot probability density functions for the following normal distributions. Make all the plots on a single page. Make sure your plots have properly labeled titles and axes, and your axes are comparable across plots.

(a) Normal Distribution with μ = 0 and σ2 = 0.4

(b) Normal Distribution with μ = 0 and σ2 = 3

(c) Normal Distribution with μ = 3 and σ2 = 3

(d) Normal Distribution with μ = 3 and σ2 = 0.4

(e) Normal Distribution with μ = −2 and σ2 = 5

(f) Normal Distribution with μ = −2 and σ2 = .25

PlotA<-density(rnorm(100000, mean=0, sd=.04^.5))

PlotB<-density(rnorm(100000, mean=0, sd=3^.5))

PlotC<-density(rnorm(100000, mean=3, sd=3^.5))

PlotD<-density(rnorm(100000, mean=3, sd=.4^.5))

PlotE<-density(rnorm(100000,mean=-2, sd=5^.5))

PlotF<-density(rnorm(100000, mean=-2, sd=.25^.5))

plot(PlotA, main="Question 7",col="red", xlab="Mean", ylab="Probability Density", ylim = c(0,2), xlim= c(-5,10))

lines(PlotB, main="B", col="purple")

lines(PlotC, main="C", col="magenta")

lines(PlotD, main="D", col="black")

lines(PlotE, main="E", col="orange")

lines(PlotF, main="F", col="blue")

Question 7 (20 points)

Peake and Eshbaugh-Soha (2008) study drug policy coverage. Their data count the number of nightly television news stories in a month focusing on drugs, from January 1977 to De- cember 1992. The dataset is in comma-separated format in the file named drugCoverage.csv. Download it from Monogan (2015)’s Dataverse. The variables in the dataset are: a character- based time index showing month and year (Year), news coverage of drugs (drugsmedia), an indicator for a speech on drugs that Ronald Reagan gave in September 1986 (rwr86), an indicator for a speech George H.W. Bush gave in September 1989 (ghwb89), the president’s approval rating (approval), and the unemployment rate (unemploy).

I downloaded the file from online and put it in my PS2 folder.

Code:

setwd("~/Documents/GitHub/QPMspring2019/problemSets/PS2")

read.csv("drugCoverage.csv")

1. (a)  Draw a histogram of the monthly count of drug-related stories.

hist(Qsev$drugsmedia, breaks= 196, main="Drug Media per Month")

pdf("PS2Q7A")

1. (b)  Draw two boxplots: One of drug-related stories and another of presidential approval. How do these figures differ and what does that tell you about the contrast between the variables?

Code:

boxplot(Qsev$drugsmedia)

boxplot(Qsev$approval)

Analysis:

The drug-related stories boxplot has many outliers and the quartiles are very uneven. The presidential approval boxplot has no outliers and very even quartiles. This tells us that the data for drug-related stories is much more skewed and spread out, but the presidential approval has more of a normal distribution.

1. (c)  Draw two scatterplots:
   * –  In the first, represent the number of drug-related stories on the vertical axis, and place the unemployment rate on the horizontal axis.

plot(Qsev$unemploy, Qsev$drugsmedia,

main = "Q 7C plot 1",

xlab = "Unemployment",

ylab = "drug media stories",

pch = 19)

* + –  In the second, represent the number of drug-related stories on the vertical axis, and place presidential approval on the horizontal axis.

plot( Qsev$drugsmedia, Qsev$unemploy,

main = "Q 7C plot 2",

xlab = "drug media stories",

ylab = "Unemploymen",

pch = 19)

* + –  How do the graphs differ? What do they tell you about the data?

In the first plot, most of the data is clustered in the lower half of the graph, and in the second plot, most of the data is clustered in the left side of the graph. In both of these cases, this means that the data is clustered closer to the Unemployment access, suggesting that drug media stories are low compared to unemployment.

1. (d)  Draw two line graphs:

– In the first, draw the number of drug-related stories by month over time. – In the second, draw presidential approval by month over time.

plot(Qsev$drugsmedia, Qsev$Year, ylab="drug media stories",xlab="Month-Year",type="l")

plot(Qsev$approval, Qsev$Year, ylab="drug media stories",xlab="Month-Year",type="l")

– What can you learn from these graphs?

But the one thing I can understand from them is that drugs media stories are highly concentrated in the first quarter of time, meaning that most of the stories occurred earlier in time, versus the approval ratings are much more evenly distributed over time.

**Question 8 (20 points)**

For this question, you will work with W-NOMINATE data to trace the policy positions of members in the U.S. House of Representatives. With the data, you will learn about po- larization (i.e. distance between the ideological positions of the Democratic Party and the Republican Party). You will also learn about the ideological cohesiveness of each party. Answer the following questions:

1. (a)  Import data on the 88th and 107th Congresses. Then, create four subsets of the data by session and party (Democratic Party in the 88th session, Democratic Party in the 107th session, Republican Party in the 88th session, and Republican Party in the 107th session).

D88 <- condataset$wnominate[condataset$congress == 88 & condataset$party== 100]

R88<- condataset$wnominate[condataset$congress == 88 & condataset$party == 200]

D107<-condataset$wnominate[condataset$congress == 107 & condataset$party == 100]

R107<-condataset$wnominate[condataset$congress == 107 & condataset$party == 200]

(b)  For the Democratic Party, calculate the median W-NOMINATE scores for two Congresses. How did the median change over time? **What does this mean?**

median(R88)

median(R107)

The median grew over time, from -0.2364 in the 88th to - 0.4719in the 107th. This means that the w-nominate score, or an individual’s party identification, distribution has shifted to the left. This means that more democrats are more strongly aligned with their party.

1. (c)  For the Republican Party, calculate the median W-NOMINATE scores for the two Congresses. How did the median change over time? What does this mean?

median(D88)

median(D107)

The median grew over time, from 0.4583 in the 88th to 0.7179 in 107th. This means that the w-nominate score, or an individual’s party identification, distribution has shifted to the right. This means that more republicans are more strongly aligned to their party.

1. (d)  For the Democratic Party, calculate the standard deviation of W-NOMINATE scores for the two Congresses. How did the standard deviation change over time? What does this mean?

The standard deviation shrank from 0.2894997 to 0.2410088, suggesting that there is less variation in the alignment between democrats now.

sd(D88)

sd(D107)

1. (e)  For the Republican Party, calculate the standard deviation of W-NOMINATE scores for the two Congresses. How did the standard deviation change over time? What does this mean?

The standard deviation shrank from 0.1703398 to 0.1320712, suggesting that there is less variation in the alignment between republicans now.

sd(R88)

sd(R107)

1. (f)  For the 88th Congress, create a plot that overlays two histograms. One histogram should plot the distribution of W-NOMINATE scores for the Democratic Party. The other histogram should plot the distribution of W-NOMINATE scores for the Republi- can Party. (Hint: to overlay two histograms, you can run two separate hist commands but include an add argument in the second hist one.)

hist(R88, xlim= c(-1,1), col="blue")

hist(D88,xlab= "Wnom", add=TRUE,col = "red",

main = "Q7 Part G")

pdf("PS2Q7F1")

1. (g)  For the 107th Congress, create a plot that overlays two histograms. One histogram should plot the distribution of W-NOMINATE scores for the Democratic Party. The other histogram should plot the distribution of W-NOMINATE scores for the Repub- lican Party.

hist(R107, xlim = c(-1,1), col = "blue")

hist(D107,xlab= "Wnom", add=TRUE,col = "red",

main = "Q7 Part G")

pdf("PS2Q7G1")

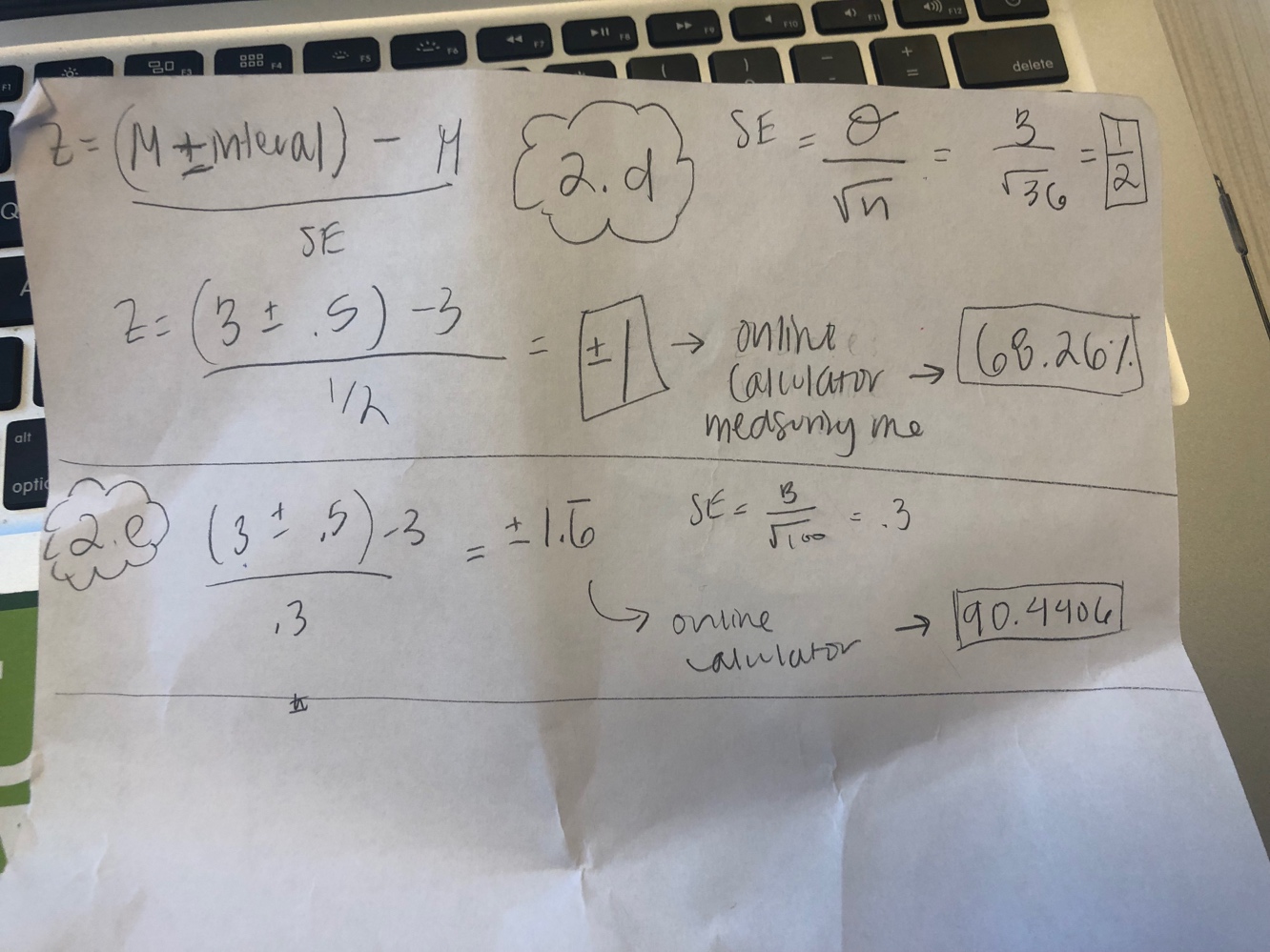
1. (h)  Based on what you have done so far, compare the 88th Congress and the 107th Congress.

– Did polarization decrease, increase, or stay the same? Are both parties responsi- ble for this or is one party responsible?

Polarization increased. This is especially visible in the plots that show the distribution of W-NOMINATE and how the data is far more polarized (clustered towards either end of the spectrum). Both parties are responsible for polarization because they both moved further outwards.

– For each party, what happened to the ideological cohesiveness of its members? Did it decrease, increase, or stay the same?

The ideological cohesiveness of members increased for both sides. In both sides, the data became clustered in much smaller areas of the graph than before.



Q2 D and E