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.

**It would not make sense to do so, since the data would not be predicting anything since the information is already available.**

# 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. set.seed(123)
2. rnorm(x, mean = 5.2, sd = 3)
4. rnorm(36, mean = 4.6, sd = 3.2)
5. vec1 <- rnorm(100000, mean=0, sd=1)
6. vec1
8. plot(density(vec1),
9. main="Distribution of vec1",
10. xlab="")
12. vec2 <- rnorm(x, mean=5.2, sd=3)
13. vec3 <- rnorm(36, mean=4.6, sd=3.2)
14. plot(density(vec2),
15. main="Tribal Society",
16. xlab="Distribution",
17. col="yellow",
18. xlim=c(0,100))
19. lines(density(vec3), lty=2, col="red")
20. Identify the population distribution. State its mean and standard deviation. Is the data skewed?
21. set.seed(123)
22. rnorm(x, mean = 5.2, sd = 3)
24. rnorm(36, mean = 4.6, sd = 3.2)
25. Identify the sample data distribution. State its mean and standard deviation. Is the data skewed?
26. set.seed(123)
27. rnorm(x, mean = 5.2, sd = 3)
29. rnorm(36, mean = 4.6, sd = 3.2)
30. Identify the sampling distribution of ¯*y*. State its mean and standard error and explain what it describes.
31. Find the probability that her sample mean falls within 0.5 of the population mean.
32. z95 <- qnorm((1 - .95)/2, lower.tail = FALSE)
33. n <- length(na.omit(x))
34. sample\_mean <- mean(x, na.rm = TRUE)
35. sample\_sd <- sd(Polviews, na.rm = TRUE)
36. lower\_95 <- sample\_mean - (z95 \* (sample\_sd/sqrt(n)))
37. upper\_95 <- sample\_mean + (z95 \* (sample\_sd/sqrt(n)))
38. confint95 <- c(lower\_95, upper\_95)
39. 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).
40. Refer to (e). If the sample were truly random, would you be surprised if the anthropologist obtained ¯*y* = 4. Why?

**No – in a random sample, any number can be obtained.**

# 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.
2. rnorm(1294, mean = 4.23, sd = 1.39)
4. z95 <- qnorm((1 - .95)/2, lower.tail = FALSE)
5. n <- length(na.omit(Polviews))
6. sample\_mean <- mean(Polviews, na.rm = TRUE)
7. sample\_sd <- sd(Polviews, na.rm = TRUE)
8. lower\_95 <- sample\_mean - (z95 \* (sample\_sd/sqrt(n)))
9. upper\_95 <- sample\_mean + (z95 \* (sample\_sd/sqrt(n)))
10. confint95 <- c(lower\_95, upper\_95)
11. Interpret the confidence interval you found in (a).
12. 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?

**Wider.**

# Question 4 (5 points)

For a normal distribution with *µ* = 50 and *σ*2 = 36, find the probability that an observation falls (Hint: type help(Normal) in R):

1. At or below the value 57.75
2. pnorm(57.25, mean=50, sd=36)
3. At or above the value of 50.45
4. pnorm(50.45, mean=50, sd=36)
5. Between the values of 52.4 and 59.4.
6. pnorm(52.4 & 59.4, mean = 50, sd = 36)

**For question four, I used the “pnorm” function to find the probability for each of the numbers, given the mean and the standard deviation.**

# 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 following 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.

1. set.seed(12345)
3. salaries <- rnorm(n=10000, mean = 40000, sd = 15000)
5. pdf("salaries.pdf", width=7, height=7)
6. plot(salaries,
7. ylab="Salaries", xlab="Plot")
8. 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.

1. Normal Distribution with *µ* = 0 and *σ*2 = 0*.*4
2. xx <- seq ( min ( x ) , max( x ) , length =100)
4. plot(xx, dnorm(xx, mean = 0, sd = 0.4) , type = "l" , xlab = " " , ylab =
5. "Probability density" , cex.lab = 1.5)
6. Normal Distribution with *µ* = 0 and *σ*2 = 3
7. xx <- seq ( min ( x ) , max( x ) , length =100)
9. plot(xx, dnorm(xx, mean = 0, sd = 3) , type = "l" , xlab = " " , ylab =
10. "Probability density" , cex.lab = 1.5)
11. Normal Distribution with *µ* = 3 and *σ*2 = 3
12. xx <- seq ( min ( x ) , max( x ) , length =100)
14. plot(xx, dnorm(xx, mean = 3, sd = 3) , type = "l" , xlab = " " , ylab =
15. "Probability density" , cex.lab = 1.5)
16. Normal Distribution with *µ* = 3 and *σ*2 = 0*.*4
17. xx <- seq ( min ( x ) , max( x ) , length =100)
19. plot(xx, dnorm(xx, mean = 3, sd = 0.4) , type = "l" , xlab = " " , ylab =
20. "Probability density" , cex.lab = 1.5)
21. Normal Distribution with *µ* = −2 and *σ*2 = 5
22. xx <- seq ( min ( x ) , max( x ) , length =100)
24. plot(xx, dnorm(xx, mean = -2, sd = 5) , type = "l" , xlab = " " , ylab =
25. "Probability density" , cex.lab = 1.5)
26. Normal Distribution with *µ* = −2 and
27. xx <- seq ( min ( x ) , max( x ) , length =100)
29. plot(xx, dnorm(xx, mean = -2, sd = 0.25) , type = "l" , xlab = " " , ylab =
30. "Probability density" , cex.lab = 1.5)

**For Question 5, I used the plot function to create six different probability densities in R, with each varying based on the means and standard deviations.**

# 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 December 1992. The dataset is in comma-separated format in the file named drugCoverage.csv. Download it from [Monogan (2015)’s Dataverse.](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/ARKOTI) The variables in the dataset are: a characterbased 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).

1. drugCoverage <- read.csv("drugCoverage.csv")
2. View(drugCoverage)
4. DrugsInMedia <- drugCoverage[drugCoverage$drugsmedia==100 & drugCoverage$may-77==88,]
5. Draw a histogram of the monthly count of drug-related stories.
6. hist(drugCoverage$approval,
7. breaks = 20,
8. main = "Drug-Related Stories by Month",
9. xlab = "Stories by Month",
10. col = "Green",
11. ylim = c(0,20))
12. Draw two boxplots: One of drug-related stories and another of presidential approval. How do these figures differ and what does that tell you about the contrast between the variables?
13. boxplot(drugCoverage)
15. quantile(drugCoverage$approval, probs=c(0, 0.25, 0.5, 0.75, 1))
17. boxplot(drugCoverage$approval, main="Boxplot", ylab="Presidential Approval", ylim=c(0,100))
19. boxplot(drugCoverage$drugsmedia, main="Boxplot", ylab="Media Coverage", ylim=c(0,100))
20. 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.
21. scatter.smooth(drugCoverage$drugsmedia, drugCoverage$unemploy, main="Scatter Plot", xlab="Unemployment", ylab="Drug-Related Stories")
    * In the second, represent the number of drug-related stories on the vertical axis, and place presidential approval on the horizontal axis.
22. scatter.smooth(drugCoverage$drugsmedia, drugCoverage$approval, main="Scatter Plot", xlab="Presidential Approval", ylab="Drug-Related Stories")
    * How do the graphs differ? What do they tell you about the data?

**The two graphs differ in that, even though the majority of the sampling distribution is on the left in both, most in the first are more toward the middle than the second.**

1. Draw two line graphs:
   * In the first, draw the number of drug-related stories by month over time.
2. line(drugCoverage$drugsmedia, main="Scatter Plot", xlab="Media Coverage", ylab="Drug-Related Stories")
   * In the second, draw presidential approval by month over time.
3. line(drugCoverage$approval, main="Scatter Plot", xlab="Presidential Approval", ylab="Drug-Related Stories")
   * What can you learn from these graphs?

**That Presidential approval and media coverage of drugs roughly coincide.**

# 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.](https://drive.google.com/open?id=1kVn-Asz8KXPQhiLSQGEabdVJh9_8_wjf) With the data, you will learn about polarization (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. 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)
2. wnominatehouse <- read.csv("wnominatehouse.csv")
3. View(wnominatehouse)
4. House88thDems <- wnominatehouse[wnominatehouse$party==100 & wnominatehouse$congress==88,]
5. View(House88thDems)
7. House107thDems <- wnominatehouse[wnominatehouse$party==100 & wnominatehouse$congress==107,]
8. View(House107thDems)
10. House88thReps <- wnominatehouse[wnominatehouse$party==200 & wnominatehouse$congress==88,]
11. View(House88thReps)
13. House107thReps <- wnominatehouse[wnominatehouse$party==200 & wnominatehouse$congress==107,]
14. View(House107thReps)
15. For the Democratic Party, calculate the median W-NOMINATE scores for two Congresses. How did the median change over time? What does this mean?
16. mean(House88thDems==100)
17. median(House88thDems==100)
18. sd(House88thDems==100)
20. mean(House107thDems==200)
21. median(House107thDems==200)
22. sd(House107thDems==200)
23. 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?
24. mean(House88thReps==100)
25. median(House88thReps==100)
26. sd(House88thReps==100)
28. mean(House107thReps==200)
29. median(House107thReps==200)
30. sd(House107thReps==200)
31. 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?
32. mean(House88thDems==100)
33. median(House88thDems==100)
34. sd(House88thDems==100)
36. mean(House107thDems==200)
37. median(House107thDems==200)
38. sd(House107thDems==200)
39. 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?
40. mean(House88thReps==100)
41. median(House88thReps==100)
42. sd(House88thReps==100)
44. mean(House107thReps==200)
45. median(House107thReps==200)
46. sd(House107thReps==200)
47. 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 Republican Party. (Hint: to overlay two histograms, you can run two separate hist commands but include an add argument in the second hist one.)
48. hist(wnominatehouse$congress,
49. breaks = 20,
50. main = "107th Congress",
51. xlab = "Democrats",
52. ylab = "Republicans",
53. col = "Blue",
54. xlim = c(0,2),
55. ylim = c(0, 2),
56. prob = TRUE)
57. points(seq(min(wnominatehouse), max(wnominatehouse), length.out=500),
58. dnorm(seq(minwnominatehouse), max(wnominatehouse), length.out=500),
59. mean(wnominatehouse), sd(wnominatehouse), type="l", col="blue")
60. For the 107th Congress, create a plot that overlays two histograms. One histogram should plot the distribution of W-NOMINATE scores for the Democratic Party. The other histogram should plot the distribution of W-NOMINATE scores for the Republican Party.
61. hist(wnominatehouse$congress,
62. breaks = 20,
63. main = "107th Congress",
64. xlab = "Democrats",
65. ylab = "Republicans",
66. col = "Blue",
67. xlim = c(0,2),
68. ylim = c(0, 2),
69. prob = TRUE)
70. points(seq(min(wnominatehouse), max(wnominatehouse), length.out=500),
71. dnorm(seq(minwnominatehouse), max(wnominatehouse), length.out=500),
72. mean(wnominatehouse), sd(wnominatehouse), type="l", col="blue")
73. 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 responsible for this or is one party responsible?

**I would say that polarization increased from the 88th Congress to the 107th.**

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

**Increased.**