Problem Set 2: Answer Key

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.

Identify the sample and population! In this case it does not make sense to construct a confidence interval since we have the population parameter. The sample and the population are the same and $\pi = 0.05$.

Question 2 (25 points)

The distribution of family size in a particular tribal society is skewed to the right, with $\mu = 5.2$ *and* $\sigma = 3$*. 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.*

(a) *Identify the population distribution. State its mean and standard deviation. Is the data skewed?*

The population distribution is skewed to the right with mean 5.2 and standard deviation 3.

(b) *Identify the sample data distribution. State its mean and standard deviation. Is the data skewed?*

The sample data distribution based on the sample of 36 families and is skewed to the right with mean 4.6 and standard deviation 3.2.

(c) *Identify the sampling distribution of $\bar{y}$. State its mean and standard error and explain what it describes.*

The sampling distribution of $\bar{y}$ is approximately normal with mean 5.2 and standard error $3/\sqrt{36} = 0.5$. This distribution describes the theoretical distribution for the sample mean.

(d) *Find the probability that her sample mean falls within 0.5 of the population mean.*

$$P(\mu - 0.5) < \bar{y} < \mu + 0.5) = P(\frac{-0.5}{3/\sqrt{36}}) < Z < \frac{0.5}{3/\sqrt{36}})$$
$$= P(-1 < Z < 1)$$

```
pnorm((-0.5/(3/sqrt(36))), lower.tail = F)
pnorm((0.5/(3/sqrt(36))), lower.tail = F)
```

This probability is 0.8413 - 0.1587 = 0.6826.

(e) *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).*

$$P(\mu - 0.5) < \bar{y} < \mu + 0.5) = P(\frac{-0.5}{3/\sqrt{100}}) < Z < \frac{0.5}{3/\sqrt{100}})$$
$$= P(-1 < Z < 1)$$

```
pnorm((-0.5/(3/sqrt(100))), lower.tail = F)
pnorm((5/(3/sqrt(100))), lower.tail = F)
```

This probability is 0.9525-0.0475=0.9050. The probability is larger than in part (d) because the standard error is smaller (since the sample size is larger).

(f) *Refer to (e). If the sample were truly random, would you be surprised if the anthropologist obtained $\bar{y} = 4$. Why?*

If the sample were truly random, then the probability that $\bar{y}$ would be 4 or less is

$$\begin{aligned} P(\bar{y} < 4) &= P(Z < \frac{4 - 5.2}{3/\sqrt{100}}) \\ &= P(Z < 4) \\ &= 0.0000317 \end{aligned}$$

```
pnorm(((4-5.2)/(3/sqrt(100))), lower.tail = T)
```

This probability is quite small. Thus, this would be a surprising result.

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

(a) *Show how to construct a 95% confidence interval from the information provided.*

$$4.23 \pm 1.96 \frac{1.39}{\sqrt{1294}} = [4.15, 4.31]$$

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

The confidence interval indicates that you are 95% confident that the population parameter μ falls in the interval from 4.15 to 4.31 because, taking repeated samples, you expect intervals created in this way to contain the parameter μ roughly 95% of the time.

(c) *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?*

(i) The 90% confidence interval would be narrower, since we do not need to include as many possible values in the confidence interval when are less confident.

(ii) The 95% confidence interval for only the string Democrats would be wider, since the standard deviation is larger and the sample size is smaller (thus causing the standard error to be larger).

Question 4 (5 points)

For a normal distribution with $\mu = 50$ *and* $\sigma^2 = 36$*, find the probability that an observation falls (Hint: type* `help(Normal)` *in* `R`*):*

(a) *At or below the value 57.75*

```
pnorm(57.75, mean=50, sd=6, lower.tail=TRUE)
```

0.9017637

(b) *At or above the value of 50.45*

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

0.4701074

(c) *Between the values of 52.4 and 59.4*

```
## create an object "a":
a <- pnorm(52.4, mean=50, sd=6, lower.tail=TRUE)

# show the stored probability in object "a"
a

# create a second object "b":
b <- pnorm(59.4, mean=50, sd=6, lower.tail=TRUE)

# show the stored probability in object "b"
b
```

```
12
13 # substract "a" from "b":
14 b-a
```

0.9414037 - 0.6554217 = 0.2859819

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 for reproducibility
2 set.seed(12345)
3 # create a sample of salaries
4 salaries <- rnorm(n=10000,mean=40000,sd=15000)
5 # plot distribution by opening up .pdf
6 pdf (file="densitySalaries.pdf" # file name
7       , width = 6  # plot width (in inches)
8       , height = 4 # plot height
9 )
10 plot(density(salaries), main="")
11 dev.off()
```

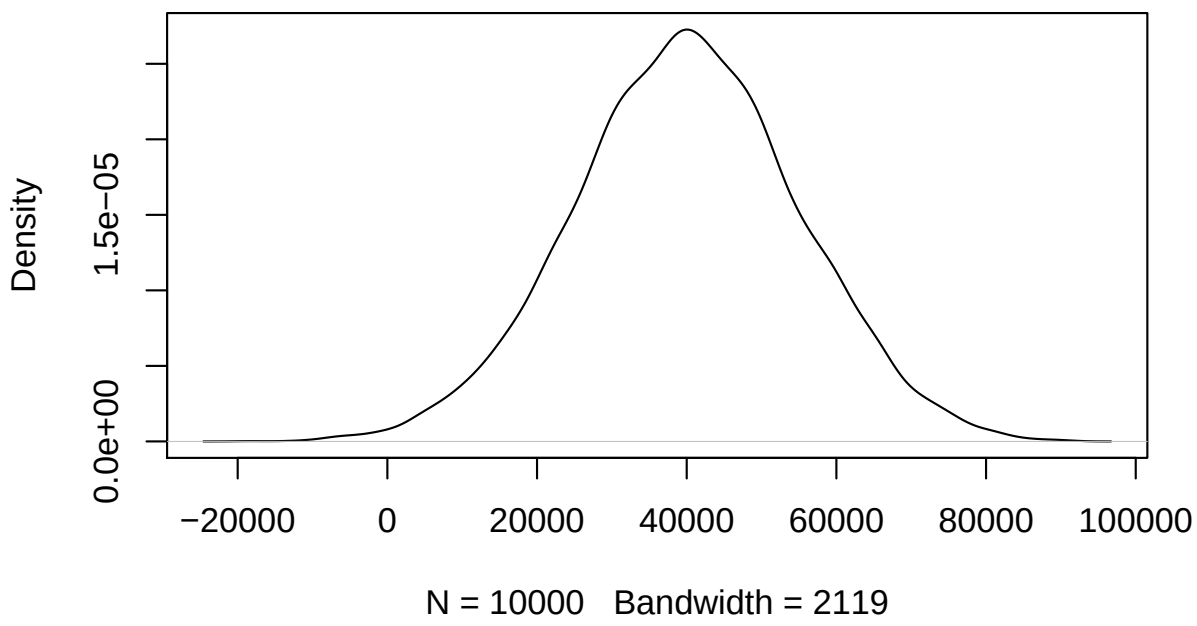
The output from the R code can be seen in Figure 1.

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* $\mu = 0$ *and* $\sigma^2 = 0.4$

Figure 1: Density distribution for simulated, random sample of salaries



(b) *Normal Distribution with* $\mu = 0$ *and* $\sigma^2 = 3$

(c) *Normal Distribution with* $\mu = 3$ *and* $\sigma^2 = 3$

(d) *Normal Distribution with* $\mu = 3$ *and* $\sigma^2 = 0.4$

(e) *Normal Distribution with* $\mu = -2$ *and* $\sigma^2 = 5$

(f) *Normal Distribution with* $\mu = -2$ *and* $\sigma^2 = \frac{1}{4}$

```r
# create 3 rows and 2 columns, so that the plots
# fit on one page
par(mfrow=c(3,2))
# set range of values for x-axis
x <- seq(-10, 10, length=100)

# Distribution A:
plot(x, dnorm(x, mean=0, sd=sqrt(0.4)), xlab="x value",
     type="l", ylim=c(0,0.8), ylab="Density",
     main=expression(paste(mu, "=0, ", sigma^2, "=0.4")))
# Distribution B:
plot(x, dnorm(x, mean=0, sd=sqrt(3)), xlab="x value",
```

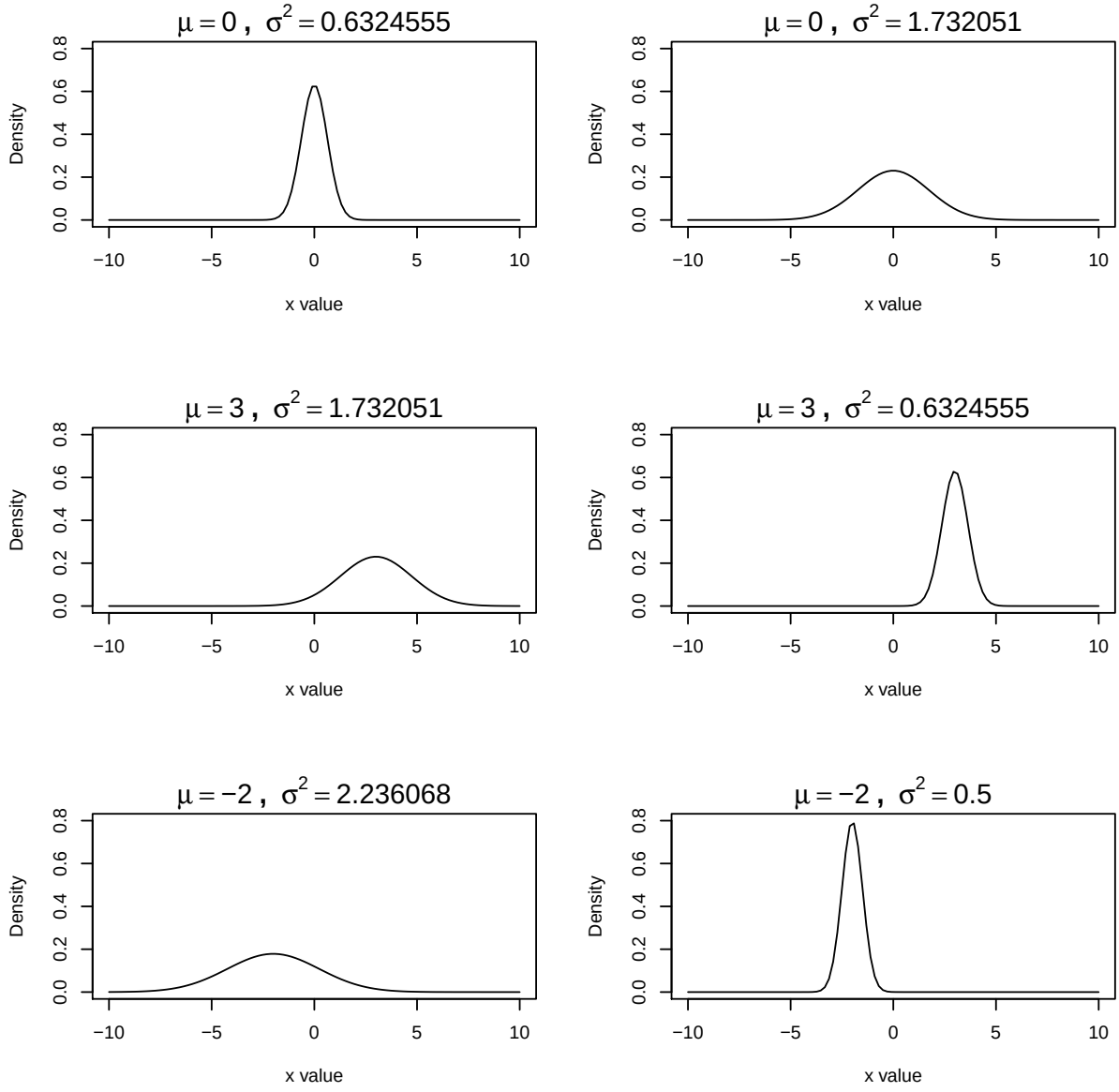
```
      type="l", ylim=c(0,0.8), ylab="Density",
      main=expression(paste(mu, "=0, ", sigma^2, "=3")))
# Distribution C:
plot(x, dnorm(x, mean=3, sd=sqrt(3)), xlab="x value",
      type="l", ylim=c(0,0.8), ylab="Density",
      main=expression(paste(mu, "=3, ", sigma^2, "=3")))
# Distribution D:
plot(x, dnorm(x, mean=3, sd=sqrt(0.4)), xlab="x value",
      type="l", ylim=c(0,0.8), ylab="Density",
      main=expression(paste(mu, "=3, ", sigma^2, "=0.4")))
# Distribution E:
plot(x, dnorm(x, mean=-2, sd=sqrt(5)), xlab="x value",
      type="l", ylim=c(0,0.8), ylab="Density",
      main=expression(paste(mu, "=-2, ", sigma^2, "=5")))
# Distribution F:
plot(x, dnorm(x, mean=-2, sd=sqrt(1/4)), xlab="x value",
      type="l", ylim=c(0,0.8), ylab="Density",
      main=expression(paste(mu, "=-2, ", sigma^2,"=",frac(1,4))))

# Easier, but more advanced:
# create function since you're doing it 6 times
pdfPlotFunction <- function(avg=0, standDiv=sqrt(0.4)){
  x <- seq(-10, 10, length=100)
  mainText <- bquote(bold(mu == .(avg) ~ ", " ~ sigma^2 == .(standDiv)))
  plot(x, dnorm(x, mean=avg, sd=standDiv), xlab="x value",
       type="l", ylim=c(0,0.8), ylab="Density",
       main="")
  mtext(mainText, side=3)
}

# open up .pdf and save
pdf("Q6.pdf")
par(mfrow=c(3,2))
pdfPlotFunction(avg=0, standDiv=sqrt(0.4))
pdfPlotFunction(avg=0, standDiv=sqrt(3))
pdfPlotFunction(avg=3, standDiv=sqrt(3))
pdfPlotFunction(avg=3, standDiv=sqrt(0.4))
pdfPlotFunction(avg=-2, standDiv=sqrt(5))
pdfPlotFunction(avg=-2, standDiv=sqrt(1/4))
dev.off()
```

The output from the R code can be seen in Figure 2.

Figure 2: PDFs for normal distribution varying mean and standard deviation



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. 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`).

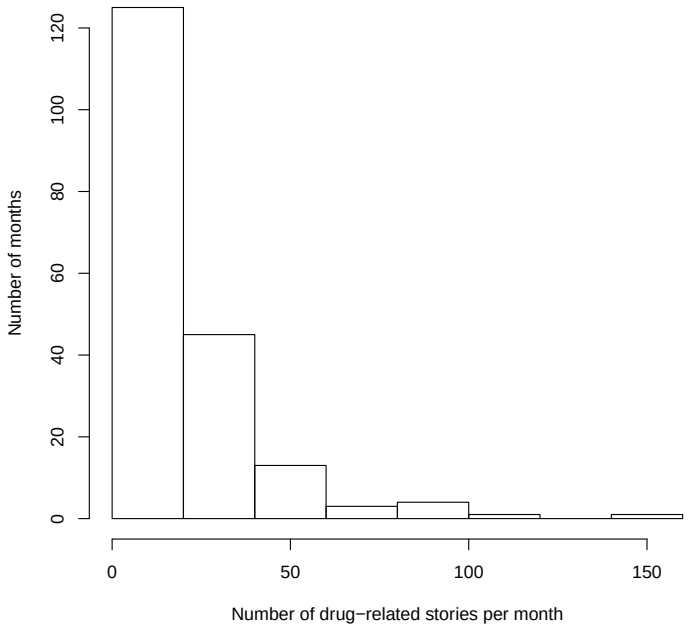
(a) *Draw a histogram of the monthly count of drug-related stories.*

```
library(date)
drug <- read.csv("drugCoverage.csv")

pdf("drugsMediaHist.pdf")
hist(drug$drugsmedia, main = "", xlab = "Number of drug-related stories
    per month", ylab="Number of months")
dev.off()
```

The output from the R code can be seen in Figure 3.

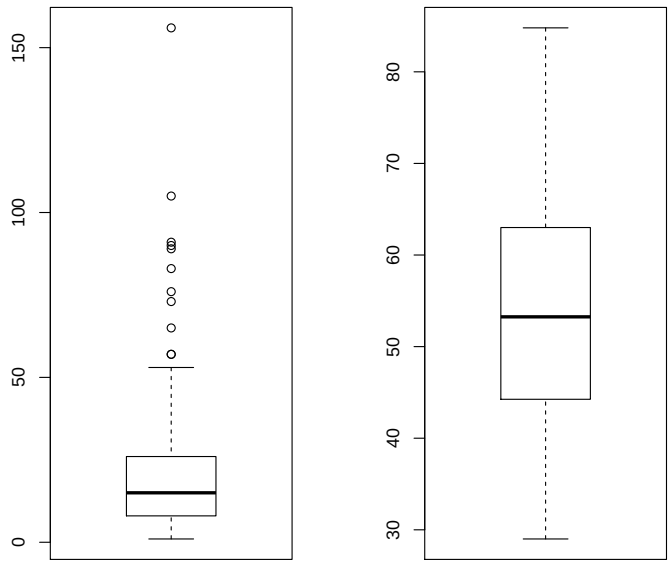
Figure 3: Histogram: Monthly count of drug-related stories.



(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?*

```
pdf("drugsBoxplot.pdf")
par(mfrow=c(1,2))
boxplot(drug$drugsmedia)
boxplot(drug$approval)
dev.off()
```

Figure 4: Boxplot: Drug-related stories and presidential approval.



The output from the R code can be seen in Figure 4. News coverage of drugs has a right-skewed distribution, whereas presidential approval rate has basically no skew. News coverage of drugs has a number of outliers. Presidential approval rate does not have any.

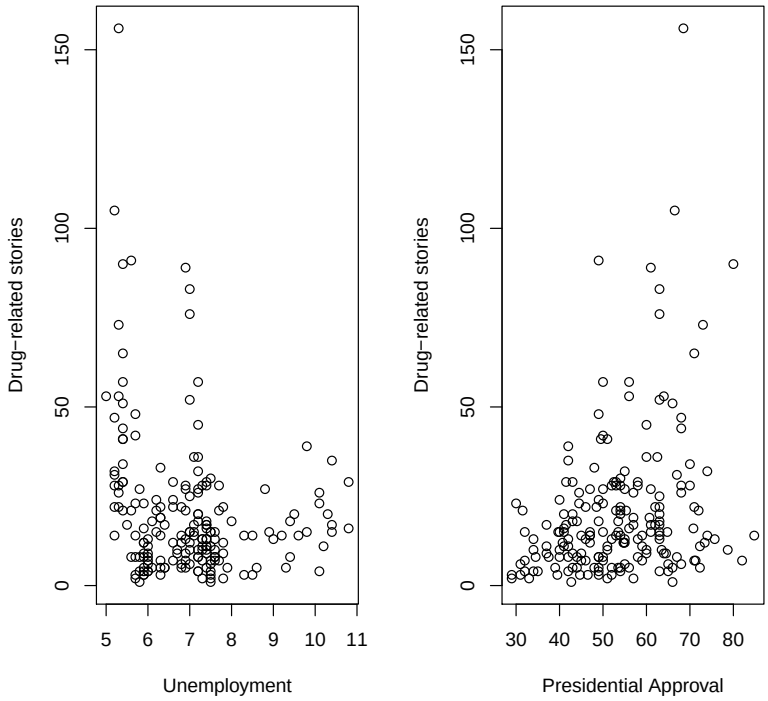
(c) *Draw two scatterplots:*

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

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

```
pdf("drugsScatterplot.pdf")
par(mfrow=c(1,2))
plot(drugsmedia~unemploy,data=drug, ylab = "Drug-related stories",
    xlab = "Unemployment")
plot(drugsmedia~approval,data=drug, ylab = "Drug-related stories",
    xlab = "Presidential Approval")
dev.off()
```

Figure 5: Scatter plot: Drug-related stories compared to unemployment and presidential approval.



The output from the R code can be seen in Figure 5. In the first plot, drug news coverage is negatively associated with unemployment rate. Higher levels of unemployment is associated with low news coverage of drugs. Maybe high unemployment leads to more news coverage of jobs and less on drugs.

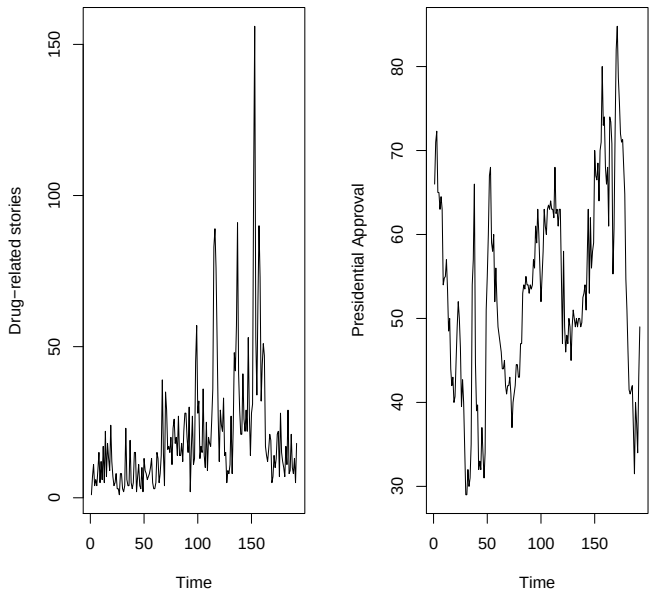
In the second plot, drug news coverage is positively associated with presidential approval rate. Higher levels of presidential approval is associated with high news coverage of drugs. Maybe if president has high approval, news do not have negative things to talk about the president, thereby covering drugs more.

(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.*
- *What can you learn from these graphs?*

```
pdf("drugsLineplot.pdf")
par(mfrow=c(1,2))
plot(drug$drugsmedia, type="l", ylab = "Drug-related stories", xlab =
    "Time")
plot(drug$approval, type="l", ylab = "Presidential Approval", xlab =
   "Time")
dev.off()
```

Figure 6: Line plot: Drug-related stories and presidential approval over time.



The output from the R code can be seen in Figure 6. Coverage of drug-related stories has been relatively stable, though it once peaked dramatically. Presidential approval rate fluctuates a lot more on a regular basis.

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 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:

(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).*

```
wnominate <- read.csv("wnominatehouse.csv")
dem2 <- wnominate[wnominate$congress==88 & wnominate$party==100,]
dem3 <- wnominate[wnominate$congress==107 & wnominate$party==100,]
rep2 <- wnominate[wnominate$congress==88 & wnominate$party==200,]
rep3 <- wnominate[wnominate$congress==107 & wnominate$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(dem2$wnominate)
median(dem3$wnominate)
```

The median decreased. The Democratic Party went farther to the left on the ideological spectrum.

(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(rep2$wnominate)
median(rep3$wnominate)
```

The median increased. The Republican Party went farther to the right on the ideological spectrum.

(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?*

```
sd(dem2$wnominate)
sd(dem3$wnominate)
```

The standard deviation decreased. The Democratic Party became more ideologically cohesive.

(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?*

```
sd(rep2$wnominate)
sd(rep3$wnominate)
```

The standard deviation decreased. The Republican Party became more ideologically cohesive.

(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 Republican Party. (Hint: to overlay two histograms, you can run two separate* **`hist`** *commands but include an* **`add`** *argument in the second* **`hist`** *one.)*

```
pdf("wnominateHist88.pdf")
hist(rep2$wnominate, xlim=c(-1,1), col=rgb(1,0,0,0.7),
     main="88th Congress", xlab = "W-NOMINATE Score")
hist(dem2$wnominate, xlim=c(-1,1), col=rgb(0,0,1,0.7), add=T)
dev.off()
```

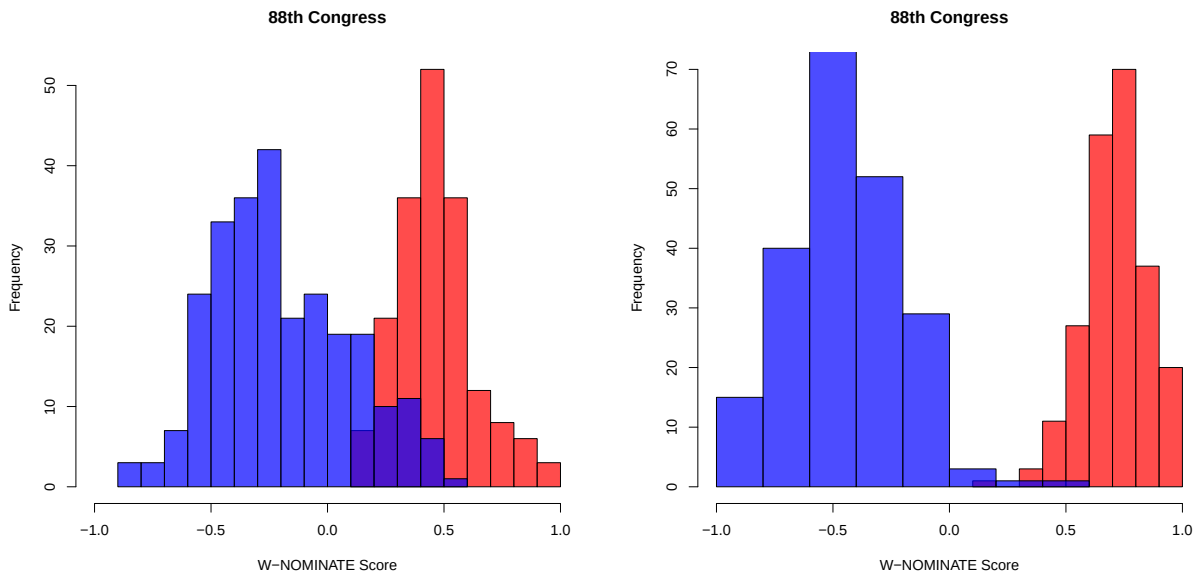
The output from the R code can be seen in left panel of Figure 7.

(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 Republican Party.*

```
pdf("wnominateHist107.pdf")
hist(rep3$wnominate, xlim=c(-1,1), col=rgb(1,0,0,0.7),
     main="88th Congress", xlab = "W-NOMINATE Score")
hist(dem3$wnominate, xlim=c(-1,1), col=rgb(0,0,1,0.7), add=T)
dev.off()
```

The output from the R code can be seen in right panel of Figure 7.

Figure 7: Histogram: W-NOMINATE scores for the Democratic and Republican Parties from the 88th and 107th Congress.



(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 responsible for this or is one party responsible?*

  Polarization increased. Both parties are responsible.

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

  Cohesiveness increased for both parties.