Problem Set 2

Quantitative Political Methodology (L32 PS 363)

Instructions

1. Print out and write your WUSTL ID at the top of **each** page, and complete each exercise in the space alloted. You may attach extra pages if the space provided is not sufficient, but please indicate that you have done so below.

Pages attached:_____

2. Please **show your work** if possible. You may lose points by simply writing in the answer. If the problem requires you to use R, please include the code you used to get your answers. If you are not sure if work needs to be shown for a particular problem, please ask a TA or post a question on Facebook.

3. The various pages of your homework should be **stapled together** (no paper clips please). If pages are lost because of a lack of a staple, no credit will be granted for that portion of the homework.

4. This problem set is **due at the beginning of class on Wednesday, September 27, 2017**. No late assignments will be accepted.

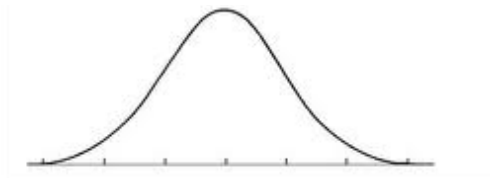5. Total available points for this homework: 70.

Question 1   (Total: 4 points)
   What proportion of a normal distribution falls in the following ranges (round to 3 decimal points)? Use Table A on the last page of Agresti and Finlay's book. Highlight the appropriate portion of the graph.
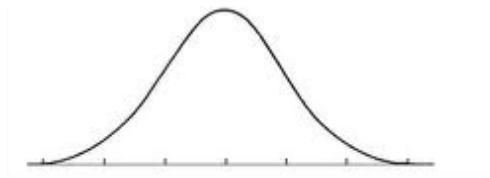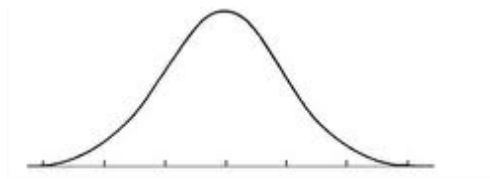
   (a) (1 point) Above a z-score of 1.96.
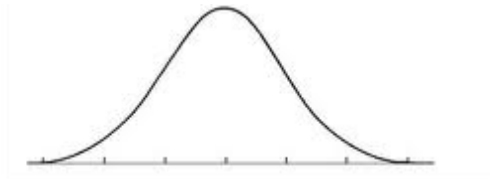
   (b) (1 point) Below a z-score of 1.96.

   (c) (1 point) Above a z-score of -1.96.

   (d) (1 point) Between the z-score of -1.96 and 1.96.

Question 2   (Total: 13 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) (1 point) Identify the population distribution. State its mean and standard deviation. Is the data skewed?

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

(c) (2 points) Identify the sampling distribution of $\bar{y}$. State its mean and standard error and explain what it describes.

(d) (3 points) Find the probability that her sample mean falls within 0.5 of the population mean.

(e) (3 points) 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).

(f) (3 points) Refer to (e). If the sample were truly random, would you be surprised if the anthropologist obtained $\bar{y} = 4$. Why?

Question 3   (Total: 9 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) (2 points) Show how to construct a 95% confidence interval from the information provided.

(b) (3 points) Interpret the confidence interval you found in (a).

(c) (4 points) 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?

For all of the following questions, use R and RStudio. At the end of the Problem Set, attach the R codes you used. If you have plots, attach them as well. Make sure you label clearly which question the codes correspond to. You will NOT receive full credit if R code is missing.

Question 4   (Total: 4 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) (1 point) At or below the value 57.75

(b) (1 point) At or above the value of 50.45

(c) (2 points) Between the values of 52.4 and 59.4

Question 5   (Total: 4 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.

Question 6   (Total: 6 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) (1 point) Normal Distribution with $\mu = 0$ and $\sigma^2 = 0.4$

(b) (1 point) Normal Distribution with $\mu = 0$ and $\sigma^2 = 3$

(c) (1 point) Normal Distribution with $\mu = 3$ and $\sigma^2 = 3$

(d) (1 point) Normal Distribution with $\mu = 3$ and $\sigma^2 = 0.4$

(e) (1 point) Normal Distribution with $\mu = -2$ and $\sigma^2 = 5$

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

Question 7   (Total: 10 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 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) (1 point) Draw a histogram of the monthly count of drug-related stories.

(b) (3 points) 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?

(c) (3 points) 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?

(d) (3 points) 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?

Question 8   (Total: 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://jmontgomery.github.io/ProblemSets/wnominatehouse.csv). 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) (2 points) 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).

(b) (2 points) For the Democratic Party, calculate the median W-NOMINATE scores for two Congresses. How did the median change over time? What does this mean?

(c) (2 points) 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?

(d) (3 points) 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?

(e) (3 points) 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?

(f) (2 points) 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.)

(g) (2 points) 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.

(h) (4 points) 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?

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