Problem Set 3

Quantitative Political Methodology (U25 363)

Due: April 3, 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 April 3, 2019. No late assignments will be accepted.*
- *Total available points for this homework is 100.*

Question 1 (5 points)

Using data on the 2008 New Hampshire Democratic Party Primary, visualize the relationship between the proportion of voters for Howard Dean in the 2004 Democratic primary and the proportion of voters for Barack Obama in the 2008 Democratic primary. To get the dataset, type:

```
install.packages("faraway")
library("faraway")
data("newhamp")
help("newhamp")
```

In addition to the relationship between the support for Dean in 2004 and the support for Obama in 2008, we are also interested in whether two different voting systems — hand-counted and machine-counted ballots — matter. At a minimum, you have to do the following things in a single plot:

- *Properly label titles and axes*
- *Set the ranges of the axes appropriately*
- *Use different colors and symbols to indicate the two ballot systems*
- *Include a legend that explains colors and symbols*

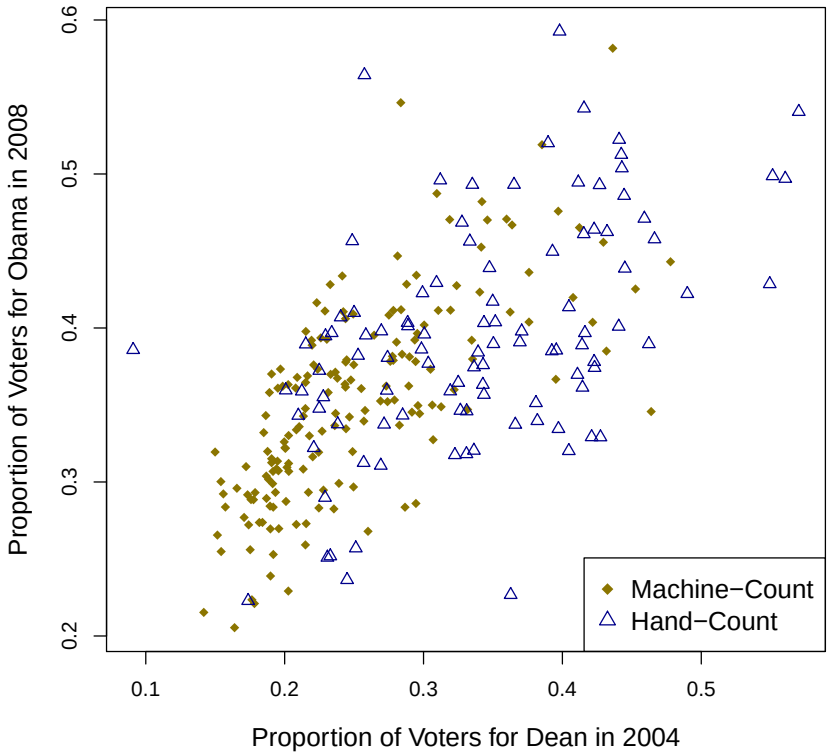
```
# install packages and load library/data
install.packages("faraway")
library("faraway")
data("newhamp")

# open up .pdf
pdf("Q1.pdf")
# plot of voters for Dean
# by voters for Obama
# this will plot only those where the votes
# were counted by machine
plot(newhamp$Dean[newhamp$votesys=="D"],
     newhamp$pObama[newhamp$votesys=="D"],
     col="gold4", pch=18,
     xlim=c(min(newhamp$Dean), max(newhamp$Dean)),
     ylim=c(min(newhamp$pObama), max(newhamp$pObama)),
     xlab="Proportion of Voters for Dean in 2004",
     ylab="Proportion of Voters for Obama in 2008",
     main="", cex.lab=1.25)

# now add points for wards with hand counting
points(newhamp$Dean[newhamp$votesys=="H"],
       newhamp$pObama[newhamp$votesys=="H"],
       col="darkblue", pch=2)

# add legend to bottom
legend("bottomright",
       pch=c(18,2),
       col=c("gold4", "darkblue"),
       legend=c("Machine-Count", "Hand-Count"),
       cex=1.25)
# close .pdf!
dev.off()
```

Figure 1: New Hampshire Democratic Party Primary 2008



Question 2 (10 points)

The shape of the t-distribution varies by a parameter call "degrees of freedom," or df for short.

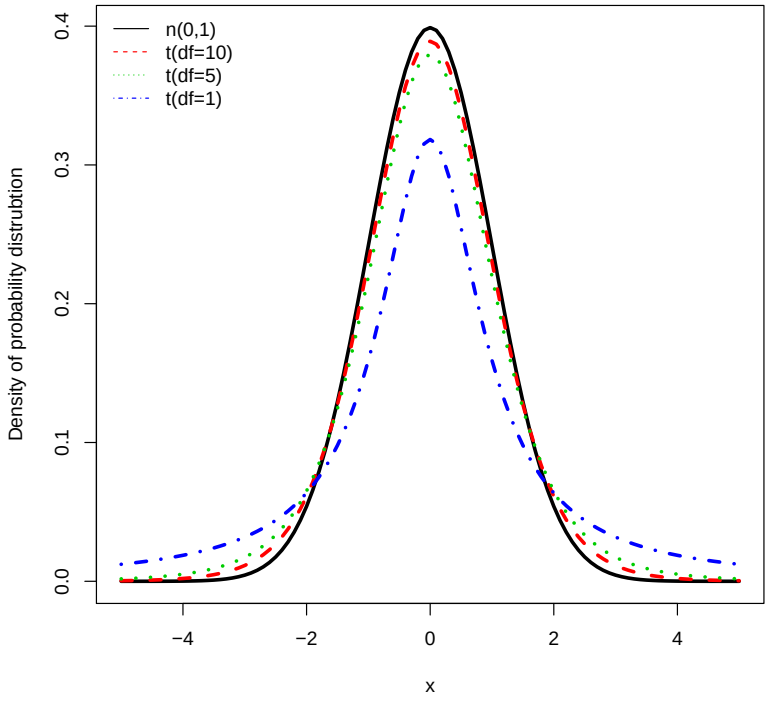
(a) *When df is large, the t-distribution approximates what other distribution?*

When *df* is large, the t-distribution approximates the normal distribution.

(b) *Use* `R` *to plot the standard normal distribution as well as three t-distributions with* $df = 20$, $df = 3$, *and* $df = 1$. *Present all plots on* the same set of axes, *print, and attach to your submitted homework. Also attach the code used to produce the plot. Give your plot a meaningful title and label your axes. Use a different color, shade of gray, or line type for each line so a reader can clearly see the difference. (You may find* `R`*'s help files useful. For example, try* `?plot`, `?lines`, `?dt`.)

```r
# set.seed() for reproducibility
set.seed(5)
# create vector of data for all values of x
x <- seq(from=-5, to=5, by=.1)
# open up .pdf
pdf("Q2.pdf")
# create line plot for the distribution function
# of the normal distribution
plot(x, dnorm(x), lwd=3, type="l", col=1, lty=1,
     ylab="Density of probability distrubtion")
# add line for t-distribution w/ DF=10
lines(x, dt(x, df=10), lwd=3, ylim=c(0, .4), col=2, lty=2)
# add line for t-distribution w/ DF=5
lines(x, dt(x, df=5), lwd=3, ylim=c(0, .4), col=3, lty=3)
# add line for t-distribution w/ DF=1
lines(x, dt(x, df=1), lwd=3, ylim=c(0, .4), col=4, lty=4)
# open legend and edit values in legend
legend("topleft",
       c("n(0,1)", "t(df=10)", "t(df=5)", "t(df=1)"),
       lty=c(1,2,3,4), col=c(1,2,3,4), bty="n")
dev.off()
```

Figure 2: Comparison of normal distribution with t distributions of varying degrees of freedom



(c) *Describe what your plot shows about the t-distribution. With reference to your plot, explain how different sample sizes might affect your estimates of population parameters.*

Figure 2 demonstrates how the t-distribution varies with increases in df, and therefore increases in sample size. t-distributions with higher degrees of freedom approach the standard normal distribution, shown on the plot in black.

We can see from Figure 2 that t-distributions with fewer degrees of freedom have fatter tails, suggesting that estimates for population parameters based on smaller samples are more likely to be further from the distribution mean, and therefore less accurate.

Question 3 (20 points)

Please find the data for this question by using the following code:

```
install.packages("Zelig")
library("Zelig")
data("voteincome")
?voteincome
```

Make sure to show all your work for parts (b) and (d) either with `R` *code you attach or by hand in the space provided. If you complete (b) and (d) in* `R`*, make sure to clearly label the code pertaining to each part of the problem.*

You would like to test the hypothesis whether the average age among American voters is different from 50.

(a) *State the null hypothesis and the alternative hypothesis.*

$$H_0 : \mu = 50$$
$$H_A : \mu \neq 50$$

(b) *Calculate the standard error, the z test-statistic and the p-value for your test.*

```r
# load data from Zelig
library("Zelig")
data("voteincome")

# how many observations are there w/o NAs
n <- length(na.omit(voteincome$age))
# get mean age (remember to remove NAs)
mean_age <- mean(voteincome$age, na.rm = T)
```

```r
# get standard dev. for age
sd_age <- sd(voteincome$age, na.rm = T)

# calculate standard error
std_error <- sd_age / sqrt(n)
# view SE
std_error
# [1] 0.4511027

# calculate t statistic
test_stat <- (mean_age - 50)/std_error
# view t statistic
test_stat
#[1] -1.637469

# double the lower tail because sample
# mean is less than hypothesized value
p_val <- 2*pnorm(test_stat, mean = 0, sd = 1, lower.tail = TRUE)
# view p-value
p_val
# [1] 0.1015326
```

(c) *What is your conclusion at $\alpha = 0.05$?*

Cannot reject null at the 95% confidence interval (we didn't observe a test statistic greater than the value we would expect to see for $\alpha = 0.05$).

(d) *Calculate the 95% confidence interval for the mean age.*

```r
# generate test statistic for CIs
# i.e. the value of the normal distribution's CDF
# for the value of x we're intersted in (95%)
# two tailed, so x=0.025
z_score <- qnorm(.025, lower.tail = FALSE)
# calculate lower bound
lwr_bound <- mean_age - z_score * (sd_age/sqrt(n))
# calculate upper bound
upr_bound <- mean_age + z_score * (sd_age/sqrt(n))
# view CIs
c(lwr_bound, upr_bound)
#[1] 48.37719 50.14548
```

(e) *How are your answers for parts (c) and (d) related?*

We cannot reject the null that the mean age equals 40 at $\alpha = 0.05$. We see that 50 is within the 95% confidence.

Question 4 (25 points)

Please show all work for this problem in the space provided.

A librarian would like to learn how many books their patrons purchase per year. In particular he wants to test the $H_0 : \mu = 10$ *against* $H_a : \mu < 10$*. He randomly selects 16 patrons and asks them how many books they purchase per year. He finds that* $\bar{y} = 9.5$ *and* $s = 1.2$*.*

(a) *Given the sample size and the fact that you do not know the population standard deviation, which test statistic would you use?*

We should use a t-statistic.

(b) *What additional assumption do you need to use the test-statistic indicated in part (a)?*

Population is normally distributed.

(c) *Calculate the test-statistic, and the p-value. What is your conclusion at significance level* $\alpha = 0.05$*?*

- Standard error: $\dfrac{1.2}{\sqrt{16}} = 0.3$
- t-statistic: $\dfrac{9.5 - 10}{0.3} = -1.666667$
- p-value: 0.05815953 (from `pt(-1.666667,df=15,lower.tail=T)` or table)
  Critical value from t table is 1.753.

We cannot reject the null hypothesis at $\alpha = 0.05$.

(d) *Assume that you know the population standard deviation* $\sigma = 1.2$*. Can you use a test statistic different from the one indicated in part (a)? If so, what is that test-statistic called?*

Yes, it is a z-statistic.

(e) *What assumption (if any) do you need to use the test-statistic indicated in part (d)?*

Population is normally distributed.

(f) *Calculate the standard error, the test-statistic and the p-value. What is your conclusion at significance level $\alpha = 0.05$?*

- Standard error and z statistic same as c.
- p-value: 0.04779032 (from `pnorm(-1.666667,lower.tail=T)`)

We can reject the null at $\alpha = 0.05$.

(g) *Compare your conclusions in parts (c) and (f). Explain the difference (if any) in the conclusions.*

We can reject the null using z statistic because we know that the standard deviation of the population decreases uncertainty.

Question 5 (5 points)

A recent poll of 698 decided voters in Pennsylvania showed 341 preferred Donald Trump and 357 preferred Hillary Clinton. Let π be the population proportion of decided Pennsylvania voters who prefer Trump.

(a) *If the voters are only given two options (Trump or Clinton) and the sample size of your survey is relatively large, what type of distribution is the population distribution? What type of distribution is the sampling distribution of your survey?*

The population distribution is binomial, while the sampling distribution is normal.

(b) *What is the value of $\hat{\pi}$, the estimate of π obtained from the survey?*

$\hat{\pi} = \frac{341}{698} = 0.489$

(c) *What is the standard error of this estimate?*

$se = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} = \sqrt{\frac{0.489(1-0.489)}{698}} = 0.0189$

(d) *Give the 95% confidence interval for the value of π.*

$\hat{\pi} \pm 1.96se = [0.452, 0.526]$

Question 6 (10 points)

Field experiments have become an important tool to understand various political phenomena. For example, "Getting Out the Vote in Local Elections: Results from Six Door-to-Door Canvassing Experiments" by Green, Gerbern, and Nickerson (2003) is an early field experimental work on political behavior. The paper is available at `http://onlinelibrary.wiley.com/doi/10.1111/1468-2508.t01-1-00126/full`. *Read the abstract (and introduction if you wish) of the paper and answer the following questions.*

(a) *As succinctly as possible, what is the causal claim being made by the authors?*

Direct canvassing increases turnout.

(b) *What is the "treatment" (or predictor) variable?*

Canvassing.

(c) *What is the outcome variable?*

Turnout.

(d) *What allows the authors to claim that their findings are causal?*

Randomization.

Question 7 (15 points)

For the 2006 GSS, a comparison of males and females on the number of hours a day that the subject watched TV gave:

Group	N	Mean	St.Dev	SE Mean
Females	1117	2.99	2.34	0.070
Males	870	2.86	2.22	0.075

(a) *Conduct all parts of a significance test to analyze whether the population means differ for females and males. Interpret the p-value, and report the conclusion for α-level = 0.05.*

Let μ_1 = mean number of hours a day that females watch TV and μ_2 = mean number of hours a day that males watch TV.

$$\begin{aligned} \bar{y_2} - \bar{y_1} &= 2.86 - 2.99 = -0.13 \\ se &= 0.10 \\ H_0 &= \mu_2 - \mu_1 = 0 \\ H_1 &= \mu_2 - \mu_1 \neq 0 \\ z &= -1.30 \end{aligned}$$

We fail to reject H_0. It appears that there is no difference in the mean number of hours a day that males and females watch TV.

(b) *If you were to construct a 95% confidence interval comparing the means, would it contain 0? (You can answer based on the result in (a), without finding the interval.)*

Since we failed to reject $H_0 = \mu_2 - \mu_1 = 0$ at the $\alpha = 0.05$ level, a 95% confidence interval would contain 0.

(c) *Do you think that the distribution of TV watching is approximately normal? Why or why not? Does this affect the validity of your inferences? Explain your answer.*

The distribution of TV watching does not appear to be normal, since the standard deviations are almost as large as their respective means. Since the sample sizes are large and the t procedures are robust against violations of normality, the inferences are probably fine despite non-normality.

Question 8 (10 points)

Imagine that the data above is changed as below (note the changed sample size). A comparison of males and females on the number of hours a day that the subject watched TV gave:

Group	N	Mean	St.Dev	SE Mean
Females	11	2.99	2.34	0.070
Males	16	2.86	2.22	0.075

Conduct all parts of a significance test to analyze whether the population means differ for females and males. Interpret the p-value, and report the conclusion for α-level = 0.05.

Let μ_1 = mean number of hours a day that females watch TV and μ_2 = mean number of hours a day that males watch TV.

$$\bar{y_2} - \bar{y_1} = 2.86 - 2.99 = -0.13$$

Since we are comparing groups with small samples, we need to calculate pooled variance

$$\hat{\sigma} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(11 - 1) * 2.34^2 + (16 - 1) * 2.22^2}{11 + 16 - 2}} = 2.27$$

$$H_0 = \mu_2 - \mu_1 = 0$$
$$H_1 = \mu_2 - \mu_1 \neq 0$$

$$\hat{\sigma}_{\bar{y_2} - \bar{y_1}} = \hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 2.27\sqrt{\frac{1}{11} + \frac{1}{16}} = 0.89$$

$$d.f = 11 + 16 - 2 = 25$$
$$t = \frac{-0.13 - 0}{0.89} \approx -0.15$$
$$p = 0.885$$

We fail to reject H_0. It appears that there is no difference in the mean number of hours a day that males and females watch TV even in the reduced sample.