Problem Set 07

WRITE YOUR NAME HERE

WRITE DATE HERE

Contents

Question 9	**10**
Question 10	**10**
Question 11	**10**

Collaboration

Please indicate who you collaborated with on this problem set:

Background

For this exercise, we will mimic the tactile sampling you did in class with virtual sampling. We will use some data from the general social survey, an annual personal-interview survey conducted in the United States. The survey is designed to monitor changes in both social characteristics and attitudes.

For this exercise, the **population** of interest will be **ALL** 2538 individuals living in a single neighborhood in 2014. As an analogy to the tactile sampling you did in class, the neighborhood is the "bowl" and the 2,538 people are the little balls.

If you get stuck as you are working through this Problem Set, it will likely be helpful to review Chapter 7 in ModernDive, in particular subsections 7.3.1 on "Terminology & notation" and 7.3.2 on "Statistical definitions". The terminology, notation, and definitions related to sampling are definitely tricky at first; the best method to master them is practice, practice, practice.

Key Symbols:

Symbol	Population Parameter	Point Estimate
Number of cases	N	n
Proportion	p	$\hat{p}$
Standard error	SE	$\widehat{SE}$

Setup

First load the necessary packages:

```
library(ggplot2)
library(dplyr)
library(forcats)
library(moderndive)
```

The GSS data we will be working with is in the `gss_cat` data frame, which comes built-in with the `forcats` package you just loaded. You can take a `glimpse()` of the `gss_cat` data set like so:

```
data(gss_cat)
glimpse(gss_cat)
```

You should also explore the data in the RStudio data viewer, and type `?gss_cat` into the **console** to see a description of the variables in this data set.

Exploratory data wrangling

This data set includes many years of data, and many variables. To start, we will restrict our analysis to only 2014, and to only the variable indicating the `marital` status of each respondent.

```
gss_14 <- gss_cat %>%
  filter(year == 2014) %>%
  select(marital)
```

The following shows all the unique responses for `marital` status:

```
gss_14  %>%
  distinct(marital)
```

```
## # A tibble: 6 x 1
##   marital
##   <fct>
## 1 Divorced
## 2 Married
## 3 Never married
## 4 Separated
## 5 Widowed
## 6 No answer
```

Setting a seed for your computer's Random Number Generator

In this problem set, will take some random samples of data using R. In order to make sure R takes the same random sample every time you run your code (so you can reproduce your work), you can do what is called "setting a seed". Do this in any code chunk where you take a random sample! Otherwise, the answers you write down might accidentally become out of sync with the output of your code when your knit your document!

You can control your computer's random number generator by providing a number to using the `set.seed` function. Any number will do - in the example below, we use 45 as our seed value.

```
set.seed(45)
```

The true population proportion p of divorced people

Again, for this exercise, the **population** of interest will be **ALL** 2,538 individuals living in this single neighborhood in 2014. Since we have data on **ALL** 2538 people living in the neighborhood, we can compute the **exact population proportion** p **of divorced people directly** using **ALL** the data:

```
gss_14 %>%
  summarize(divorced = sum(marital == "Divorced"),
            N = n()) %>%
  mutate(p = divorced / N)
```

```
## # A tibble: 1 x 3
##   divorced     N     p
##      <int> <int> <dbl>
## 1      411  2538 0.162
```

> Note that we used N (the size of the full population, 2,538 people) and computed p (not $\hat{p}$). And, no inference from sample to the population is needed. This is because we're working with the **entire population** of interest. We do not need to *estimate* the true proportion, or infer

> something about the true population proportion of divorced people in this neighborhood in 2014, because in this case, we can compute it directly (just like counting all red balls in the bowl). Thus, we know that p is exactly 0.16. In other words, this situation is not a realistic reflection of a real life problem.

For the rest of this problem set, we will be *simulating* the act of sampling from this neighborhood population to understand and study how factors like sample size influence **sampling variation**.

Demo: Sampling 50 people in the neighborhood

Estimating $\hat{p}$ from a single sample

We are first going to use random sampling to **ESTIMATE** the true **population** proportion p of the neighborhood that are divorced with only a **sample** of 50 people.

> This will represent a situation of only having the resources to knock on 50 doors to get responses from people in this neighborhood!

```
set.seed(42)

n50_1rep <- gss_14 %>%
  rep_sample_n(size = 50, reps = 1)
```

Be sure to look at the results in the viewer. And remember, you can set the seed to whatever value you like.

Next, let's calculate the **sample proportion** $\hat{p}$ of people who identified as `Divorced` in our sample of 50 people.

```
n50_1rep %>%
  summarize(divorce_count = sum(marital == "Divorced"),
            n = n()) %>%
  mutate(p_hat = divorce_count/ n)
```

```
## # A tibble: 1 x 4
##   replicate divorce_count     n p_hat
##       <int>         <int> <int> <dbl>
## 1         1             8    50  0.16
```

This sample proportion $\hat{p}$ is an **ESTIMATE**; it's our **best guess** of what the **true population** proportion p of `Divorced` people is in this neighborhood, based on a sample of only 50 people. It is reasonably close to the true population proportion $p = 0.16$ we calculated from the full population.

Question 1a

Modify the code below to take 3 samples of 50 people instead of just 1 sample, and then compute an estimate of the proportion of Divorced people in the entire population from each sample individually.

```
set.seed(18)

n50_3rep <- gss_14 %>%
  rep_sample_n(size = 50, reps = 1)
```

Answer:

Question 1b

How do the $\hat{p}$ estimates from different samples compare? Why was your estimate different each time?

Answer:

Estimating $\widehat{SE}$ from a single sample

Typically we only have the opportunity to collect **one sample** for our study, and so we have to use the amount of variability in our **single sample** as an estimate of the amount of variability we might expect in our results if we had taken a random sample of 50 different people. The $\widehat{SE}_{\hat{p}}$ serves as an **ESTIMATE** of **sampling variability** if you only have a **single sample**. The formula for estimating the standard error of $\hat{p}$ is the following:

$$\widehat{SE}_{\hat{p}} \approx \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}}$$

> Note that we use n to represent the size of the sample and we that use $\hat{p}$ to represent the proportion of divorced people because we are ESTIMATING a proportion based on only a sample. Likewise, the SE "wears a hat" because we are ESTIMATING the true standard error based on a sample.

The standard error of $\hat{p}$ can be estimated in R like so:

```
n50_1rep %>%
  summarize(divorce_count = sum(marital == "Divorced"),
            n = n()) %>%
  mutate(p_hat = divorce_count/ n,
         se_hat = sqrt(p_hat * (1 - p_hat) / n))
```

```
## # A tibble: 1 x 5
##   replicate divorce_count     n p_hat se_hat
##       <int>         <int> <int> <dbl>  <dbl>
## 1         1             8    50  0.16 0.0518
```

Demo: Generating a sampling distribution of $\hat{p}$

If you ran the code chunk that takes a random sample of 50 people a thousand more times, and wrote down every $\hat{p}$ you got, you would have constructed a "sampling distribution" of the proportion of divorced people.

> A sampling distribution shows every (or nearly every!) possible value a point estimate can take on, along with how likely each value is to be observed, for samples **of a given size** from a population.

Sampling distribution of $\hat{p}$ for n = 50

Instead of running the sampling code chunk for n = 50 over and over, we can "collect" 1000 samples of n = 50 really easily in R. The following code chunk takes 1000 **different** samples of n = 50 and stores them in the data frame `n50_1000rep`:

```
set.seed(19)

n50_1000rep <- gss_14 %>%
  rep_sample_n(size = 50, reps = 1000)
```

Be sure to look at `n50_rep1000` in the data viewer to get a sense of these 1000 samples look like.

Question 2a

What is the name of the column in the `n50_1000rep` data frame that identifies which of the 1000 samples an observation belongs to?

Answer:

Question 2b

What is the sample size n for each of the 1000 samples we took? (i.e. how many people are sampled each time)?

Answer:

The following code chunk calculates the sample proportion $\hat{p}$ of people who reported they were divorced for each of the **1000 samples**

```
p_hat_n50_1000rep <- n50_1000rep %>%
  group_by(replicate) %>%
  summarize(divorce_count = sum(marital == "Divorced"),
            n = n()) %>%
  mutate(p_hat = divorce_count / n)
```

Take a look at the first five rows of the results:

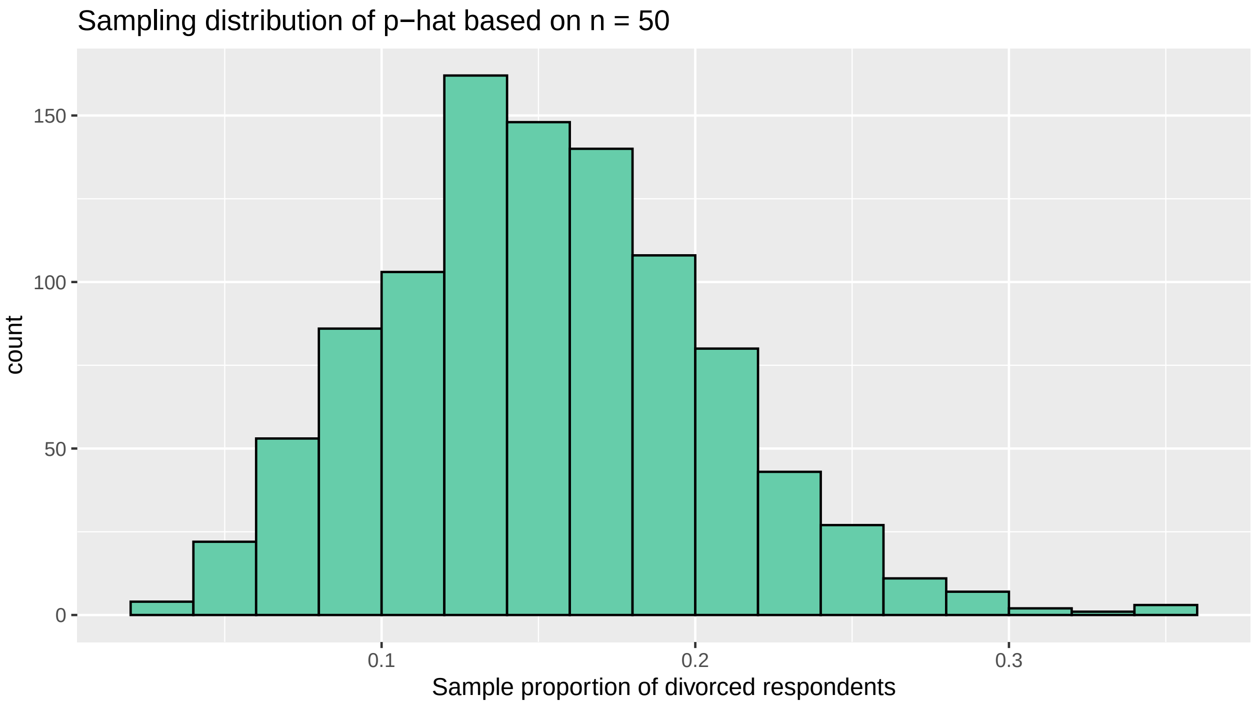
```
p_hat_n50_1000rep %>%
  slice(1:5)
```

```
## # A tibble: 5 x 4
##   replicate divorce_count     n p_hat
##       <int>         <int> <int> <dbl>
## 1         1             4    50  0.08
## 2         2             5    50  0.1
## 3         3             8    50  0.16
## 4         4             9    50  0.18
## 5         5            10    50  0.2
```

Visualizing the sampling distribution of $\hat{p}$ for n = 50

We can plot the **sampling distribution** of these 1000 $\hat{p}$ estimates of divorced respondents with a histogram, like so:

```
ggplot(p_hat_n50_1000rep, aes(x = p_hat)) +
  geom_histogram(binwidth = 0.02, color = "black", fill = "aquamarine3", boundary=0) +
  labs(x = "Sample proportion of divorced respondents",
       title = "Sampling distribution of p-hat based on n = 50")
```



Question 3

Based on the histogram above, what appeared to be a very common (i.e., likely) value of $\hat{p}$? What was a very uncommon (i.e., unlikely) value?

Answer:

Mean and standard error of the sampling distribution of $\hat{p}$ for n = 50

Finally we can estimate the mean of the sampling distribution by calculating the mean of all 1000 $\hat{p}$ estimates, and the standard error of the sampling distribution by calculating the standard deviation of all 1000 $\hat{p}$ values like so:

```r
p_hat_n50_1000rep %>%
  summarize(M_p_hat = mean(p_hat),
            SE_p_hat = sd(p_hat))
```

```
## # A tibble: 1 x 2
##   M_p_hat SE_p_hat
##     <dbl>    <dbl>
## 1   0.162   0.0522
```

Basically, we treat the 1000 point estimates of the population proportion just like any other sample of numbers.

Question 4

How do the population proportion and standard error estimates computed by taking the mean and standard deviation of the 1000 simulated sample proportions compare to the estimates of $\hat{p}$ and $\widehat{SE}$ based on your

single sample of 50 people earlier in this Problem Set?

Answer:

Question 5a

Use the `rep_sample_n` function to collect 1000 virtual samples of size $n = 15$. **BE SURE TO NAME YOUR SAMPLE SOMETHING NEW, TO ENSURE YOU CAN DISTINGUISH IT FROM THE n = 50 SAMPLE ABOVE!**

Answer:

```
set.seed(910)
```

Question 5b

Calculate sample proportion $\hat{p}$ of people who reported they were `Divorced` for each replicate of your n = 15 sampling

Answer:

Question 5c

Visualize the sampling distribution of $\hat{p}$ from your n = 15 sampling with a histogram

Answer:

Question 5d

Calculate the mean of the n = 15 sampling distribution, and the standard error of the n = 15 sampling distribution

Answer:

Question 6a

How does the standard error of the n= 15 sampling distribution compare to the standard error of the n = 50 sampling distribution?

Answer:

Question 6b

Explain any observed differences from Question 6a

Answer:

Question 7a

Use the `rep_sample_n` function to collect 1000 virtual samples of size $n = 600$. **Note: BE SURE TO NAME YOUR SAMPLE SOMETHING NEW, TO ENSURE YOU CAN DISTINGUISH IT FROM THE n = 50, and n = 15 SAMPLES ABOVE!**

Answer:

```r
set.seed(84)
```

Question 7b

Calculate the proportion $\hat{p}$ of people who reported they were `Divorced`for each replicate of your n = 600 sampling

Answer:

Question 7c

Calculate the mean of the n = 600 sampling distribution, and the standard error of the n = 600 sampling distribution.

Answer:

Question 7d

Was there more **variability** from sample to sample when we took a sample size of 600 or when we took a sample size of 50? **Explain what evidence you have for assessing this**

Answer:

Question 8

Which sampling distribution looked more normally distributed (bell shaped and symmetrical); the one built on n = 15, 50 or 600? **Why?**

Answer:

Estimating $\hat{p}$ and the standard error of $\hat{p}$ from a single sample (revisited)

In most instances, we do not have access to the full population as we did in this GSS data; instead we have to take a **sample** to try to say something about the **larger population**. Furthermore, in the real world, we typically only take a **single** sample from the population, due to time or money constraints.

So how do we **ESTIMATE** a $\hat{p}$ and a standard error of $\hat{p}$ when we only have a single sample, and not 1000 repeated samples? As demonstrated at the very beginning of the Problem Set we:

- estimate $\hat{p}$ from the sample
- use the formula for the standard error of $\hat{p}$ below, to estimate SE based on a single sample

$$\widehat{SE}_{\hat{p}} \approx \sqrt{\frac{\hat{p} \times (1-\hat{p})}{n}}$$

Question 9

Imagined we collected only a single small sample of 15 respondents like so:

```
set.seed(53)

n15_1rep <- gss_14 %>%
  rep_sample_n(size = 15, reps = 1)
```

Following the example from the beginning of the Problem Set (roughly line 140), estimate

1. The sample proportion $\hat{p}$ of people who identified as `Divorced` based on `n15_1rep`.
2. The standard error of $\hat{p}$

   You should get a value reasonably close to the estimate we made earlier from our simulated sampling distribution for n = 15! Note that when you must estimate a standard error from **only a single sample**, the formula **contains the sample size, n**. The larger the sample size n, the larger the number in the denominator of the SE formula.

Question 10

Fill in the R Markdown table below with all the standard errors you computed for this problem set. In other words:

1. Replace `x` with the standard error you obtained by taking the standard deviation of the n = 15 sampling distribution
2. Replace `a` with the standard error you obtained fora single sample of n = 15 using the mathematical formula
3. etc., etc.

When you are done, make sure all the `|` in the table still line up so your results print out in a table!

Sample size n	SE via sd of sampling distribution	SE via one sample and formula
15	x	a
50	y	b

Question 11

Based on what you observed for Question 10, **IF** you collected a single sample from 600 respondents, do you think the standard error will be smaller or larger than the one you calculated for n = 15. **Explain your reasoning** (Note: if you are not sure you can collect a sample and calculate the standard error)

Answer: