Problem Set 3

Due date: 2 October

Please upload your completed assignment to the ELMs course site (under the assignments menu). Remember to include an annotated script file for all work with R and show your math for all other problems (if applicable, or necessary). Please also upload your completed assignment to the Github repository that you have shared with us. *We should be able to run your script with no errors.*

Total points: 30

Question 1

Points: 10

Suppose I want to examine the hypothesis that consuming political news coverage from cable network channels (e.g., CNN, MSNBC, Fox News, etc.) causes citizens to have less trust in government than if they were to consume it from mainstream print media (e.g., New York Times, Wall Street Journal, etc.). What is more, suppose that I believe that this effect is greatest among "young" citizens. Design the contours of an experiment to test this proposition (and be sure to sketch the key elements of the experimental design). Second, briefly describe any significant shortcomings that the experimental design might have.

Experimental Design

Pre-treatment questions

- Ask for respondents demographics and what kind of news media or just media consumption the respondents take part in

    - Podcast, social media, news outlets, news channels
    - This will hopefully let us know if a respondent has any sort of pre-treatment biases and is an outlier

1. Sample collection

- I would make sure my data sample population to be at least 200 respondents so that my sample size is strong, balanced across the selected age groups and test the effect. Ideally I'd prefer 400-1,500 respondents.

- Ideally, I am using Prolific or some other organization to gather my sample so it is randomized and I can get a nationally representative data. While it doesn't ensure my data will be balanced across the board but it is better than trying to grab a convince sample

- Experiment will randomly assign individuals to a treatment vs control category

(100 respondents min) Young citizens = 18-35

(100 respondents min) Older citizens = 36+

2. Data collection

Experimental survey study

Control: Have respondents read mainstream print media

Treatment: Watch mainstream cable network news channels

IV: How an individual consumes news

DV: An individuals trust in government

- I would ensure that internal and external validity would prevail

- To ensure that, I would modulate a fake story on a mainstream print article, blur the name of the mainstream journal and make the article look like a real article with a fake news story. (obviously I will let the respondents know it is a fabricated story at the end)

- Next, I would test and check the effects of my treatment and control for both age groups (young and older citizens)

- Then I would specifically look at how the young citizens age group and compare the effects to older to see if my second hypothesis, that cable network news creates a larger effect of distrust in younger adults, and see if my hypothesis was true

3. Short comings of this experimental design would be that we could get response bias, some individuals may be included or excluded more in the sample than others

- People could tell that there is a right or wrong answer in a survey

- Wording effects, some questions could push people to answer in a certain way

- Potential bias, if the sample not representative of population

4. Post-treatment questions

- See if the respondents trust the government after getting the control or treatment

Question 2

Points: 10

Below are four hypothetical research designs to determine the effects of democratic governance on the funding of educational institutions. Briefly evaluate each sampling design in terms of its: (1) sampling variability; and, (2) sources of potential bias.

   A. A sample of educational funding levels for every country in Europe for each of the last 100 years.

   1. The sampling variability

This sample does not have a lot of variability, it is low. Looking at all the countries educational funding in the world vs. JUST the countries in Europe, will have sampling bias.

   2. Potential bias:

The sample only covers countries in Europe, the way education in general can be very similar formats and so the funding levels may also be similar due to the sample all being from the same continent. There could be some differences but overall, they'll likely be similar in one way or another.

B. A sample of educational funding levels in five randomly picked countries drawn from the global population of countries for each of the last 10 years.

   1. Sample Variability

The sample variability is moderately high due to its random selection of countries from the GLOBAL population of countries from the last 10 years.

   2. Potential bias

The potential bias is lower due to the random selection of the countries GLOBALLY not just from one continent/country. This will allow for a better evaluation of democratic governance of educational funding because you will receive a larger variation of countries and bias is cut down due to the random selection

C. A survey of educational administrators that yields responses that allow for a comparison of educational funding levels in Russia and the United States for the six-year period from 2010 to 2015.

   1. Sample Variability

The sample variability for this survey is low because we are only looking at two countries, not randomly selected, and have a survey of educational administrators responses to compare their educational funding levels

- I don't think a survey is the best way to evaluate the educational FUNDING because of potential response bias

2. Potential Bias

Response bias can be found in these surveys, if a educational admin suspects they need to respond in a certain way.

- Selection/sample bias, since these educational administrators are NOT randomly selected (or it isn't explicitly said so), some individuals might be more likely to be included or excluded than others

- Convenience sampling: Just ask whoever is around (e.g college students). Potential bias: Opinions limited to individuals present. Likely to have sampling bias, sample not representative of population - unclear on how they gathered their educational admin responses

- Voluntary Response Sampling: Individuals choose to be involved. Bias: sample designs systematically favors a particular outcomes. Some individuals more likely to be measured than others - unclear on how they gathered their educational admin responses

D. A sample of 75 randomly selected countries' educational funding levels for each of the past 25 years.

1. Sample Variability

The sample variability is very high due to its random selection of 75 countries from the GLOBAL population of countries from the last 25 years.

2. Potential Bias

The potential bias is lower due to the random selection of the countries GLOBALLY not just from one continent/country. This will allow for a better evaluation of democratic governance of educational funding because you will receive a larger variation of countries and bias is cut down due to the random selection. No one group should be over-represented

Question 3

Points: 10

Using the `hanmerKalkanANES` dataset (posted on ELMs), construct a cross-tab with `presvote` (dichotomous variable coded with a value label) as the dependent variable and `ageCategorical` (1 = 18-30; 2 = 31-40; 3 = 41-50; 4 = 51-60; and 5 = 61-over) as the independent variable. Create a table with the cross-tab results and interpret them – i.e., what is the apparent association, if any, between age and U.S. presidential vote in 2004?

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ----------------------- tidyverse 2.0.0 --
v dplyr     1.1.3     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.3     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.0
v purrr     1.0.2
-- Conflicts ------------------------------------------- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(modelsummary)
load("~/Desktop/GVPT 622 Lab/hanmerKalkanANES.RData")
library(dplyr)

# Data cleaning and handling
library(tidyverse)

# Reading data files
library(rio)

# Web-scraping
library(httr2)
library(rvest)
```

```
Attaching package: 'rvest'

The following object is masked from 'package:readr':
```

presvote		1	2	3	4	5	All
Kerry	N	42	30	30	30	55	187
	% row	22.5	16.0	16.0	16.0	29.4	100.0
Bush	N	31	36	40	44	45	196
	% row	15.8	18.4	20.4	22.4	23.0	100.0
All	N	73	66	70	74	100	1000
	% row	7.3	6.6	7.0	7.4	10.0	100.0

```
guess_encoding
```

```r
x <- x |>
  mutate(ageCategorical =case_when(
    18<=age& age<=30 ~ "1",
    31<=age& age<=40 ~ "2",
    41<=age& age<=50 ~ "3",
    51<=age& age<=60 ~ "4",
    61<=age ~ "5"
  ))

summary(x$age)
```

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
18.00   34.00   49.00   48.01   61.00   88.00    617
```

```r
 #cross tab visual
datasummary_crosstab(presvote ~ ageCategorical, data = x)


#The conditional distribution
#how to see how a variable behaves when you know the conditions of another variable

#Kerry conditional distribution
x %>%
  summarize(prob = sum(ageCategorical == "3" & presvote =="Kerry", na.rm = TRUE)/sum(ageCa
```

```
      prob
1 0.4285714
```

```
#Bush conditonal distribution
x %>%
  summarize(prob = sum(ageCategorical == "3" & presvote =="Bush", na.rm = TRUE)/sum(ageCat
```

```
       prob
1 0.5714286
```

Next, what is:

A. The conditional distribution of respondents who are 41-50 that voted for Kerry?

- The conditional distribution of respondents who are 41-50 that voted for Kerry is 0.4285714

B. The conditional distribution of respondents who are 41-50 that voted for Bush?

- The conditional distribution of respondents who are 41-50 that voted for Bush is .5714286