Statistics in R – Data Analysis Project

**Behavioral Risk Factor Surveillance System (BRFSS) - 2013**

# Part 1: Data

## A: 2 pt for correct reasoning for generabizability

Answer should discuss whether random sampling was used. Learners might discuss any reservations, those should be well justified.

BRFSS did use randomization by using “Random Digit Dialing” (Centers for Disease Control and Prevention[CDC], n.d.), which called both random cell phones and landlines. Thus, the results can be generalized to the following population: non-institutionalized, adults (18 years and older) residing in the U.S. with access to a landline or cell phone.

I have reservations about the survey under representing Black and Latino men because of the “non-institutionalized” population. “Black men are six times as likely to be incarcerated as white men, and Hispanic men are more than twice as likely to be incarcerated as non-Hispanic white men”. (The Sentencing Project, n.d.)

Also, individuals without access to a landline or cell phone cannot be a population generalized to. These may include homeless adult populations in the U.S or homes without the means to pay for a landline or cell phone.

References

Centers for Disease Control and Prevention. (n.d.). *BRFSS Frequently Asked Questions (FAQs)*

. <https://www.cdc.gov/brfss/about/brfss_faq.htm>

The Sentencing Project. (n.d.) *Criminal Justice Facts*. <https://www.sentencingproject.org/criminal-justice-facts/>

Population:

* Non-institutionalized, adult population (18 years and older) residing in US.
* Started with 15 states collect information – now all 50 states, DC, PR, Guam and US Virgin Islands, American Samoa, Microneia, Palau
* Monthly telephone interviews

What is collects

* Preventative health practices
* Risk behaviors linked to chronic diseases, injuries and preventable infectious diseases of adult population

About data

* All missing values are coded NA
* Categorical values are factors in R

# B: 1 pt for correct reasoning for causality

Answer should discuss whether random assignment was used.

While the BRFSS did use randomization by using “Random Digit Dialing” (Centers for Disease Control and Prevention[CDC], n.d.), the data was collected as a retrospective observational study, thus establishing correlation, but not causation.

References

Centers for Disease Control and Prevention. (n.d.). *BRFSS Frequently Asked Questions (FAQs)*

. <https://www.cdc.gov/brfss/about/brfss_faq.htm>

# Part 2: Research Questions

## A: 3 pts for each research question (x 3), 2 pts for at least two of the questions involving three variables

* Should be phrased in a non-causal way (1 pt)
* Should be well defined / not vague (1 pt)
* Is clear why this is of interest to the author / audience (1 pt)

**Question 1:**

Are non-institutionalized adults residing in the US with some college education and currently married have a higher probability of having health care coverage. Using only completed interviews.

**Variables**:

* **\_age65yr: Reported Age In Two Age Groups Calculated Variable**
  + 1 Age 18 to 64 [age between 18 and 64]
  + 2 Age 65 or older [age 65 and over]
    - Filter these out of initial result set
  + Need to remove/ignore NA
* **\_hcvu651 = Respondents Aged 18-64 With Health Care Coverage**
  + 1 Have health care coverage [age between 18 and 64, and hlthplan = 1]
  + 2 Do not have health care coverage [age between 18 and 64, and hlthplan = 2]
  + Remove NA – this removes people older than 64
    - Filter these from initial result set
* **\_educag = Computed Level Of Education Completed Categories**
  + 1 Did not graduate High School [educa coded 1, 2, or 3]
  + 2 Graduated High School [educa coded 4]
  + 3 Attended College or Technical School [educa coded 5]
  + 4 Graduated from College or Technical School [educa coded 6]
  + Need to remove/ignore NA
    - Filter these from initial result set
* **Marital = Marital Status**
  + 1 Married
  + 2 Divorced
  + 3 Widowed
  + 4 Separated
  + 5 Never Married
  + 6 A member of an unmarried couple
  + Need to remove/ignore NA
    - Filter these from initial result set
* **dispcode: Final Disposition**
  + 1100 Completed Interview
  + Need to filter from result set any rows where dispcode != 1100

**Question 2:**

Are non-institutionalized adults residing in the US more likely to have had a flu shot in the last 12 months if they have had some college education? Using only completed interviews.

**Variables**

* + **\_educag = Computed Level Of Education**

Compl Value Value Label Frequency Percent Cumulative Frequency

1 Did not graduate High School [educa coded 1, 2, or 3] 42,132 8.57 8.57

2 Graduated High School [educa coded 4] 142,953 29.07 37.64

3 Attended College or Technical School [educa coded 5] 134,242 27.30 64.93

4 Graduated from College or Technical School [educa coded 6] 170,173 34.60 99.54

NA Don’t know/Not sure/Missing [educa coded 9 or missing] 2,273 0.46 100.00

Total 491,773 100.00

Variable type: categorical Missing values: 9 eted Categories

* **flushot6: Adult Flu Shot/Spray Past 12 Mos**

Value Value Label Frequency Percent Cumulative Frequency

1 Yes 208,387 42.37 42.37

2 No 242,939 49.40 91.78

NA Don’t know/Not Sure 959 0.20 91.97

NA Refused 2,791 0.57 92.54

NA [Missing] 36,697 7.46 100.00

Total 491,773 100.00

Variable type: categorical Missing values: 7, 9, BLANK

* + **dispcode: Final Disposition**
    - 1100 Completed Interview
    - Need to filter from result set any rows where dispcode != 1100

**Question 3:**

Are non-institutionalized adults residing in the US less likely to have exercised in the last 30 days if they have reported binge drinking in the last 30 days? Does the likelihood increase if they report binge drinking more than 10 days in the last 30 days? Attempting to find binge drinkers who must have participated in binge drinking on a weekday vs. only weekends. Using only completed interviews.

**Variables**:

* **exerany2: Exercise In Past 30 Days**

Value Value Label Frequency Percent Cumulative Frequency

1 Yes 332,429 67.60 67.60

2 No 125,314 25.48 93.08

NA Don’t know/Not Sure 561 0.11 93.19

NA Refused 2,154 0.44 93.63

NA [Missing] 31,315 6.37 100.00

Total 491,773 100.00

Variable type: categorical Missing values: 7, 9, BLANK Notes: If 2, 7, or 9 is indicated, skip to strength. Refer to \_totinda for another version of this question.

* **\_rfbing5\*\*: Binge Drinking Calculated Variable**

Value Value Label Frequency Percent Cumulative Frequency

1 No 409,209 83.21 83.21

2 Yes 58,831 11.96 95.17

NA Don’t know/Refused/Missing [alcday5 coded 777 or 999,or drnk3ge5 coded 7 or 9] 23,733 4.83 100.00

Total 491,773 100.00

Variable type: categorical Missing values: 9

* **drnk3ge5:Binge Drinking**

Value Value Label Frequency Percent Cumulative Frequency

[1 - 76] Number of drinks 58,831 25.07 25.07

NA Don’t know/Not Sure 2,560 0.01 25.08

0 None 172,592 0.35 25.43

NA Refused 667 0.00 25.43

NA [Missing] 257,123 0.52 25.95

Total 491,773 100.00

Variable type: continuous Missing values: 77, 99, BLANK Recoded to zero: 88 Notes: alcday is coded 888, 777, or 999. Refer to \_rfbing5 for another version of this question.

* **dispcode: Final Disposition**
  + 1100 Completed Interview
  + Need to filter from result set any rows where dispcode != 1100

# Part 3: Exploratory Data Analysis (EDA) – 30 points

* 3 pts for plots
  + Plots should address the research questions (1 pt)
  + Plots should be constructed correctly (1 pt)
  + Plots should be formatted well – size not too large, not too small, etc. (1 pt)
* 3 pts for summary statistics
  + Summary statistics should address the research questions (1 pt)
  + Summary statistics should be calculated correctly (1 pt)
  + Summary statistics should be formatted well – not taking up pages and pages, etc. (1 pt)
* 4 pts for narrative
  + Each plot and/or R output should be accompanied by a narrative (1 pt)
  + Narrative should interpret the visuals / R output correctly (2 pts)
  + Narrative should address the research question (1 pts)

# Overall – 6 points

* Uploaded HTML document resulting from the Rmd template: 1 pt
* Organization: 1 pts
* Readability of the text: 2 pts
* Readability of the code: 2 pts