Some Basics of Descriptive Inference

POST 8000 – Foundations of Social Science Research for Public Policy

Steven V. Miller

Department of Political Science

Goal for Today

Introduce you to basic descriptive analysis in R.

https://github.com/svmiller/post8000/tree/master/lab-scripts

First Things First

```
# new data set. Install it again.
devtools::install_github("svmiller/post8000r")

library(post8000r)
library(tidyverse)

# Check out documentation.
?gss_spending
?pwt_sample
?usa_justifbribe
```

Defining and Measuring Variables

1. Nominal
2. Ordinal
3. Interval
    - You can also toss ratio into this.

Correct classification will condition how we can *describe* variables.

Central Tendency

The most common description of interest is the **central tendency**.

- This is the variables “typical”, or “average” value.
- This takes on different forms contingent on variable type.

Think of what follows as a “tool kit” for researchers.

- More precise variables allow for more precise measures.
- Use the right tool for the job, if you will.

Mode

The **mode** is the most basic central tendency statistic.

- It identifies the most frequently occurring value.

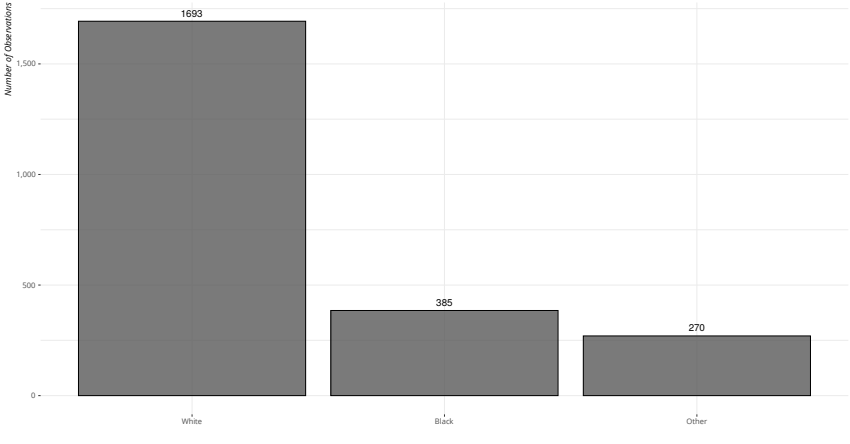
```r
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}

getmode(gss_spending$race)
```

```
## [1] 1
```

A graph will bring this to life as well (see R script).

A Bar Chart of Respondent Race in the General Social Survey (2018)

There's a clear mode for white people.



Data: General Social Survey (2018)

Mode

If I randomly grabbed a respondent from that sample and guessed “White”, I would be right 72 times out of 100 (on average).

- No other guess, on average, would be as good.

This is the only central tendency statistic for nominal variables.

Median

The **median** is the middlemost value.

- It's the most precise statistic for ordinal variables.
- It's a useful robustness check for interval variables too.

Formally, a median *m* exists when the following equalities are satisfied.

$$P(X \leq m) \geq \frac{1}{2} \text{ and } P(X \geq m) \geq \frac{1}{2} \tag{1}$$

Finding the Median

Order the observations from lowest to highest and find what value lies in the exact middle.

- The median is the point where half the values lie below and half are above.
- We can do this when our variables have some kind of “order”.
- Medians of nominal variables are nonsensical.

Alternatively, use R:

```
# The median is those with a HS ed or less.
median(gss_spending$degree, na.rm=T)
```

```
## [1] 1
```

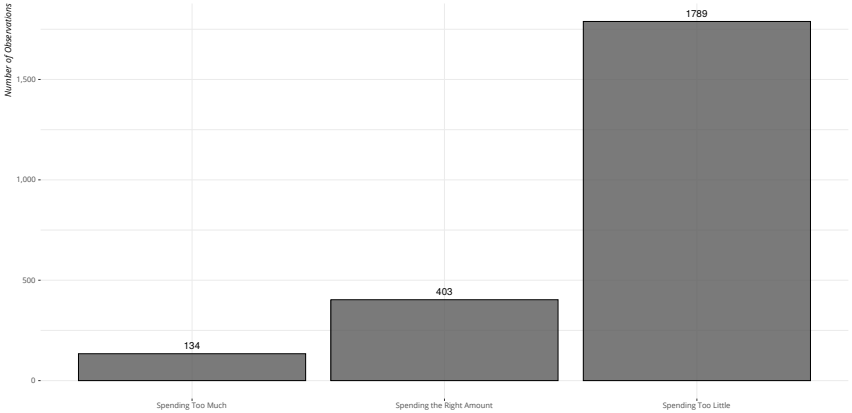
```
# The median is 17 ($35k-$39k)
# https://gssdataexplorer.norc.org/variables/6168/vshow
median(gss_spending$rincom16, na.rm=T)
```

```
## [1] 17
```

Graphing helps too...

A Bar Chart of Attitudes Toward Education Spending in the General Social Survey (2018)

A clear majority (median and mode) of Americans think we are spending too little on improving the American education system in 2018.



Data: General Social Survey (2018).
Prompt: "We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount."

Mean

The arithmetic **mean** is used only for interval variables.

- This is to what we refer when we say “average”.

Formally, *i* through *n*:

$$\frac{1}{n}\Sigma x_i \tag{2}$$

We can always describe interval variables with mode and median.

- We cannot do the same for ordinal or nominal with the mean.

Mean Real GDP, by Year

From our `pwt_sample` data frame from last week.

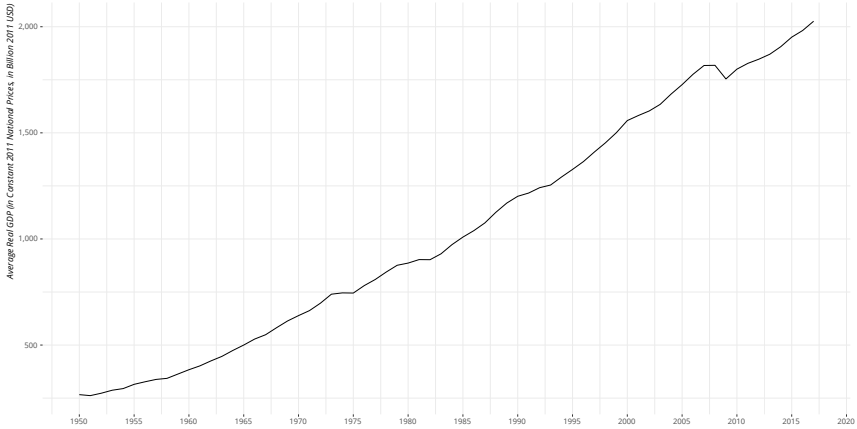
```
pwt_sample %>%
  group_by(year) %>%
  summarize(meanrgdpna = mean(rgdpna, na.rm=T)) %>%
    head(5)
```

```
## # A tibble: 5 x 2
##    year meanrgdpna
##   <dbl>      <dbl>
## 1  1950    266360.
## 2  1951    261707.
## 3  1952    273036.
## 4  1953    287153.
## 5  1954    294913.
```

You can also graph this too.

Average Real GDP for 21 Rich Countries, 1950–2017

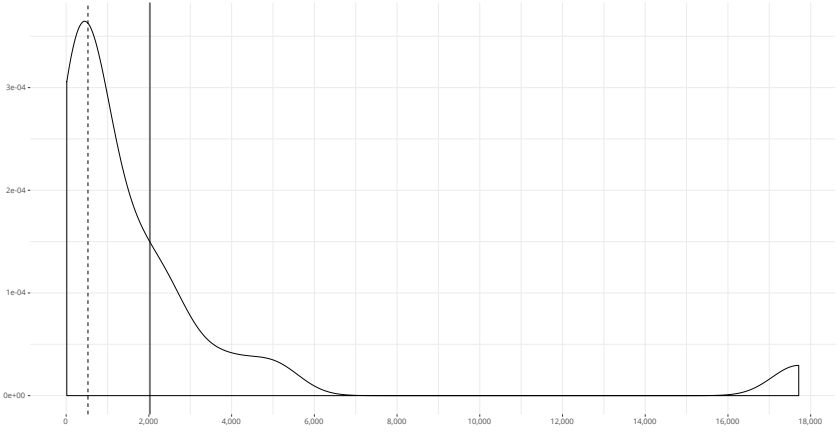
The average real GDP in 2017 was over 2 trillion dollars, which should seem super sketchy.



Data: Penn World Table (v. 9.1.)

A Density Plot of Real GDP (in Billions 2011 USD) in 2017 for 21 Rich Countries

Suddenly, 'average' doesn't look so 'average', certainly because of a high-leverage case like the U.S.



Data: Penn World Table (v. 9.1). Median noted in the dashed line while the mean is noted in the solid line.

Standard Deviation

The standard deviation is a measure of how “average” is “average.”

1. Subtract μ from every value in the population.
2. Square that deviation for every observation.
    - If you didn’t, the sum of deviations would equal zero.
3. Add those squared deviations together.
    - This is the sum of squared deviations.
4. Calculate arithmetic mean for sum of squared deviations.
    - This is an important statistic called the **variance**.
5. Take the square root of the variance.

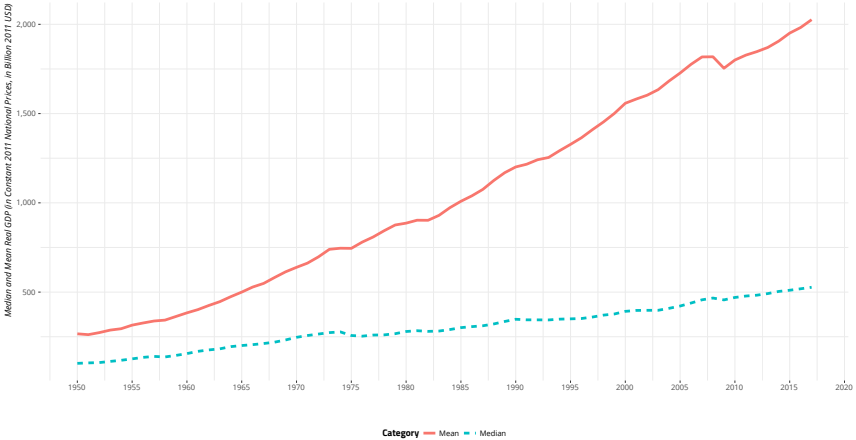
Or Just Use R...

```
pwt_sample %>%
  group_by(year) %>%
  mutate(rgdpb = rgdpna/1000) %>%
  summarize(meanrgdpb = mean(rgdpb, na.rm = T),
            sd = sd(rgdpb,na.rm=T)) %>%
    head(5)
```

```
## # A tibble: 5 x 3
##    year meanrgdpb    sd
##   <dbl>     <dbl> <dbl>
## 1  1950      266.  502.
## 2  1951      262.  520.
## 3  1952      273.  542.
## 4  1953      287.  568.
## 5  1954      295.  566.
```

Median and Mean Real GDP for 21 Rich Countries, 1950-2017

A mean that further separates from the mean over the course of the data suggests a worsening skew problem (here: the U.S.).



Data: Penn World Table (v. 9.1.)

Some Wrinkles

- Dummy variables
- Ordinal/interval

Dummy Variables

A variable with just two values is called a **dummy variable**.

- Some type of phenomenon is either present or absent.
- Typically coded as 1 or 0, respectively.

Gender is the most common and intuitive dummy variables.

- We typically code women as 1, men as 0.

We don't try to explain variations in gender (seriously, don't), but gender may explain phenomena of interest.

- e.g. support for parental leave policies in Europe, support for contraceptive coverage in the U.S.

Dummy Variables

The wrinkles:

- Some dummies may imply order, but they are technically nominal (again: “there” or “not there”).
- You can use mode/median/mean on any dummy.
  - The mode and median will be the same.
  - The mean will be the proportion of 1s.

```r
gss_spending %>%
  # where sex = 1 for women, 0 for men
  summarize(mode = getmode(sex),
            median = median(sex),
            mean = mean(sex))
```

```
## # A tibble: 1 x 3
##    mode median  mean
##   <dbl>  <dbl> <dbl>
## 1     1      1 0.552
```

Is It Ordinal or Interval?

The difference between ordinal and interval is mostly intuitive, but there is a gray area sometimes.

- Do we know if a guy who earns $50,001 is exactly one dollar richer than a guy who makes $50k even?
  - We may have an issue of cents.
- Is the person who is 21 exactly one year older than a 20-year-old?
  - We may have an issue of days and months.

How would you know when it's ordinal or interval?

A Rule of Thumb

We love to treat technically ordinal variables as interval when we can.

- Especially true for age and income.

We asks ourselves two questions.

1. How many different values are there?
2. How are the data distributed?

A Rule of Thumb

- If it has seven or more different values, you can *start* to think of it as interval.

However, check to see how the data are distributed.

- Is it bimodal? Is there a noticeable skew?
- If so, *don't* treat it as interval.

```
gss_spending %>%
  summarize(mean = mean(age, na.rm=T),
            median = median(age, na.rm=T),
            distinct = n_distinct(age))
```

```
## # A tibble: 1 x 3
##    mean median distinct
##   <dbl>  <dbl>    <int>
## 1  49.0     48       73
```

Age has tyipcally over 70 different categories in survey data.

- Differences are granular
- Decimals would make some sense.

```
gss_spending %>%
  summarize(mean = mean(sumnat, na.rm=T),
            median = median(sumnat, na.rm=T),
            distinct = n_distinct(sumnat))
```

```
## # A tibble: 1 x 3
##    mean median distinct
##   <dbl>  <dbl>    <int>
## 1  6.37      7       33
```

Same basically holds for my attitudes toward government spending variable.

- Differences are granular.
- Decimals may make some sense.
    - esp. if you center it on zero.

A Rule of Thumb

Some variables are just too damn ugly/information-poor.

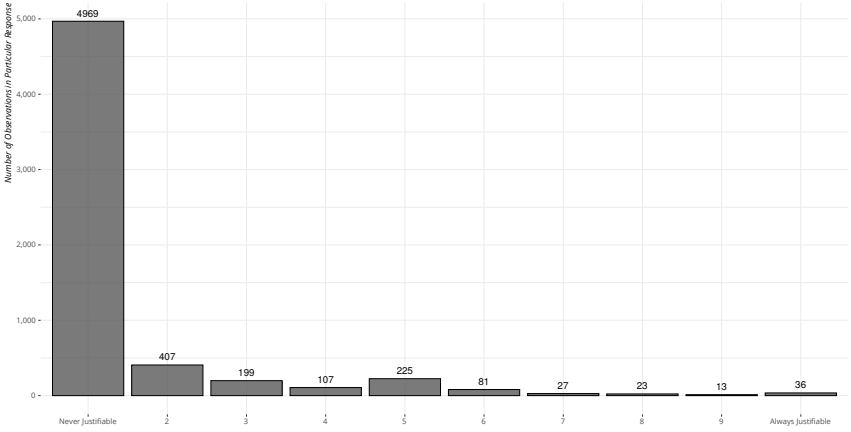
- Consider this prompt from the WVS about the justifiability of accepting a bribe on a 1-10 scale.

```
usa_justifbribe %>%
  na.omit %>%
  summarize(mean = mean(justifbribe, na.rm=T),
            median = median(justifbribe, na.rm=T),
            distinct = n_distinct(justifbribe))
```

```
## # A tibble: 1 x 3
##    mean median distinct
##   <dbl>  <dbl>    <int>
## 1  1.52      1       10
```

The Justifiability of Taking a Bribe in the U.S., 1995-2011

There are just 10 different responses in this variable with a huge right skew. I wouldn't ask for a mean of this.



Data: World Values Survey, 1995-2011

Recoding Variables

Think of variables as varying by levels of precision.

- You can always toss out information, but you can't add it back.
- Take care when you first devise your instrument.

```r
gss_spending %>%
  select(age) %>%
  na.omit %>%
  mutate(agegroups = cut(age, breaks = c(-Inf, 30, 40,
                                         50, 65, Inf),
                         labels=c("18-30","31-40","41-50",
                                  "51-65","66-89")),  # create five age
    # Create a 30 or younger dummy
    age30ory = ifelse(age <= 30, 1, 0)) -> agevars
```

```r
# Let's see what this did
agevars %>%
  na.omit %>%
  group_by(agegroups) %>%
  summarize(min = min(age),
            max = max(age))
```

```
## # A tibble: 5 x 3
##   agegroups   min   max
##   <fct>     <dbl> <dbl>
## 1 18-30        18    30
## 2 31-40        31    40
## 3 41-50        41    50
## 4 51-65        51    65
## 5 66-89        66    89
```

Table of Contents