Defining and Measuring Variables

POSC 3410 – Quantitative Methods in Political Science

Steven V. Miller

Department of Political Science

Goal for Today

Discuss the definition and measurement of variables.

Introduction

The previous lecture discussed that we start with a broader concept that interests us.

- e.g. “political tolerance”, “war”, or “state development”

However, proceeding with political *science* requires a tangible measure of the concept in question.

- Measurement really is the heart of science.

Once we have that measure, we have, in essence, a **variable**.

Variables

What is a **variable**?

- It is the empirical measurement of a characteristic.
- It's also a numeric array of data that has at least two separate values.

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.

Levels of Measurement

There are three levels of precision in a variable.

1. Nominal
2. Ordinal
3. Interval

Nominal Variables

A **nominal variable** has the lowest level of precision.

- This is also called a “categorical variable”.

The numeric values in these variables code differences *and nothing else*.

Nominal Variables

What does this mean? Take our gender example.

- i.e. women = 1 and men = 0.
- We need to substitute these numeric values for labels in order to do any statistical analysis.

Numerically, we know 1 > 0.

- That does not mean we are saying that women are “better” than men.

We are not saying that 1 > 0, but that 1 != (i.e. does not equal) 0.

- All binary variables are, by design, nominal variables.

Nominal Variables

There are other examples of nominal variables with plenty of different values. Examples:

- State of origin (e.g. Alabama, Alaska, Arizona...)
- Country of Origin (e.g. USA, Canada, Bahamas...)
- Race (e.g. white, black, etc...)
- Religion (e.g. Protestant, Catholic, Muslim, etc...)
- Party affiliation (e.g. Democrat, Republican, Independent, etc...)

Again, values in these variables simply code differences.

Ordinal Variables

Ordinal variables capture rank, or order, within the numeric values.

- They often (but do not always) look like Likert items.

Likert items make a statement and prompt a level of agreement with the statement.

- e.g. “People who sell cannabis should always be prosecuted”
  - Actual question from British Social Attitudes Survey
- Answers: Strongly agree, agree, neutral, disagree, strongly disagree.
- Corresponding values: 2, 1, 0, -1, -2.
  - Alternatively: 1, 2, 3, 4, 5. It’s just good to have a zero, though.

Ordinal Variables

Since the variable captures degree of agreement, we can say that 2>1 and 1>-2.

- People who respond “agree” are more in agreement with the statement than those who “strongly disagree”.
- However, this variable does not precisely say much.

An ordinal variable captures order and rank, but only captures *relative* difference.

What about Party ID?

Consider this wrinkle in how to measure party identification. From GSS:

1. Strong Democrat
2. Not strong Democrat
3. Independent, near Democrat
4. Independent
5. Independent, near Republican
6. Not strong Republican
7. Strong Republican
8. Other party

Is this ordinal or nominal?

What about Party ID?

"Other Party" makes this a nominal variable.

- Its inclusion automatically eliminates a natural semblance of order.

One "solution": treat it is as missing. Why?

- Statistically: "Other party" is usually no more than 3% of the data.
- Theoretically: most "independents" are closet partisans anyway.
- What remains: an ordinal measure of "partisanship" and not a nominal variable of party ID

Interval Variables

An **interval variable** captures *exact* differences.

- It's our most precise level of measurement.

Perhaps the most common interval measure we observe is age in years.

- i.e. someone who is 34 is 13 years older than someone who is 21.
- Notice the difference is no longer relative, but exact and precise.

Age is an easy way of thinking of interval variables, but we have others too.

- Political economy researchers have a glut of interval variables.
- e.g. gross national income, GDP per capita, kilowatt hours consumed per capita, consumer price index.

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.

- e.g. financial satisfaction on a 10-point scale.
- e.g. justifiability of bribe-taking on a 10-point scale.

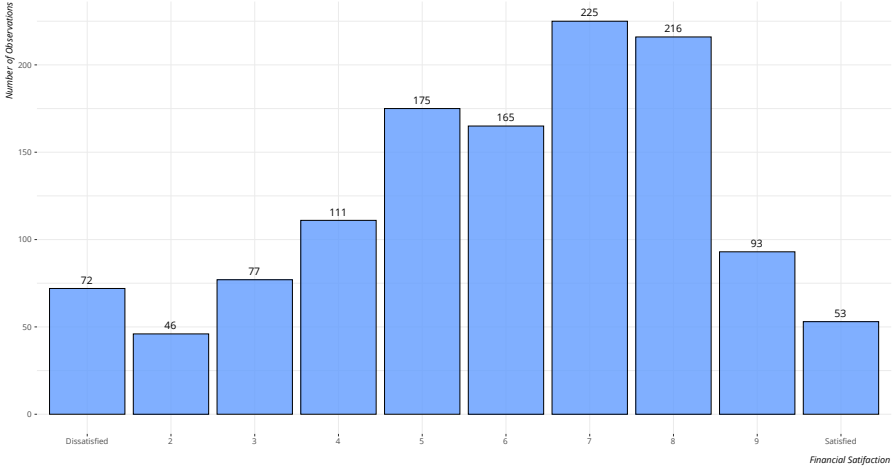
However, check to see how the data are distributed.

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

We'll be using two examples from the 2000 wave of World Values Survey.

The Distribution of Financial Satisfaction in the U.S. in 2006

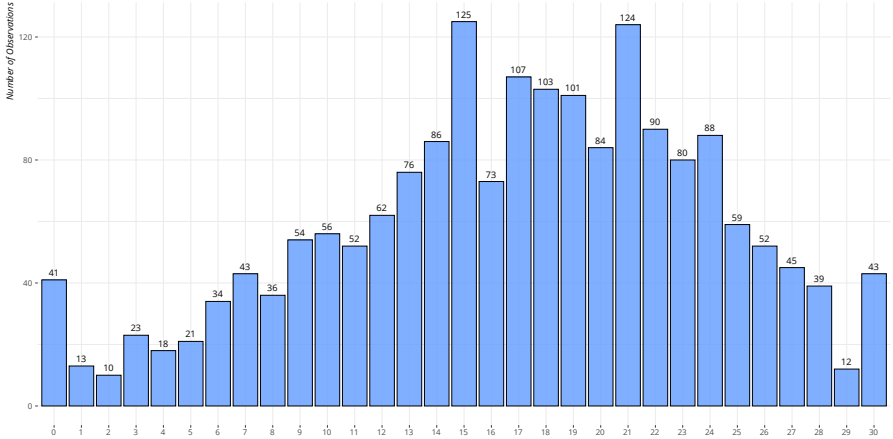
Data are limited to a 1-10 scale, but are sufficiently spaced out with no heaping. You could treat this as interval for convenience.



Data: World Values Survey (United States, 2006)

A Bar Chart of Pro-Immigration Sentiment in the United Kingdom from the ESS Data (Round 9)

There's a natural heaping of 0s and 30s but I've seen worse variables treated as interval for an OLS model or summarized by means.

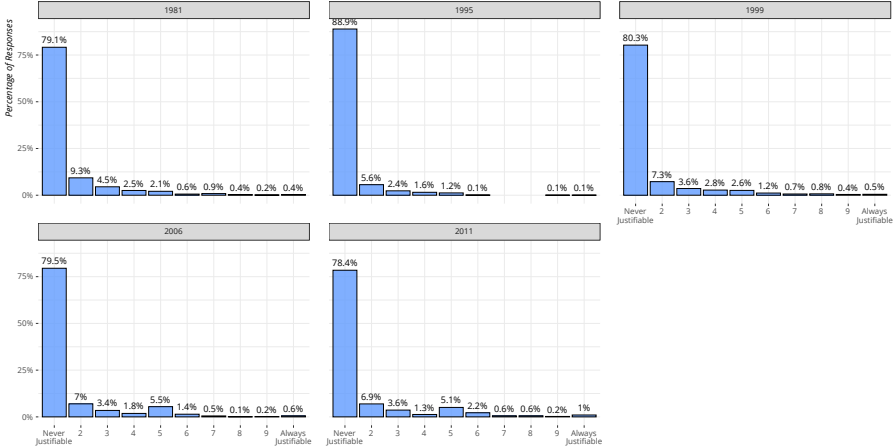


Value of the Pro-Immigration Sentiment Variable

Data: European Social Survey, Round 9 in the United Kingdom
Blog post: http://svmiller.com/blog/2020/03/what-explains-british-attitudes-toward-immigration-a-pedagogical-example/

The Justifiability of Taking a Bribe in the United States, 1981-2011

There is a clear right skew with a natural heaping at 0. *Don't* treat this as interval and don't ask for a mean of it.

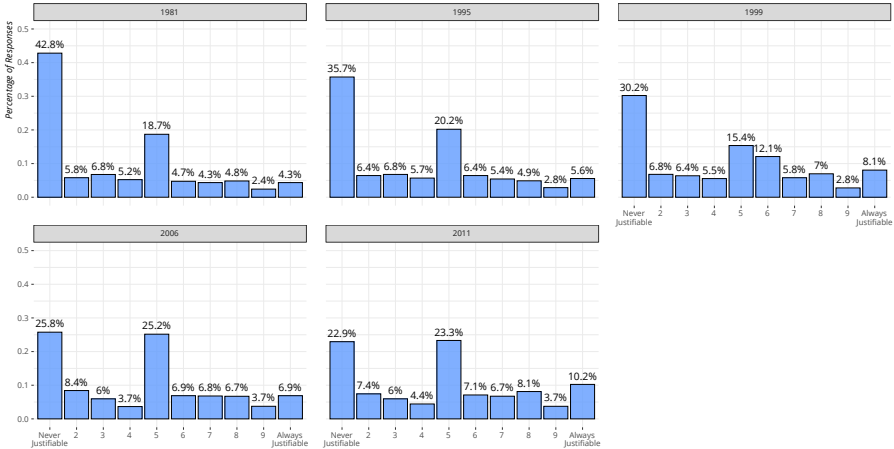


Justifiability of Taking a Bribe

Data: World Values Survey (United States, 1981-2011)

The Justifiability of an Abortion in the United States, 1981-2011

You're observing clear clumping/heaping in these data for which an "average" wouldn't look so average.



Justifiability of an Abortion

Data: World Values Survey (United States, 1981-2011)

Condensing Interval to Nominal

You can always condense a measure to lower levels of precision, but cannot add levels of precision. Take income, for example.

- **Interval**: income in dollars.
  - This will likely have a right skew, though.
- **Ordinal**: 0-$25k, $25k-$50k, $50k-$75k, $75k-$100k, $100k and above
- **Nominal**: low income earners (i.e. < $25k) and not low income earners.

An Example in World Values Survey Data, USA 2006

"In what group your household is, counting all wages, salaries, pensions and other incomes":

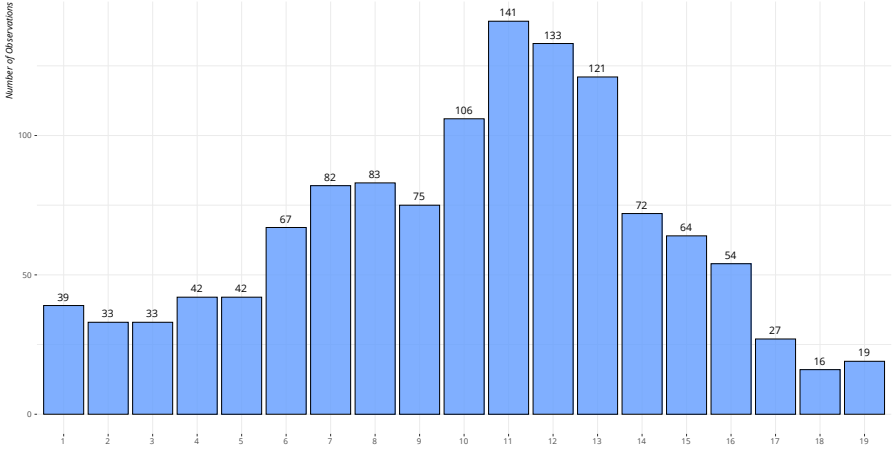
- 840041 US: Less than $5,000
- 840042 US: $5,000 to $7,499
- 840043 US: $7,500 to $9,999
- 840044 US: $10,000 to $12,499
- 840045 US: $12,500 to $14,999
- 840046 US: $15,000 to $19,999
- 840047 US: $20,000 to $24,999
- 840048 US: $25,000 to $29,999
- 840049 US: $30,000 to $34,999
- 840050 US: $35,000 to $39,999
- 840051 US: $40,000 to $49,999
- 840052 US: $50,000 to $59,999
- 840053 US: $60,000 to $74,999
- 840054 US: $75,000 to $84,999
- 840055 US: $85,000 to $99,999
- 840056 US: $100,000 to $124,999
- 840057 US: $125,000 to $149,999
- 840058 US: $150,000 to $174,999
- 840059 US: $175,000 or more

Example Code

```
USA %>% filter(s002 == 5) %>%
    select(s020, x047cs) %>%
    mutate(inccat = x047cs - 840040) %>%
  haven::zap_labels() -> USA2006

USA2006 %>%
  count(inccat) %>%
  ggplot(.,aes(as.factor(inccat), n)) +
  geom_bar(stat="identity", alpha=0.8, fill="#619cff", color="black")

# 5 income categories:
# 1 = $19,999 and below; 2 = between $20k and 39,999; 3 = between $40k and 74,999
# 4 = between $75k and 99,999; 5 = $100k and above
USA2006 %>% # note: carr comes in my stevemisc package
  mutate(cat5 = carr(inccat, "1:6=1; 7:10=2; 11:13=3;
                     14:15=4; 16:19=5")) %>%
  group_by(cat5) %>% tally() %>%
  ggplot(.,aes(cat5, n)) +
  geom_bar(stat="identity", alpha=0.8, fill="#619cff", color="black")

USA2006 %>% # just those making $100k or more
  mutate(highincome = ifelse(inccat >= 16, 1, 0)) %>%
  count(highincome) %>%
  ggplot(.,aes(as.factor(highincome), n)) +
  geom_bar(stat="identity", alpha=0.8, fill="#619cff", color="black")
```

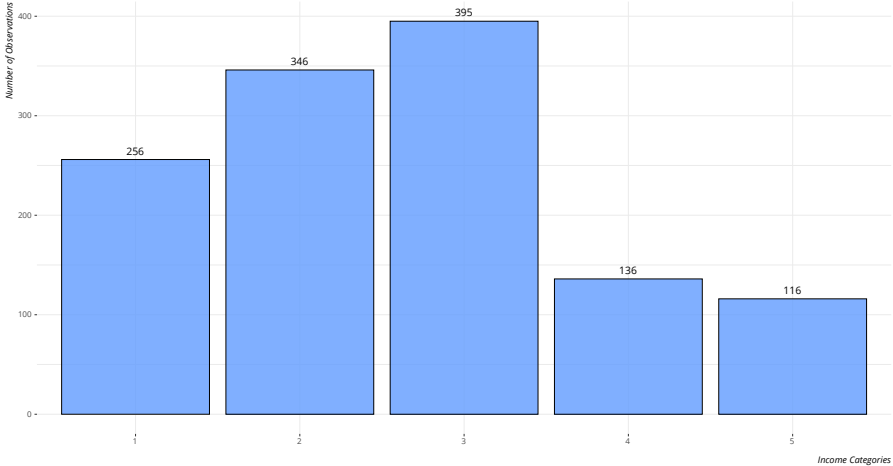
The Distribution of Self-Reported Household Incomes in the 2006 U.S. Wave of World Values Survey Data

The data are neatly distributed in 19 distinct income categories.



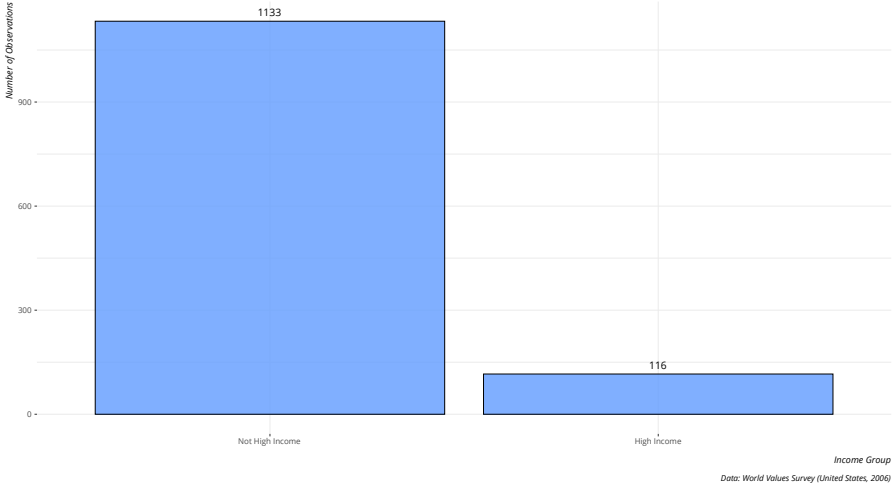
The Distribution of Self-Reported Household Incomes in the 2006 U.S. Wave of World Values Survey Data

The data here are collapsed to five categories, for which a mean would not be useful information.



The Distribution of Self-Reported Household Incomes in the 2006 U.S. Wave of World Values Survey Data

The data here are collapsed to two categories, which makes it a dummy variable.



Conclusion

This lecture focused on describing variables by their precision.

- Variables are nominal, ordinal, or interval.
- We have intuitive means to classify them.

Correctly classifying them is important.

- This will condition our choice of tools for descriptive and inferential statistics.

Table of Contents