Basic Descriptive Statistics

EH6127 – Quantitative Methods

Steven V. Miller

Department of Economic History and International Relations

Goal(s) for Today

1. Define basic levels of measurement, or how you should think about them.
2. Discuss basic descriptive statistics (i.e. central tendency and dispersion).
3. Do some preliminary bivariate analysis.

Levels of Measurement

The classic typology, a la Stanley Smith Stevens.

1. Nominal
2. Ordinal
3. Interval
4. Ratio

My Preferred Derivation of this Framework

1. Does it have just two values?
2. Would an arithmetic mean ("average") make sense if (when) it has decimal points?

Dummy Variables

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

- Some type of phenomenon is either present or absent.
- A statistical analysis will typically impose 0s and 1s on these variables, even if you see names/labels.

Gender is among the most common and intuitive dummy variables.

- We typically code women as 1, men as 0.
- Caveat: this might be changing the more we understand/unpack gender.

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.

Does an Arithmetic Mean Make Sense?

...especially if it had decimals? If no:

- Nominal (“unordered-categorical”)
- Ordinal (“ordered-categorical”)
- Both have a finite set of values that can occur.

If yes, you can call it “continuous” (or “interval”, or whatever).

- Counts, integers, percentages, proportions, ratio, real numbers, to name a few.
- All these are much more granular, have far more possible values.

Nominal Variables

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

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

Nominal Variables

What does this mean? Take gender, for 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:

- County of origin (e.g. Stockholm, Västerbotten, Norrbotten..)
- Country of Origin (e.g. USA, Canada, Bahamas...)
- Race (e.g. white, black, etc...)
- Religion (e.g. Protestant, Catholic, Muslim, etc...)
- Party vote choice (e.g. Vänsterpartiet, Socialdemokraterna, 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. "[I would be] ashamed if close family member gay or lesbian"
- Answers: Strongly agree, agree, neutral, disagree, strongly disagree.
- Corresponding values: 1, 2, 3, 4, 5

Since the variable captures degree of (dis)agreement, we can say that 2>1 and 5>2.

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

“Continuous” Variables

“Continuous” variables captures *exact* differences.

- It’s our most precise level of measurement.

Perhaps the most common continuous 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 continuous variables, but we have others too.

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

What's the Cutoff?

The difference between ordinal and continuous 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 continuous?

A Rule of Thumb

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

- Certainly true for age and income.

Ask yourselves two questions.

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

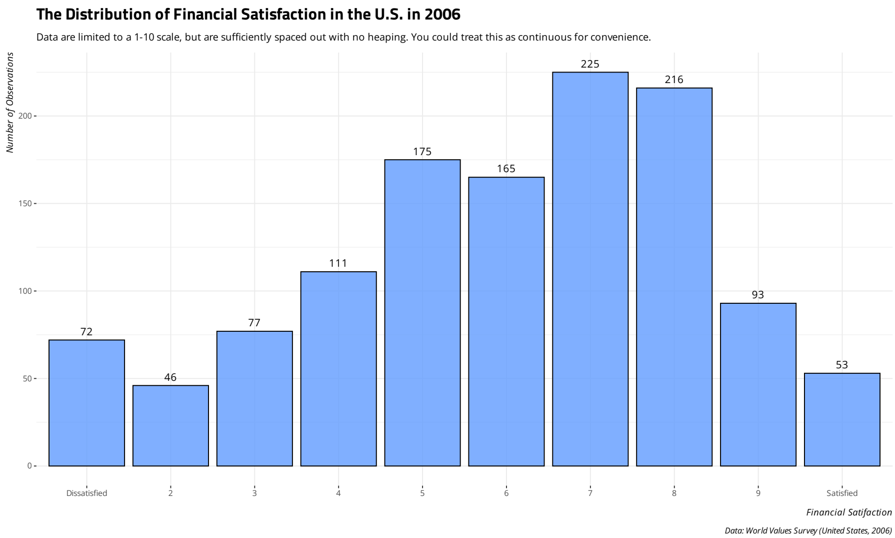
A Rule of Thumb

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

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

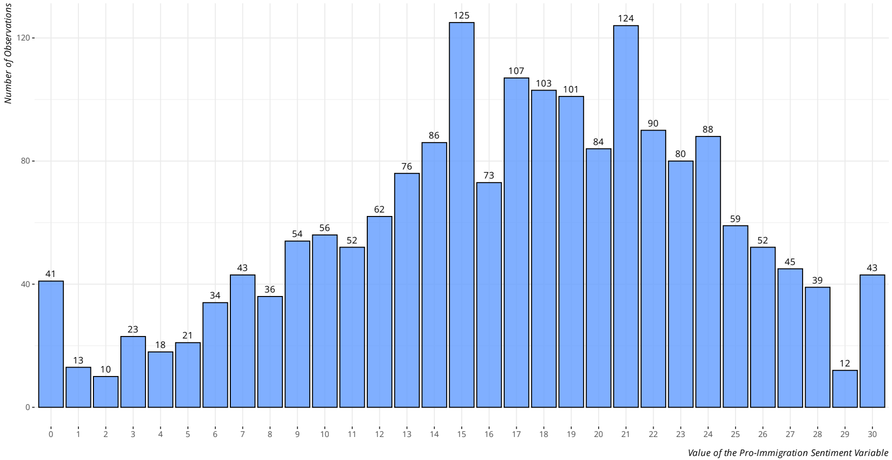
However, check to see how the data are distributed.

- Is it bimodal? Is there a noticeable skew?
- If so, *resist the urge* to treat it as contiunous.


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 continuous for convenience.
Number of Observations
0
50
100
150
200
72
46
77
111
175
165
225
216
93
53
Dissatisfied
2
3
4
5
6
7
8
9
Satisfied
Financial Satifaction
Data: World Values Survey (United States, 2006)

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

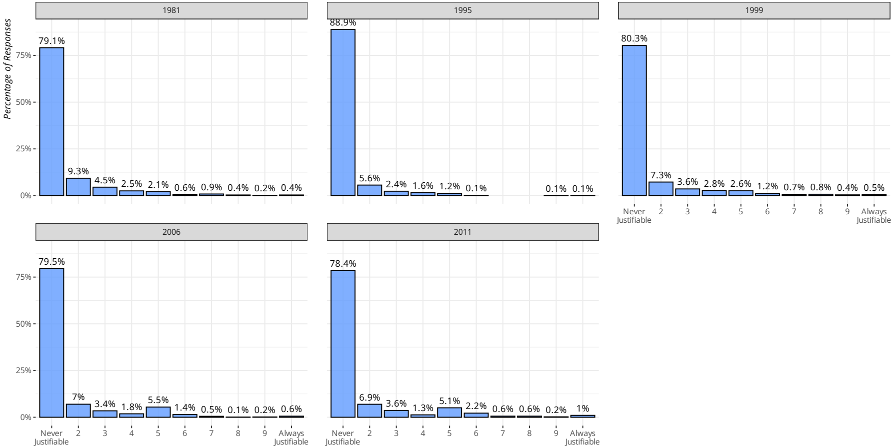
There's a natural heaping of 0s and 30s but I've seen worse variables treated as continuous for an OLS model or summarized by means.



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 continuous and don't ask for a mean of it.

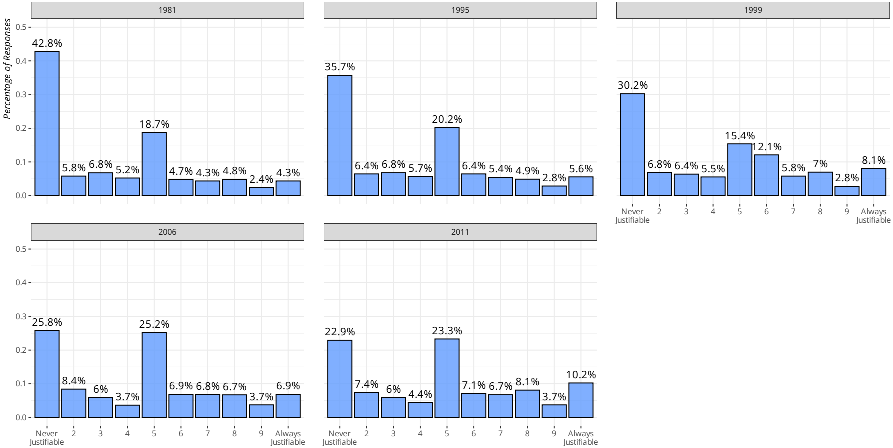


Justifiability of Taking a Bribe

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 Continuous to Nominal

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

- **Continuous**: 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.

Central Tendency

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

- Mode: most commonly occurring value
- Median: middlemost value
- Mean: arithmetic average

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.

These statistics may be dissatisfying because they don't tell you much.

- Then again, your data aren't telling you much.

Basic inferential takeaway: absent any other information, no other guess about an unordered-categorical variable would be as good, on average, as the mode.

Median

The **median** is the middlemost value.

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

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.
    - For an even number of observations, take the two that straddle the middle and get the midway point of those two.
- We can do this when our variables have some kind of "order".
- Medians of nominal variables are nonsensical.

Mean

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

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

Formally, *i* through *n*:

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

We can always describe continuous variables with the median.

- We cannot do the same for ordinal or nominal with the mean.
- For really granular data, there is likely no real proper “mode” to report.

A Comment on Dummy Variables

Dummy variables behave curiously in measures of central tendency.

- Mode: most frequently occurring value (as it is nominal).
- Median: also the mode.
- Mean: the proportion of 1s.

Dispersion

We also need to know variables by reference to its **dispersion**.

- i.e. "how average is 'average'?"
- How far do variables deviate from the typical value?
- If they do, measures of central tendency can be misleading.

In a lot of applications, you can just visualize this or look for a table.

- If you have continuous data, you can get a precise measure: the **standard deviation**.
  - i.e. the square root of the sum of squared deviations for each observation from the mean.
- For less precise data: just eye-ball it.
  - You could ask for an inter-quartile range, but, again, eye-ball it.

How to Calculate a Standard Deviation

Table 1: Calculating the Mean and Standard Deviation of Ten People's Age

age	mean	dvtn	sum_dvtn	dvtn2	sum_dvtn2	variance	sd
41	36.3	4.7	0	22.09	266.1	29.567	**5.438**
32	36.3	-4.3	0	18.49	266.1	29.567	**5.438**
31	36.3	-5.3	0	28.09	266.1	29.567	**5.438**
32	36.3	-4.3	0	18.49	266.1	29.567	**5.438**
34	36.3	-2.3	0	5.29	266.1	29.567	**5.438**
40	36.3	3.7	0	13.69	266.1	29.567	**5.438**
30	36.3	-6.3	0	39.69	266.1	29.567	**5.438**
35	36.3	-1.3	0	1.69	266.1	29.567	**5.438**
44	36.3	7.7	0	59.29	266.1	29.567	**5.438**
44	36.3	7.7	0	59.29	266.1	29.567	**5.438**

Alternatively:

```r
sd(x)
```

```
## [1] 5.437524
```

A Frequency Table

Table 2: A Frequency Table of the Region of Swedish Respondents (ESS, 2018/19)

Region	N	Percentage
Stockholms län	336	21.83%
Östra Mellansverige	225	14.62%
Småland med Öarna	133	8.64%
Sydsverige	218	14.17%
Västsverige	304	19.75%
Norra Mellansverige	142	9.23%
Mellersta Norrland	66	4.29%
Övre Norrland	115	7.47%

Note:

Data: European Social Survey v. 9 [edition: 3.1].

A Cumulative Percentage Table

Table 3: How Often Do Americans Say They Attend Religious Services?

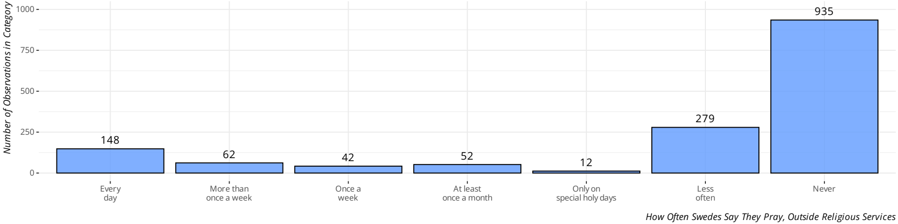
	N	Percentage	Cumulative Percentage
Never or Less Than Once a Year	853	36.58%	36.58%
Once a Year	300	12.86%	49.44%
Several Times a Year	239	10.25%	59.69%
Once a Month	146	6.26%	65.95%
2-3 Times a Month	186	7.98%	73.93%
Nearly Every Week	88	3.77%	77.7%
Every Week or More	520	22.3%	100%

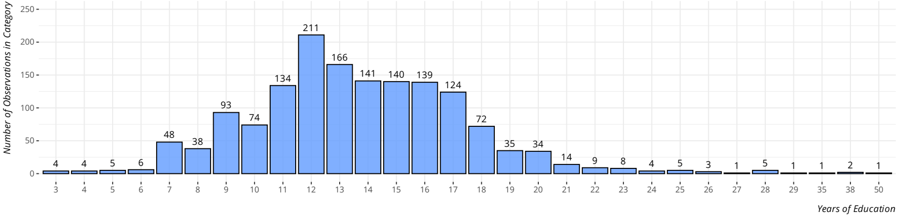
Note:

Data: General Social Survey (2018)

Nothing Beats Looking at Your Data for Rudimentary Diagnostics

Bar charts like these may point to unusual heaping patterns or extreme/anomalous observations.

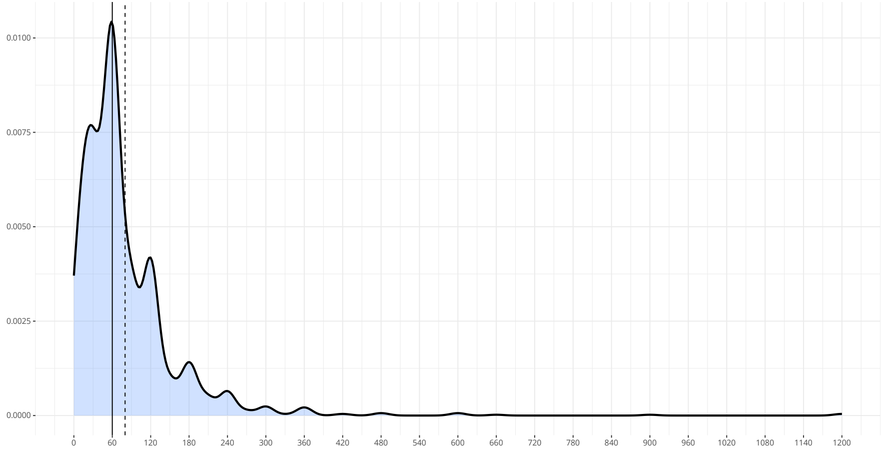




Data: Swedish respondents in the European Social Survey v. 9 [edition: 3.1].

A Density Plot of News Consumption About Politics in Sweden

There is a pretty obvious right skew for the politics addicts in these data.



Data: European Social Survey v. 9 [edition: 3.1]. Vertical line added for median (60) and mean (80) in these data.

Tools for Bivariate Analysis

We're building toward regression, but let's start simple.

- Correlation
- Scatterplots

Correlation

Correlation is a measure of how closely two things travel together.

- **Pearson's correlation coefficient** (or **Pearson's *r***) will tell us how strongly two things travel together.

Pearson's r

$$\frac{\sum\left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right)}{n-1}$$

...where:

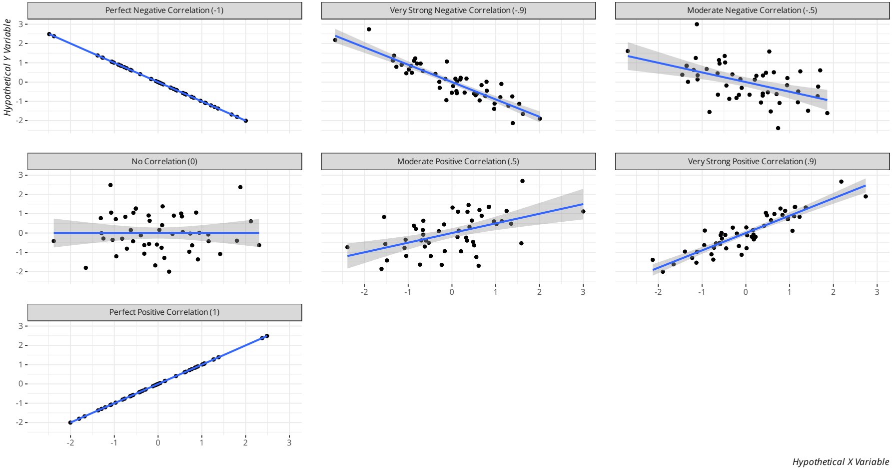
- x_i, y_i = individual observations of *x* or *y*, respectively.
- $\overline{x}$, $\overline{y}$ = means of *x* and *y*, respectively.
- s_x, s_y = standard deviations of *x* and *y*, respectively.
- *n* = number of observations in the sample.

Properties of Pearsons r

1. Pearson's *r* is symmetrical.
2. Pearson's *r* is bound between -1 and 1.
3. Pearson's *r* is standardized.

Various Linear Patterns You Could Deduce from a Scatterplot

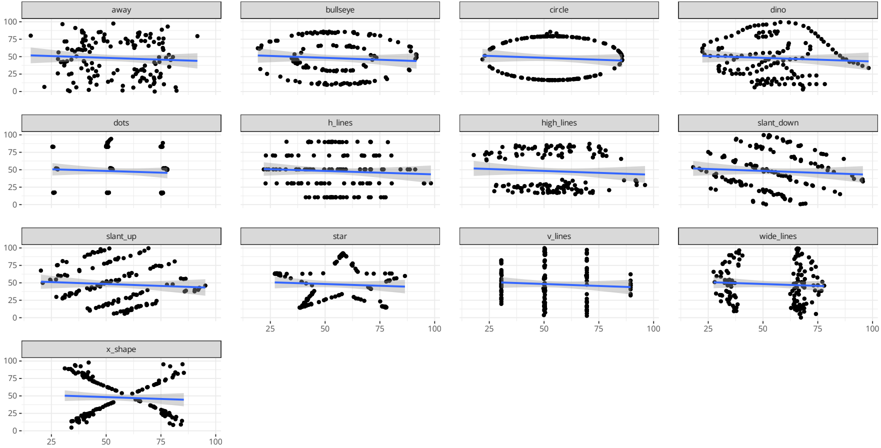
Do note: you can describe these correlations however you want. There is no formal metric, beyond direction, perfection, and zero.



Data: Simulated with smvrnorm() in {stevemisc} package.

Beware the Datasaurus!

No, seriously. Look at your damn data and never trust a summary statistic without looking at it.



Data: Cairo (2016) and Matejka and Fitzmaurice (2017). Note: all these data sets have the same means and standard deviations for x and y, along with the same correlation.

Conclusion

On levels of measurement:

- Dummy variables are a special class of nominal variables.
- You can think of ordinal as continuous if there are enough values and no weird clumping for finite responses.
- A "continuous" measure: iff (sic) a mean would make sense (whether or not it's the best measure of central tendency).

On central tendency and dispersion:

- *Look at your damn data.*
- "Average" might not look so "average."
  - There's a reason a lot of economic data are summarized in medians.
- *Look at your damn data.*
  - No seriously: never trust a summary statistic without first looking at it.

Table of Contents