Central Tendency and Dispersion

POSC 3410 – Quantitative Methods in Political Science

Steven V. Miller

Department of Political Science

Goal for Today

Describe variables by reference to central tendency and dispersion.

Defining and Measuring Variables

Last lecture focused on a typology of variables.

1. Nominal
2. Ordinal
3. Interval

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.

Suppose I have a random sample of 50 students and measured party affiliation.

- Democrats: 26; Republicans: 20; Others: 4

What's the modal category?

Mode

If I randomly grabbed a student from that sample and guessed “Democrat”, I would be right 26 times of 50 (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.

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.

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.

The interval variable with no dispersion problem is one in which the mode, median, and mean are the same value.

- This will not happen when there is a significant **skew**, or a **bimodal** distribution.

Frequency Distribution

A **frequency distribution** is a summary of a variable's values.

Table 1: Region of Residence (General Social Survey, 2018)

Region	**Frequency**	**Percentage**
Midwest	2815	20.38%
Northeast	1440	10.43%
South	6377	46.17%
West	3179	23.02%
Total	*13811*	*100%*

Cumulative Percentage

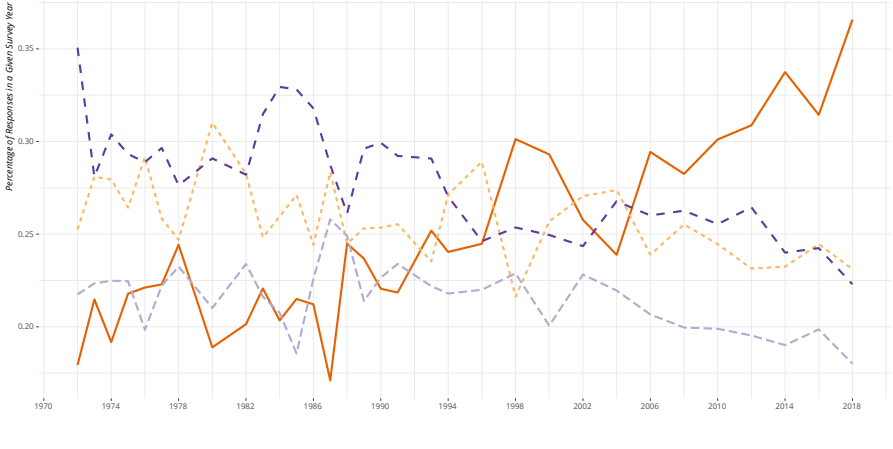
A **cumulative percentage** is the percentage of cases at or below a given value.

Table 2: Attendance at Religious Services (General Social Survey, 2018)

Attendance	Frequency	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%

The Variation of Self-Reported Church Attendance in the United States, 1972-2018

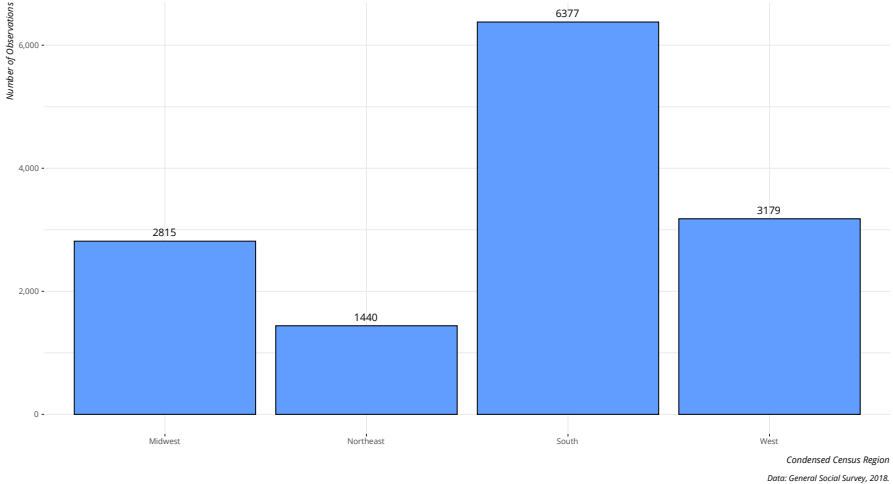
In these truncated categories, the "never" group went from being the smallest in 1972 to the clear largest in 2018.



Data: General Social Survey, 1972-2018

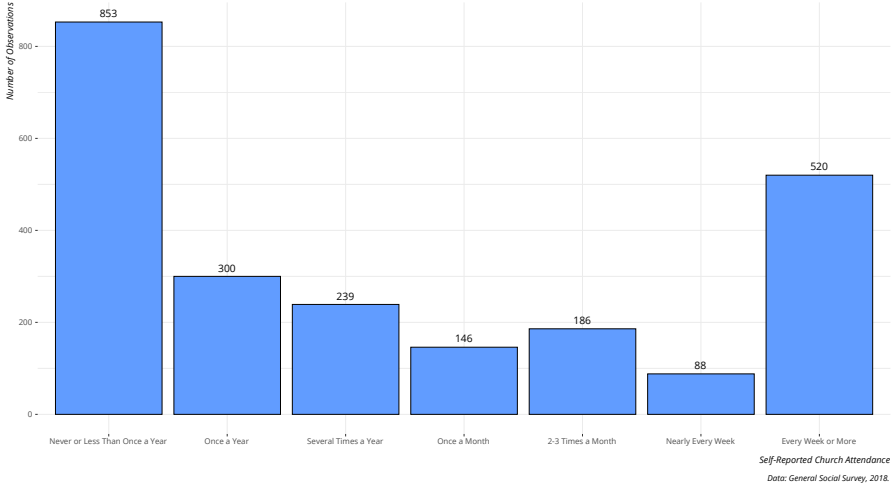
A Bar Chart of Region of Residence in the General Social Survey (2018)

Your mileage may vary, but I think there's more value in bar charts for basic descriptive stuff.



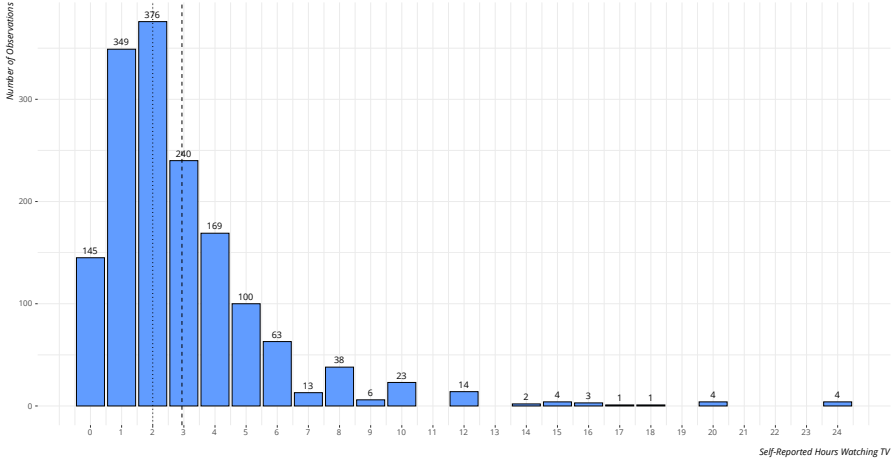
A Bar Chart of Self-Reported Church Attendance in the General Social Survey (2018)

A simple bar chart helps less with cumulative percentages, but it'll better point you in the direction of potential bimodality.



Self-Reported Hours Watching TV (General Social Survey, 2018)

A simple bar chart will also help visualize skew in an intuitive way.



Data: General Social Survey, 2018. No one is watching 24 hours of TV. You're not fooling anyone.
Mode and median (2) captured in dotted line. Mean in dashed line.

Conclusion

Here are some final thoughts.

- There is a reason we discuss “median income” and not the “average income”.
- The mean of a dummy variable communicates the percentage of 1s, divided by 100.
- Skew is mostly a problem of interval variables, and a problem of degree.

Always look carefully at your data!

Table of Contents