POLS3316: Statistics for Political Scientists

Lecture 2: Variables, Units of Observation, Variable Types, Measures of Central Tendency

Instructor: Tom Hanna, Fall 2025

2026-01-27

Table of contents

1 Descriptive Statistics

2 Why do we use descriptive statistics?

- Explore the data
- See patterns in the data

- Communicate about the data

3 What are the most basic things we need to know?

- What are the variables?
- What is the scope of the data (time, geography, cases)?
- What is the unit of observation?

4 What are variables?

- In math (algebra) - a symbol that represents a number
- In science - a characteristic or attribute that can vary across units of observation
- In statistics - a characteristic or attribute that can vary across units of observation
- In R coding - a column in a *data frame*
- In all of the above - something that can take on different values

5 What is a unit of observation?

- In science - the entity that is being measured
- In statistics - the entity that is being measured
- In R coding - a row in a *data frame*
- In all of the above - the thing that has the variable measured on/for it

6 Putting it together

- Variables are characteristics or attributes that vary across units of observation

7 Example

Titanic data

```r
# Load Titanic dataset

titanic_data <- as.data.frame(Titanic)

# merge the first three colnames to create a new rowname "Passenger_Type" that combines Clas

titanic_data <- titanic_data %>%
  unite("Passenger_Type", Class, Sex, Age, sep = "-")

# where survived is No, create a variable called number of deaths that is equal to Freq, othe

titanic_data <- titanic_data %>%
  mutate(number_of_deaths = ifelse(Survived == "No", Freq, 0))

# drop the Freq column and Survived column

titanic_data <- titanic_data %>%
  select(-Freq, -Survived)

# View the first 8 rows of the dataset
head(titanic_data, 8)
```

```
     Passenger_Type number_of_deaths
1    1st-Male-Child                0
2    2nd-Male-Child                0
3    3rd-Male-Child               35
4   Crew-Male-Child                0
5  1st-Female-Child                0
6  2nd-Female-Child                0
7  3rd-Female-Child               17
8 Crew-Female-Child                0
```

- units of observation
- variables: columns

8 Wide format

BEWARE! Not all data is formatted this way! Sometimes you have to think “is this a variable or a unit of observation?”

For example, the following data on Scandinavian temperatures:

```r
# From https://sejdemyr.github.io/r-tutorials/basics/wide-and-long/
# created by SIMON EJDEMYR

# Create long dataset
country_long <- data.frame(
    expand.grid(country = c("Sweden", "Denmark", "Norway"), year = 1994:1996),
    avgtemp = round(runif(9, 3, 12), 0)
    )

# Create wide dataset
country_wide <- data.frame(
    country = c("Sweden", "Denmark", "Norway"),
    avgtemp.1994 = country_long$avgtemp[1:3],
    avgtemp.1995 = country_long$avgtemp[4:6],
    avgtemp.1996 = country_long$avgtemp[7:9])
country_wide
```

```
  country avgtemp.1994 avgtemp.1995 avgtemp.1996
1  Sweden           10            9           11
2 Denmark            6            6            3
3  Norway            6            5            8
```

It looks like the unit of observation is country and the variable is a combination of year and temperature.

9 Long Format

If we look at it in long format, it's a little clearer:

```r
country_long
```

```
  country year avgtemp
1  Sweden 1994      10
2 Denmark 1994       6
3  Norway 1994       6
4  Sweden 1995       9
5 Denmark 1995       6
6  Norway 1995       5
7  Sweden 1996      11
```

```
8 Denmark 1996       3
9  Norway 1996       8
```

The variable is average temperature.

10 Country-year format

```
# change the country_long dataset to a wide dataframe with country-year as the row name and c
# combine country and year to make the row names
# avgtemp is still the variable
# 9 observations
# do not pivot wider

country_wide2 <- country_long %>%
  unite("country_year", country, year, sep = "-") %>%
  pivot_wider(names_from = country_year, values_from = avgtemp)

df_long <- country_wide2 |>
  pivot_longer(cols = everything(),
               names_to = "country_year",
               values_to = "avg_temp")

df_long <- as.data.frame(df_long)

## Create two variables - high_temp and low_temp by adding a random number between 1 and 5 to

set.seed(123) # for reproducibility
df_long <- df_long %>%
  mutate(high_temp = avg_temp + sample(1:5, n(), replace = TRUE),
         low_temp = avg_temp - sample(1:5, n(), replace = TRUE))

# sort by country_year

df_long <- df_long %>%
  arrange(country_year)

df_long
```

```
  country_year avg_temp high_temp low_temp
```

```
1 Denmark-1994        6         9        1
2 Denmark-1995        6         9        5
3 Denmark-1996        3         4        2
4  Norway-1994        6         8        3
5  Norway-1995        5        10        1
6  Norway-1996        8        10        3
7  Sweden-1994       10        13        7
8  Sweden-1995        9        11        6
9  Sweden-1996       11        15       10
```

But the unit of observation is no longer country - it is country-year.

Beware again: This is *wide format* in statistical terms. This is *long format* in R *tidyverse* nomenclature.

11 Measures of Central Tendency

Measures of central tendency help us:

- reveal patterns
- find the typical measurement
- find the center

12 Measures of Central Tendency

A few numbers that can summarize the center of measurement

- Mean
- Median
- Mode

13 Mean

- Symbol: $\bar{x}$
- Not the middle value
- Not the most common
- The center of mass - the sum above equals the sum below
- Formula is $\bar{x} = \frac{\sum X_i}{n}$

- Read that: The mean of X equals the sum of the observations (i) of X divided by the number (n) of observations.

14 Example A:

A. What is the mean of 1,5,7,9,10,12,18

15 Example A:

A. What is the mean of 1,5,7,9,10,12,18

```r
A <- c(1,5,7,9,10,12,18)

answer_a1 <- (sum(A)/7)

answer_a2 <- mean(A)

# label this as "produced by sum function" and "produced by mean function"

paste("Mean produced by sum function:", answer_a1)
```

```
[1] "Mean produced by sum function: 8.85714285714286"
```

```r
paste("Mean produced by mean function:", answer_a2)
```

```
[1] "Mean produced by mean function: 8.85714285714286"
```

16 Example A:

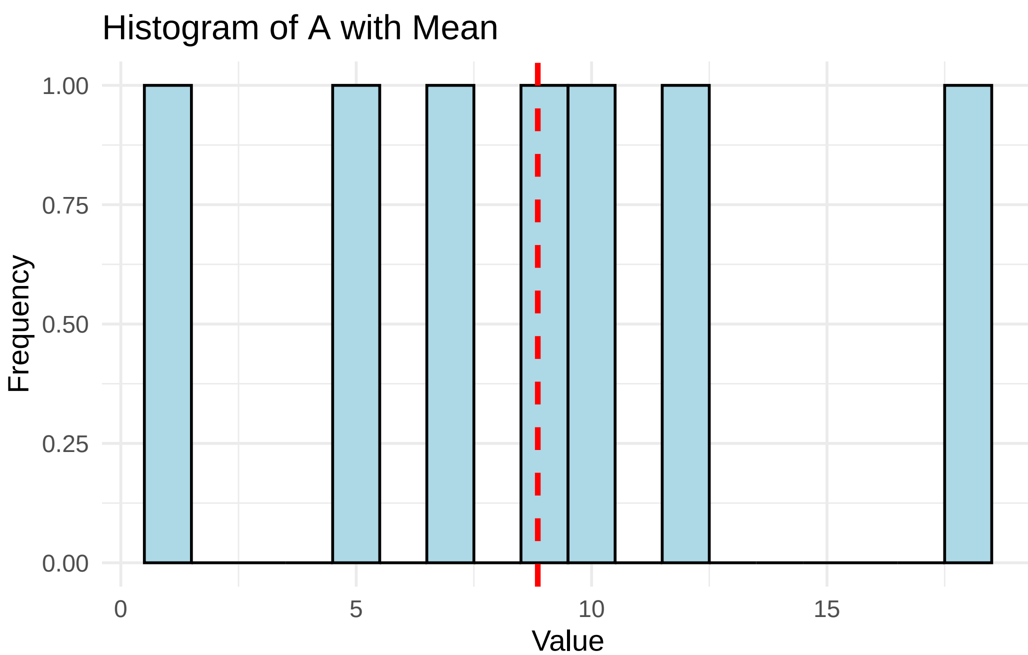
A. What is the mean of 1,5,7,9,10,12,18

Center of mass

```r
# demonstrate how the mean is the center of mass with ggplot

library(ggplot2)
df_a <- data.frame(x = A)
ggplot(df_a, aes(x = x)) +
  geom_histogram(binwidth = 1, fill = "lightblue", color = "black") +
  geom_vline(aes(xintercept = mean(x)), color = "red", linetype = "dashed", size = 1) +
  labs(title = "Histogram of A with Mean", x = "Value", y = "Frequency") +
  theme_minimal()
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```



17 Example B

B. What is the mean of 10,20,25,30,35,40,45,50,75

18 Example B

B. What is the mean of 10,20,25,30,35,40,45,50,75

```r
B <- c(10,20,30,35,40,45,50,75,25)

mean_b1 <- (sum(B)/9)

mean_b2 <- mean(B)

mean_b1
```

```
[1] 36.66667
```

```r
mean_b2
```

```
[1] 36.66667
```

19 Median

- Midpoint
- Half observations are greater, half are lower
- Just count
- Even observations - midpoint between middle two

20 Example A

A - 1,5,7,9,10,12,18

21 Example A:

A. What is the mean of 1,5,7,9,10,12,18

```r
# label this as "median of A"

paste("Median of A:", median(A))
```

```
[1] "Median of A: 9"
```

22 Example B

B - 10,20,25,30,35,40,45,50,75

23 Example B

B - 10,20,25,30,35,40,45,50,75

```
# label this as "median of B"

paste("Median of B:", median(B))
```

```
[1] "Median of B: 35"
```

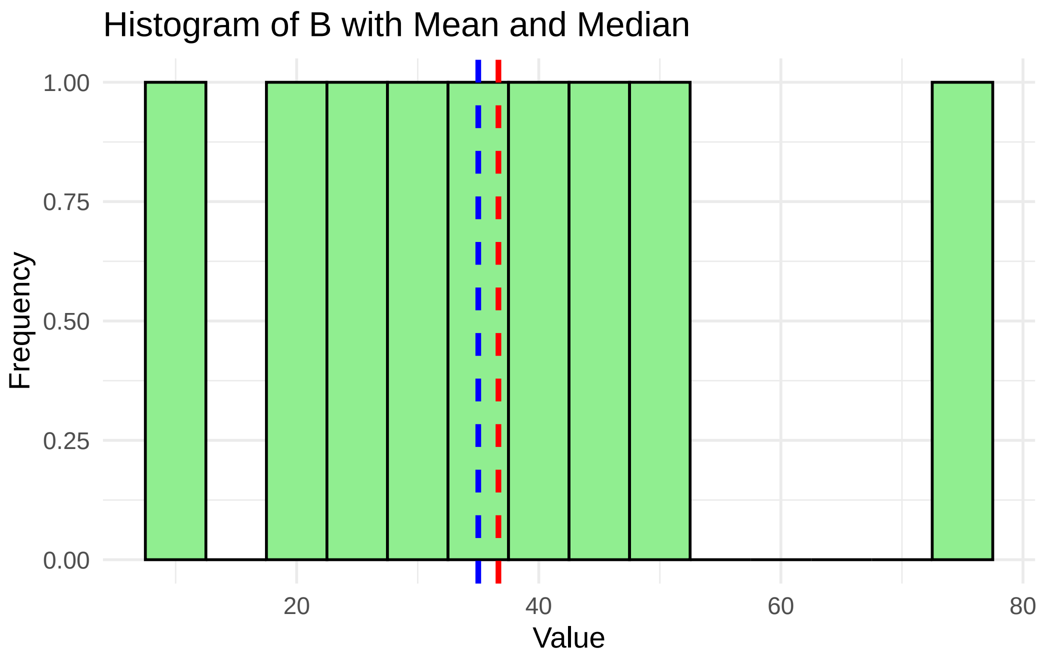
24 Example B

B - 10,20,25,30,35,40,45,50,75

Midpoint and Center of Mass

```
# demonstrate how the midpoint and center of mass are different with ggplot

library(ggplot2)
df_b <- data.frame(x = B)
ggplot(df_b, aes(x = x)) +
  geom_histogram(binwidth = 5, fill = "lightgreen", color = "black") +
  geom_vline(aes(xintercept = mean(x)), color = "red", linetype = "dashed", size = 1) +
  geom_vline(aes(xintercept = median(x)), color = "blue", linetype = "dashed", size = 1) +
  labs(title = "Histogram of B with Mean and Median", x = "Value", y = "Frequency") +
  theme_minimal()
```



25 Keep in mind for later

In both of our examples, the mean and median were close but not the same. That isn't always the case.

26 Mode

- Most common value
- Just count

27 Examples:

C. 1,2,3,4,4,5,6,7

Answer:

D. 10,20,30,30,40,40,40,50,50,60,70

Answer:

28 Advantages and disadvantages

- Median isn't affected by outliers
- Mean gives the broader picture because it includes the outliers.
- Mode is the only option for *categorical variables*.

29 Variable Types

- Categorical (nominal, ordinal)
- Numerical (interval, ratio)

30 Variable Type Examples: Categorical

- Nominal (Order is meaningless)

  ```
  - Gender
  - Race
  - Religion
  - Democrat vs Republican (also binary)
  ```

- Ordinal (ORDer matters)

  ```
  - Education level
  - Income brackets
  - Likert scale responses
  ```

- Binary

  ```
  - Yes/No
  - 0/1
  - True/False
  ```

31 Variable Type Examples: Numerical

- Interval (equal intervals, no true zero) *

```
- Temperature (Celsius, Fahrenheit)
- IQ scores
- Calendar years
```

- Ratio (equal intervals, true zero) *

```
- Height
- Weight
- Age
- Income
- Kelvin temperature
```

- Discrete (countable values)

```
- Number of children
- Number of countries in a trade agreement
- Battle deaths in a conflict
```

32 Social Science examples

- V-Dem v2x_libdem → ratio (0-1, true zero = no polyarchy) *
- Polity2 score (-10 to 10) → interval (0 not "no democracy") *
- Country GDP → ratio *
- Year → interval
- Battle deaths (COW) → ratio/discrete *
- is_autocracy (0/1) → binary/categorical
- regime_type (dem/au/other) → nominal categorical
- freedom_level (low/med/high) → ordinal categorical

33 Skewed distribution - when mean and median are different

The three numbers are often different for the same sample or population.

Example:

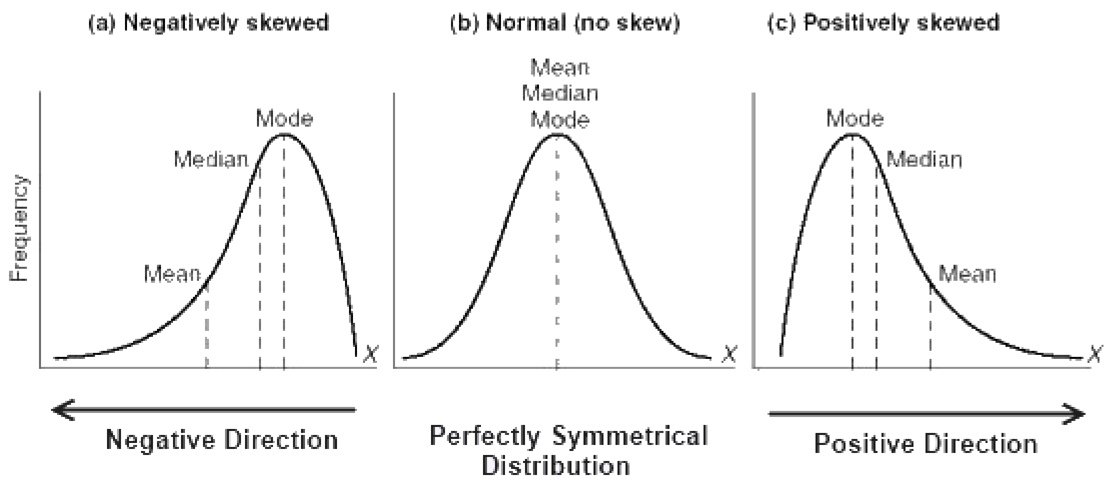


Figure 1: Negatively skewed, Normal, and Positively Skewed distributions

34 Authorship and License

- Author: Tom Hanna
- Website: tomhanna.me

skewed distribution source: Statistics by Jim