

Chapter 2: Exploring Data

Yu Yang

School of Statistics
University of Minnesota

September 9, 2022

Types of Data

A *variable* is any characteristic of a subject in a population.

ex: height, IQ, income, # of hot dogs eaten last year, gender, eye color

1. Categorical (Qualitative) Variable:

Classifies subjects as belonging to a certain group/category.

ex: gender, eye color, car make, race, major, area code

2. Quantitative Variable:

Takes on numerical values that represent different magnitudes.

- 2.1 Discrete : The possible values of a discrete quantitative variable form a set of separate numbers (i.e. can be listed).

ex: # of hot dogs eaten, # of t.v.'s, # of accidents/day

- 2.2 Continuous : The possible values of a continuous quantitative variable form an interval. That is, there is an infinite continuum of possible values.

ex: height, blood pressure, amount of rainfall

Numerical Summaries of Categorical Data

Frequency Table

A *frequency table* is a listing of possible values for a variable, together with the number of observations for each value. (Note that we can also construct frequency tables for quantitative variables.)

Proportion

A *proportion* of observations that fall in a certain category is the count of observations in that category divided by the total number of observations. (NOTE: $\text{percent} = 100 \times \text{proportion}$)

Frequency Table

| Social Media | Frequency | Proportion | Percent |
|--------------|-----------|------------|---------|
| Facebook | 18 | .050 | 5% |
| Instagram | 172 | .480 | 48% |
| Twitter | 45 | .126 | 12.6% |
| YouTube | 82 | .229 | 22.9% |
| others | 41 | .115 | 11.5% |
| Total | 358 | 1 | 100% |

Figure 1: Example Frequency Table

Numerical Summaries of Quantitative Data

Notations

1. n = the number of observations in a sample
2. x_i = the i th observation of a sample (so the list of observations is x_1, x_2, \dots, x_n)
3. \sum = summation

$$\sum x_i = x_1 + x_2 + \dots + x_n$$

Measures of Center

1. **mean** (\bar{x}) = the average of all observations

$$\bar{x} = \frac{\sum x_i}{n}$$

2. **median** (M) = the middle number when measurements are ordered from smallest to largest

When n is odd, $M =$ **the middle value.**

When n is even, $M =$ **the average of the middle two values.**

3. mean vs. median

Measures of Spread I

Range

The *range* is the difference between the largest and smallest observations.
That is,

$$\text{range} = \text{maximum} - \text{minimum}$$

Measures of Spread II

Percentile

The p th *percentile* of a distribution is the value below which $p\%$ of the observations fall.

Interquartile Range (IQR)

The *interquartile range* is the difference between the first and third quartiles. That is,

$$\text{IQR} = Q3 - Q1$$

1. **First Quartile (Q1)** = 25th percentile
The lowest 25% of the data lies below Q1.
2. **Second Quartile (Q2)** = 50th percentile = median
50% of the data are below and 50% are above M
3. **Third Quartile (Q3)** = 75th percentile
The highest 25% of the data lies above Q3.

Measures of Spread III

Sample Variance (s^2)

The *sample variance* of a set of observations is the "average" of the squared deviations from the mean.

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}$$

Sample Standard Deviation (s)

The *standard deviation* is the square root of the sample variance:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$

Measure of Spread III (cont.)

Properties of s

1. Interpretation: distance that a “typical” observation falls from the mean
2. s is measured in the same units as the original observations
3. Use s in conjunction with mean
4. $s \geq 0$
5. The larger s is, the greater the spread of the data.
6. $s = 0 \Rightarrow$ there is no variation in the data.
7. s is also not robust to outliers and skewness.

Measure of Spread III (cont.)

Interpreting the Magnitude of s

Unless the data set is extremely skewed or has extreme outliers, nearly all of the observations will fall within $3s$ of the mean \bar{x} .

$$[\bar{x} - 3s, \bar{x} + 3s] \quad (1)$$

5-number Summary

The *5-number summary* is a brief numerical description of the center *and* spread of a distribution:

minimum Q1 *M* Q3 maximum

Choosing the Proper Numerical Summaries

- mean vs. median
- range vs. IQR vs. s

Graphical Summaries of Categorical Data

Pie Chart

A circle is drawn with a “slice of pie” representing each category’s % of observations.

Bar Graph

A bar is drawn for each category with the bar’s height representing the % or count of observations.

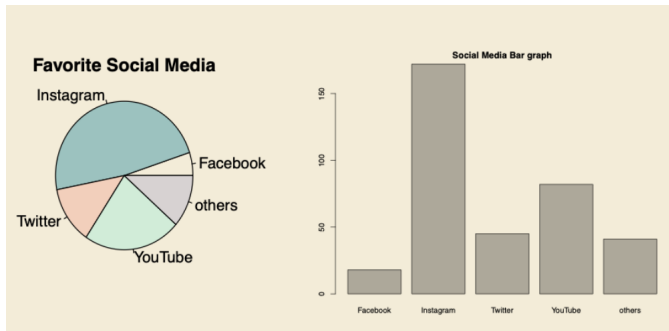


Figure 2: Pie Chart and Barplot

Graphical Summaries of Categorical Data (cont.)

Pie Charts vs. Bar Graphs

1. Pie charts emphasize a category's relation to the whole, but make it difficult to compare categories to each other.
2. Bar graphs compare the sizes of each group of a categorical variable (not in relation to the whole).
3. Bar graphs are easier to read and more flexible than pie charts.

Graphical Summaries of Quantitative Data

Major Focuses

1. overall shape
2. center and spread
3. outliers

Alternatives

1. Stem-and-Leaf Plot
2. Histogram
3. Boxplot

To Be Continued