Week 2: Data Summary and Visualization

4. Numerical variables II

Stat 140 - 04

Mount Holyoke College

Dr. Shan Shan

Slides posted at *http://sshanshans.github.io/stat140*

Outline

1. Today: Summary statistics

2. Main ideas
    1. Measures of Central Tendency
    2. Measures of Dispersion

3. Summary

Summary statistics

Given a visualization of the distribution of a numerical variable, one needs to find ways to summarize the information that facilitate understanding and insight. Two standard tools for this are:

1. Measures of Central Tendency (mean, median)
2. Measures of Dispersion (standard deviation, range, IQR)

Outline

1. Today: Summary statistics

2. Main ideas
    1. Measures of Central Tendency
    2. Measures of Dispersion

3. Summary

Outline

1. Today: Summary statistics

2. Main ideas
    1. Measures of Central Tendency
    2. Measures of Dispersion

3. Summary

Suppose we observe n numbers, $x_1, \ldots, x_n$.

There are two commonly used statistics used to summarize the **center** of the distribution of these values:

- The **mean** is the average of these values (add them up and divide by n). We use $\bar{x}$ to denote the mean:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + \cdots + x_n}{n}$$

- The **median** is the middle value when you arrange them in order. (If the sample size n is even, you take the average of the middle two values)

Example

Suppose one observes the following data:

$$1, 0, 2, -2, 1, -2, 5, -1$$

The mean is $\bar{X} = \frac{1}{8}(1 + 0 + 2 - 2 + 1 - 2 + 5 - 1) = 0.5$.

Let's order the data,

$$-2, -2, -1, 0, 1, 1, 2, 5$$

The median is the average of 0 and 1, or 0.5.

Example

Poll question

How do the mean and median of the following two datasets compare?
Dataset 1: 30, 50, 70, 90
Dataset 2: 30, 50, 70, 1000

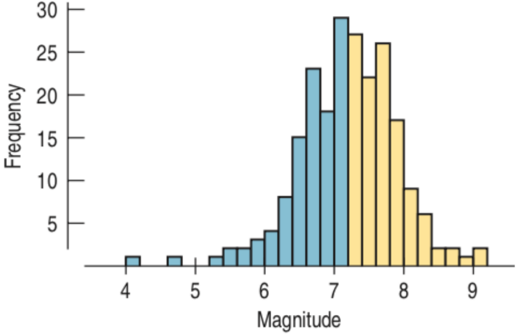
- a $\bar{x}_1 = \bar{x}_2$, $\text{median}_1 = \text{median}_2$
- b $\bar{x}_1 < \bar{x}_2$, $\text{median}_1 = \text{median}_2$
- c $\bar{x}_1 < \bar{x}_2$, $\text{median}_1 < \text{median}_2$
- d $\bar{x}_1 > \bar{x}_2$, $\text{median}_1 < \text{median}_2$
- e $\bar{x}_1 > \bar{x}_2$, $\text{median}_1 = \text{median}_2$

Median is more robust

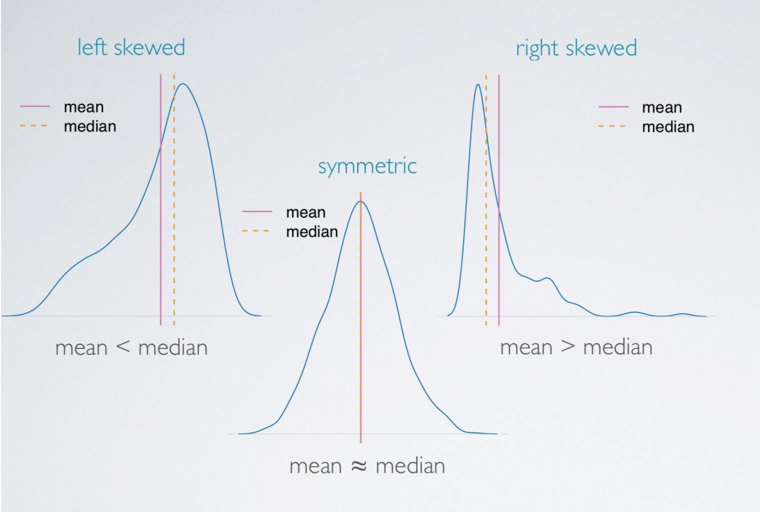
The mean can be pulled in misleading directions if there are outliers. A single large or small datum will have a large influence on the mean, but not on the median.

An outlier is an incorrect or unrepresentative observation that is very different from the others in the sample.

Median is the center of a histogram. Half of the data are less than the median and half are greater than the median.



The mean can be pulled towards the tail if the distribution is skewed.



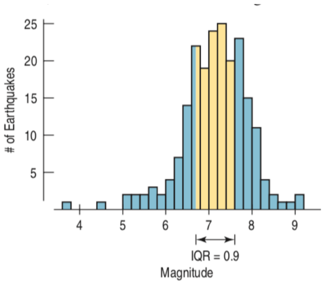
1. Today: Summary statistics

2. Main ideas
    1. Measures of Central Tendency
    2. Measures of Dispersion

3. Summary

The **inter-quartile range** (IQR):

$$\text{IQR} = \text{Q3 - Q1} = \text{75th percentile - 25th percentile}$$



The IQR is the width of an interval covering the middle half of the data.

1. The **variance** is (almost) the average squared difference of each observation from the mean.

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

2. The **standard deviation** is the square root of the variance. Intuitively, you can think of it as the average distance of the data points from the mean (although technically, that's not exactly right).

$$\text{Standard Deviation} = \sqrt{\text{Variance}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Outline

1. Today: Summary statistics

2. Main ideas
    1. Measures of Central Tendency
    2. Measures of Dispersion

3. Summary

Summary of main ideas

1. Measures of Central Tendency
2. Measures of Dispersion

Message:
Mean, Variance, and Standard deviation are sensitive to outliers and skewness. They should only be used when a distribution looks "nice" (unimodal, symmetric, no outliers). Otherwise, use median and IQR to summarize center and spread.