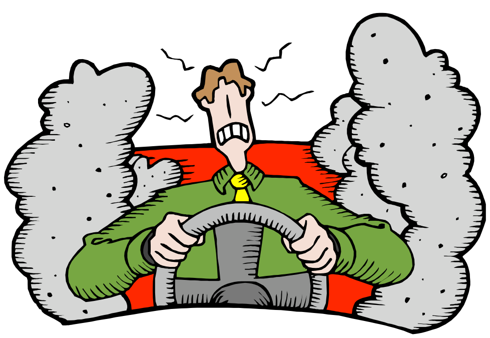
Measures of Central Tendency - the Mean

Dr Tom Ilvento
Department of Food and Resource Economics
University of Delaware

Overview

- We will begin looking at various measures of the center of the data - think of it as a “typical” value
- We will start with the **mean**
- I will also talk about some math symbols we will need and use to work with data, especially continuous data.
- And you will start to see how we uses statistics and graphs to tell a story.
- I will use two data sets to demonstrate the mean, one of which is marriage rates for the 50 states and Washington D.C.

I won’t drive with students!!



The fastest speed from past classes was 100.1 mph

Fastest Speed

Stem and Leaf of Fastest Speed

Stem	Leaf	Count
18	6	1
17	02	2
16	1	1
15	0	1
14	005	3
13	0007	4
12	0000000005555	13
11	0000000000000000024555567	25
10	00000000000000455555	20
9	00000000000000000155555555555555555555555788	44
8	00000000000000255555555555555555	32
7	0555555	7
6	5558	4

6|5 represents 65

Some Math Tools

- Sigma Notation
  - The Greek symbol Σ
  - Stands for summation

$$\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + x_4 + x_5 \ldots x_n$$

Alternatives to Math Formulas

- At times it is difficult to use the proper math symbols in Power Point, Word, Quizzes, or e-mails

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- As an alternative, we will use (and you can as well in assignments), the following symbols

Mean = Sum(X)/n

- Many of these follow the same usage in Excel or other spreadsheets

Alternatives to Math Formulas

- **Multiplication** *
  - 5*3 = 15
- **Power** ^
  - 5^3 = 5^3 = 125
- **Square Root** **SQRT** or **^.5**
  - $\sqrt{25}$ = **SQRT**(25) = $(25)^{.5}$ = (25)**^.5** = 5
- **Summation** **Sum**
  - $\sum_{i=1}^{n} x_i$ = **Sum**(x)

Alternative Math Symbols

- $\bar{x}$ **Mean(x)**
- μ **mu**
- σ **sigma**
- Standard Error **SE**

Central Tendency

- The central tendency of a variable is the tendency of the data to cluster or center about certain numerical values
- You might also think of this as a "typical" value
- The variability is the spread of the data
- For central tendency we will focus on the mean, the mode, and the median

The Mean or Arithmetic Average

- The arithmetic mean or mean is the sum of the measurements divided by the number of measurements contained in the data set
- For a sample, a statistic, we use x with a bar over it $\overline{x}$
- For a population, a parameter, we use the Greek **μ**

There are two ways to express the mean

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

The sum of all the values, divided by the number of values

$$\overline{x} = \sum_{i=1}^{n} (x_i / n)$$

The sum of each value weighted by the number of values - a mathematical expectation

Fastest Speed example

- n = 157
- Sum(x) = 15,718.0
- Mean = 15,718.0/157 = 100.1
- On average, the fastest speed of dr ilvento's students is 100.1 mph

Stem and Leaf of Fastest Speed

Stem	Leaf	Count
18	6	1
17	02	2
16	1	1
15	0	1
14	005	3
13	0007	4
12	0000000005555	13
11	0000000000000000024555567	25
10	00000000000000455555	20
9	00000000000000000015555555555555555555555788	44
8	00000000000000025555555555555555	32
7	0555555	7
6	5558	4

6|5 represents 65

Suggestion for significant digits: for calculated statistics, use one more decimal place than the original data.

As a measure of central tendency, the mean has several advantages:

- The mean uses **information of all the values** in a variable
  - We are adding all the values together, and then dividing by the sample size
  - Using more information is usually better
- The mean has two important mathematical properties:
  1. **The sum of the deviations about the mean equals zero**
  2. **The sum of squared deviations about the mean is a minimum**

Sum of deviations about the mean equal zero

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x}$$

$$= \sum_{i=1}^{n} x_i - n \cdot \frac{\sum_{i=1}^{n} x_i}{n}$$

$$= \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0$$

Sum of squared deviations about the mean is a minimum

- This is called the ***Least Squares*** property

$$\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- **There is no other value or constant we could substitute in the equation for the mean that would result in a lower sum of squares.**

Other Properties of the Mean

- We can make **inferences** from a sample to a population for the mean
- The mean forms the basis for a number of other statistics known as **Product Moment Statistics**
- But, the mean is **sensitive to outliers and extremes** in the data. It is not as **resistant** as other measures of central tendency

The effect of an outlier - Marriage Rate data

- Marriage rate data set
  - **n = 51** (50 states and Washington D.C.)
  - **sum(x) = 441.7**
  - **mean = 441.7/51 = 8.66**

Stem and Leaf

Stem	Leaf	Count
6	1	1
5		
5		
4		
4		
3		
3		
2		
2	3	1
1		
1	0013	4
0	55666666666677777777777777777777888888889999	44
0	4	1

0|4 represents 4

Removing the outliers on the Marriage Rate Data

- Revised Marriage rate data set
  - **n = 49**
  - **sum(x) = 358.22**
  - **mean = 358.22/49 = 7.31**
  - **About a 15.1% decrease from 8.66**

Stem & Leaf of Marriage Rate

Stem	Leaf	Count
4	2 7	2
5	0 5 5 8 9 9	6
6	1 1 1 3 3 4 5 6 6 7 7 8 9 9	14
7	0 0 0 0 3 3 3 4 4 4 7 9	12
8	1 1 2 3 3 4 6 8 9 9	10
9	4 5	2
10	3 5	2
11		0
12	6	1

4|2 = 4.2

Let's look at the effect of outliers on the Student Speed Data

Student Speed Data
for MPH
Stem unit: 10 in 100

Stem	Leaf
6	5 5 5 8
7	0 5 5 5 5 5 5
8	0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
9	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 7 8 8
10	0 0 0 0 0 0 0 0 0 0 0 0 0 0 4 5 5 5 5 5 5
11	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 4 5 5 5 6 7
12	0 0 0 0 0 0 0 0 0 5 5 5 5
13	0 0 0 7
14	0 0 5
15	0
16	1
17	0 2
18	6

	MPH
Mean	100.11
Standard Error	1.62
Median	95.00
Mode	95.00
Standard Deviation	20.28
Sample Variance	411.47
Kurtosis	3.00
Skewness	1.35
Range	121.00
Minimum	65.00
Maximum	186.00
Sum	15718.00
Count	157.00

The mean only changed slightly by removing the top five scores, from 100.11 to 97.89, about a 2% decrease.

Closing thoughts on the mean and Outliers

- **Key point: in and of themselves, outliers are not wrong or bad.**
  - They should be examined to determine in they are not part of the population,
  - Or if they are a mistake in coding or measurement.
- I will present you with a strategy for accessing what is an outlier and the impact of outliers on measures of central tendency.
  - Based on a probability framework
  - And the standard deviation