## Measures of Variability

#### **Dr Tom Ilvento**

Department of Food and Resource Economics



## **Measures of Variability**

- Now we will shift to the spread of the data
- Variability is the key to most of statistics
  - Why is there variation?
  - Why do groups differ?
- This lecture will focus on the
  - Range
  - Inter-quartile Range (IQR)
  - Variance (Var, σ², s²)
  - Standard Deviation (Std Dev, σ, s)

2

# Central Tendency only tells part of the story

- Imagine two data sets
  - Data set 1 has a mean, median, and mode of 5
  - Data set 2 has a mean, median, and mode of 5

### Two data sets

- Data set 1
  - $\{2, 3, 4, 5, 5, 6, 7, 8\}$   $\Sigma x = 40 \text{ n=8}$
  - mean = 5; median = 5; mode =5
- Data set 2
  - $\{5, 5, 5, 5, 5, 5, 5, 5\}$   $\Sigma x = 40$  n=8
  - mean = 5; median = 5; mode =5
- We need something more to help describe a variable – the spread or the variability

3

### **The Range**

- Let's start with the Range
- The range is the difference between the largest measurement and the smallest measurement
- To calculate the range we need
  - Minimum Value
  - Maximum Value

5

### Issues with the Range

- It is an order statistic
- Note that the range depends upon the two most extreme values.
- and may be seriously influenced by outliers or unusual cases.

6

## The Range for the Marriage Data

- Marriage data for 2005
  - Minimum is 4.2
  - Maximum is 61.0
  - Range is 61.0 4.2 = 56.8
- Without Nevada in the data set, the range is
  - Range is 22.5 4.2 = 18.3

Quantiles		
100.0%	maximum	61.000
99.5%		61.000
97.5%		49.450
90.0%		10.140
75.0%	quartile	8.300
50.0%	median	7.000
25.0%	quartile	6.300
10.0%		5.560
2.5%		4.350
0.5%		4.200
0.0%	minimum	4.200

# An alternative to the range - the IQR

- The Inter-Quartile Range (IQR)
- Based on the difference between the Third Quartile (Q3 or the 75 Percentile) and the First Quartile (Q1 or the 25 Percentile)
  - This is a positional measure
  - As long as we can order the data, we can find a value for any percentile.
- IQR is less sensitive to the extreme values in a data set than Range

7

### The IQR for the Marriage Data

- Marriage data for 2005
  - Q1 is 6.3
  - Q3 is 8.3
  - IQR is 8.3 6.3 = 2.0
- Without Nevada in the data set, the IQR is

of the values of the variable.

- IQR is 8.3 6.3 = 2.0

NOTHING CHANGED!! The IQR shows the range of the middle 50%

Quantiles		
100.0%	maximum	61.000
99.5%		61.000
97.5%		49.450
90.0%		10.140
75.0%	quartile	8.300
50.0%	median	7.000
25.0%	quartile	6.300
10.0%		5.560
2.5%		4.350
0.5%		4.200
0.0%	minimum	4.200

## **Excel and the Range/IQR**

- Excel will find the max and min values and the quartiles
  - =MIN(B5:B104)
  - =MAX(B5:B104)
  - =QUARTILE(B5:B104,1) for first quartile
  - =QUARTILE(B5:B104,3) for third quartile
- But you have to calculate the ranges yourself by subtracting the values

10

## What about using the mean to calculate a measure of spread

- The concept of deviations around the mean can be intuitively appealing.
- If the mean is a good measure of central tendency. then it is reasonable to ask how different (or how far away) is a particular value of X from its mean.
- The mean deviation might be a summary measure
  - But this won't work! Remember, the sum of deviations around the mean always equals zero!
- The Mean Absolute Difference might work, but it doesn't have all the properties we might want.

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

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

#### The Variance

- Another approach would be to square the differences from the mean
- The square will always give positive values
- This is called the variance

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

ш

# Note: Population versus the sample

- When we are dealing with a population we use the Greek term  $\sigma^2$  (sigma squared)
- When we are dealing with a sample we use s<sup>2</sup>
- And, we use n-1 in the denominator
  - This has to do with degrees of freedom
  - Which has to do with making inferences from a sample to the population.
  - Using n in the formula for s<sup>2</sup> tends to underestimate σ<sup>2</sup>

**Sample Variance** 

$$s^{2} = \frac{\sum_{i=1}^{n} (x_{i} - \overline{x})^{2}}{(n-1)}$$

14

### A closer look at the Variance

- The numerator is called the Total Sum of Squares
- It is the sum of squared deviations about the mean
- And when we divide by n, or n-1, we have the Mean Squared Deviation
- $\sum_{i=1}^{n} (x_i \overline{x})^2$

## Computational formula for the Variance

$$s^{2} = \frac{\sum_{i=1}^{n} x_{i}^{2} - \frac{\left(\sum_{i=1}^{n} x_{i}\right)^{2}}{n}}{n-1}$$

15

# Computation formula for the Variance

$$s^{2} = \frac{\sum_{i=1}^{n} x_{i}^{2} - \frac{\left(\sum_{i=1}^{n} x_{i}\right)^{2}}{n}}{n-1}$$

$$s^{2} = \frac{\sum_{i=1}^{4} x_{i}^{2} - \frac{\left(\sum_{i=1}^{4} x_{i}\right)^{2}}{4}}{4-1} = \sum_{i=1}^{n} x_{i}$$

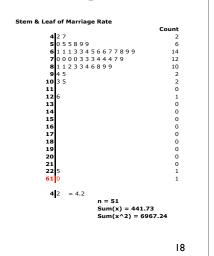
$$s^{2} = \frac{\sum_{i=1}^{4} x_{i}^{2} - \frac{\left(\sum_{i=1}^{4} x_{i}\right)^{2}}{4}}{4-1} = \frac{\left(54 - \frac{14^{2}}{4}\right)}{3}$$

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

$$\sum_{i=1}^{4} x_{i}^{2} - \frac{\left(\sum_{i=1}^{4} x_{i}\right)^{2}}{4} + \frac{\left(54 - \frac{14^{2}}{4}\right)}{3} + \frac{54 - 49}{3} + \frac{67}{3}$$

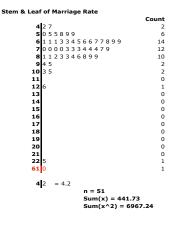
## You try it with the Marriage Data

Calculate the Variance



## You try it with the Marriage Data

- Calculate the Variance
- $s^2 = [6967.24 (441.73)^2/51]/(51-1)$
- $s^2 = [6967.24 3825.99]/(50)$
- $s^2 = [3141.25]/(50)$
- $s^2 = 62.83$



17

19

**Standard Deviation** 

- One problem with the variance is that it is expressed in squared units and can be difficult to interpret
- If you take it the square root of the variance we bring it back to original units
- This is called the Standard Deviation
  - s for a sample
  - σ for a population

### **Standard Deviation**

- The standard deviation (Std Dev) is the average deviation of the values from the mean or the average spread
- It is always positive
- The Std Dev is a basic building block for analyzing our data

21

- It provides insights into identifying outliers
- It is important in inference
- For the Marriage Rate data,
  - s = SQRT(Var) = SQRT(62.83) = 7.93

## **Summary**

- Our main measure of variability in the data is the variance – in reference to deviations about the mean
- We focus on squared deviations because of the property of the mean - Variance
- But then take the square root to bring it back to regular terms - Standard Deviation
- For samples we use n-1 as the denominator referred to as degrees of freedom