

Measures of Dispersion I

Tom Ilvento
STAT 200

Central Tendency 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} $S_x = 40$ $n=8$
mean = 5
- Data set 2
 - {5, 5, 5, 5, 5, 5, 5, 5} $S_x = 40$ $n=8$
mean = 5
- We need something more to help describe a variable – the **variability**

Variability

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

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.

Data Example with the range

- **Marriage data**
 - **Minimum** is 5.8
 - **Maximum** is 88.2
 - **Range is** $88.2 - 5.8 = 82.4$
- Without Nevada in the data set, the range is
 - $16.4 - 5.8 = 10.6$

An alternative to the range

- **The Inter-Quartile Range**
- Based on the difference between the **Third Quartile** (Q3 or the 75 Percentile) and the **First Quartile** (Q1 or the 25 Percentile)
- Less sensitive to the extreme values in a data set

Finding the Quartiles

- Example: N=50
 - Q1 is the 13th case **7.6**
 - Q3 is the 39th case **10.1**
- **The Inter-Quartile Range is**
 - **$10.1 - 7.6 = 2.5$**

Recent use of Quartiles

- U.S. News and World Report publishes data on colleges and universities
- One indicator used the 1st and 3rd Quartiles of SAT/ACT scores of students who enrolled in 1999
- Why was the Inter-quartile range used?
 - For UD it was **1040 to 1240**
 - For the top university (Harvard) it was **1400 to 1590**

What about using the mean to help measure variability?

- 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 the mean of X.
- The **mean deviation** might be a summary measure

Mean Deviation

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

- However, mean deviation will not work because the numerator will always be zero.
- **Remember:** the sum of deviations around the mean is always zero

Absolute Mean deviation

- One approach would be to sum the absolute differences and divide by n – the **mean absolute difference**

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

The Variance

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

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

NOTE: Population versus Sample

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

Sample Variance

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

A closer look at the Variance

- The numerator is called the
- **Total Sum of Squares**

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

A closer look at the Variance

- And when we divide by n-1 we have the
- **Mean Squared Deviation**

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

Computational formula for s^2

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

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

Computational formula for s^2

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

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

Computational formula for s^2

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

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

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

Here's the extra steps for the numerator for s^2

$$\begin{aligned} &= \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + \sum_{i=1}^n \bar{x}^2 \\ &= \sum_{i=1}^n x_i^2 - \frac{2}{n} \sum_{i=1}^n x_i \sum_{i=1}^n x_i + n\bar{x}\bar{x} = \sum_{i=1}^n x_i^2 - \frac{2}{n} \sum_{i=1}^n x_i \sum_{i=1}^n x_i + \frac{n \sum_{i=1}^n x_i}{n} \frac{\sum_{i=1}^n x_i}{n} \\ &= \sum_{i=1}^n x_i^2 - \frac{2}{n} \sum_{i=1}^n x_i \sum_{i=1}^n x_i + \sum_{i=1}^n x_i \sum_{i=1}^n x_i = \sum_{i=1}^n x_i^2 - \frac{1}{n} \sum_{i=1}^n x_i \sum_{i=1}^n x_i \\ &= \sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n} \end{aligned}$$

Computational formula for s^2

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

Computational formula for s^2

- So, if I gave you

- n
- $\sum x$
- $\sum x^2$

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

- You could calculate the variance!!**

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
 - S for a population

Marriage Data Example

- $n = 50$
- $Sx = 523.36$
- $Sx^2 = 11,892.45$

Marriage Data Example

- $n = 50$
- $Sx = 523.36$
- $Sx^2 = 11,892.45$

- $s^2 = [11,892.45 - (523.36)^2/50]/(50-1)$

Marriage Data Example

- $n = 50$
- $Sx = 523.36$
- $Sx^2 = 11,892.45$
- $s^2 = [11,892.45 - (523.36)^2/50]/(50-1)$
- $= [11,892.45 - 5,478.11]/49$
- $= 6,414.34/49$
- $= 130.90$
- $s = 11.44$

A Note of Caution

- The variance and the standard deviation are very sensitive to outliers (extreme values)
- When you square large numbers you get **much larger** numbers
- Look what happens when we remove Nevada from the Marriage Rate data

Marriage Data Example

- $n = 49$
- $Sx = 435.12$
- $Sx^2 = 4,104.66$
- $s^2 = [4,104.66 - (435.12)^2/49]/(49-1)$
- $= [4,104.66 - 3,863.87]/48$
- $= 240.79/48$
- $= 5.02$
- $s = 2.24$

Comparisons with Nevada

Statistic	With Nevada
Sum x	536.9
Sum x ²	12,045.01
Mean	10.5
Median	8.5
Mode	8.7
Maximum	88.1
Minimum	6.1
Variance	127.87
Std Deviation	11.31

Comparisons with and without Nevada

Statistic	With Nevada	W/O Nevada
Sum x	523.36	435.12
Sum x ²	11,892.45	4,104.66
Mean	10.47	8.88
Median	8.4	8.4
Mode	NA	NA
Maximum	88.3	16.4
Minimum	5.8	5.8
Variance	130.90	5.02
Std Deviation	11.44	2.24

Excel Commands for Measures of Dispersion

Sum	=SUM(B5:B104)	3,699.40
Count	=COUNT(B5:B104)	100.00
Mean	=AVERAGE(B5:B104)	36.99
Minimum	=MIN(B5:B104)	30.00
Maximum	=MAX(B5:B104)	44.90
Median	=MEDIAN(B5:B104)	37.00
Mode	=MODE(B5:B104)	37.00
Range	subtract the max and min	14.90
First Quartile	=QUARTILE(B5:B104,1)	35.68
Third Quartile	=QUARTILE(B5:B104,3)	38.33
Inter-Quartile Range	subtract Q3 minus Q1	2.65
Variance	=VAR(B5:B104)	5.85
Std Deviation	=STDEV(B5:B104)	2.42

Descriptive Statistics of EPA data using Excel

MPG	
Mean	36.99
Standard Error	0.24
Median	37.00
Mode	37.00
Standard Deviation	2.42
Sample Variance	5.85
Kurtosis	0.77
Skewness	0.05
Range	14.90
Minimum	30.00
Maximum	44.90
Sum	3699.40
Count	100.00

Use the menu in Excel with the following sequence

Tools
Data Analysis
Descriptive Statistics

The Standard Deviation and the Range

- A quick approximation of the standard deviation is
- **Range/4**
- **EPA MPG Data Example**
 - **Approximation** $(44.90 - 30.0)/4 = 3.73$
 - **Whereas, $s = 2.42$**
- **It's just an approximation!!!**

Additional Problems

- Small sample of 41 bones on an archeological dig. The size of the bones are measured in inches.
 - **Count** 41
 - **Sum x** 379.56
 - **Sum x²** 3571.74
- Solve for the mean, variance, standard deviation

■ Mean = $379.46/41 = 9.26$



Variance =

$$s^2 = [3571.74 - [379.56]^2/41]/40$$

$$s^2 = 1.4485$$

Standard Deviation =

$$s = 1.20$$

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



Interpreting the Standard Deviation

- We can use the standard deviation to express the proportion of cases that might fall within one or 2 standard deviations from the mean.
- We can use two theorems to help
 - **Chebyshev's Rule**
 - **Empirical Rule**



Chebyshev's Rule

- Is based on a mathematical theorem for any data
- At least 3/4 of the measurements will fall within **± 2 standard deviations** from the mean
- At least 8/9 of the measurements will fall within **± 3 standard deviations** from the mean

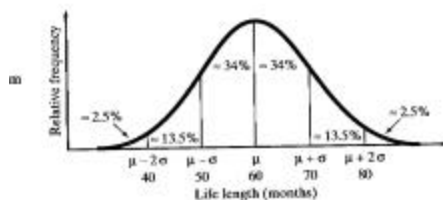


Empirical Rule

- Based on a symmetrical distribution where the mean, median, and the mode are similar – the EPA mpg data fits this



A Symmetrical Curve



Empirical Rule cont.

- Approximately 68% of the measurements will be **± 1 standard deviation** from the mean
- Approximately 95% of the cases fall between **± 2 standard deviations** from the mean



Empirical Rule

- Approximately 99.7% of the cases will fall within ± 3 standard deviations from the mean
 - This means it will be very rare to be more than 3 standard deviations from the mean when dealing with a symmetrical distribution



Empirical Rule

- For the EPA mpg data we would expect that 68% of the cases would fall between
 - 36.99 ± 2.42 or
 - Between **34.57 to 38.41**



Empirical Rule and EPA mpg Data

- **1 Std Dev**
 - 36.99 ± 2.42 34.57 to 38.41
- **2 Std Dev**
 - 36.99 ± 4.84 32.15 to 41.83
- **3 Std Dev**
 - 36.99 ± 7.26 29.73 to 44.25



A Symmetrical Curve

This example has a mean = 60

And a standard deviation of 10

