STA 1013 : Statistics through Examples

Lecture 8: Measure of Variation

Hwiyoung Lee

September 18, 2019

Department of Statistics, Florida State University

Overview

1. Measure of Variation
2. Range
3. Interquartile Range
4. Standard Deviation

Measure of Variation

Why variation matters ?

Imagine customers waiting in line for tellers at two different banks

- **Bank A** can enter any one of three different lines leading to three different tellers
- **Bank B** also has three tellers, but all customers wait in a single line and are called to the next available teller

Why variation matters ?

The following values are waiting times, in minutes, for 11 customers at each bank. The times are arranged in ascending order

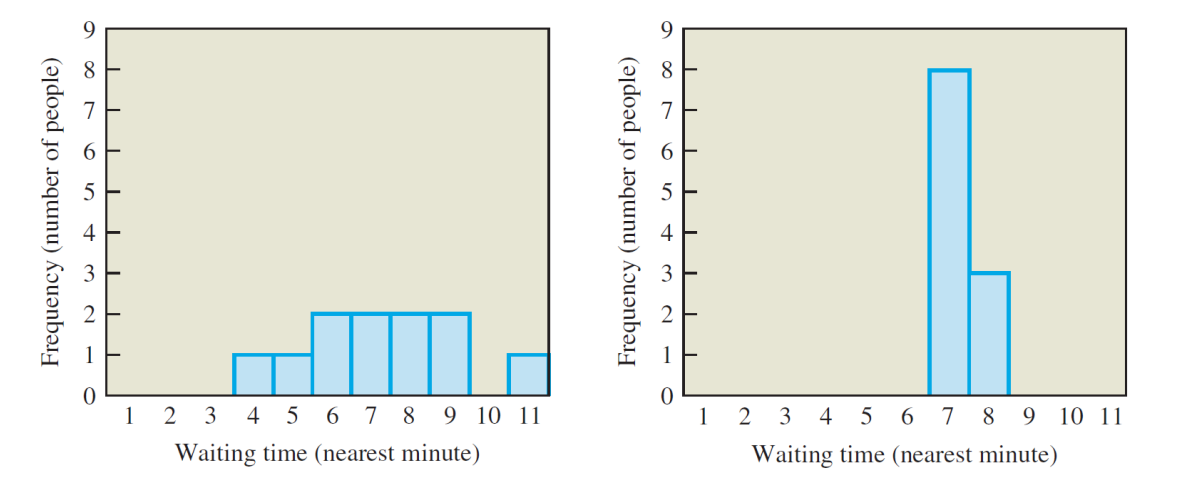
Bank A	4.1	5.2	5.6	6.2	6.7	7.2	7.7	7.7	8.5	9.3	11.0
Bank B	6.6	6.7	6.7	6.9	7.1	7.2	7.3	7.4	7.7	7.8	7.8

- Mean A : _____ , Median A : _____
- Mean B : _____ , Median B : _____

You'll probably find more unhappy customers at **Bank A** than at **Bank B**, but this is not because the average wait is any longer

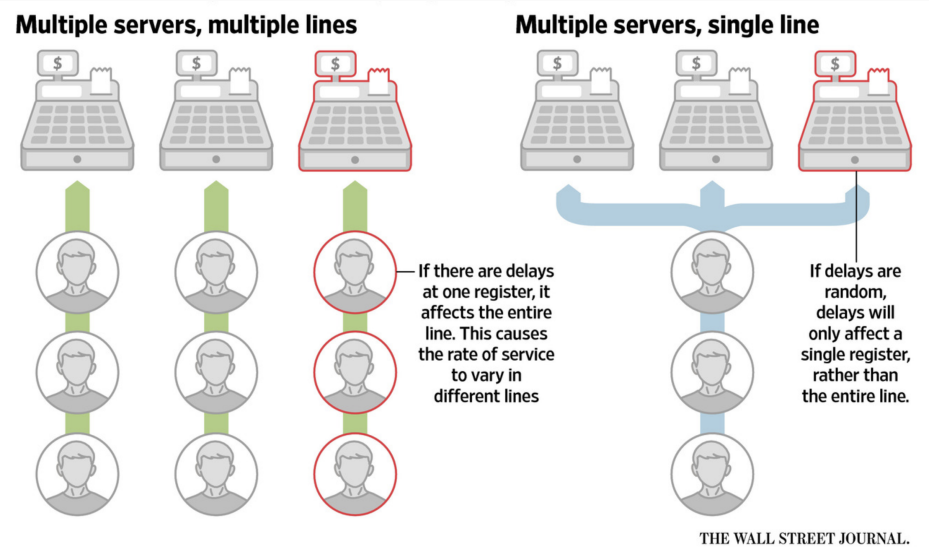
Why variation matters ?

- The difference in customer satisfaction comes from the variation at the two banks
- The waiting times at **Bank A** vary over a fairly wide range, so a few customers have long waits and are likely to become annoyed.
- In contrast, the variation of the waiting times at **Bank B** is small, so all customers feel they are being treated roughly equally.



Why variation matters ?

- Multiple lines, Multiple tellers (Bank A) → Large variation
  - Picked the fastest line → short waiting time
  - Picked the slowest line → Long waiting time
- Single line, Multiple tellers (Bank B) → Small variation
  - May not be in the slowest one, also are not in the fastest



Range

Range

Range

The range of a set of data values is the difference between its highest and lowest data values:

$$\textbf{Range} = \textbf{max} - \textbf{min}$$

Bank data

Bank A	4.1	5.2	5.6	6.2	6.7	7.2	7.7	7.7	8.5	9.3	11.0
Bank B	6.6	6.7	6.7	6.9	7.1	7.2	7.3	7.4	7.7	7.8	7.8

- Range of Bank A : _____
- Range of Bank B : _____

Range

Consider the following two sets of quiz scores for nine students.

Quiz 1	1	10	10	10	10	10	10	10	10
Quiz 2	2	3	4	5	6	7	8	9	10

Which set has the greater range?

- Range of Quiz 1 : _____
- Range of Quiz 2 : _____

Would you also say that this set has the greater variation?

Range

- Aside from a single low score (1 : outlier), Quiz 1 has no variation at all because every other student got a 10.
- In contrast, no two students got the same score on Quiz 2, and the scores are spread throughout the list of possible scores.
- Quiz 2 therefore has greater variation even though Quiz 1 has greater range.

Interquartile Range

Quartile

- A better way to describe variation is to consider a few intermediate data values in addition to the high and low values.
- A common way involves looking at the quartiles, or values that divide the data distribution into quarters.

Quartile

- The first quartile (or lower quartile) :

$$\mathbf{Q_1} = \frac{n+1}{4} \text{ th observation}$$

- The second quartile (or middle quartile) is the **median** :

$$\mathbf{Q_2} = 2 \times \frac{n+1}{4} \text{ th observation}$$

- The third quartile (or upper quartile) :

$$\mathbf{Q_3} = 3 \times \frac{n+1}{4} \text{ th observation}$$

- Interquartile Range :

$$\mathbf{IQR} = \mathbf{Q_3} - \mathbf{Q_1}$$

Interquartile Range

Quiz data

Quiz 1	1	10	10	10	10	10	10	10	10
Quiz 2	2	3	4	5	6	7	8	9	10

- IQR of Quiz 1 : _____

  Q_1 : _____ , Q_2 : _____ , Q_3 : _____

- IQR of Quiz 2 : _____

  Q_1 : _____ , Q_2 : _____ , Q_3 : _____

Problem of IQR

Obs	1	2	3	4	5	6	7	8	9	10	11
Data 1	-100	-10	-5	-3	-1	0	1	3	5	10	100
Data 2	-5.2	-5.1	-5	-3	-1	0	1	3	5	5.1	5.2

- Mean of Data 1 = Mean of Data 2
- Median of Data 1 = Median of Data 2
- IQR of the Data 1 :
- IQR of the Data 2 :

Would you say that data set 1 and set 2 have the same variation?

Problem of IQR

- This is basically the range of the central 50% of the observations in the distribution
- **Problem** : The interquartile range does not take into account the variability of the total data (only the central 50%).
- We are "throwing out" half of the data.

Another measure which takes into account the whole data is needed. $\rightarrow$ **Standard deviation, Variance**

Five-Number Summary

The five-number summary for a data distribution consists of the following five numbers:

low value, **lower quartile**, **median**, **upper quartile**, **high value**

Eample : Five-Number Summary

Bank A	4.1	5.2	5.6	6.2	6.7	7.2	7.7	7.7	8.5	9.3	11.0
Bank B	6.6	6.7	6.7	6.9	7.1	7.2	7.3	7.4	7.7	7.8	7.8

Bank A

- min
- Q_1
- Q_2
- Q_3
- max

Bank B

- min
- Q_1
- Q_2
- Q_3
- max

Exercise

Data (n=16): 1, 1, 2, 2, 2, 2, 3, 3, 4, 4, 5, 5, 6, 7, 8, 10

- $\min$:
- Q_1 :
- Q_2 :
- Q_3 :
- $\max$:
- IQR :
- Range :

Outliers

- Identifying Outliers in a Sample of Real Data
- First calculate the lower and upper fences:

$$\text{LF} : Q_1 - 1.5 \times \text{IQR}$$
$$\text{UF} : Q_3 + 1.5 \times \text{IQR}$$

- Outliers are values that lie outside of the fences
  - Low outliers are values that are less than the lower fence
  - High outliers are values that are greater than the upper fence

Example

Example(n=15)

0, 0, 2, 3, 4, 7, 9, 12, 17, 18, 20, 22, 45, 56, 98

- LF : _____
- UF : _____
- Outlier : _____

Exercise : Identifying Outliers

- **Space Shuttle Flights** : Listed below are the durations (in hours) of a sample of all flights of NASA's Space Transport System (space shuttle):

  0 73 95 165 191 192 221 235
  235 244 259 262 331 376 381

  Answer : _____

- **blood Alcohol** : Blood alcohol concentrations of drivers involved in fatal crashes:

  0.01 0.12 0.13 0.14 0.16 0.16
  0.17 0.24 0.27 0.29 0.46

  Answer : _____

Standard Deviation

Standard Deviation

- The five-number summary characterizes variation well
- Statisticians prefer to describe variation with a single number
- The single number most commonly used to describe variation is called the **standard deviation**
- The standard deviation is a measure of **how widely data values are spread around the "mean" of a data set**
- To calculate a standard deviation,
  - First find the mean
  - Calculate how much each data value "deviates" from the mean

Deviation ?

Average of Deviation :

$$\frac{\text{Sum of (Data value - mean)}}{\text{The number of observations}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n}$$

- The above is not a good measure
- Why can't we simply compute the average deviation about the mean ?
- The summation in the numerator is always 0

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

- The average deviation will not work.

Example

Obs	1	2	3	4	5	6	7	8	9	10	11
Data 1	-100	-50	-10	-5	-1	**0**	1	5	10	50	100
Data 2	-5	-4	-3	-2	-1	**0**	1	2	3	4	5

- Data 1 has larger variation than Data 2
- Data 1 : $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$
- Data 2 : $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$

Average of deviation $(x_i - \bar{x})$ shouldn't be used !!!

Example

The means of the Bank A and Bank B are **7.2**

Bank A		**Bank B**	
x_i	$x_i - \bar{x}$	x_i	$x_i - \bar{x}$
4.1	-3.1	6.6	-0.6
5.2	-2	6.7	-0.5
5.6	-1.6	6.7	-0.5
6.2	-1	6.9	-0.3
6.7	-0.5	7.1	-0.1
7.2	0	7.2	0
7.7	0.5	7.3	0.1
7.7	0.5	7.4	0.2
8.5	1.3	7.7	0.5
9.3	2.1	7.8	0.6
11.0	3.8	7.8	0.6
$\sum x_i - \bar{x}$	0	$\sum x_i - \bar{x}$	0

Standard Deviation

Standard Deviation

$$s = \sqrt{\frac{\text{Sum of (Data value - mean)}^2}{\text{The number of observations-1}}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\mathbf{n-1}}}$$

- Use Squared Deviation $(x_i - \bar{x})^2$
- By squaring the deviation, we eliminate the problem of the deviations summing to zero
- Note we divide by $\mathbf{(n-1)}$, **not** $\mathbf{n}$
- $(n-1)$ is called **the degree of freedom (df)**

Variance

Variance

$$s^2 = \frac{\text{Sum of (Data value - mean)}^2}{\text{The number of observations-1}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\mathbf{n-1}}$$

- The variance is the standard deviation squared
- Conversely **standard deviation** $= \sqrt{\textbf{variance}}$

Example

The means of the Bank A and Bank B are **7.2**

Bank A			**Bank B**		
x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
4.1	-3.1	9.61	6.6	-0.6	0.36
5.2	-2	4	6.7	-0.5	0.25
5.6	-1.6	2.56	6.7	-0.5	0.25
6.2	-1	1	6.9	-0.3	0.09
6.7	-0.5	0.25	7.1	-0.1	0.01
7.2	0	0	7.2	0	0
7.7	0.5	0.25	7.3	0.1	0.01
7.7	0.5	0.25	7.4	0.2	0.04
8.5	1.3	1.69	7.7	0.5	0.25
9.3	2.1	4.41	7.8	0.6	0.36
11.0	3.8	14.44	7.8	0.6	0.36
SS	$\sum(x_i - \bar{x})^2$	38.46	$\sum(x_i - \bar{x})^2$		1.98
Variance (s^2)	$\frac{\sum(x_i-\bar{x})^2}{n-1}$	3.846	Variance (s^2)	$\frac{\sum(x_i-\bar{x})^2}{n-1}$	0.198
S.D (s)	$\sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}$	1.961	S.D (s)	$\sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}$	0.444

Exercise

body Temperatures : Body temperatures (in degrees Fahrenheit) of randomly selected normal and healthy adults:

Data	Deviation $(x_i - \bar{x})$	Deviation^2 $\left((x_i - \bar{x})^2\right)$
98.6		
98.6		
98.0		
98.0		
99.0		
98.4		
98.4		
98.4		
98.4		
98.6		

- Standard Deviation :

Note : Population version

Population Standard Deviation

$$\sigma = \sqrt{\frac{\text{Sum of (Data value - population mean)}^2}{\text{Population size}}} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{\mathbf{N}}}$$

- Divide by N
- Rarely used (We usually don't know the whole population)
  - μ :usually unknown
  - N : usually unknown
- Population variance (σ^2) :

$$\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{\mathbf{N}}$$