My reservoir is like the distance between Aden and Amman in al-Balqa

Al-Balqa Applied University

Faculty of Medicine

Epidemiology and Biostatistics

الوبائيات والإحصاء الحيوي (31505204)

Lecture 3

Descriptive statistics

Measures of variability

Graphical display: looking at data

25-6-2019

Methods of description (Descriptive statistics)

- **Two most common methods of description:**

1. Measures of **location** (Central Tendency).

2. Measures of **spread** (variation or dispersion).

Measures of Dispersion (Variability)

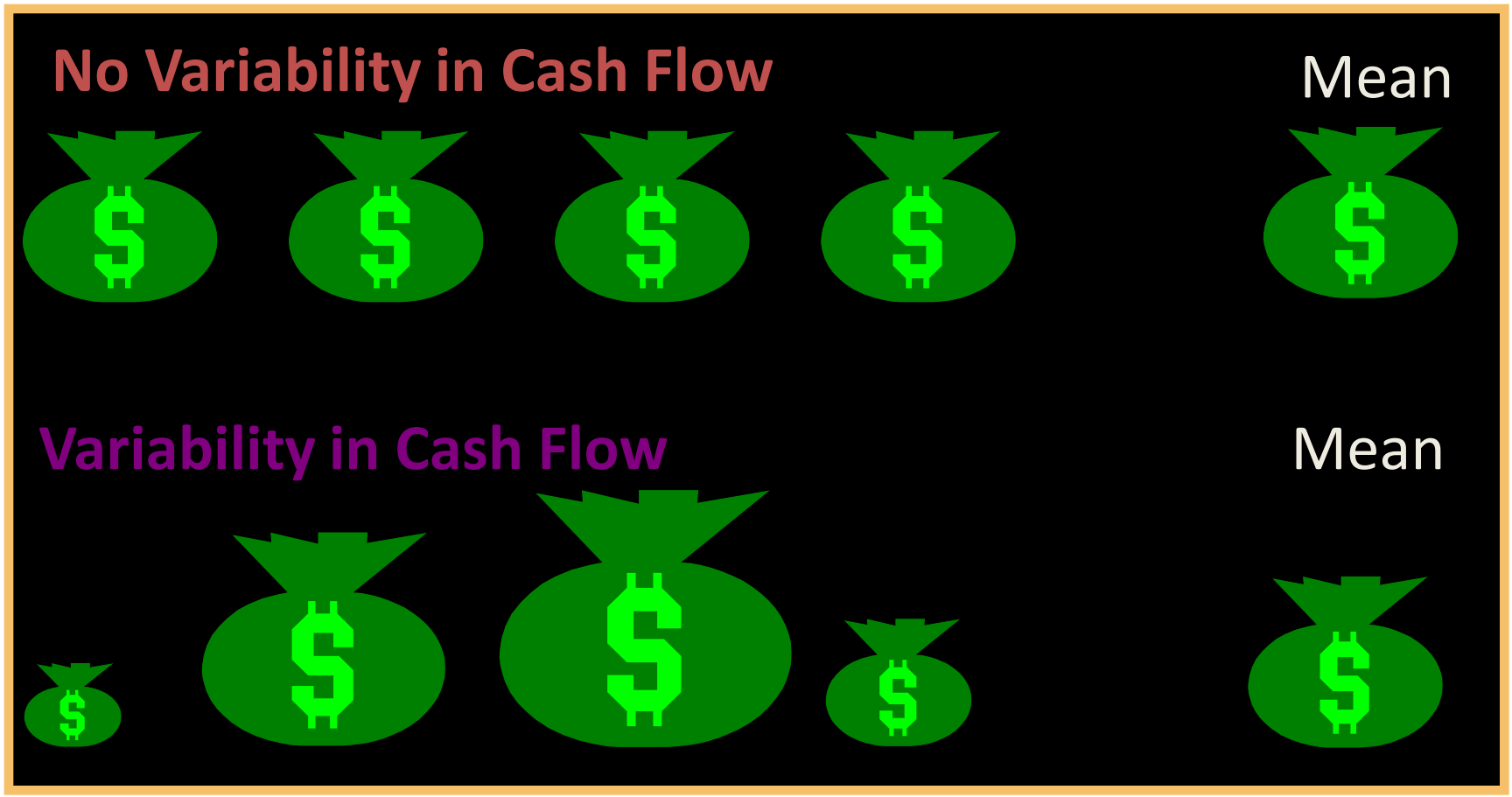
- A **measure** of **variability** is a summary statistic that represents the amount of **dispersion** in a dataset.
- It refers to **how spread out the scores are.**

- In other words, how similar or different participants are from one another on the variable. **It is either homogeneous or heterogeneous sample.**

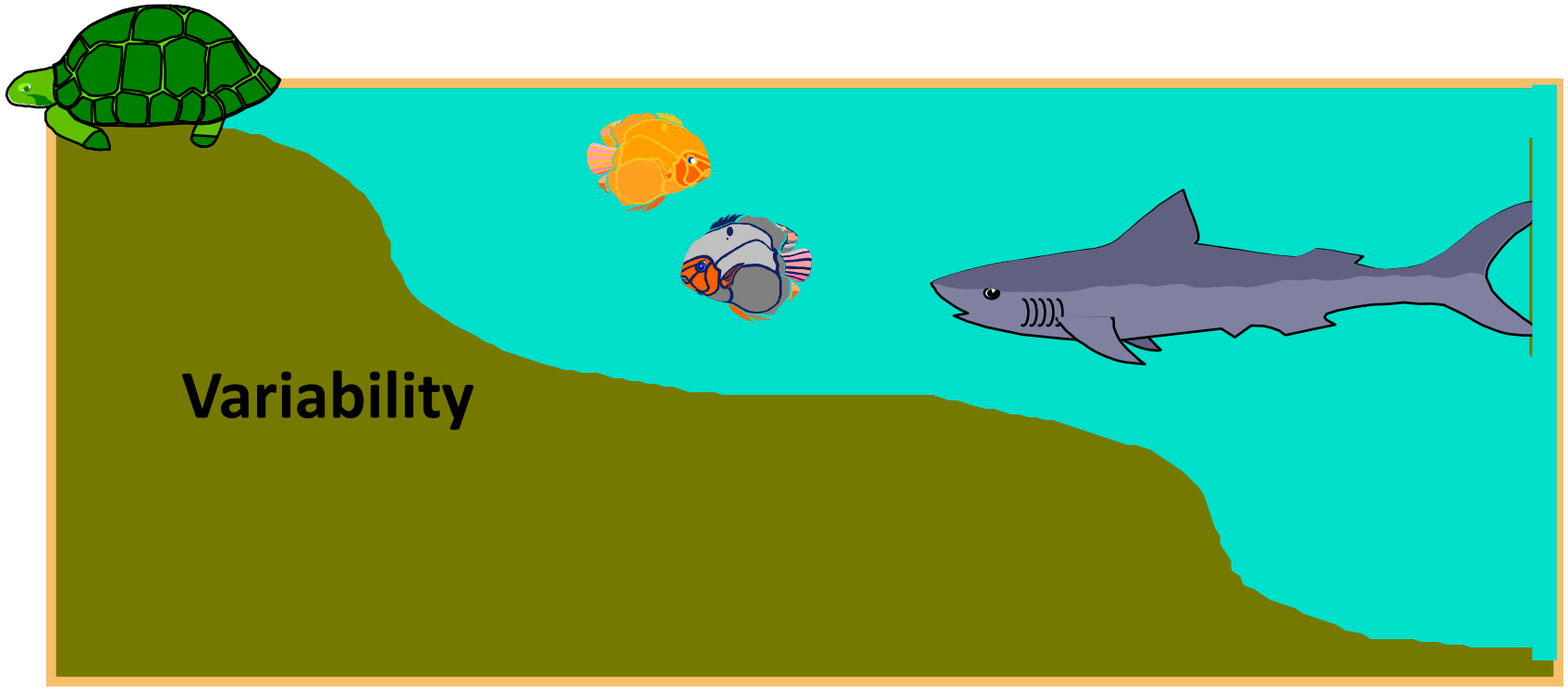
➢ A **low dispersion** indicates that the data points tend to be clustered tightly around the center.

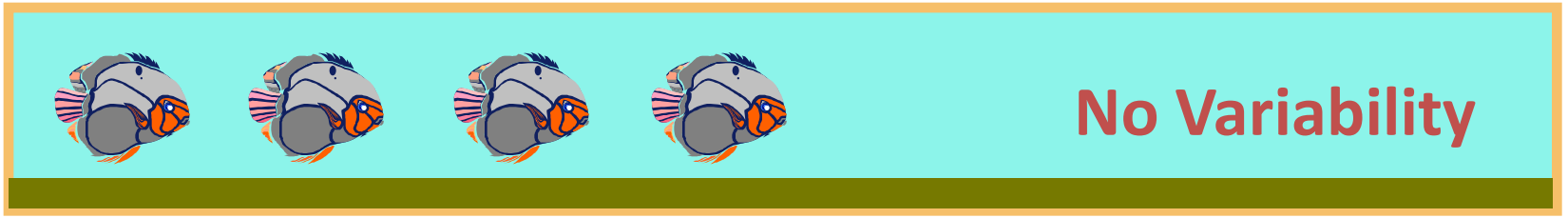
➢ **No Variability –- No Dispersion.**

Variability



Variability

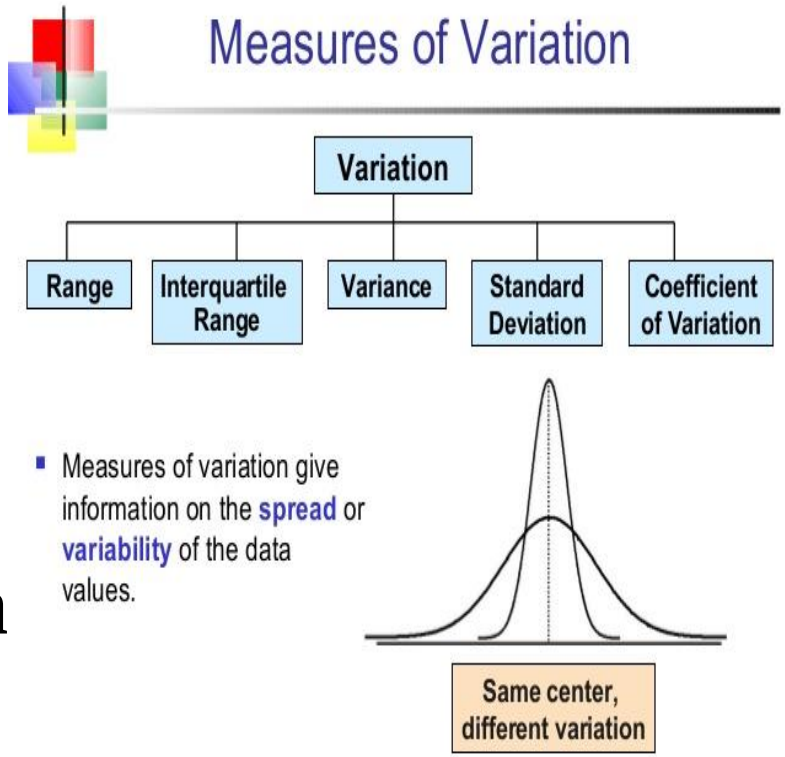




Measures of Dispersion

❑ We will study these **five measures of dispersion:**

- **Range**
- **Interquartile Range**
- **Variance**
- **Standard Deviation**
- **Coefficient of Variation**



Why do we need to look at measures of dispersion?

❑ **Why is it important to measure the spread of data?**

- A measure of spread gives us an idea of how well the mean.
- **For example**, represents the data. If the spread of values in the data set is **large**, the **mean** is **not as representative** of the data as if the spread of data is small.
- This is because a **large spread** indicates that there are probably **large differences** between individual scores.

Example / Why is it Important? ..

- **Example:** You want to choose the best brand of paint for your house. You are interested in how long the paint lasts before it fades and you must repaint. The choices are narrowed down to 2 different paints. The results are shown in the chart. Which paint would you choose?

The chart indicates the number of months a paint lasts before fading.

Paint A	Paint B
10	35
60	45
50	30
30	35
40	40
20	25
210	**210**

Does the Average Help?

- ***Paint A:*** Avg = 210/6 = 35 months

- ***Paint B:*** Avg = 210/6 = 35 months

- They both last 35 months before fading.
- No help in deciding which to buy.

Consider the Spread

- ***Paint A:*** Spread = 60 – 10 = 50 months

- ***Paint B:*** Spread = 45 – 25 = 20 months

- Paint B has a smaller ***variance*** which means that it performs more consistently.

➢ Choose paint B.

The Range

- The range of a data set is the **difference** between the **highest score** and the **lowest score** in the distribution.

- It is the simplest measure of variability.

- It provides a quick summary of a distribution's variability.

- It also provides useful information about a distribution when there are **extreme values.**

The Range …

Range = $X_{largest} - X_{smallest}$

- **Example:**
  - **1**, 2, 3, 4, 5, 8, 9, 21, 25, **30**
  - Answer: Range = 30 – 1 = 29.

- **Pros**:
  - **Easy to calculate**

- **Cons:**
  - Value of range is only determined by **two values.**

  - One problem with the range is **that it is influenced by extreme values at either end.**

The Range …

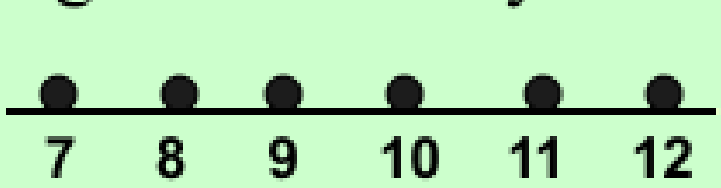
- If we have two extreme values **one very low** and the **other very high** then the **range** is **useless.**

- **Example:** If we have a set of marks for some students , one student got **zero** and another one got **100**, however the other students marks were between **60-70.**

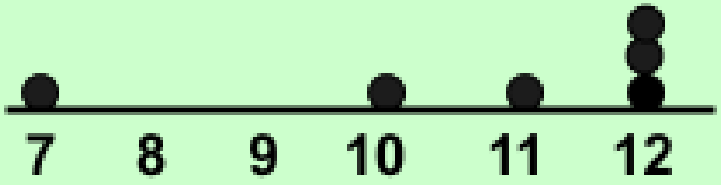
- **Theoretically,** the range should be **10**; however the **actual range** is **100** because of the two extreme values, so , in this case it is useless.

The Range …

- Ignores the way in which data are distributed



Range = 12 - 7 = 5



Range = 12 - 7 = 5

- Sensitive to outliers

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,**5**

Range = 5 - 1 = 4

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,**120**

Range = 120 - 1 = 119

Standard Deviation

- The **standard deviation** is the **square root of the variance**.
- **For sample** $s = \sqrt{s^2}$

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

- It is measured in the **same units** as the data , making it more easily comparable , than the variance, to the mean.

- **Example:** If you are measuring **weight**, then the **SD** for the data would be in **Kg**.

➢ Standard deviation takes the units of measure of the data that it represents.

- Standard deviation tells us about how the data is distributed about the mean value.

Steps for Calculating Standard Deviation

1. Calculate the difference between each value and the mean.
2. Square each difference.
3. Add the squared differences.
4. Divide this total by n-1 to get the sample variance.
5. Take the square root of the sample variance to get the sample standard deviation.

Sample Data (X_i) : 10 12 14 15 17 18 18 24

$n = 8$ Mean $= \overline{X} = 16$

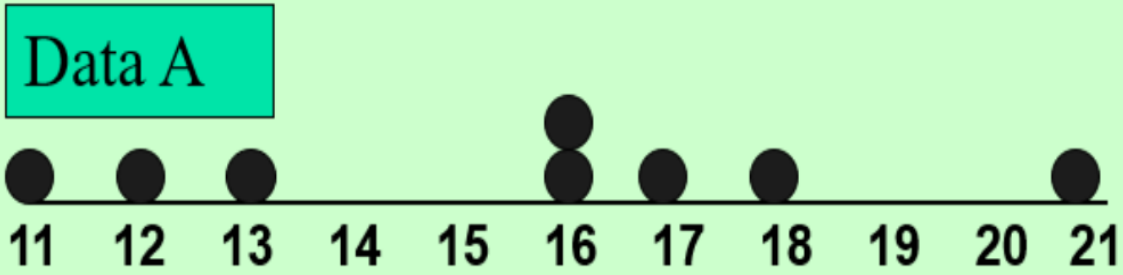
$$S = \sqrt{\frac{(10-\overline{X})^2 + (12-\overline{X})^2 + (14-\overline{X})^2 + \cdots + (24-\overline{X})^2}{n-1}}$$

$$= \sqrt{\frac{(10-16)^2 + (12-16)^2 + (14-16)^2 + \cdots + (24-16)^2}{8-1}}$$

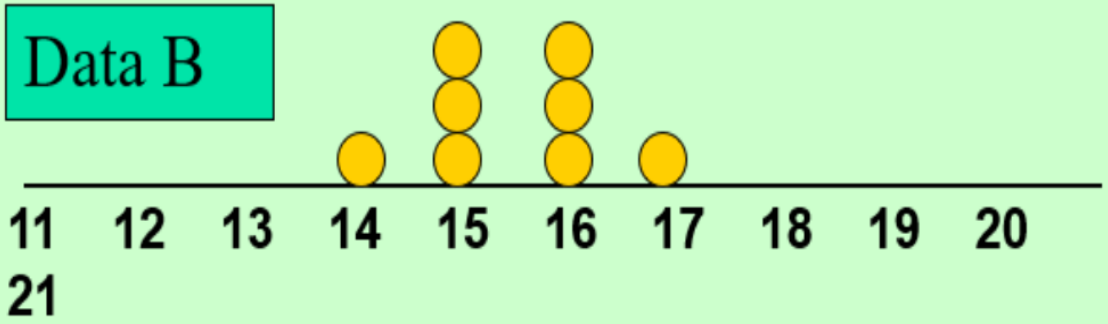
$$= \sqrt{\frac{130}{7}} = 4.3095$$

→ A measure of the "average" scatter around the mean

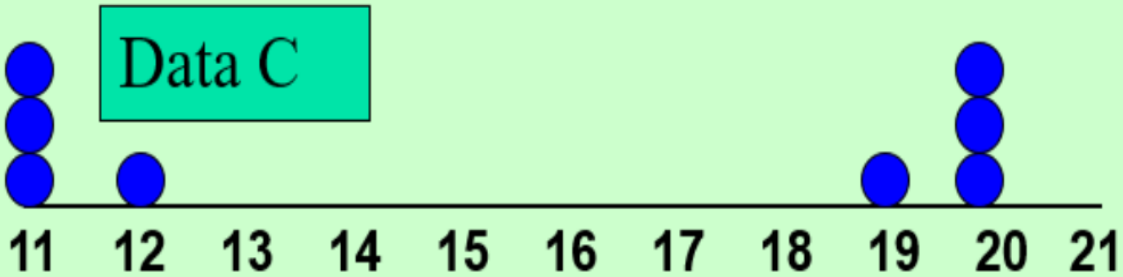
Comparing Standard Deviations



Mean = 15.5
S = 3.338

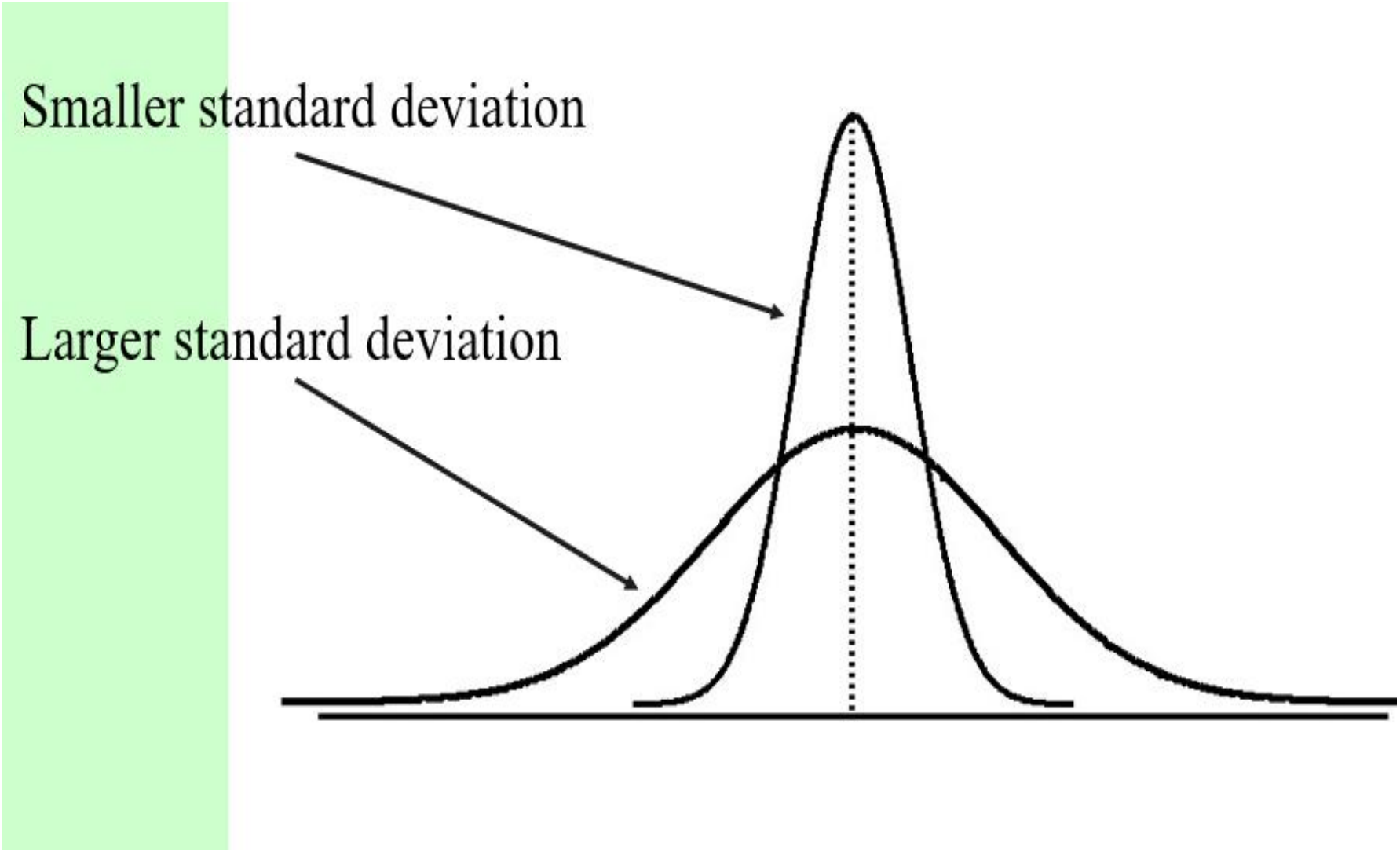


Mean = 15.5
S = 0.926



Mean = 15.5
S = 4.570

Comparing Standard Deviations



Standard Deviation ..

- Instead, we use:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$$

- This is the "definitional formula" for standard deviation.
- The standard deviation has lots of nice properties, including:
  - By squaring the deviation, we eliminate the problem of the deviations summing to zero.
  - In addition, this sum is a minimum. No other value subtracted from X and squared will result in a smaller sum of the deviation squared. This is called the "least squares property."
- Note we divide by (n−1), not n. This will be referred to as a loss of one degree of freedom.

Standard Deviation …

- **Standard Deviation** is a standardized measure of dispersion of the data around the mean, mathematically the standard deviation is the square root of the variance.
  - Interval, and ratio data.

Body Temperature	
Patient Name	Temp.
001	**37**
002	37
003	38
004	38.5
005	38.5
Mean	**37.8**

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

$$s = \sqrt{0.575} = 0.758$$

Standard Deviation..

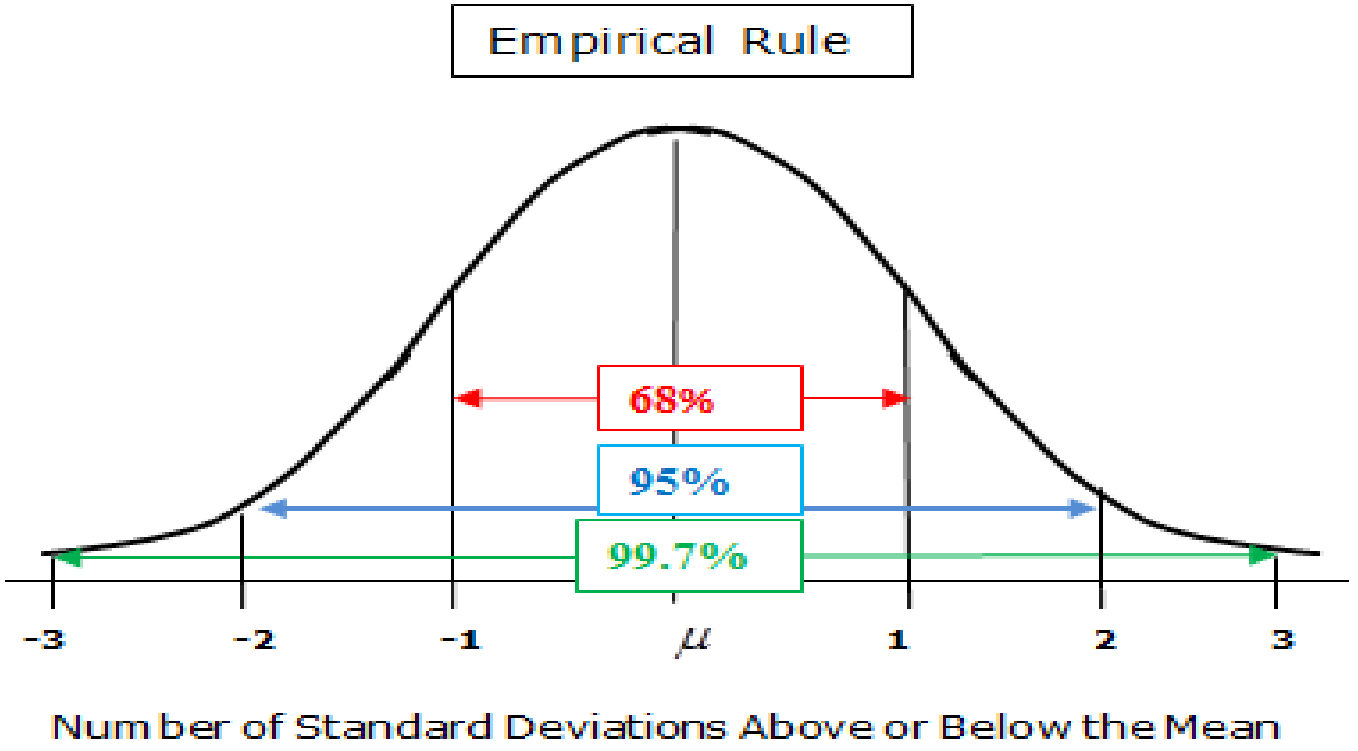
- The **smaller the standard deviation**, **the better is the mean as the summary of a typical score**.

- **Example 1:** 10 people weighted 150 kg, the SD would be zero, and the mean of 150 would communicate perfectly accurate information about all the participants wt.

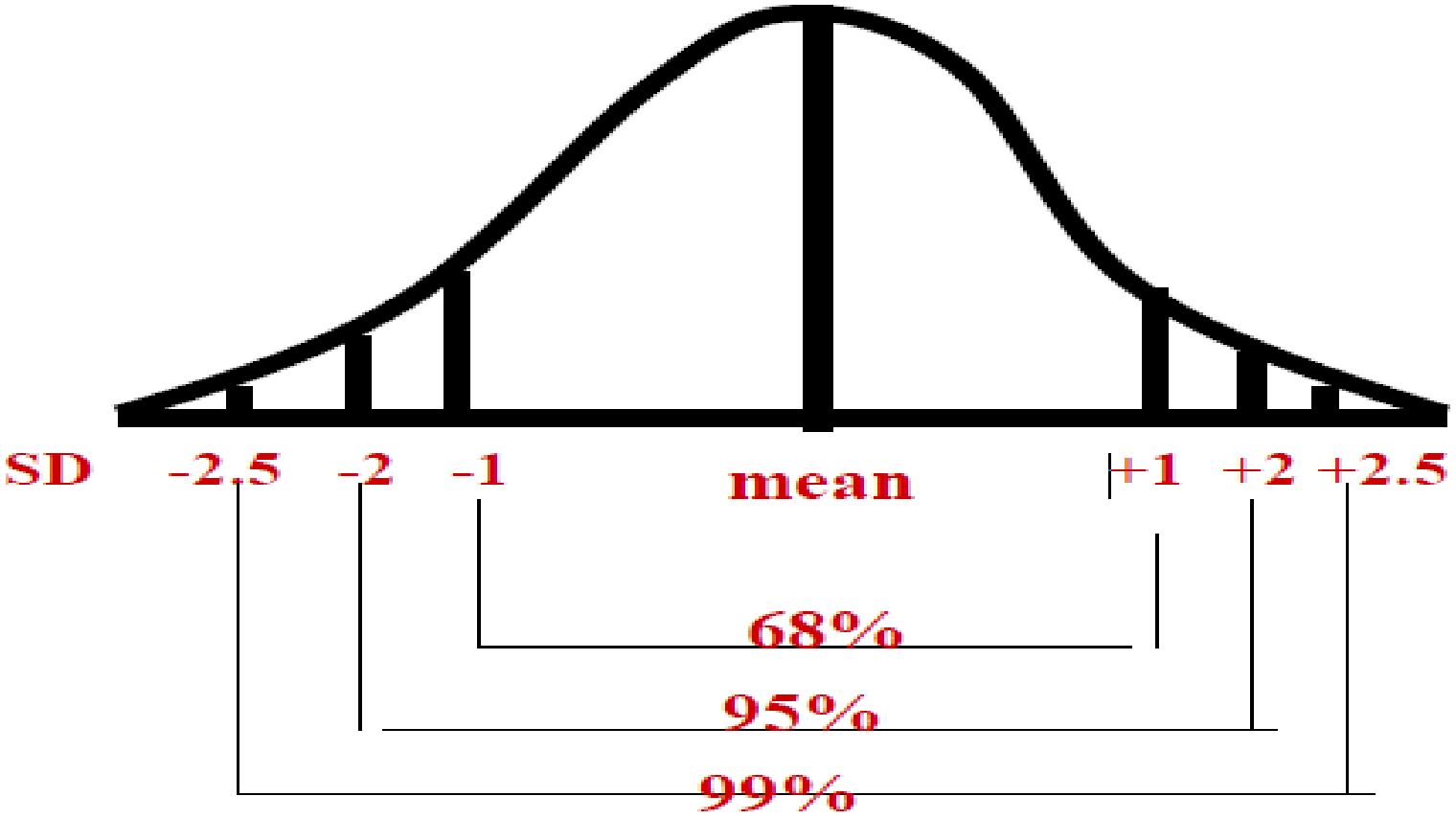
- **Example 2 :** Would be a heterogeneous sample 5 people 100 kg and another five people 200 kg. The mean still 150, but the SD would be 52.7.

Standard Deviation…

- In **normal distribution** there are **3 SDs above** the mean and **3 SDs below** the mean.



Relationship between SD and frequency distribution



Standard Deviation..

Example. Two data sets, X and Y. Which of the two data sets has greater variability? Calculate the standard deviation for each.

We note that both sets of data have the same mean:

$\overline{X} = 3$

$\overline{Y} = 3$

X_i	Y_i
1	0
2	0
3	0
4	5
5	10

(continued…)

Standard Deviation …

$$S_X = \sqrt{\frac{10}{4}} = 1.58$$

X	$\bar{X}$	$(X-\bar{X})$	$(X-\bar{X})^2$
1	3	-2	4
2	3	-1	1
3	3	0	0
4	3	1	1
5	3	2	4
		$\Sigma=0$	10

$$S_Y = \sqrt{\frac{80}{4}} = = 4.47$$

Y	$\bar{Y}$	$(Y-\bar{Y})$	$(Y-\bar{Y})^2$
0	3	-3	9
0	3	-3	9
0	3	-3	9
5	3	2	4
10	3	7	49
		$\Sigma=0$	80

[Check these results with your calculator.]

Standard Deviation: N vs. (n-1)

Note that $\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N}}$ and $s = \sqrt{\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-1}}$

- You divide by N only when you have taken a census and therefore know the population mean. This is rarely the case.
- Normally, we work with a sample and calculate sample measures, like the sample mean and the sample standard deviation:
- The reason we divide by n−1 instead of n is to assure that *s* is an *unbiased* estimator of σ.
  - We have taken a shortcut: in the second formula we are using the sample mean, $\overline{X}$, a statistic, in lieu of μ, a population parameter. Without a correction, this formula would have a tendency to understate the true standard deviation. We divide by n−1, which increases *s*. This makes it an *unbiased estimator* of σ.
  - We will refer to this as "losing one degree of freedom" (to be explained more fully later on in the course).

Variance

- **The variance** is the average is the squared differences between each data value and the mean.

Average (approximately) of squared deviations of values from the mean

- Sample variance:

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$$

Where $\overline{X}$ = arithmetic mean

n = sample size

X_i = i^{th} value of the variable X

Variance…

- **Variance** is a **measure of dispersion of the data around the mean, mathematically the variance is the average squared deviation from the mean.**
  - Interval, and ratio data.

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

Body Temperature	
Patient Name	Temp.
001	**37**
002	37
003	38
004	38.5
005	38.5
Mean	**37.8**

$$s^2 = \frac{(37-37.8)^2 + (37-37.8)^2 + (38-37.8)^2 + (38.5-37.8)^2 + (38.5-37.8)^2}{5-1}$$

$$s^2 = \frac{2.3}{4} = 0.575$$

Variance…

- It is **no actual unit** because of the square.

- **(n-1)**: is an indirect measure of variability , its called the **degree of freedom.**

- **Why do we square the result?**

➢ If we calculate the total differences between the values and the mean it will be **zero** ; so we square it to get the actual values.

Range, variance, and standard deviation.

- The **more** the data are spread out, the **greater** the range, variance, and standard deviation.

- The **less** the data are spread out, the **smaller** the range, variance, and standard deviation.

- If the values are **all the same** (no variation), all these measures will be zero.

Coefficient of Variation (C.V)

- **The coefficient of variation** is a **measure of comparing of two dispersions or more,** mathematically the standard deviation is divided by the mean.
- It is **relative variability** around the mean.
- This can be used to compare two distributions directly to see which has more dispersion because it **does not depend on units of the distribution.**

$$C.V. = \frac{s}{\bar{x}} * (100)$$

	Sample 1	Sample 2
Age	**25 years**	**11 Years**
Mean Weight	145	80
SD	10	10
C.V.	**6.9**	**12.5**

Properties of the Coefficient of Variation

1. Useful for **comparing the variability** of two or more variables (compare between different things).

2. It is **independent of unit** of measurement.

3. Measures the **relative variation**.

4. Always in **percentage (%).**

Coefficient of Variation..

- Stock A:
  - Average price last year = $50
  - Standard deviation = $5

$$CV_A = \left(\frac{S}{\bar{X}}\right) \cdot 100\% = \frac{\$5}{\$50} \cdot 100\% = 10\%$$

- Stock B:
  - Average price last year = $100
  - Standard deviation = $5

$$CV_B = \left(\frac{S}{\bar{X}}\right) \cdot 100\% = \frac{\$5}{\$100} \cdot 100\% = 5\%$$

Both stocks have the same standard deviation, but stock B is less variable relative to its price

Example: Stock Prices

Which stock is more volatile?

Closing prices over the last 8 months:

$$CV_A = \frac{\$1.62}{\$1.70} \times 100\% = 95.3\%$$

$$CV_B = \frac{\$11.33}{\$188.88} \times 100\% = 6.0\%$$

	Stock A	Stock B
JAN	$1.00	$180
FEB	1.50	175
MAR	1.90	182
APR	.60	186
MAY	3.00	188
JUN	.40	190
JUL	5.00	200
AUG	.20	210
Mean	$1.70	$188.88
s^2	2.61	128.41
s	$1.62	$11.33

Answer: The standard deviation of B is higher than for A, but A is more volatile:

Inter-Quartile Range-IQR

- **The interquartile rang** of a data set is the difference between the **third quartile** and the **first quartile.**

- The Interquartile range (IQR) is the score at the **75th percentile or 3rd quartile (Q3) minus the score at the 25th percentile or first quartile (Q1).** Are the most used to define **outliers.**

- ***It is not sensitive to extreme values.***

Inter-Quartile Range-IQR..

- IQR = Q3 – Q1
- **Example:**

➢ (n = 15): **0, 0, 2, 3, 4, 7, 9, 12, 17, 18, 20, 22, 45, 56, 98**

✓ Q1 = 3, Q3 = 22

✓ **IQR** = 22 – 3 = 19 (Range = 98).

- 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.

Exercise: Test Scores

Data (n=10): 0, 0, 40, 50, 50, 60, 70, 90, 100, 100

Compute the mean, median, mode, quartiles (Q1, Q2, Q3), range, interquartile range, variance, standard deviation, and coefficient of variation. We shall refer to all these as the descriptive (or summary) statistics for a set of data.

Answer. First order the data:

0, 0, 40, 50, 50 ¦ 60, 70, 90, 100, 100

- Mean: $\Sigma X_i = 560$ and $n = 10$, so $\overline{X} = 560/10 = 56$.
  Median $= Q_2 = 55$
- $Q_1 = 40$; $Q_3 = 90$ (Note: Excel gives these as $Q_1 = 42.5$, $Q_3 = 85$.)
- Mode = 0, 50, 100
  Range = 100 – 0 = 100
- IQR = 90 – 40 = 50
- $s^2 = 11{,}840/9 = 1315.5$
- $s = \sqrt{1315.5} = 36.27$
- CV = (36.27/56) x 100% = 64.8%

Relative Standing

- It provides information about the **position of an individual score value within a distribution scores.**

- **Two types**:

➢ Percentile Ranks.

➢ Standard Scores

Percentile Ranks

- It is the **percentage of scores** in the distribution that **fall at or below a given value.**

- P = Number of scores less than a given score divided by total number scores X 100.

- **Example:**

✓ Suppose you received a score of 90 on a test given to a class of 50 people.
✓ Of your classmates, 40 had scores lower than 90.
✓ P = 40/50 X 100 = 80.
✓ YOU achieved a higher score than 80% of the people who took the test, which also means that almost 20% who took the test did better than you.

- Percentiles are symbolized by the **letter P**, with a subscript indicating the percentage below the score value.

- Hence, P60 refers to the 60th percentile and stands for the score below which 60% of values fall.

Percentile Ranks

- The statement P40= 55 means that 40% of the values in the distribution fall below the score 55.

- The are several interpercentile measures of variability. The most common being the Interquartile range (IQR).

Standard Scores

- There are scores that are expressed in **terms of their relative distance from the mean**.

- It provides information **not only about rank but also distance between scores.**

- **It often called Z-score.**

Z Score

- Is a standard score that indicates how many SDs from the mean a particular values lies.

- Z = Score of value – mean of scores divided by standard deviation.

Standard Normal Scores

- How many standard deviations away from the mean are you?

Standard Score (Z) = $\frac{\textit{Observation} - \textit{mean}}{\textit{Standard deviation}}$

"Z" is normal with mean 0 and standard deviation of 1.

Standard Normal Scores

- Example: Male Blood Pressure, mean = 125, s = 14 mmHg
  - BP = 167 mmHg
  - BP = 97

$$Z = \frac{167 - 125}{14} = 3.0$$

$$Z = \frac{97 - 125}{14} = -2.0$$

What is the Usefulness of a Standard Normal Score?

- It tells you **how many SDs** (s) an observation is from the mean.
- Thus, it is a way of quickly **assessing how "unusual" an observation** is.

Example: Suppose the mean BP is 125 mmHg, and standard deviation = 14 mmHg

- Is 167 mmHg an unusually high measure?
- If we know Z = 3.0, does that help us?

Standardizing Data: Z-Scores

To compute the Z-scores:

$$Z = \frac{X - \bar{X}}{s}$$

Example.

Data: 0, 2, 4, 6, 8, 10

$\bar{X} = 30/6 = 5$; s = 3.74

x	→	z
0	$\frac{0-5}{3.74}$	-1.34
2	$\frac{2-5}{3.74}$	-.80
4	$\frac{4-5}{3.74}$	-.27
6	$\frac{6-5}{3.74}$	.27
8	$\frac{8-5}{3.74}$	.80
10	$\frac{10-5}{3.74}$	1.34