Chapter 3
Describing Data Using Numerical Measures

Chapter Goals:

After completing this chapter, you should be able to:

- Compute and interpret the mean, median, and mode for a set of data.
- Compute the range, variance, and standard deviation and know what these values mean.
- Construct and interpret a box and whiskers plot.
- Compute and explain the coefficient of variation and z scores.
- Use numerical measures along with graphs, charts, and tables to describe data.

- Measures of Centeral Tendency (or Location).
  - ✓ Mean, median, mode, geometric mean, midrange.
- Other measures of Location.
  - ✓ Weighted mean, percentiles, quartiles.
- Measures of Variation.
  - ✓ Range, interquartile range, variance and standard deviation, coefficient of variation.

3.1 : Measures of Center:

Depending on whether we work with a population or sample, a numerical measure is either a parameter or statistic.

Parameter: A measure computed from the entire population. As long as the population does not change, the value of the parameter will not change.

Statistic: A measure computed from a sample that has been selected from a population. The value of the statistic will depend on which sample is selected.

Measures of Center (or Location):

1. **The Mean (Average)**:

It is the arithmetic average of data values.

a. Population mean:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

Where: μ: the population mean (mu)

N: population size

x_i: ith individual value of variable x

b. Sample mean:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Where: $\bar{x}$: the sample mean (x-bar)

n: sample size

Example 3.1 page 76: Compute the population mean.

$x_1 = \$42,000 \quad x_2 = \$23,900 \quad x_3 = \$115,600 \quad x_4 = \$13,800 \quad x_5 = \$7,900$

$x_6 = \$41,000 \quad x_7 = \$52,900 \quad x_8 = \$76,100 \quad x_9 = \$5,800 \quad x_{10} = \$33,200$

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{42000 + 23900 + \cdots + 33200}{10} = \frac{412200}{10} = 41220$$

The mean sales in spain is \$41,220

Example Consider a sample of bottle bursting strength data of a set of 5 soft drink bottles

251 255 254 253 252

Find the sample mean.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{251 + \cdots + 252}{5} = \frac{1265}{5} = 253$$

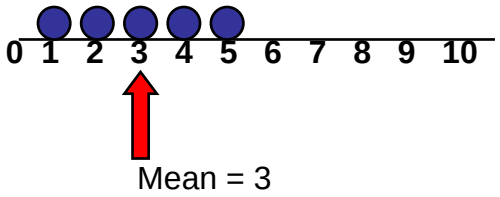
Example: Consider the following humidity readings rounded to the nearest percent:

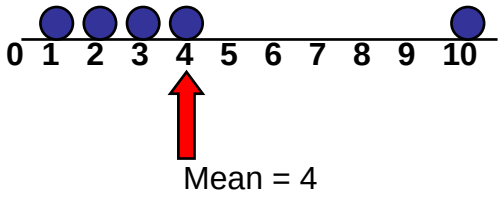
29	44	12	53	21	34	39	25	48	23
17	24	27	32	34	15	42	21	28	37

Find the *sample* mean.

Notes:

- ✓ The mean is the most common measure of central tendency.
- ✓ Mean = sum of values divided by the number of values.
- ✓ Affected by extreme values (outliers).
- ✓ The sum of deviations from the mean is **ZERO**.





$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$ $\qquad$ $\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$

And the sum of deviation

$= \sum (x_i - \bar{x}) = (1-3) + (2-3) + (3-3) + (4-3) + (5-2) = 0$

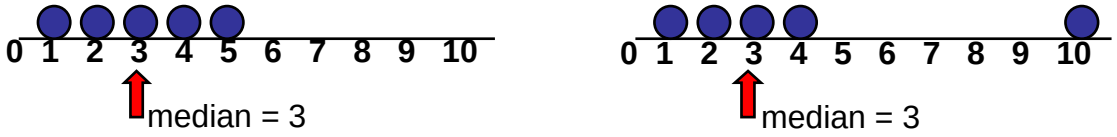
2. **Median:**

The median of a sample of n observations $x_1, x_2, ..., x_n$ is the *middle observation when the observations are arranged* in ascending or descending order if the number of observations is odd. If the number of observations is even, it is the average of the middle two observations. In other words, for any sample of size n, the sample median (M_d) is given by

$$M_d = \begin{cases} \left(\frac{n+1}{2}\right)\text{th observation} & \text{, if } n \text{ is odd} \\ \dfrac{\left(\frac{n}{2}\right)\text{th observation} + \left(\frac{n+2}{2}\right)\text{th observation}}{2} & \text{, if } n \text{ is even} \end{cases}$$

Notes:

1. The population median denoted by $(\tilde{\mu})$.
2. Not affected by extreme values.
3. Computed from the center of the values.
4. Does not use information from all the data.



Example: Marks obtained by 6 students in STAT 211 are given below

81 82 98 83 80 85.

Find the median.

Example: Marks obtained by 7 students in STAT 211 are given by

81 82 98 83 80 85 82.

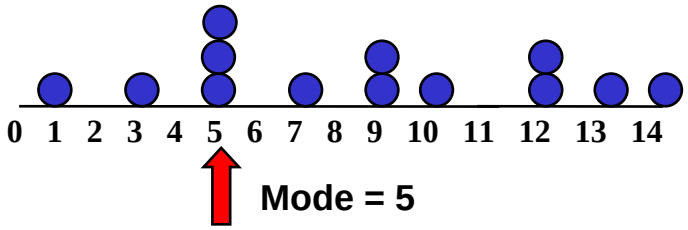
Find the median.

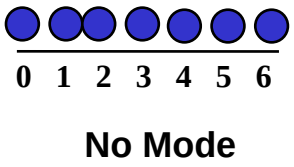
3. **Mode:**

The mode of a sample is the observation occurring the maximum number of times i.e. the observations with the *largest frequency*.

Notes:

- A measure of central tendency.
- Value that occurs most often.
- Not affected by extreme values.
- Used for either numerical or categorical data.
- There may be no mode.
- There may be several modes.





Example: The following samples provide prices, in Saudi Riyals (SR), of a computer monitor.

a) 1200, 1000, 1500, 1200, 1000, 1200
b) 1300, 1200, 1000

What is the modal price for each part?

Exampl:e The following table shows the hourly wages in SR earned by the employees of a small company and the number of employees who earn each wage.

Wages/hour	6	8	10	13
No. of employees	3	5	4	4

Then find the modal **wage** per hour.

Example: Assume that the grades for 8 students are A+, A, D+, B, B, D+,F, D+, find the mode.

4. **Weighted Mean**:

Used when values are grouped by *frequency* or *relative importance*.

a. Population mean:

$$\mu_w = \frac{\sum w_i x_i}{\sum w_i}$$

w_i: the weight of i^{th} value

b. Sample mean:

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

Example: For a sample of 26 Repair Projects, find the mean (the average number of days needed to complete a project).

Days to complete x_i	No. of projects w_i	$w_i\ x_i$
5	***4***	***20***
6	***12***	***72***

7	*8*	*56*
8	*2*	*16*
Total	**26**	**164**

Exampl:e The following table shows the hourly wages in SR earned by the employees of a small company and the number of employees who earn each wage.

Wages/hour x_i	6	8	10	13	**Total**
No. of employees w_i	3	5	4	4	**16**
$\mathbf{x_i\ w_i}$	18	40	40	52	**150**

The find the **mean wage** per hour.

Which measure of location is the "best"?

1. The ***mean:*** is generally used, unless extreme values (outliers) exist.
2. The ***median:*** is often used, since the median is not sensitive to extreme values.

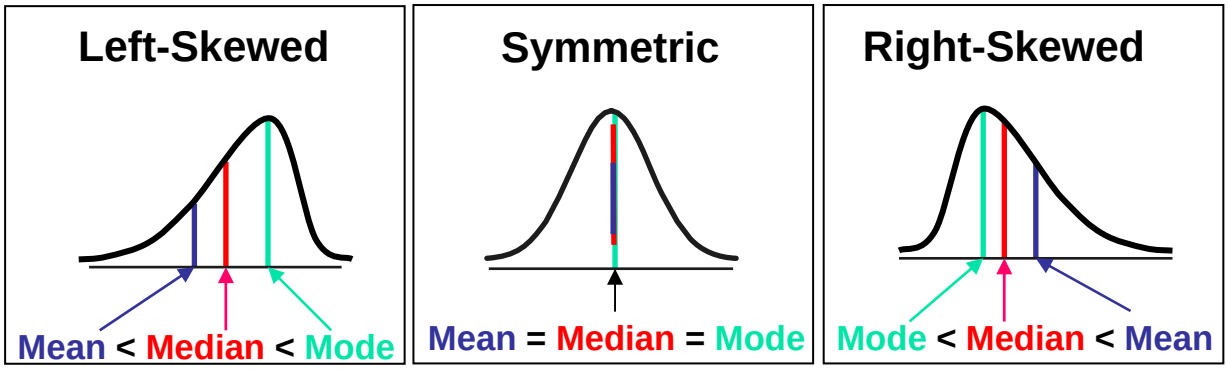
The Shape of a Distribution:

Describes how data is distributed, Symmetric or skewed.

Symmetric data: the data spread, uniformly or regularly, around the center.
Skewed data: the data that are not symmetric.

Left – skewed: $\bar{x} < M_d < Mode$

Right – skewed: $\bar{x} > M_d > Mode$



(Longer tail extends to left)

(Longer tail extends to right)

Other Location Measures:

1. **Pecentiles:** The α^{th} percentile (P_α) is the value that exceeds $\alpha\%$ of the data, and is obtained by the following steps:

Step 0: **Sort**: Order the observations in an ascending manner.

Step 1: **Locate**, determine the Rank (***the location***) of, the percentile:

$$R_\alpha = \alpha\ (n+1)/100, \quad \alpha = 1,\ 2,\ \cdots, 99.$$

Step 2: **Separate** i (the largest integer not exceeding R_α) and the decimal part (d) of R_α and write $R_\alpha = i + d$. If R_α is an integer then the α^{th} percentile is the ith observation $P_\alpha = X_{(i)}$.

Step 3: **Calculate** the α^{th} percentile is then given by

$$P_\alpha = x_{(i)} + d\left(x_{(i+1)} - x_{(i)}\right) = (1-d)\,x_{(i)} + d\,x_{(i+1)}, \qquad \alpha = 1, 2, ..., 99,$$

2. **Quartiles:** Those values that divided the data set into four equal size qroup. where $x_{(i)}$ is the ith observation after ordering the observations ascendingly.

The 25th percentile is called the 1st quartile and is denoted by Q_1.

The 50th percentile is called the 2nd quartile and is denoted by Q_2.

The 75th percentile is called the 3rd quartile and is denoted by Q_3.

Note:

- ✓ The 50th percentile is called the 2nd quartile Q_2 is equal to the median
- ✓ The αth percentile in a data array, where $(0 \le \alpha \le 100)$
  - ➢ $\alpha\%$ are less than or equal to this value.
  - ➢ $(1-\alpha)\%$ are greater than or equal to this value

Example: An independent consumer group tested radial tires from a major brand to determine expected tread life. The data (in thousands of miles) are given below:

50	54	52	47	56	51	51
48	56	53	43	56	58	42

Find the 1st, 2nd , 3rd quartiles, the 90th percentile and D_2.

Solution: 42 43 47 48 50 51 51 52 53 54 56 56 56 58.

Q_1 = 47.75, Q_2 = 51.5, Q_3 = 56.

Data level Issues:

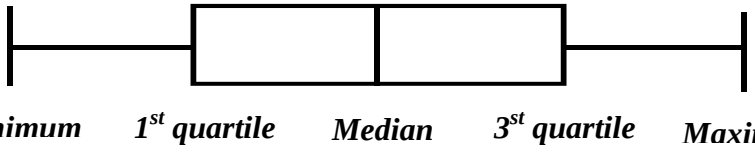
You need to be very aware of the level of data you are working with before computing the numerical measures.

1. We can find the mean and the percentiles for ratio and interval data level.
2. We can find the median for ratio, interval and ordinal data level.
3. We can find the mode for ratio, interval, ordinal and nominal data level.

Box and Whisker Plot (5-number summary)

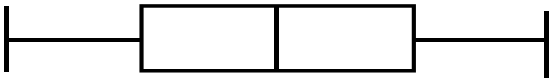
A Graphical display of data using:

Minimum -- Q_1 -- Median -- Q_3 -- Maximum



Shape of Box and Whisker Plots

- ✓ The Box and the line inside are *centered* between the endpoints if the data are symmetric around the median.



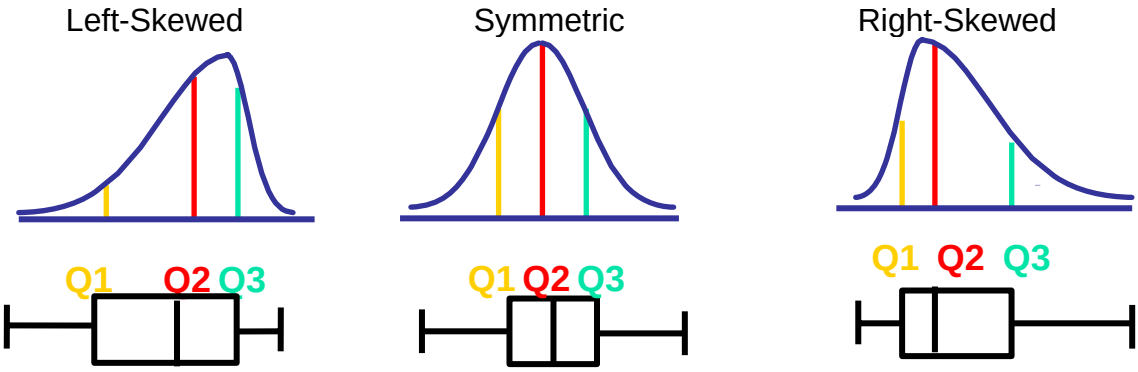
- ✓ A Box and Whisker plot can be shown in either vertical or horizontal format
- ✓ The Inter Quartile Range $IQR = Q_3 - Q_1$

To construct a box and whisker plot, follow the following steps

Step 1: Find Q1, Q2, Q3.

Step 2: Extend the line to both directions down to the minimun and up to the maximun, any value left outside is considered as an *outlier*.

Distribution Shape and Box and Whisker Plot:



Example: Refer to the last example, develop a box and whisker plot.

Example Q3.12 page 93:

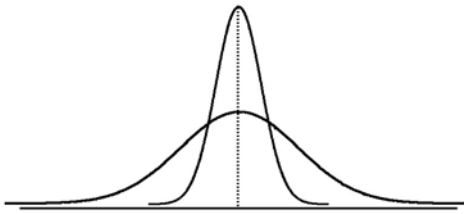
Examine the following data:

23	65	45	19	35	28	39	100	50	26	25	27
24	17	12	106	23	19	39	70	20	18	44	31

a. Compute the quartiles.
b. Calculate the 90th percentile.
c. Develop a box and whisker plot.
d. Calculate the 20th and the 30th percentiles.

3.2 : Measures of Variation (Dispersion):

- ✓ The variation happenes when all the data are not the same.
- ✓ Measures of variation give information on the **spread** or **variability** of the data values.



Same center , different variation

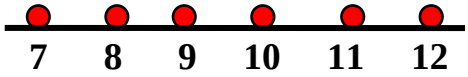
1. Range:

- ✓ Simplest measure of variation
- ✓ Difference between the largest and the smallest observations:

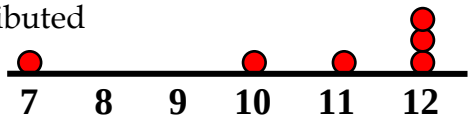
$$Range = Max - Min$$

Disadvantages of the Range

1. Ignores the way in which data are distributed



Range = 12 - 7 = 5



Range = 12 - 7 = 5

2. Sensitive to outliers

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

Range = 5 - 1 = 4

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

Range = 120 - 1 = 119

2. Interquartile Range:

- ✓ It is a measure of variation.
- ✓ Interquartile range = 3rd quartile – 1st quartile.
- ✓ Can eliminate some outlier problems by using the interquartile range .
- ✓ IQR will not change if the max and the min are changed.

Example Q2.12 page 93: Find IQR.

3. Variance:

It is the average of squared deviations of the values from the mean.

a. Population Variance:

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

b. Sample Variance:

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

Note: the variance has a squared unit of the data.

4. Standard Deviation (StDev):

It is the positive square root of the variance. It measures the variation in a set of data. It give an indication how spread out a distribution.

a. Population Standard Deviation:

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

b. Sample Standard Deviation:

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

Note:

- ✓ Most commonly used measure of variation.
- ✓ Shows variation about the mean.
- ✓ Has the same unit as the original data.

Example

Sample data (x_i) : 10 12 14 15 17 18 18 24

Compute the variance and the standard deviation.

Example: You are given the following data for the number of times a population of six families dined out during the previous month

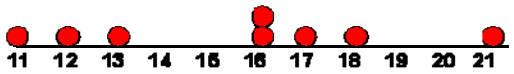
4 6 9 4 5 7

a) Compute the range.
b) Compute the variance and the standard deviation.
c) Assume that these data represented a sample rather than a population. Compute the variance and the standard deviation.

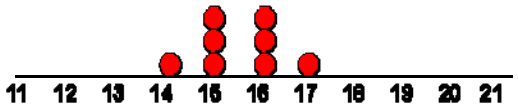
Comparing Standard Deviations

Assume that there is three samples

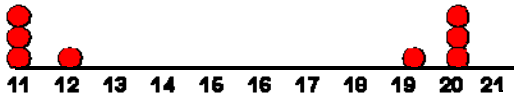
Sample A: the mean = 15.5, and the std = 3.338



Sample B: the mean = 15.5, and the std = 0.9285



Sample C: the mean = 15.5, and the std = 4.57



Then Sample B has the smallest std, which means that the observations are closer to each other more than the other samples.

3.3 : Using the mean and the standard deviation together:

If two sets of data have the same mean, then the set with larger standard deviation has the greater relative spread. But if the two sets of data have different means, then relative variation cannot be determined by comparing standard deviation.

Coefficient of Variation (C.V):

It is the ratio of the standard deviation to the mean. It is used to measure the *relative variation* for distributions with different means.

When C.V of one set of data is more than the other one, this mean that the data has greater relative variation (relative spread).

So, coefficient of variation is:

- ✓ Measures relative variation.
- ✓ Always in percentage (%).
- ✓ Shows variation relative to mean.
- ✓ It is UNITLESS.
- ✓ Is used to compare two or more sets of data measured in different units

a. Population C.V:

$$CV = \left(\frac{\sigma}{\mu} \right) \cdot 100\%$$

b. Sample C.V:

$$CV = \left(\frac{s}{\bar{x}} \right) \cdot 100\%$$

Example:

Assume that there is two type of stocks.
Stock A: Average price last year = \$50 and Standard deviation = \$5
Stock B: Average price last year = \$100 and Standard deviation = \$5
Which one have less variation to its price?

Solution:

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

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

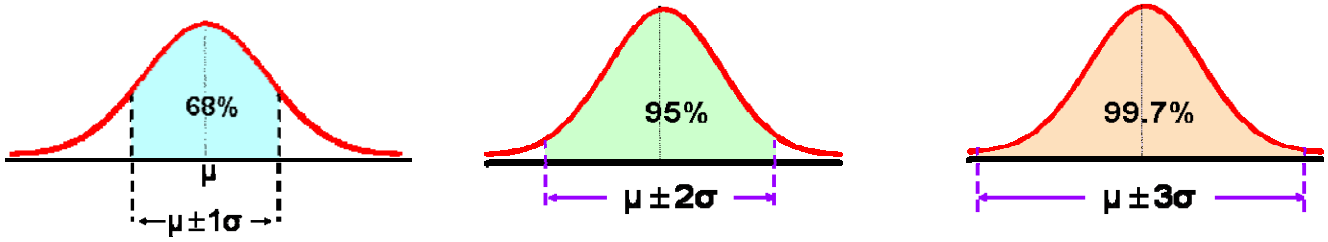
Both stocks have the same standard deviation, but stock B is less variable relative to its price

Example Q3.35 page 110:

The Empirical Rule:

If the data distribution is *bell-shaped*, then the interval:

1. $\mu \pm 1\sigma$ contains about 68% of the values in the population or the sample.
2. $\mu \pm 2\sigma$ contains about 95% of the values in the population or the sample
3. $\mu \pm 3\sigma$ contains about 99.7% of the values in the population or the sample.



Tchebysheff's Theorem

Regardless of how the data are distributed, *at least* $\left(1-\frac{1}{k^2}\right)$ *of the values will fall within k standard deviations of the mean* (fall within [μ±kσ]), i.e.

If k = 1: at least $(1 - 1/1^2) = 0\%$ of the data will be within (μ ± 1σ).
If k = 2 at least $(1 - 1/2^2) = 75\%$ of the data will be within (μ ± 2σ).
If k = 3 at least $(1 - 1/3^2) = 89\%$ of the data will be within (μ ± 3σ).

Example Q3.34 page 110

Standardized Data Values

A standardized data value refers to the *number of standard deviations a value is from the mean*. Standardized data values are sometimes referred to as z-scores N-scores (normal scores).

- It is used to *compare* two or more distributions when the data scales are different.
- If $z > 0$: then x is *above* μ by z standard deviations.
- If $z < 0$: then x is *below* μ by z standard deviations.
- If $z = 0$: then x the same as μ.

Standardized Population Values:

$$z = \frac{x - \mu}{\sigma}$$

Where:

- ✓ x = original data value
- ✓ μ = population mean
- ✓ σ = population standard deviation
- ✓ z = standard score (number of standard deviations x is from μ)

Standardized Sample Values

$$z = \frac{x - \bar{x}}{s}$$

where:

- ✓ x = original data value
- ✓ $\bar{x}$ = sample mean
- ✓ s = sample standard deviation
- ✓ z = standard score (number of standard deviations x is from $\bar{x}$)

Example: Two distributions have the following characteristics:
Distribution A: μ = 45,600 and σ = 6,333
Distribution B: μ = 33.4 and σ = 4.05
If a value from distribution A is 50,000 and a value from distribution B is 40, indicate which one is *relatively closer to its respective mean*.
Solution:

Distribution A: $z_A = \frac{x - \mu_A}{\sigma_A} = \frac{50,000 - 45,600}{6,333} = 0.695$

Distribution B: $z_B = \frac{x - \mu_B}{\sigma_B} = \frac{40 - 33.40}{4.05} = 1.63$

The smaller the z value, the relatively closer the x value is to its respective mean.

1. The 50,000 value is .6948 standard deviations from the mean of distribution A.
2. The value 40 is 1.6296 standard deviations from the mean of distribution B.

Therefore, the value from distribution A is *relatively closer* to its mean.