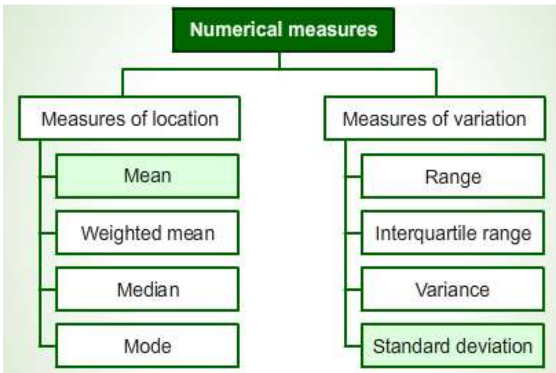
Learning Outcome

By the end of this lecture, you will be able to:

- Compute and explain the coefficient of variation and the Z scores
- Examine the data shape
- Use Tchebysheff's Theorem

Introduction

Recall the measures of location and measures of variation.



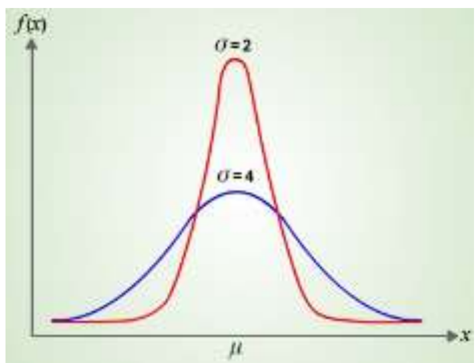
Measures of location and Measures of variation

In this lecture, let's explore the use of mean and the standard deviation together.

Using the Mean and Standard Deviation Together

If two sets of data have the same mean, then the set with the larger standard deviation has the greater relative spread.

However, if the two sets of data have different means, then relative variation cannot be determined by comparing the standard deviations. In this case, we use the coefficient of variation.



Distribution of data with same mean and different standard deviation

Coefficient of Variation

The coefficient of variation is the ratio of the standard deviation to the mean. It has no units and is always given in percentage (%).
It is used to
- Compare two or more sets of data measured in different units
- Measure the relative variation for distributions with different means

The coefficient of variation shows variation relative to the mean. The greater the coefficient of variation, the greater is the relative variation or relative spread.

Types of coefficient of variation

Population coefficient of variation	$CV = \left(\frac{\sigma}{\mu}\right).100\%$ μ is the population mean σ is the Population standard deviation
Sample coefficient of variation	$CV = \left(\frac{S}{\bar{x}}\right).100\%$ $\bar{x}$ is the sample mean S is the Sample standard deviation

Example

Assume the average annual percentage rates of return over the past 10 years for two mutual funds are 14.8% and 10.7% with the standard deviations 16.74% and 9.97%.

Which fund would you classify as having the higher relative risk?

Solution:

Coefficient of variation for the first fund is = 16.74/14.8 = 1.1313=113.13%

Coefficient of variation for the second fund is 9.97/10.7 = 0.9316=93.16%

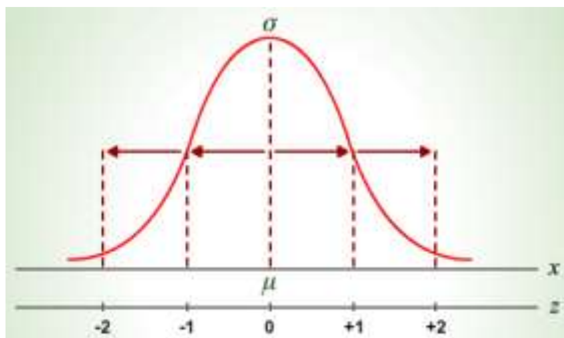
So, the first fund has higher relative risk.

Standardized Data Values

A Standardized data value refers to the number of standard deviations a value is away from the mean. Standardized data values are sometimes referred to as z-scores.

Z-scores are used to:

- Compare two or more distributions when the data scales are different
- Identify outliers:
    - The larger the z – score, the greater is the distance of the value from the mean.
    - If the value of the z – score is less than – 3 or greater than + 3, then the value is an outlier



Standardized data values

If $z > 0$	x is above μ by z standard deviations
If $z < 0$	x is below μ by z standard deviations
If $z = 0$	x is equal to the μ

Standardized Data Values ...Contd.

Standardized data values are two types. They are standardized population values and standardized sample values.

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 ...Contd.

Two distributions have the given 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 30, indicate which one is relatively closer to its respective mean.

Solution:

Distribution A:

$$z_A = \frac{x_A - \mu_A}{\sigma_A} = \frac{50{,}000 - 45{,}600}{6{,}333} = 0.695$$

Distribution B:

$$z_B = \frac{x_B - \mu_B}{\sigma_B} = \frac{30 - 33.40}{4.05} = -0.8395$$

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

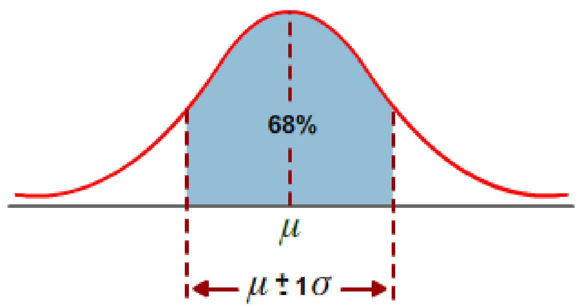
1. The 50,000 value is 0.6948 standard deviations above the mean of distribution A.
2. The value 30 is 0.8395 standard deviations below the mean of distribution B.

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

The Empirical Rule

The empirical rule characterizes the data distribution shape of a population or a sample. If the data distribution is bell-shaped or mound-shaped, then the intervals can be analyzed in three ways:

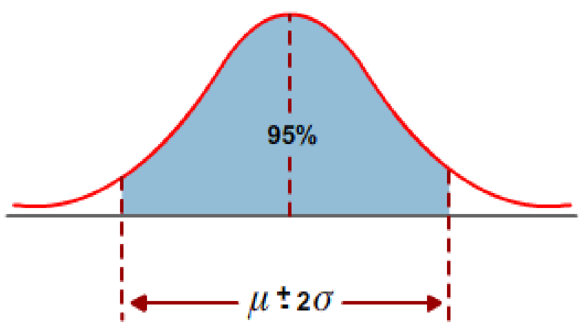
Case 1:



The empirical rule

- $\mu \pm 1\sigma$ contains about 68% of the values in the population
- Use $\bar{x} \pm 1s$ for the sample

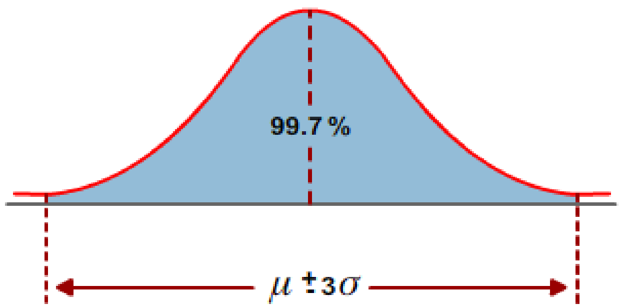
Case 2:



The empirical rule

- $\mu \pm 2\sigma$ contains about 95% of the values in the population
- Use $\bar{x} \pm 2s$ for the sample

Case 3:



The empirical rule

- $\mu \pm 3\sigma$ contains about 99.7% of the values in the population
- Use $\bar{x} \pm 3s$ for the sample

Tchebysheff's Theorem

As per 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 or fall within $[\mu \pm k\sigma]$.

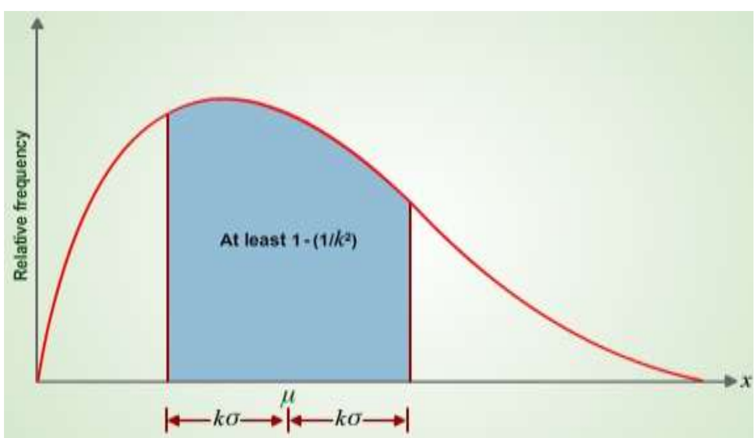


Illustration of tchebysheff's theorem

If $k= 1$:at least $(1 - 1/1^2) = 0\%$ of the data will be within $(\mu \pm 1\sigma)$
If $k= 2$:at least $(1 - 1/2^2) = 75\%$ of the data will be within $(\mu \pm 2\sigma)$
If $k= 3$:at least $(1 - 1/3^2) = 89\%$ of the data will be within $(\mu \pm 3\sigma)$

Example

The scores of 25 students on a STAT examination are listed on the given stem–and–leaf plot.

Stem–and–leaf plot

Stem	Leaf
5	0 3 5 8
6	2 2 4 5 7
7	0 0 3 5 5 8 9
8	3 4 4 6 6 9
9	0 0 4

a. Determine the sample mean and the sample standard deviation.
b. From the shape of the plot, what percentage of the data values would you expect to be within one standard deviation from the mean? Two standard deviations?
c. Find the actual percentages for the intervals given in *b*.

Solution:

a) Mean = 73.68
Standard deviation = 12.79

b) From the shape:
   i. Within one standard deviation about 68% of the data
   ii. Within two standard deviation about 95% of the data

c) Within one standard deviation:

$$[\bar{X} \pm 1S] = [60.88117, 86.47883]$$

Number of observations in the interval = 17
The percentage of observation in the interval = 68%

Within two standard deviations:

$$[\bar{X} \pm 2S] = [48.08234, 99.27766]$$

Number of observation in the interval = 25
The percentage of the observation in the interval = 100%

Recap

In this lecture, you have learned that:

- The coefficient of variation is the ratio of the standard deviation to the mean
- Standardized data values or *Z*-scores refer to the number of standard deviations a value is away from the mean
- As per 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