Topic 3: Descriptive Statistics

1. Measures of Central Tendency.
2. Measures of Variation.
3. Other Measures.
4. Excel Output.

This topic focuses on selected measures used to describe a distribution or data set.

1. Measures of Central Tendency.

0. The measures of central tendency considered here are the following:

   1. Arithmetic mean;
   2. Median;
   3. Mode;
   4. Weighted mean;
   5. Geometric mean.
3.2

1. Arithmetic mean:

1. Sample mean: \( \bar{X} = \frac{\sum X_i}{n} = \frac{X_1 + X_2 + X_n}{n} \)

where: \( \bar{X} \) = sample mean;
\( \sum X_i \) = sum of all the observations in the sample;
\( n \) = sample size; number of observations in the sample.

2. Population mean: \( \mu = \frac{\sum X_i}{N} = \frac{X_1 + X_2 + X_N}{N} \)

where: \( \mu \) = population mean;
\( \sum X_i \) = sum of all the observations in the population;
\( N \) = population size; number of observations in the population.

Interpretation: The arithmetic mean is widely used in various analyses. However, it has the limitation of being sensitive to outliers (extreme values) in a distribution.
2. Median (Mdn):

The median is the middle value in an ordered data set; it is the value above which 50% of the observations fall and below which 50% of the observations fall.

In the case of an odd number of observations, the median is the middle value;
In the case of an even number of observations, the median is the average (arithmetic mean) of the two middle values.

The median position in a data set is given by the following expression:

\[ \text{Median position} = X\left(\frac{n+1}{2}\right) \]

Interpretation: The median is not sensitive to outliers (extreme values) in a distribution. For this reason, the median rather than the arithmetic mean is commonly used in the analysis of income data.
3.4

3. Mode:

The mode is the most frequently observed value in a data set.

Distributions can have no mode, or they can be unimodal, bimodal, or multimodal.

Interpretation: Analysis of modal data is common in the clothing industry where frequent size classifications are important.
3.5

4. Weighted mean:

\[ X_w = \frac{\sum (Xw)}{\sum w} \]

where: \( X_w \) = weighted mean;
\( X \) = an observation;
\( w \) = the weight (appropriate to the observation).

Interpretation: The arithmetic mean assumes that all observations are of equal importance; in this case the implied weight is 1. The weighted mean does not assume that all observations are of equal importance; the weighted mean permits assignment of different weights, and thus different levels of importance, to the observations in the data set.
5. Geometric mean:

The geometric mean gives the average percent change in a percentage change data set.

The geometric mean is found by taking the $n^{th}$ root of the product of $n$ numbers. Thus,

$$GM = \sqrt[n]{X_1 \times X_2 \ldots X_n}$$

where: $GM =$ geometric mean;

$n =$ number of observations in the data set;

$X =$ an observation.

Interpretation: All data in the data set must be positive; data cannot be negative.
3.7

6. Central tendency in sum.
   1. Central tendency as the center of a distribution.
   2. Central tendency in distributions.
      1. The case of a symmetric distribution: $\bar{X} = \text{Mdn} = \text{Mode}$
      2. The case of asymmetric distributions: $X \neq \text{Mdn} \neq \text{Mode}$
2. Measures of Variation.

0. Measures of variation considered here are the following:

1. Range;
2. Variance;
3. Standard deviation;
4. Coefficient of variation;
5. Z-Value.
1. Range:

The range is the difference between the highest and the lowest values in a data set.

Range = Highest value - Lowest value.

Interpretation: The range gives the analyst a quick sense of the spread in a data set. However, it ignores all observations except the two in its definition.
3.10

2. Variance:

1. Sample variance: 

\[ s^2 = \frac{\sum(X - \bar{X})^2}{n-1} \]

where: 
- \( s^2 \) = sample variance; 
- \( X \) = an observation in the sample; 
- \( \bar{X} \) = sample mean; 
- \( n \) = sample size; number of observations in the sample; 
- \( n-1 \) is the degrees of freedom.

2. Population variance: 

\[ \sigma^2 = \frac{\sum(X - \mu)^2}{N} \]

where: 
- \( \sigma^2 \) = population variance; 
- \( X \) = an observation in the population; 
- \( \mu \) = population mean; 
- \( N \) = population size; number of observations in the population.

The respective numerators, \( \sum(X - \bar{X})^2 \) and \( \sum(X - \mu)^2 \), are the sum of the squared deviations from the mean, sum of squares of \( X \), or simply the sum of squares.

Interpretation: The variance is the average of the squared deviations of the observations from their mean.

1. The variance is a measure of average deviation.

2. The variance is expressed in original measurement units squared.
3.11

3. Standard deviation:

1. Sample standard deviation: $s = \sqrt{s^2} = \sqrt{\frac{\sum (X - \bar{X})^2}{n-1}}$

   where: $s$ = sample standard deviation;
   $s^2$ = sample variance;
   $X$ = an observation in the sample;
   $\bar{X}$ = sample mean;
   $n$ = sample size; number of observations in the sample; $n-1$ is the degrees of freedom.

2. Population standard deviation: $\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (X - \mu)^2}{N}}$

   where: $\sigma$ = population standard deviation;
   $\sigma^2$ = population variance;
   $X$ = an observation in the population;
   $\mu$ = population mean;
   $N$ = population size; number of observations in the population.

Again, the respective numerators, $\sum (X - \bar{X})^2$ and $\sum (X - \mu)^2$, are the sum of the squared deviations from the mean, sum of squares of $X$, or simply the sum of squares.

Interpretation: The standard deviation is the average of the deviations of the observations from their mean.

1. The standard deviation is a measure of average deviation.

2. The standard deviation is expressed in original measurement units.
4. Coefficient of variation:

\[ CV = \frac{s}{\bar{X}} \] (100)

where: CV = coefficient of variation;
\[ s = \text{sample standard deviation}; \]
\[ \bar{X} = \text{sample mean}. \]

The coefficient of variation is a measure of relative variation; it is the standard deviation expressed as a percent of the mean.

Interpretation: The CV is useful in comparing two different distributions with respect to variation.
5. Z-Value:

1. Sample Z-value: \[ Z = \frac{X - \bar{X}}{s} \]
   where: \( Z \) = standard normal variable; \( X \) = an observation in the sample; \( \bar{X} \) = sample mean; \( s \) = sample standard deviation.

2. Population Z-value: \[ Z = \frac{X - \mu}{\sigma} \]
   where: \( Z \) = standard normal variable; \( X \) = an observation in the population; \( \mu \) = population mean; \( \sigma \) = population standard deviation.

Interpretation: The standard normal variable, \( Z \), expresses an observation in standard deviation units. A \( Z \) distribution has a mean of 0 and a standard deviation of 1.
6. Variation in sum.
   1. Variation as the spread of a distribution.
   2. Applied variation analysis.
      0. Two cases:
         1. The empirical rule.
         2. Chebyshev's (Tchebyshev's) Theorem.
1. The case of the empirical rule.

The empirical rule is applicable to normal distributions, that is symmetric bell-shaped distributions.

The empirical rule: The empirical rule states the relative frequency (percentage) of observations, within a normal distribution, which fall within a given number of standard deviations of the mean.

The one, two, and three standard deviation cases:

- The interval, $\mu \pm 1\sigma$, includes 68.3% of the observations.
- The interval, $\mu \pm 2\sigma$, includes 95.5% of the observations.
- The interval, $\mu \pm 3\sigma$, includes 99.7% of the observations.

The empirical rule illustrated.

Applications.
2. The case of Chebyshev's Theorem.

Chebyshev's theorem is applicable to any type of distribution, normal or otherwise.

Chebyshev's theorem: For any distribution, the minimum proportion (percentage) of observations that fall within \( k \) standard deviations of the mean is given by:

\[
1 - \left(\frac{1}{k^2}\right)
\]

where: \( k = \) the number of standard deviations from the mean;
\( k > 1 \)

The term, \( 1/k^2 \), gives the maximum proportion (percentage) of observations that fall beyond \( k \) standard deviations of the mean.

<table>
<thead>
<tr>
<th>Interval</th>
<th>Minimum proportion of observations within the interval, ( 1 - \left(\frac{1}{k^2}\right) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( k )</td>
<td>( \mu \pm k\sigma )</td>
</tr>
<tr>
<td>2</td>
<td>( \mu \pm 2\sigma ) \hspace{1cm} ( 1 - (1/2^2) = .75 ) or 75%</td>
</tr>
<tr>
<td>3</td>
<td>( \mu \pm 3\sigma ) \hspace{1cm} ( 1 - (1/3^2) = .89 ) or 89%</td>
</tr>
<tr>
<td>4</td>
<td>( \mu \pm 4\sigma ) \hspace{1cm} ( 1 - (1/4^2) = .94 ) or 94%</td>
</tr>
</tbody>
</table>

Chebyshev's theorem illustrated.

Applications.
3. Other Measures.

0. Other measures considered here are the following:

   1. Skewness.
   2. Kurtosis.
   3. Correlation.
1. Skewness:

Skewness centers on the matter of asymmetry in distributions.

Pearson coefficient of skewness:

\[
Sk = \frac{3(\bar{X} - \text{Mdn})}{s}
\]

where: 
- \(Sk\) = coefficient of skewness;
- \(\bar{X}\) = sample mean;
- \(\text{Mdn}\) = sample median;
- \(s\) = sample standard deviation.

Two types of distributions with respect to asymmetry:

1. Positively skewed distribution:
   Distribution in which the tail is on the right;

   In this case, \(\bar{X} > \text{Mdn}\) \(\Rightarrow\) \(Sk > 0\) (positive).

   Illustrate: Age distribution at a rock concert.

2. Negatively skewed distribution:
   Distribution in which the tail is on the left;

   In this case, \(\bar{X} < \text{Mdn}\) \(\Rightarrow\) \(Sk < 0\) (negative).

   Illustrate: Age distribution at a classical concert.
2. Kurtosis:

Kurtosis centers on the matter of the vertical shape of distributions, that is, their peakedness or flatness.

Three types of distributions with respect to kurtosis:

1. Leptokurtic distribution: Peaked distribution; $K > 0$ (positive).

2. Platykurtic distribution: Flat distribution; $K < 0$ (negative).

3. Mesokurtic distribution: Normal distribution; $K = 0$.

Illustrate each distribution.
3. Correlation:

The correlation coefficient, $r$, is a measure of the relationship between two variables.

The range of values that the correlation coefficient can take on is given by the following:

$$-1 \leq r \leq +1$$

The correlation coefficient, $r$, contains two pieces of information:

1. The type of relationship: direct (+); inverse (-).
2. The strength of the relationship: stronger ($\rightarrow |1|$); weaker ($\rightarrow 0$).

Various illustrations.

1. Direct: strong, moderate, weak.
2. Inverse: strong, moderate, weak.
3. No relationship.

Examples.
4. Excel Output.

Descriptive Statistics in the Analysis Toolpak.