Module 2 - Averages and tons of formulas

Created	
Topics	

- In order to analyse the data that has been organised in tables and graphical form, we need to calculate certain magnitudes known as numerical measures
- They may be classified in two groups: measures of central tendency (averages) and measures of dispersion

Module Notes

Four Numerical Measures of Central Tendency

1. Arithmetic mean
2. Geometric mean
3. Median
4. Mode

Arithmetic Mean

Arithmetic Mean for Population

- calculated by adding the values of the observations and dividing by the total number of observations
- where X represents any particular value in the population, N is the number of items in the population

$$\mu = \frac{1}{N} \sum X$$

- The $\sum$ symbol is the summation symbol

and the numerator gives the sum of all the X values

Arithmetic Mean of a Sample

$$\bar{X} = \frac{\sum X}{n}$$

Sum of deviation from the mean is equal to naught

$$\sum (X - \bar{X}) = 0$$

Geometric Mean (GM)

$$GM = \sqrt[n]{(X_1)(X_2)\dots(X_n)}$$

- geometric mean is always smaller than or equal to the arithmetic mean
- In some exercises we need to find the average percentage increase over a period. In such cases, rather than the previous formula we should instead apply

$$GM = \sqrt[n]{V_2/V_1 - 1}$$

- where V1 (V2) is the value at the beginning (end) of the period under consideration

Weighted Mean

- found by multiplying each observation by its corresponding weight. In this case some values contribute more than others |they have more `weight'. If each number X has a weight w then the weighted mean

$$\bar{X} = \frac{\sum (wX)}{\sum w}$$

Video Notes

Calculate mean, median, and mode for grouped data

- Frequency Mean = Sum of (frequencies * midpoints) / sum of frequencies
- Frequency Median = Cumulative frequency / 2
- Mode = Interval that appears the most up the most (has the highest frequency

The Geometric Mean (GM)

- Multiply two numbers together and take the square root of it

$$GM = \sqrt[n]{(X_1)(X_2)\dots(X_n)}$$

- The ratio between the smallest number and the geometric mean will also be equal to the ratio between the geometric mean and the largest number

Textbook Notes - Chapter 3

Introduction

- Qualitative data goes into a frequency table and gets charted in a bar chart or pie chart
- Quantitative data goes into a frequency distribution table and gets portrayed in a histogram or frequency polygon
- **Dispersion**: The variation or the spread in the data
- To describe dispersion, we consider the range, the mean deviation, the variance, and the standard deviation

Population Mean

- Examples: average house price, average salary, average number of overtime hours worked

$$\text{Population mean} = \frac{\text{Sum of all the values in the population}}{\text{Number of values in the population}}$$

$$\mu = \frac{\sum x}{N}$$

- μ represents the population mean
- N is the number of items in the population
- x represents any particular value
- Σ indicates the operation of adding
- $\sum x$ is the sum of the x values in the population

- any measurable characteristic of a population is called a parameter
- The mean of a population is a parameter

Sample Mean

- The mean of a sample and the mean of a population are computed in the same way, but the shorthand notation used is different

$$\text{Sample mean} = \frac{\text{Sum of all the values in the sample}}{\text{Number of values in the sample}}$$

$$\bar{x} = \frac{\sum x}{n}$$

- x̄ is the sample mean (read as "x bar")
- n is the number in the sample
- x represents any particular value
- Σ indicates the operation of adding
- $\sum x$ is the sum of the x values in the sample

Properties of the Arithmetic Mean

1. To compute mean, the data must be measured at the interval or ratio level
2. All the values are included in computing the mean
3. The mean is unique (there is only one mean in a set of data)
4. The sum of the deviations of each value from the mean is zero

- The mean may not be an appropriate average to represent the data if there are extreme values in the calculation

The Weighted Mean

- Convenient way to compute the arithmetic mean when there are several observations of the same value
- Weighted mean is referred to as $\bar{x}_w$ "x bar sub w"

$$\bar{x}_w = \frac{w_1x_1 + w_2x_2 + w_3x_3 + \cdots + w_nx_n}{w_1 + w_2 + w_3 + \cdots + w_n}$$

$$\bar{x}_w = \frac{\sum(wx)}{\sum w}$$

- The denominator of a weighted mean is always the sum of the weights

The Median

- The middle of a group of values
- The median is unaffected by extremely low or high prices
- Median is found by averaging the two middle values

The Mode

- Especially useful in summarizing nominal-level data
- The mode is the value that appears at the highest frequency

The Relative Positions of the Mean, Median, and Mode

- For any symmetric unimodal distribution, the mode, median, and mode are located at the centre and are always equal
- If a distribution is nonsymmetric or skewed, the relationship amoung the three measures changes.
- In a positively skewed distribution:
  - Arithmetic mean is the largest of the three measures because the mean is influenced more than the other measures by a few extremely high values
  - The median is generally the next largest measure (second biggest)
  - The mode is the smallest of the three
- In a negatively skewed distribution:

- The mean is the smallest (is influenced heavily by a few smaller observations)
- The median is greater than the arithmetic mean (middle value)
- The mode is the greatest

The Geometric Mean

- Useful in finding the average change of percentages, ratios, indexes, or growth rates over time
- The geometric mean of a set of n positive numbers is defined as the nth root of the product of the n values

$$GM = \sqrt[n]{(x_1)\,(x_2)\cdots(x_n)}$$

- Will always be less than or equal to the arithmetic mean
- All data values must be positive

The rate of increase

$$GM = \sqrt[n]{\frac{\text{Value at end of period}}{\text{Value at beginning of period}}} - 1$$

Why Study Dispersion?

- Measures of location (mean, median, and mode) only describe the center of data and does not tell us anything about the spread of the data
- A small value for a measure of dispersion indicates that the data are clustered closely around the arithmetic mean
  - The mean is then considered representative of the data
- A large measure of dispersion indicates that the mean is not reliable

Range

- $Range = MaxValue - MinValue$
- Disadvantage is that the range does not take into consideration all of the values

Mean Deviation

- Takes all values into consideration
- It's a measure of the average distance between an observation and the mean of the observation
- The mean deviation is the mean of the differences between individual observations and the arithmetic mean
- **Advantages**
  - Uses all the values in the computation
  - It's easy to understand it - it is the average amount by which values deviate from the mean

$$MD = \frac{\sum |x - \bar{x}|}{n}$$

- x is the value of each observation
- x̄ is the arithmetic mean of the values
- n is the number of observations in the sample
- $||$ indicates the absolute value

Disadvantages

- use of absolute values

Variance and Standard Deviation

Population Variance

- σ is the symbol for population variance
- x is the value of each observation in the population
- μ is the arithmetic mean of the population
- N is the number of observations in the population

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

Steps to compute:

1. Find the mean
2. Find the difference between each observation and the mean
3. Square the difference
4. Sum all the squared differences
5. Divide the sum of the squared differences by the number of observations in the population

Population Standard Deviation

- The square root of the population variance is the population standard deviation

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$$

Sample Variance

- $\bar{x} = \frac{\sum x}{n}$

Sample Variance Deviation Formula

- s^2 is the sample variance
- x is the value of each observation in the sample
- $\bar{x}$ is the mean of the sample
- n is the number of observations in the sample

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

Sample Variance Direct Formula

$$s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n - 1}$$

Sample Standard Deviation Direct Formula

$$s = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n - 1}}$$

Interpretation and Uses of The Standard Deviation

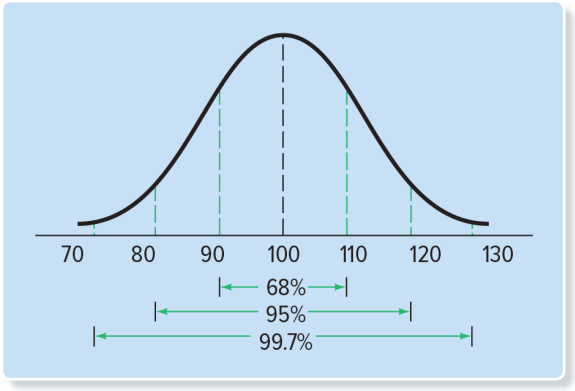
Chebyshev's Theorem (1821-1894) - Minimum proportion of the values that lie within a specified number of standard deviations of the mean

- At least 75% of values must lie between the mean plus 2 standard deviations and the mean minus 2 standard deviations
- At least 88.9% of values will lie between plus 3 standard deviations and minus 3 standard deviations of the mean

- 96% of values will lie between plus and minus 5 standard deviations of the mean
- This relationship applies regardless of the shape of the distribution

The Empirical Rule / Normal Rule

- Applies to symmetric, bell-shaped distribution

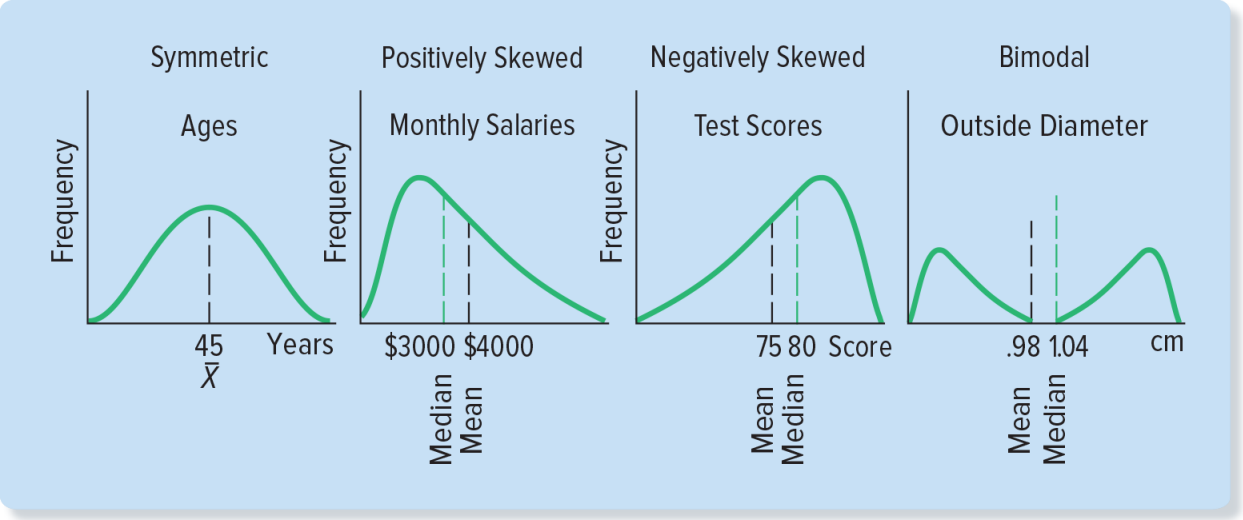


Coefficient of Variation (CV)

- It's a very useful measure when:
  1. The data are in different units (such as dollars and days absent)
  2. The data are in the same units, but the means are far apart (such as the incomes of the top executives and the incomes of the unskilled employees)

$$CV = \frac{s}{x}(100)$$ ← Multiplying by 100 converts the decimal to a percent

Skewness



- **Symmetric:** The mean and median are equal and the data values are evenly spread around those values. The values below the mean and median are a mirror image of those above
- **Positively Skewed (to the right):** The mean is larger than the median
- **Negatively Skewed (to the left):** The mean is smaller than the median
- **Bimodal Distribution:** Will have two or more peaks. Often the case when values are from two or more populations

Pearson's Coefficient of Skewness

$$sk = \frac{3\,(\bar{x} - \text{Median})}{s}$$

- The coefficient of skewness can range from -3 up to 3.
- A value near -3 indicates considerable negative skewness
- A value such as 1.63 indicates moderate positive skewness
- A value of 0 will occur when the mean and median are equal, indicates the distribution is symmetric and that there is no skewness present

Measures of Position

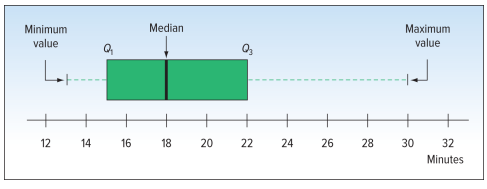
- **Quartiles**: divide a set of observations into four equal parts
  - The first quartile (Q_1) is the value below the 25% mark

- The second quartile (Q_2) is the median
- The third quartile (Q_3) is the value below which 75% of the observations occur

- **Deciles:** divide a set of observations into 10 equal parts
- **Percentiles:** divide a set of observations into 100 equal parts

Box Plots

- is a graphical display, based on quartiles that helps us picture a set of data
- To construct a box plot, we only need 5 statistics:
  - minimum value
  - Q_1 (the first quartile)
  - the median
  - Q_3 the third quartile
  - maximum value
- The distance between the ends of the box is the interquartile range (distance between the first and third quartiles)
  - It shows the spread or dispersion of the middle 50% of deliveries

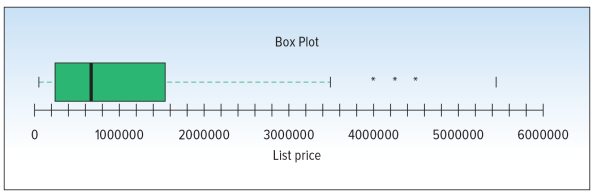


Process:

- Create a scale along the horizontal axis
- Draw a box at that starts at Q_1 and ends at $Q3$
- Place a vertical line at the median
- Add "whiskers" (vertical lines) at the max and min values and extend the horizontal lines from the box to them

Outliers

- An outlier is a value that is more than 1.5 times the interquartile range larger than Q_3 or smaller than Q_1
- Outliers are indicated by small circles on the box plot



Arithmetic Mean of Grouped Data

- A mean or standard deviation from grouped data is an ESTIMATE of the corresponding actual values
- $\bar{x}$ is the designation for the sample mean
- x is the midpoint of each class
- f is the frequency in each class
- fx is the frequency in each class times the midpoint of the class
- $\sum fx$ is the sum of these products
- n is the total number of frequencies

$$\bar{x} = \frac{\sum fx}{n}$$

The Median of Grouped Data

- only an estimation
- L is the lower limit of the median class
- N is the size of the population
- f is the frequency of the median class
- f_c is the cumulative frequencies up to but excluding the median class
- i is the class with of the median class

$$\text{Median} = L + \frac{\frac{N}{2} - f_c}{f}(i)$$

Standard Deviation of Grouped Data

- s symbol for the sample standard deviation
- x is the midpoint of a class

- f is the class frequency
- n is the total number of sample observations

$$s = \sqrt{\frac{\sum fx^2 - \frac{(\sum x)^2}{n}}{n-1}}$$

Steps to find the Standard Deviation

1. Each class frequency is multiplied by its class midpoint $(f * x)$
2. Calculate fx^2
3. Sum the fx and fx^2 columns