Statistical Analysis

JOHNS HOPKINS
CAREY BUSINESS SCHOOL

Some of the slides in this presentation are from Bowerman, B. L., O'Connell, R. T., & Murphree, E.S., (2010). *Business Statistics in Practice* (6th Ed.), Copyright © McGraw-Hill Education.

McGraw-Hill makes no representations or warranties as to the accuracy of any information contained in the McGraw-Hill Material, including any warranties of merchantability or fitness for a particular purpose. In no event shall McGraw-Hill have any liability to any party for special, incidental, tort, or consequential damages arising out of or in connection with the McGraw-Hill Material, even if McGraw-Hill has been advised of the possibility of such damages.

Descriptive Statistics: Numerical Methods

1. Describing Central Tendency
2. Measures of Variation
3. Percentiles and Quartiles
4. Weighted Means and Grouped Data
5. The Geometric Mean

Describing Central Tendency

In addition to describing the shape of a distribution, we want to describe the data set’s central tendency

- A measure of central tendency represents the center or middle of the data
- Population mean (μ) is average of the population measurements

Population parameter: a number calculated from all the population measurements that describes some aspect of the population

Sample statistic: a number calculated using the sample measurements that describes some aspect of the sample

Bowerman, B. L., O'Connell, R. T., & Murphree, E. S., (2010). Business Statistics in Practice (6th Ed.), Copyright © McGraw-Hill Education

Measures of Central Tendency

Mean, μ

The average or expected value

Median, M_d

The value of the middle point of the ordered measurements

Mode, M_o

The most frequent value

Bowerman, B. L., O'Connell, R. T., & Murphree, E. S., (2010). Business Statistics in Practice (6th Ed.), Copyright © McGraw-Hill Education

The Sample Mean

For a sample of size n, the **sample mean** $(\bar{x})$ is defined as

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

and is a point estimate of the population mean

It is the value to expect, on average and in the long run

Bowerman, B. L., O'Connell, R. T., & Murphree, E. S., (2010). Business Statistics in Practice (6th Ed.), Copyright © McGraw-Hill Education

Car Mileage Case *example*

Sample mean for five car mileages

30.8, 31.7, 30.1, 31.6, 32.1

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i}{5} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5}$$

$$\bar{x} = \frac{30.8 + 31.7 + 30.1 + 31.6 + 32.1}{5} = \frac{156.3}{5} = 31.26$$

Bowerman, B. L., O'Connell, R. T., & Murphree, E. S., (2010). Business Statistics in Practice (6th Ed.), Copyright © McGraw-Hill Education

The Median

The **median M_d** is a value such that 50% of all measurements, after having been arranged in numerical order, lie above (or below) it

1. If the number of measurements is odd, the median is the middlemost measurement in the ordering
2. If the number of measurements is even, the median is the average of the two middlemost measurements in the ordering

Bowerman, B. L., O'Connell, R. T., & Murphree, E. S., (2010). Business Statistics in Practice (6th Ed.), Copyright © McGraw-Hill Education

Car Mileage Case *example*

Sample of five car mileages: 30.8, 31.7, 30.1, 31.6, 32.1

In order: 30.1, 30.8, 31.6, 31.7, 32.1

There is an odd number of observations, so median is the observation in the middle, or *31.6*

Bowerman, B. L., O'Connell, R. T., & Murphree, E. S., (2010). Business Statistics in Practice (6th Ed.), Copyright © McGraw-Hill Education

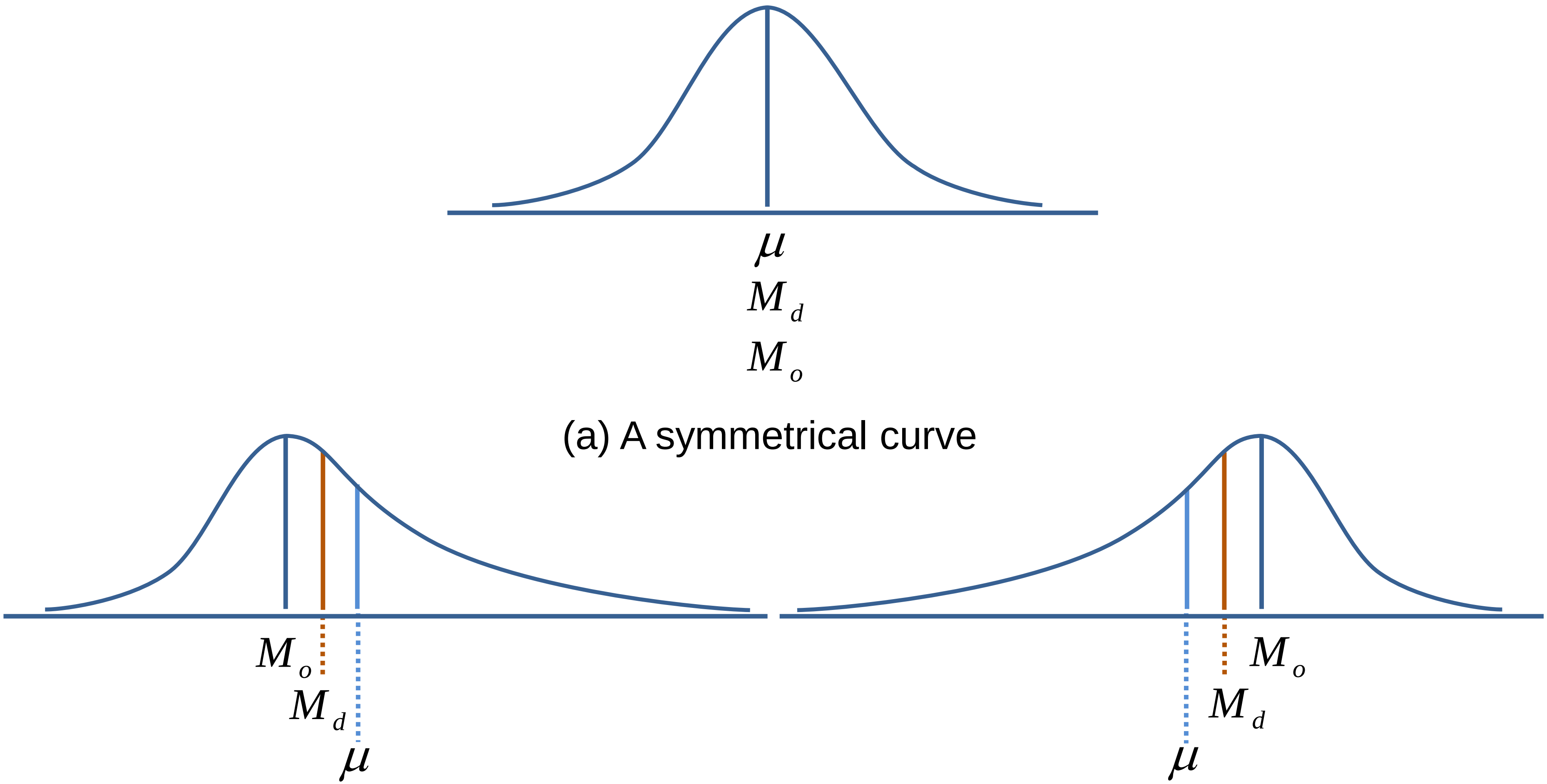
The Mode

The **mode** $\boldsymbol{M_o}$ of a population or sample of measurements is the measurement that occurs most frequently

- Modes are the values that are observed “most typically”
- Sometimes higher frequencies at two or more values
  - If there are two modes, the data is bimodal
  - If more than two modes, the data is multimodal
- When data are in classes, the class with the highest frequency is the modal class
  - The tallest box in the histogram

Bowerman, B. L., O'Connell, R. T., & Murphree, E. S., (2010). Business Statistics in Practice (6th Ed.), Copyright © McGraw-Hill Education

Relationships Among Mean, Median and Mode



(a) A symmetrical curve

(b) A curve skewed to the right

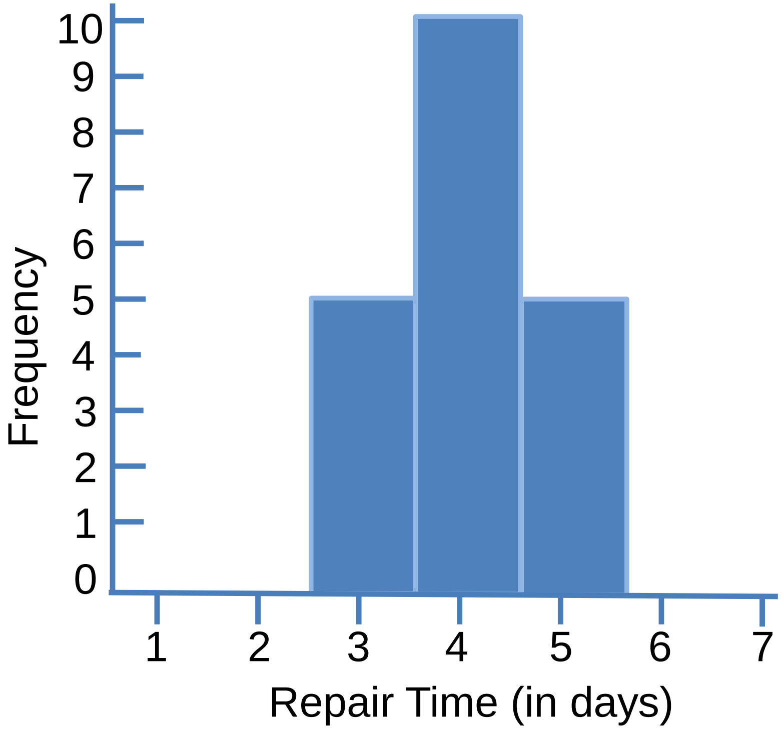
(c) A curve skewed to the left

Adapted from Bowerman, B. L., O'Connell, R. T., & Murphree, E. S., (2010). Business Statistics in Practice (6th Ed.), Copyright © McGraw-Hill Education

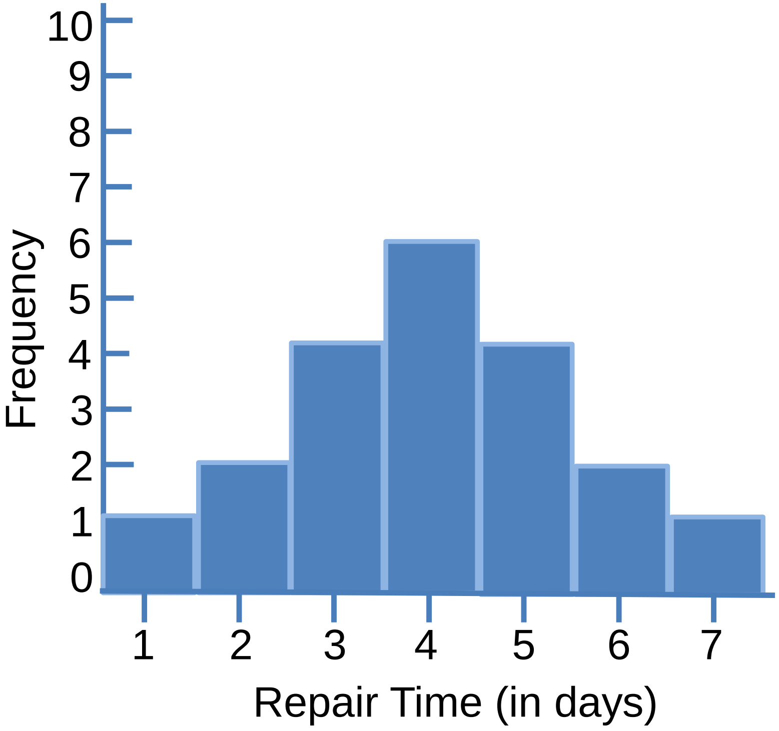
Measures of Variation

Knowing the measures of central tendency is not enough

Both of the distributions below have identical measures of central tendency



American Service Center



National Service Center

Adapted from Bowerman, B. L., O'Connell, R. T., & Murphree, E. S., (2010). Business Statistics in Practice (6th Ed.), Copyright © McGraw-Hill Education

Measures of Variation

Range

Largest minus the smallest measurement

Variance

The average of the squared deviations of all the population measurements from the population mean

Standard Deviation

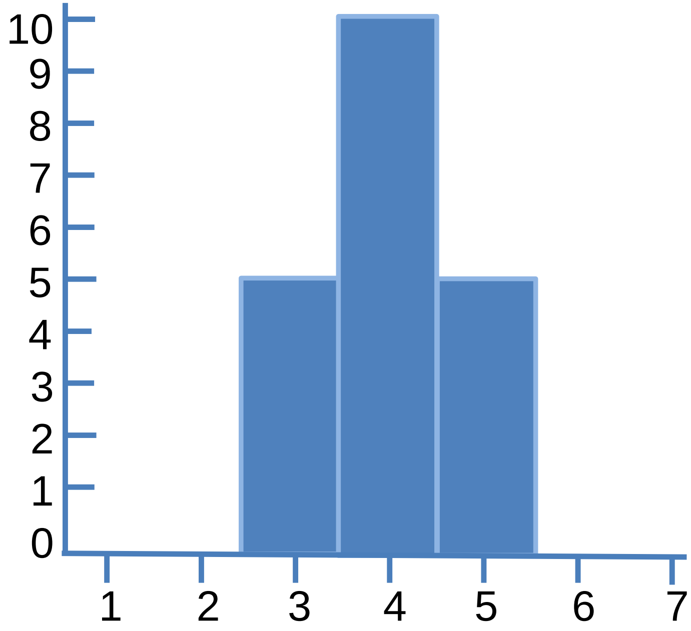
The square root of the population variance

Bowerman, B. L., O'Connell, R. T., & Murphree, E. S., (2010). Business Statistics in Practice (6th Ed.), Copyright © McGraw-Hill Education

The Range

Largest observation minus smallest observation *measures the interval spanned by all the data*

In this example, largest value is 5 and smallest is 3
Range is (5 – 3) = 2 days



Adapted from Bowerman, B. L., O'Connell, R. T., & Murphree, E. S., (2010). Business Statistics in Practice (6th Ed.), Copyright © McGraw-Hill Education

Population Variance and Standard Deviation

The *population variance* (σ^2) is the average of the squared deviations of the individual population measurements from the population mean (μ)

The *population standard deviation* (σ) is the positive square root of the population variance

Bowerman, B. L., O'Connell, R. T., & Murphree, E. S., (2010). Business Statistics in Practice (6th Ed.), Copyright © McGraw-Hill Education

Variance

For a population of size N, the population variance σ^2 is:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N} = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_N - \mu)^2}{N}$$

For a sample of size n, the sample variance s^2 is:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}$$

Bowerman, B. L., O'Connell, R. T., & Murphree, E. S., (2010). Business Statistics in Practice (6th Ed.), Copyright © McGraw-Hill Education

Standard Deviation

Population standard deviation (σ)

$$\sigma = \sqrt{\sigma^2}$$

Sample standard deviation (s)

$$s = \sqrt{s^2}$$

Bowerman, B. L., O'Connell, R. T., & Murphree, E. S., (2010). Business Statistics in Practice (6th Ed.), Copyright © McGraw-Hill Education

Sample Variance and Standard Deviation *example*

- Sample of five car mileages are 30.8, 31.7, 30.1, 31.6, 32.1
- The sample mean is 31.26
- The variance and standard deviation are as follows

$$s^2 = \frac{\sum_{i=1}^{5}\left(x_i - \bar{x}\right)^2}{5-1}$$

$$= \frac{(30.8-31.26)^2 + (31.7-31.26)^2 + (30.1-31.26)^2 + (31.6-31.26)^2 + (32.1-31.26)^2}{4}$$

$$= \frac{2.572}{4} = 0.643$$

$$s = \sqrt{s^2} = \sqrt{0.643} = 0.8019$$

Bowerman, B. L., O'Connell, R. T., & Murphree, E. S., (2010). Business Statistics in Practice (6th Ed.), Copyright © McGraw-Hill Education

The Empirical Rule for Normal Populations

If a population has mean μ and standard deviation σ and is described by a normal curve, then

- 68.4 % of the population measurements lie within one standard deviation of the mean: $[\mu - \sigma, \mu + \sigma]$
- 95.5% of the population measurements lie within two standard deviations of the mean: $[\mu - 2\sigma, \mu + 2\sigma]$
- 99.7% of the population measurements lie within three standard deviations of the mean: $[\mu - 3\sigma, \mu + 3\sigma]$

Bowerman, B. L., O'Connell, R. T., & Murphree, E. S., (2010). Business Statistics in Practice (6th Ed.), Copyright © McGraw-Hill Education

z Scores

For any x in a population or sample, the associated z score is

$$z = \frac{x - \text{mean}}{\text{standard deviation}}$$

The z score is the number of standard deviations that x is from the mean

- A *positive* z score is for x *above* (greater than) the mean
- A *negative* z score is for x *below* (less than) the mean

Bowerman, B. L., O'Connell, R. T., & Murphree, E. S., (2010). Business Statistics in Practice (6th Ed.), Copyright © McGraw-Hill Education

Coefficient of Variation

Measures the size of the standard deviation relative to the size of the mean

$$\text{Coefficient of variation} = \frac{\text{Standard deviation}}{\text{Mean}} \times 100\%$$

Used to:

- Compare the relative variabilities of values about the mean
- Compare the relative variability of populations or samples with different means and different standard deviations
- Measure risk

Bowerman, B. L., O'Connell, R. T., & Murphree, E. S., (2010). Business Statistics in Practice (6th Ed.), Copyright © McGraw-Hill Education

Percentiles and Quartiles

For a set of measurements arranged in increasing order, the p^{th} percentile is a value such that p percent of the measurements fall at or below the value and $(100 - p)$ percent of the measurements fall at or above the value

- The **first quartile Q_1** is the 25th percentile
- The **second quartile** (median) is the 50th percentile
- The **third quartile Q_3** is the 75th percentile
- The **interquartile range IQR** is $Q_3 - Q_1$

Bowerman, B. L., O'Connell, R. T., & Murphree, E. S., (2010). Business Statistics in Practice (6th Ed.), Copyright © McGraw-Hill Education

Calculating Percentiles

1. Arrange the measurements in increasing order
2. Calculate the index $i = (p/100)n$ where p is the percentile to find

   (a) If i is not an integer, round up and the next integer greater than i denotes the p^{th} percentile

   (b) If i is an integer, the p^{th} percentile is the average of the measurements in the i and $i+1$ positions

Bowerman, B. L., O'Connell, R. T., & Murphree, E. S., (2010). Business Statistics in Practice (6th Ed.), Copyright © McGraw-Hill Education

Percentile *example*

$$i = (10/100)12 = 1.2$$

Not an integer so round up to 2

10th percentile is in the second position so 11,070

$$i = (25/100)12 = 3$$

Integer so average values in positions 3 and 4

25th percentile (18,211+26,817)/2 or 22,514

7,524	11,070	18,211	26,817	36,551	41,286
49,312	57,283	72,814	90,416	135,540	190,250

Bowerman, B. L., O'Connell, R. T., & Murphree, E. S., (2010). Business Statistics in Practice (6th Ed.), Copyright © McGraw-Hill Education

Five Number Summary

1. The smallest measurement
2. The first quartile, Q_1
3. The median, M_d
4. The third quartile, Q_3
5. The largest measurement

Adapted from Bowerman, B. L., O'Connell, R. T., & Murphree, E. S., (2010). Business Statistics in Practice (6th Ed.), Copyright © McGraw-Hill Education

Outliers

Outliers are measurements that are very different from other measurements

- They are either much larger or much smaller than most of the other measurements

Outliers lie beyond the fences of the box-and-whiskers plot

- Measurements between the inner and outer fences are mild outliers
- Measurements beyond the outer fences are severe outliers

Bowerman, B. L., O'Connell, R. T., & Murphree, E. S., (2010). Business Statistics in Practice (6th Ed.), Copyright © McGraw-Hill Education

Weighted Means and Grouped Data

Sometimes, some measurements are more important than others

- Assign numerical “weights” to the data
- Weights measure relative importance of the value

Calculate weighted mean as

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

where w_i *is the weight assigned to the* i^{th} *measurement* x_i

Adapted from Bowerman, B. L., O'Connell, R. T., & Murphree, E. S., (2010). Business Statistics in Practice (6th Ed.), Copyright © McGraw-Hill Education

Weighted Average *example*

A grocery store carries five brands of pickles. Each brand has a different profit and different volume of each sold per day.

Pickle Brand	Profit per unit (X)	Volume per day (W)	XW
Star	$2.00	3	$6.00
Mt. Olive	$3.50	7	$24.50
Heinz	$5.00	15	$75.00
Giant	$7.50	12	$90.00
Supreme	$6.00	15	$90.00
	$24.00	**52**	**$285.50**

Arithmetic average = $24/5 = $4.80
Weighted average = $285.50/52 = $5.49

Descriptive Statistics for Grouped Data

- Data already categorized into a frequency distribution or a histogram is called grouped data

- Can calculate the mean and variance even when the raw data is not available

- Calculations are slightly different for data from a sample and data from a population

Bowerman, B. L., O'Connell, R. T., & Murphree, E. S., (2010). Business Statistics in Practice (6th Ed.), Copyright © McGraw-Hill Education

Descriptive Statistics for Grouped Data *(Sample)*

Sample *mean* for grouped data

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

Sample *variance* for grouped data

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

f_i is the frequency for class $\boldsymbol{i}$

M_i is the midpoint of class $\boldsymbol{i}$

$N = \sum f_i =$ sample size

Adapted from Bowerman, B. L., O'Connell, R. T., & Murphree, E. S., (2010). Business Statistics in Practice (6th Ed.), Copyright © McGraw-Hill Education

Descriptive Statistics for Grouped Data *(Population)*

Population *mean* for grouped data

$$\mu = \frac{\sum f_i M_i}{\sum f_i} = \frac{\sum f_i M_i}{N}$$

Population *variance* for grouped data

$$\sigma^2 = \frac{\sum f_i \left(M_i - \bar{x}\right)^2}{N}$$

f_i is the frequency for class $\boldsymbol{i}$

M_i is the midpoint of class $\boldsymbol{i}$

$N = \sum f_i =$ population size

Adapted from Bowerman, B. L., O'Connell, R. T., & Murphree, E. S., (2010). Business Statistics in Practice (6th Ed.), Copyright © McGraw-Hill Education

The Geometric Mean

For rates of return of an investment, use the geometric mean to give the correct wealth at the end of the investment

Suppose the rates of return (expressed as decimal fractions) are $R_1, R_2, \ldots, R_n$ for periods 1, 2, …, n

The mean of all these returns is the calculated as the geometric mean:

$$R_g = \sqrt[n]{(1+R_1)\times(1+R_2)\times\cdots\times(1+R_n)}-1$$

Bowerman, B. L., O'Connell, R. T., & Murphree, E. S., (2010). Business Statistics in Practice (6th Ed.), Copyright © McGraw-Hill Education

Geometric Mean *example*

Revenues for Computer Nerd Company

Year	Revenue	Percentage of Previous Year
2005	$50,000	—— ——
2006	$55,000	55 / 50 = 1.10
2007	$66,000	66 / 55 = 1.20
2008	$60,000	60 / 66 = 0.91
2009	$78,000	78 / 60 = 1.30

$$GM = \sqrt[4]{(1.10)(1.20)(0.91)(1.30)} = \sqrt[4]{1.56} = 1.1179$$

Arithmetic Mean = $(1.1 + 1.2 + 0.91 + 1.3)/4 = 1.1275$

Using Arithmetic Mean

Arithmetic Mean = $(1.1+1.2+0.91+1.3)/4=1.1275$

$50,000	x	1.1275	=	$56,375
$56,375	x	1.1275	=	$63,563
$63,563	x	1.1275	=	$71,667
$71,667	x	1.1275	=	$80,805

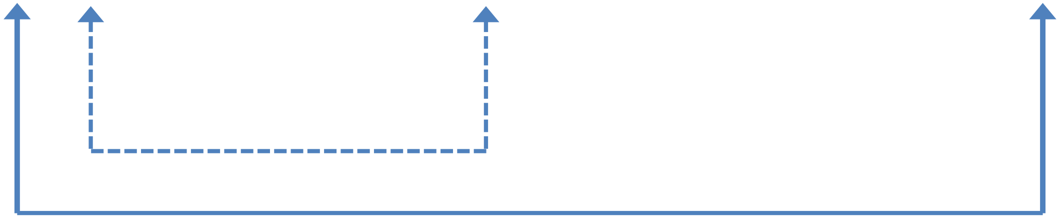
Using Geometric Mean

$$GM = \sqrt[4]{(1.10)(1.20)(0.91)(1.30)} = \sqrt[4]{1.56} = 1.1179$$

\$50,000	x	1.1179	=	\$55,895
\$55,895	x	1.1179	=	\$62,485
\$62,485	x	1.1179	=	\$69,852
\$69,852	x	1.1179	=	\$78,088 ≈ \$78,000

Using Geometric and Arithmetic Mean

Year	Revenue	Calculation Using Arithmetic Mean	Calculation Using Geometric Mean
2005	$50,000	—	—
2006	$55,000	$56,375	$55,895
2007	$66,000	$63,563	$62,485
2008	$60,000	$71,667	$69,852
2009	**$78,000**	**$80,805**	**$78,088 ≈ $78,000**

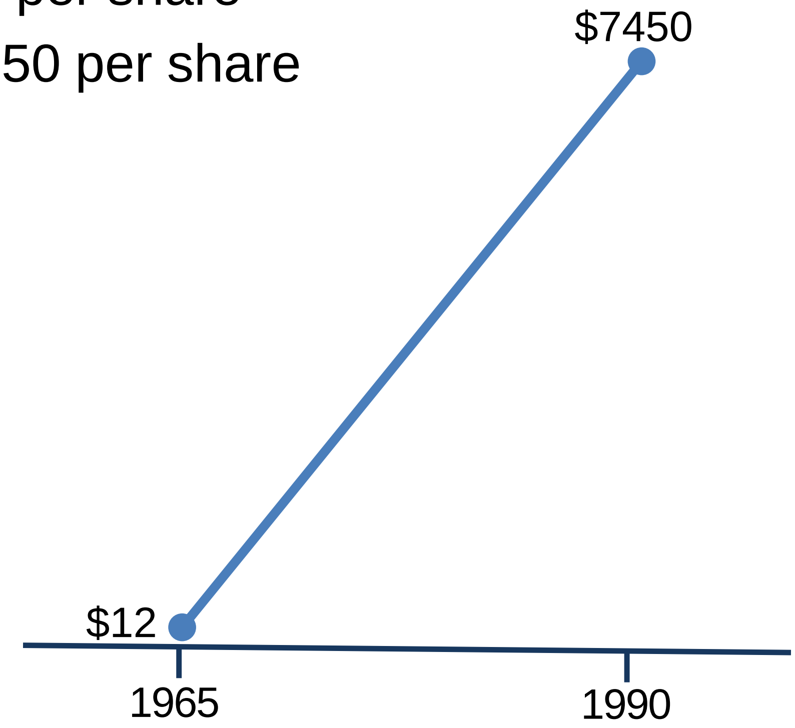


Application of Geometric Mean

Warren Buffett amassed considerable wealth in textile manufacture purchased in 1965:

1965 ► $12 per share

1990 ► $7450 per share



Application of Geometric Mean

Financial News

In 1990 Courts assessed tax owed based on annual increase in his holdings using arithmetic mean.

Buffet argued using the geometric mean to compute the percentage increase which is always less than arithmetic mean.

As a result Buffet paid less tax.