Statistical Analysis in Fin Mkts

MSF 502

Li Cai

3 Numerical Descriptive Measures

C H A P T E R

Chapter 3 Learning Objectives (LOs)

LO 3.1: Calculate and interpret the arithmetic mean, the median, and the mode.

LO 3.2: Calculate and interpret percentiles and a box plot.

LO 3.3: Calculate and interpret a geometric mean return and an average growth rate.

LO 3.4: Calculate and interpret the range, the mean absolute deviation, the variance, the standard deviation, and the coefficient of variation.

Chapter 3 Learning Objectives (LOs)

LO 3.5: Explain mean-variance analysis and the Sharpe ratio.

LO 3.6: Apply Chebyshev’s Theorem and the empirical rule.

LO 3.7: Calculate the mean and the variance for grouped data.

LO 3.8: Calculate and interpret the covariance and the correlation coefficient.

Investment Decision

- As an investment counselor at a large bank, Rebecca Johnson was asked by an inexperienced investor to explain the differences between two top-performing mutual funds:
  - Vanguard's Precious Metals and Mining fund (Metals)
  - Fidelity's Strategic Income Fund (Income)
- The investor has collected sample returns for these two funds for years 2000 through 2009. These data are presented in the next slide.

Investment Decision

Year	Metals	Income	Year	Metals	Income
2000	−7.34	4.07	2005	43.79	3.12
2001	18.33	6.52	2006	34.30	8.15
2002	33.35	9.38	2007	36.13	5.44
2003	59.45	18.62	2008	−56.02	−11.37
2004	8.09	9.44	2009	76.46	31.77

- **Rebecca would like to**
  1. **Determine the typical return of the mutual funds.**
  2. **Evaluate the investment risk of the mutual funds.**

3.1 Measures of Central Location

LO 3.1 Calculate and interpret the arithmetic mean, the median, and the mode.

- **The arithmetic mean is a primary measure of central location.**
  - **Sample Mean** $\bar{x}$

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

  - **Population Mean** μ

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

LO 3.1

3.1 Measures of Central Location

Example: Investment Decision

- **Use the data in the introductory case to calculate and interpret the mean return of the Metals fund and the mean return of the Income fund.**

$$\text{Metals fund mean return} = \frac{-7.34 + 18.33 + \cdots + 76.46}{10} = \frac{246.54}{10} = 24.65\%$$

$$\text{Income fund mean return} = \frac{4.07 + 6.52 + \cdots + 31.77}{10} = \frac{85.14}{10} = 8.51\%$$

LO 3.1

3.1 Measures of Central Location

- **The mean is sensitive to outliers.**
- **Consider the salaries of employees at**

Title	Salary
Administrative Assistant	$40,000
Research Assistant	40,000
Computer Programmer	65,000
Senior Research Associate	90,000
Senior Sales Associate	145,000
Chief Financial Officer	150,000
President (and owner)	550,000

$$\mu = \frac{\Sigma x_i}{N} = \frac{40{,}000 + 40{,}000 + \cdots + 550{,}000}{7} = \$154{,}286.$$

- **This mean does not reflect the typical salary!**

LO 3.1

3.1 Measures of Central Location

- **The median is another measure of central location that is not affected by outliers.**
- **When the data are arranged in ascending order, the median is**
  - **the middle value if the number of observations is odd, or**
  - **the average of the two middle values if the number of observations is even.**

LO 3.1

3.1 Measures of Central Location

- **Consider the sorted salaries of employees at Acetech (*odd number*).**

Position:		3 values below		4		3 values above	
Value:	$40,000	40,000	65,000	90,000	145,000	150,000	550,000

Median = 90,000

- **Consider the sorted data from the Metals funds of the introductory case study (*even number*).**

Position:	1	2	3	4	5	6	7	8	9	10
Value:	−56.02	−7.34	8.09	18.33	33.35	34.30	36.13	43.79	59.45	76.46

- **Median = (33.35 + 34.30) / 2 = 33.83%.**

LO 3.1

3.1 Measures of Central Location

- **The mode is another measure of central location.**
  - **The most frequently occurring value in a data set**
  - **Used to summarize qualitative data**
  - **A data set can have no mode, one mode (unimodal), or many modes (multimodal).**

Position:	1	2	3	4	5	6	7
Value:	$40,000	40,000	65,000	90,000	145,000	150,000	550,000

The mode is $40,000 since this value appears most often.

3.2 Percentiles and Box Plots

LO 3.2 Calculate and interpret percentiles and a box plot.

- **In general, the *p*th percentile divides a data set into two parts:**
  - **Approximately *p* percent of the observations have values less than the *p*th percentile;**
  - **Approximately (100 – *p*) percent of the observations have values greater than the *p*th percentile.**

LO 3.2

3.2 Percentiles and Box Plots

- **Calculating the *p*th percentile:**
  - **First arrange the data in ascending order.**
  - **Locate the position, L_p, of the *p*th percentile by using the formula:**

$$L_p = (n+1)\frac{p}{100}$$

  - **We use this position to find the percentile as shown next.**

LO 3.2

3.2 Percentiles and Box Plots

- **Consider the sorted data from the introductory case.**

Position:	1	2	3	4	5	6	7	8	9	10
Value:	−56.02	−7.34	8.09	18.33	33.35	34.30	36.13	43.79	59.45	76.46

- **For the 25th percentile, we locate the position:**

$$L_{25} = (n+1)\frac{p}{100} = (10+1)\frac{25}{100} = 2.75$$

- **Similarly, for the 75th percentile, we first find:**

$$L_{75} = (n+1)\frac{p}{100} = (10+1)\frac{75}{100} = 8.25$$

LO 3.2

3.2 Percentiles and Box Plots

Calculating the *p*th percentile

- Once you find L_p, observe whether or not it is an integer.
  - If L_p is an integer, then the L_pth observation in the sorted data set is the *p*th percentile.
  - If L_p is not an integer, then interpolate between two corresponding observations to approximate the *p*th percentile.

LO 3.2

3.2 Percentiles and Box Plots

- Both $L_{25} = 2.75$ and $L_{75} = 8.25$ are not integers, thus
  - The 25th percentile is located 75% of the distance between the second and third observations, and it is

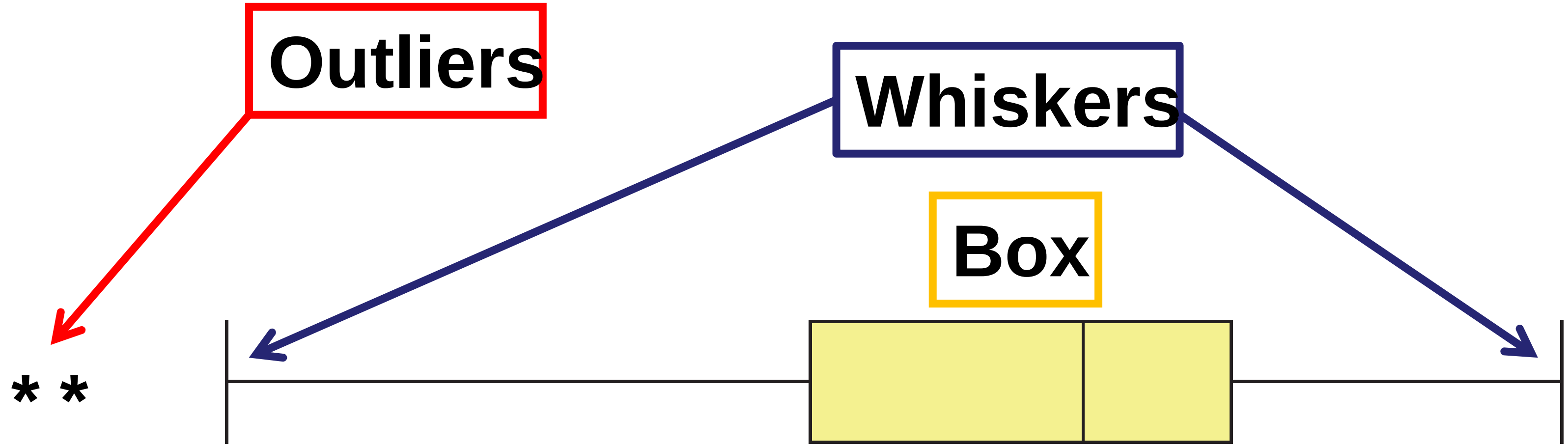
$$-7.34 + 0.75(8.09 - (-7.34)) = 4.23$$

  - The 75th percentile is located 25% of the distance between the eighth and ninth observations, and it is

$$43.79 + 0.25(59.45 - 43.79) = 47.71$$

LO 3.2 3.2 Percentiles and Box Plots

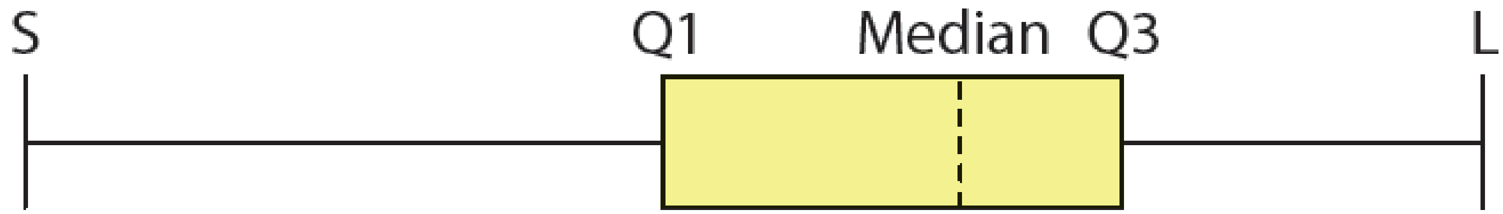
- **A box plot allows you to:**
  - **Graphically display the distribution of a data set.**
  - **Compare two or more distributions.**
  - **Identify outliers in a data set.**



LO 3.2

3.2 Percentiles and Box Plots

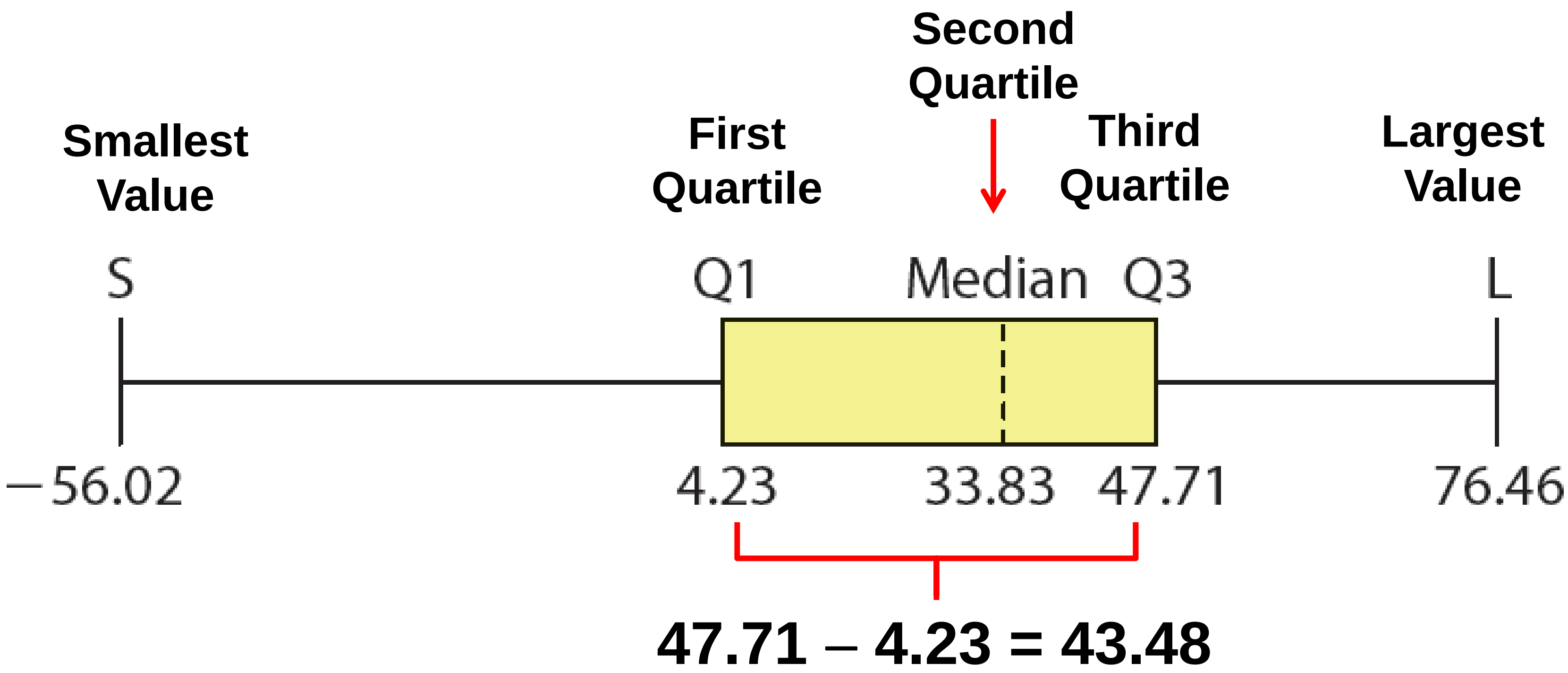
- **The box plot displays 5 summary values:**
  - **S = smallest value**
  - **L = largest value**
  - **Q1 = first quartile = 25th percentile**
  - **Q2 = median = second quartile = 50th percentile**



LO 3.2

3.2 Percentiles and Box Plots

- **Using the results obtained from the Metals fund data, we can label the box plot with the 5 summary values:**



- **Note that IQR = Q3 – Q1 =43.48**

3.2 Percentiles and Box Plots

Detecting outliers

- Calculate IQR = 43.48
- Calculate 1.5 × IQR, or 1.5 × 43.48 = 65.22

There are outliers if

- Q1 – S > 65.22, or if
- L – Q3 > 65.22

- There are no outliers in this data set.

3.3 The Geometric Mean

LO 3.3 Calculate and interpret a geometric mean return and an average growth rate.

- **Remember that the arithmetic mean is an additive average measurement.**
  - **Ignores the effects of compounding.**
- **The geometric mean is a multiplicative average that incorporate compounding. It is used to measure:**
  - **Average investment returns over several years,**
  - **Average growth rates.**

LO 3.3

3.3 The Geometric Mean

- **For multiperiod returns $R_1, R_2, \ldots, R_n$, the geometric mean return G_R is calculated as:**

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

where n is the number of multiperiod returns.

LO 3.3

3.3 The Geometric Mean

- **Using the data from the Metals and Income funds, we can calculate the geometric mean returns:**

$$\text{Metals Fund: } G_R = \sqrt[10]{(1-0.0734)(1+0.1833)\cdots(1+0.7646)} - 1$$
$$= (5.1410)^{1/10} - 1 = 0.1779, \text{ or } 17.79\%.$$

$$\text{Income Fund: } G_R = \sqrt[10]{(1+0.0407)(1+0.0652)\cdots(1+0.3177)} - 1$$
$$= (2.1617)^{1/10} - 1 = 0.0801, \text{ or } 8.01\%.$$

LO 3.3

3.3 The Geometric Mean

- **Computing an average growth rate**
  - **For growth rates $g_1, g_2, \ldots, g_n$, the average growth rate G_g is calculated as**

$$G_g = \sqrt[n]{(1+g_1)(1+g_2)\cdots(1+g_n)} - 1$$

where n is the number of multiperiod growth rates.

  - **For observations $x_1, x_2, \ldots, x_n$, the average growth rate G_g is calculated as**

$$G_g = \sqrt[n-1]{\frac{x_n}{x_{n-1}} \frac{x_{n-1}}{x_{n-2}} \frac{x_{n-2}}{x_{n-3}} \cdots \frac{x_2}{x_1}} - 1 = \sqrt[n-1]{\frac{x_n}{x_1}} - 1$$

where n–1 is the number of distinct growth rates.

LO 3.3

3.3 The Geometric Mean

- **For example, consider the sales for Adidas (in millions of €) for the years 2005 through 20**

Year	2005	2006	2007	2008	2009
Sales	6,636	10,084	10,299	10,799	10,381

e:

- 2005–2006: $\frac{10{,}084 - 6{,}636}{6{,}636} = 0.5196$
- 2006–2007: $\frac{10{,}299 - 10{,}084}{10{,}084} = 0.0213$
- 2007–2008: $\frac{10{,}799 - 10{,}299}{10{,}299} = 0.0485$
- 2008–2009: $\frac{10{,}381 - 10{,}799}{10{,}799} = -0.0387$

$$G_g = \sqrt[4]{(1 + 0.5196)(1 + 0.0213)(1 + 0.0485)(1 - 0.0387)} - 1 = 0.1184, \text{ or } 11.84\%$$

- **The average growth rate using the simplified formula is:**

$$G_g = \sqrt[n-1]{\frac{x_n}{x_1}} - 1 = \sqrt[5-1]{\frac{10{,}381}{6{,}636}} - 1$$

$$= 1.5643^{1/4} - 1 = 0.1184, \text{ or } 11.84\%$$

3.4 Measures of Dispersion

LO 3.4 Calculate and interpret the range, the mean absolute deviation, the variance, the standard deviation, and the coefficient of variation.

- **Measures of dispersion gauge the variability of a data set.**
- **Measures of dispersion include:**
  - **Range**
  - **Mean Absolute Deviation (MAD)**
  - **Variance and Standard Deviation**
  - **Coefficient of Variation (CV)**

LO 3.4

3.4 Measures of Dispersion

- **Range**

$$\text{Range} = \text{Maximum Value} - \text{Minimum Value}$$

  - **It is the simplest measure.**
  - **It is focusses on extreme values.**

- **Calculate the range using the data from the Metals and Income funds**

Metals fund: 76.46% − (–56.02%) = 132.48%

Income fund: 31.77% − (–11.37%) = 43.14%

LO 3.4

3.4 Measures of Dispersion

- **Mean Absolute Deviation (MAD)**
  - **MAD is an average of the absolute difference of each observation from the mean.**

$$\text{Sample MAD} = \frac{\sum |x_i - \bar{x}|}{n}$$

$$\text{Population MAD} = \frac{\sum |x_i - \mu|}{N}$$

3.4 Measures of Dispersion

- Calculate MAD using the data from the Metals fund.

x_i	$x_i - \bar{x}$	$\lvert x_i - \bar{x} \rvert$
-7.34	$-7.34 - 24.65 = -31.99$	31.99
18.33	$18.33 - 24.65 = -6.32$	6.32
⋮	⋮	⋮
76.46	$76.46 - 24.65 = 51.81$	51.81
	Total = 0 (approximately)	Total = 271.12

$$\text{MAD} = \frac{\sum \lvert x_i - \bar{x} \rvert}{n} = \frac{271.12}{10} = 27.11.$$

LO 3.4

3.4 Measures of Dispersion

- **Variance and standard deviation**

- **For a given sample,**

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} \quad \text{and} \quad s = \sqrt{s^2}$$

- **For a given population,**

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N} \quad \text{and} \quad \sigma = \sqrt{\sigma^2}$$

3.4 Measures of Dispersion

- Calculate the variance and the standard deviation using the data from the Metals fund.

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
−7.34	−7.34 − 24.65 = −31.99	$(-31.99)^2 = 1{,}023.36$
18.33	18.33 − 24.65 = −6.32	$(-6.32)^2 = 39.94$
⋮	⋮	⋮
76.46	76.46 − 24.65 = 51.81	$(51.81)^2 = 2{,}684.28$
	Total = 0 (approximately)	Total = 12,407.44

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1} = \frac{12{,}407.44}{10-1} = 1{,}378.60(\%)^2.$$

$$s = \sqrt{1{,}378.60} = 37.13(\%).$$

LO 3.4 3.4 Measures of Dispersion

- **Coefficient of variation (CV)**
  - **CV adjusts for differences in the magnitudes of the means.**
  - **CV is unitless, allowing easy comparisons of mean-adjusted dispersion across different data sets.**

$$\text{Sample CV} = \frac{s}{\bar{x}}$$

$$\text{Population CV} = \frac{\sigma}{\mu}$$

LO 3.4

3.4 Measures of Dispersion

- **Calculate the coefficient of variation (CV) using the data from the Metals fund and the Income fund.**

- **Metals fund: C**$= \frac{s}{\bar{x}} = \frac{37.13\%}{24.65\%} = 1.51$

- **Income fund: C**$= \frac{s}{\bar{x}} = \frac{11.07\%}{8.51\%} = 1.30$

Synopsis of Investment Decision

- Mean and median returns for the Metals fund are 24.65% and 33.83%, respectively.
- Mean and median returns for the Income fund are 8.51% and 7.34%, respectively.
- The standard deviation for the Metals fund and the Income fund are 37.13% and 11.07%, respectively.
- The coefficient of variation for the Metals fund and the Income fund are 1.51 and 1.30, respectively.

3.5 Mean-Variance Analysis and the Sharpe Ratio

LO 3.5 Explain mean-variance analysis and the Sharpe Ratio.

- **Mean-variance analysis:**
  - **The performance of an asset is measured by its rate of return.**
  - **The rate of return may be evaluated in terms of its reward (mean) and risk (variance).**
  - **Higher average returns are often associated with higher risk.**
- **The Sharpe ratio uses the mean and variance to evaluate risk.**

LO 3.5

3.5 Mean-Variance Analysis and the Sharpe Ratio

- **Sharpe Ratio**
  - **Measures the extra reward per unit of risk.**
  - **For an investment I, the Sharpe ratio is computed as:**

$$\text{Sharpe Ratio} = \frac{\bar{x}_I - \bar{R}_f}{s_I}$$

where $\bar{x}_I$ is the mean return for the investment
$\bar{R}_f$ is the mean return for a risk-free asset
s_I is the standard deviation for the investment

LO 3.5

3.5 Mean-Variance Analysis and the Sharpe Ratio

- **Sharpe Ratio Example**
  - **Compute the Sharpe ratios for the Metals and Income funds given the risk free return of 4%.**
  - $$\text{Metals fund: } \frac{\bar{x}_I - \bar{R}_f}{s_I} = \frac{24.65 - 4}{37.13} = 0.56$$
    $$\text{Income fund: } \frac{\bar{x}_I - \bar{R}_f}{s_I} = \frac{8.51 - 4}{11.07} = 0.41.$$
  - **Since 0.56 > 0.41, the Metals fund offers more reward per unit of risk as compared to the Income fund.**

3.6 Chebyshev's Theorem and the Empirical Rule

LO 3.6 Apply Chebyshev's Theorem and the empirical rule.

- **Chebyshev's Theorem**
  - **For any data set, the proportion of observations that lie within *k* standard deviations from the mean is at least $1–1/k^2$, where *k* is any number greater than 1.**

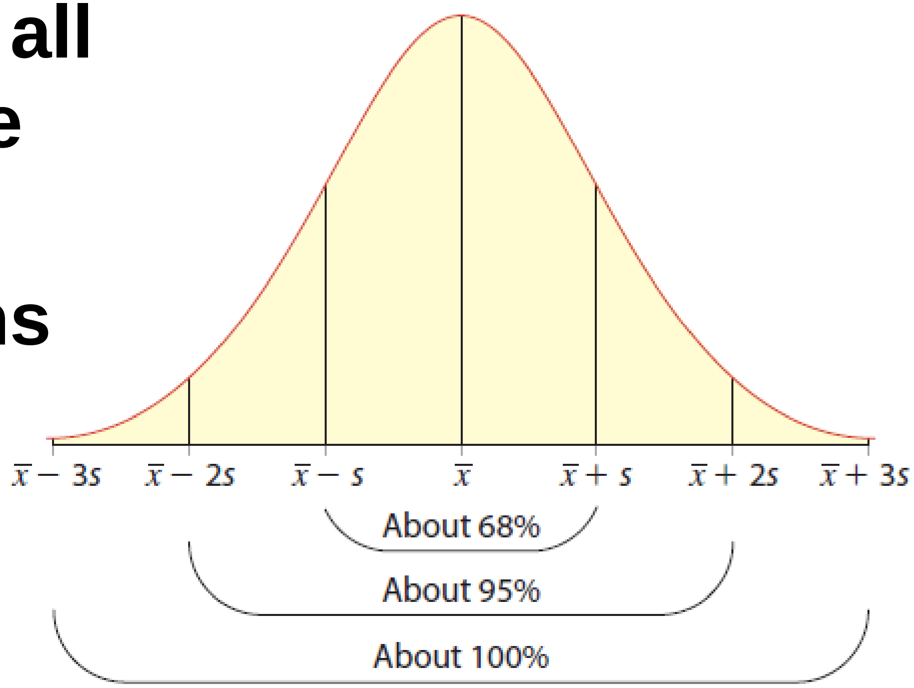
- **Consider a large lecture class with 280 students. The mean score on an exam is 74 with a standard deviation of 8. At least how many students scored within 58 and 90?**

  With $k = 2$, we have $1–1/2^2 = 0.75$. At least 75% of 280 or 210 students scored within 58 and 90.

LO 3.6

3.6 Chebyshev's Theorem and the Empirical Rule

- **The Empirical Rule:**
  - **Approximately 68% of all observations fall in the interval** $\bar{x} \pm s$ **.**
  - **Approximately 95% of all observations fall in the interval** $\bar{x} \pm 2s$ **.**
  - **Almost all observations fall in the interval** $\bar{x} \pm 3s$ **.**



LO 3.6

3.6 Chebyshev's Theorem and the Empirical Rule

- **Reconsider the example of the lecture class with 280 students with a mean score of 74 and a standard deviation of 8. Assume that the distribution is symmetric and bell-shaped. Approximately how many students scored within 58 and 90?**
  - **The score 58 is two standard deviations below the mean while the score 90 is two standard deviations above the mean.**
  - **Therefore about 95% of 280 students, or 0.95(280) = 266 students, scored within 58 and 90.**

3.7 Summarizing Grouped Data

LO 3.7 Calculate the mean and the variance for grouped data.

- **When data are grouped or aggregated, we use these formulas:**

$$\text{Mean: } \bar{x} = \frac{\sum m_i f_i}{n}$$

$$\text{Variance: } s^2 = \frac{\sum (m_i - \bar{x})^2 f_i}{n-1}$$

$$\text{Standard Deviation: } s = \sqrt{s^2}$$

where m_i and f_i are the midpoint and frequency of the ith class, respectively.

LO 3.7

3.7 Summarizing Grouped Data

- **Consider the frequency distribution of house prices.**
- **Calculate the average house price.**

Class (in \$1,000s)	f_i	m_i	$m_i f_i$	$(m_i - \bar{x})^2 f_i$
300 up to 400	4	350	1,400	$(350 - 522)^2 \times 4 = 118{,}336$
400 up to 500	11	450	4,950	$(450 - 522)^2 \times 11 = 57{,}024$
500 up to 600	14	550	7,700	$(550 - 522)^2 \times 14 = 10{,}976$
600 up to 700	5	650	3,250	$(650 - 522)^2 \times 5 = 81{,}920$
700 up to 800	2	750	1,500	$(750 - 522)^2 \times 2 = 103{,}968$
Total	36		18,800	372,224

- **For the mean, first multiply each class's midpoint by its respective frequency.**
- **Finally, sum the fourth column and divide by the sample size to obtain the mean = 18,800/36 = 522 or \$522,000.**

LO 3.7

3.7 Summarizing Grouped Data

- **Calculate the sample variance and the standard deviation.**

Class (in $1,000s)	f_i	m_i	$m_i f_i$	$(m_i - \bar{x})^2 f_i$
300 up to 400	4	350	1,400	$(350 - 522)^2 \times 4 = 118{,}336$
400 up to 500	11	450	4,950	$(450 - 522)^2 \times 11 = 57{,}024$
500 up to 600	14	550	7,700	$(550 - 522)^2 \times 14 = 10{,}976$
600 up to 700	5	650	3,250	$(650 - 522)^2 \times 5 = 81{,}920$
700 up to 800	2	750	1,500	$(750 - 522)^2 \times 2 = 103{,}968$
Total	36		18,800	372,224

- **First calculate the sum of the weighted squared differences from the mean.**
  - **Dividing this sum by (n–1) = 36–1 = 35 yields a variance of 10.635($)^2.**
  - **The square root of the variance yields a standard deviation of $103.13.**

LO 3.7

3.7 Summarizing Grouped Data

- **Weighted Mean**

- **Let $w_1, w_2, \ldots, w_n$ denote the weights of the sample observations $x_1, x_2, \ldots, x_n$ such that** $w_1 + w_2 + \ldots + w_n = 1$ **, then**

$$\bar{x} = \sum w_i x_i$$

LO 3.7

3.7 Summarizing Grouped Data

- **A student scores 60 on Exam 1, 70 on Exam 2, and 80 on Exam 3. What is the student's average score for the course if Exams 1, 2, and 3 are worth 25%, 25%, and 50% of the grade, respectively?**

- **Define w_1 = 0.25, w_2 = 0.25, and w_3 = 0.50.**

$$\bar{x} = \sum(w_i x_i) = 0.25(60) + 0.25(70) + 0.5(80) = 72.50$$

- **The unweighted mean is only 70 as it does not incorporate the higher weight given to the score on Exam 3.**

3.8 Covariance and Correlation

LO 3.8 Calculate and interpret the covariance and the correlation coefficient.

- **The covariance (s_{xy} or σ_{xy}) describes the direction of the linear relationship between two variables, *x* and *y*.**

- **The correlation coefficient (r_{xy} or ρ_{xy}) describes both the direction and strength of the relationship between *x* and *y*.**

LO 3.8

3.8 Covariance and Correlation

- **The sample covariance s_{xy} is computed as**

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- **The population covariance σ_{xy} is computed as**

$$\sigma_{xy} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N}$$

LO 3.8

3.8 Covariance and Correlation

- **The sample correlation r_{xy} is computed as**

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

- **The population correlation ρ_{xy} is computed as**

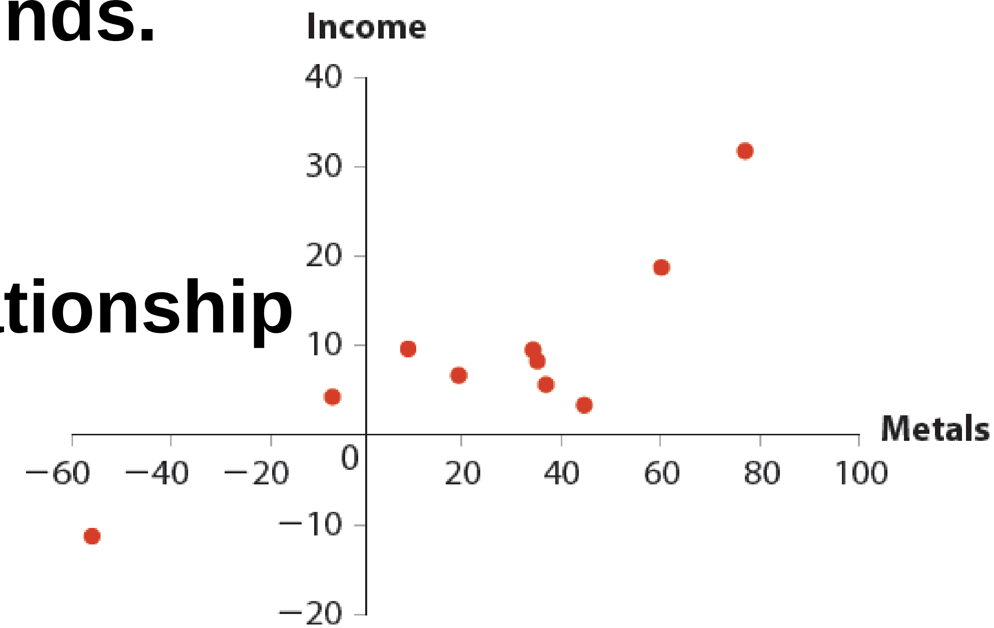
$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

- **Note, $-1 \le r_{xy} \le +1$ or $-1 \le \rho_{xy} \le +1$**

LO 3.8

3.8 Covariance and Correlation

- **Let' s calculate the covariance and the correlation coefficient for the Metals (*x*) and Income (*y*) funds.**



- **Positive Relationship**

$$\bar{x} = 24.65,\ s_x = 37.13,\ \bar{y} = 8.51,\ s_y = 11.07$$

- **Also recall:**

LO 3.8

3.8 Covariance and Correlation

- **We use the following table for the calculations.**

x_i	y_i	$(x_i - \bar{x})(y_i - \bar{y})$
−7.34	4.07	(−7.34 − 24.65)(4.07 − 8.51) = 142.04
18.33	6.52	(18.33 − 24.65)(6.52 − 8.51) = 12.58
⋮	⋮	⋮
76.46	31.77	(76.46 − 24.65)(31.77 − 8.51) = 1,205.10
		Total = 3,165.55

- **Covariance:** $s_{xy} = \dfrac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \dfrac{3{,}165.55}{10 - 1} = 351.73$

- **Correlation** $r_{xy} = \dfrac{s_{xy}}{s_x s_y} = \dfrac{351.73}{(37.13)(11.07)} = 0.86$

End of Chapter