Statistical Analysis in Fin Mkts

MSF 502

Li Cai

7 Sampling and Sampling Distributions

CHAPTER

Chapter 7 Learning Objectives (LOs)

LO 7.1: Differentiate between a population parameter and a sample statistic.

LO 7.2: Explain common sample biases.

LO 7.3: Describe simple random sampling.

LO 7.4: Distinguish between stratified random sampling and cluster sampling.

LO 7.5: Describe the properties of the sampling distribution of the sample mean.

Chapter 7 Learning Objectives (LOs)

LO 7.6: Explain the importance of the central limit theorem.

LO 7.7: Describe the properties of the sampling distribution of the sample proportion.

LO 7.8: Use a finite population correction factor.

LO 7.9: Construct and interpret control charts for quantitative and qualitative data.

Marketing Iced Coffee

- In order to capitalize on the iced coffee trend, Starbucks offered for a limited time half-priced Frappuccino beverages between 3 pm and 5 pm.
- Anne Jones, manager at a local Starbucks, determines the following from past historical data:
  - 43% of iced-coffee customers were women.
  - 21% were teenage girls.
  - Customers spent an average of $4.18 on iced coffee with a standard deviation of $0.84.

Marketing Iced Coffee

- One month after the marketing period ends, Anne surveys 50 of her iced-coffee customers and finds:
  - 46% were women.
  - 34% were teenage girls.
  - They spent an average of $4.26 on the drink.
- Anne wants to use this survey information to calculate the probability that:
  - Customers spend an average of $4.26 or more on iced coffee.
  - 46% or more of iced-coffee customers are women.
  - 34% or more of iced-coffee customers are teenage girls.

7.1 Sampling

LO 7.1 Differentiate between a population parameter and sample statistic.

- **Population**—consists of all items of interest in a statistical problem.
  - **Population Parameter** is unknown.
- **Sample**—a subset of the population.
  - **Sample Statistic** is calculated from sample and used to make inferences about the population.
- **Bias**—the tendency of a sample statistic to systematically over- or underestimate a population parameter.

7.1 Sampling

LO 7.2 Explain common sample biases.

- **Classic Case of a "Bad" Sample: The *Literary Digest* Debacle of 1936**
  - **During the1936 presidential election, the *Literary Digest* predicted a landslide victory for Alf Landon over Franklin D. Roosevelt (FDR) with only a 1% margin of error.**
  - **They were wrong! FDR won in a landslide election.**
  - **The *Literary Digest* had committed *selection bias* by randomly sampling from their own subscriber/ membership lists, etc.**
  - **In addition, with only a 24% response rate, the *Literary Digest* had a great deal of non-response bias.**

LO 7.2

7.1 Sampling

- **Selection bias—a systematic exclusion of certain groups from consideration for the sample.**
  - **The *Literary Digest* committed selection bias by excluding a large portion of the population (e.g., lower income voters).**
- **Nonresponse bias—a systematic difference in preferences between respondents and non-respondents to a survey or a poll.**
  - **The *Literary Digest* had only a 24% response rate. This indicates that only those who cared a great deal about the election took the time to respond to the survey. These respondents may be atypical of the population as a whole.**

7.1 Sampling

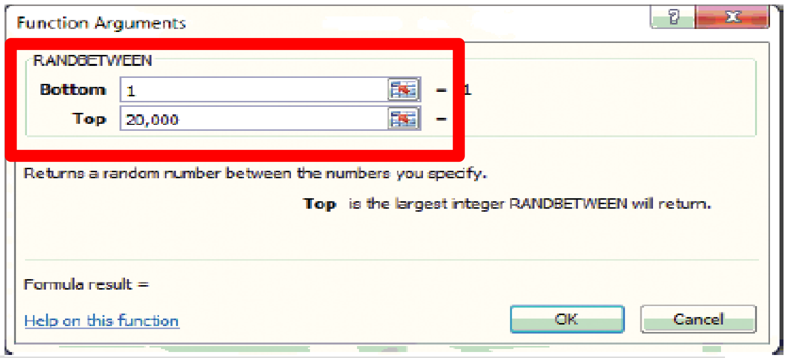
LO 7.3 Describe simple random sampling.

- **Sampling Methods**
  - **Simple random sample is a sample of *n* observations which has the same probability of being selected from the population as any other sample of *n* observations.**
    - **Most statistical methods presume simple random samples.**
    - **However, in some situations other sampling methods have an advantage over simple random samples.**

LO 7.3

7.1 Sampling

- **Example: In 1961, students invested 24 hours per week in their academic pursuits, whereas today's students study an average of 14 hours per week.**
  - **A dean at a large university in California wonders if this trend is reflective of the students at her university. The university has 20,000 students and the dean would like a sample of 100. Use Excel to draw a simple random sample of 100 students.**
  - **In Excel, choose Formulas > Insert function > RANDBETWEEN and input the values shown here.**



7.1 Sampling

LO 7.4 Distinguish between stratified random sampling and cluster sampling.

- **Stratified Random Sampling**
  - **Divide the population into mutually exclusive and collectively exhaustive groups, called strata.**
  - **Randomly select observations from each stratum, which are proportional to the stratum's size.**
  - **Advantages:**
    - **Guarantees that the each population subdivision is represented in the sample.**
    - **Parameter estimates have greater precision than those estimated from simple random sampling.**

LO 7.4

7.1 Sampling

- **Cluster Sampling**
  - **Divide population into mutually exclusive and collectively exhaustive groups, called clusters.**
  - **Randomly select clusters.**
  - **Sample every observation in those randomly selected clusters.**
  - **Advantages and disadvantages:**
    - **Less expensive than other sampling methods.**
    - **Less precision than simple random sampling or stratified sampling.**
    - **Useful when clusters occur naturally in the population.**

LO 7.4

7.1 Sampling

- **Stratified versus Cluster Sampling**
  - **Stratified Sampling**
    - **Sample consists of elements from each group.**
    - **Preferred when the objective is to increase precision.**
  - **Cluster Sampling**
    - **Sample consists of elements from the selected groups.**
    - **Preferred when the objective is to reduce costs.**

7.2 The Sampling Distribution of the Means

LO 7.5 Describe the properties of the sampling distribution of the sample mean.

- **Population is described by parameters.**
  - **A *parameter* is a *constant*, whose value may be unknown.**
  - **Only one population.**
- **Sample is described by statistics.**
  - **A *statistic* is a random variable whose value depends on the chosen random sample.**
  - **Statistics are used to make *inferences* about the population parameters.**
  - **Can draw multiple random samples of size *n*.**

LO 7.5

7.2 The Sampling Distribution of the Sample Mean

- **Estimator**
  - **A statistic that is used to estimate a population parameter.**
  - **For example, $\bar{X}$, the mean of the sample, is an estimator of μ, the mean of the population.**
- **Estimate**
  - **A particular value of the estimator.**
  - **For example, the mean of the sample $\bar{x}$ is an estimate of μ, the mean of the population.**

LO 7.5

7.2 The Sampling Distribution of the Sample Mean

- **Sampling Distribution of the Mean $\bar{x}$**
  - **Each random sample of size *n* drawn from the population provides an estimate of μ—the sample mean $\bar{x}$.**
  - **Drawing many samples of size *n* results in many different sample means, one for each sample.**
  - **The sampling distribution of the mean is the frequency or probability distribution of these sample means.**

LO 7.5

7.2 The Sampling Distribution of the Sample Mean

- **Example**

One simple random sample drawn from the population—a single *distribution of values* of *X*.

	Random Variable				
	X_1	X_2	X_3	X_4	Mean of X
	6	10	8	4	5.57
	5	10	4	3	5.71
	1	8	4	3	6.36
	4	1	6	2	4.07
	6	6	8	4	
	7	7	8	6	
	1	5	10	5	
	5	5	9	1	
	4	6	4	2	
	7	4	9	5	
	8	5	8	6	
	9	2	7	7	
	9	1	2	3	
	6	10	2	6	
Means	5.57	5.71	6.36	4.07	5.43

A *distribution of means* from each random draw from the population—a *sampling distribution*.

Means from each distribution (random draw) from the population.

LO 7.5

7.2 The Sampling Distribution of the Sample Mean

- **The Expected Value and Standard Deviation of the Sample Mean**
  - **Expected Value**
    - **The expected value of *X*,**

      $$E(X) = \mu$$

    - **The expected value of the mean,**

      $$E(\bar{X}) = E(X) = \mu$$

LO 7.5

7.2 The Sampling Distribution of the Sample Mean

- **The Expected Value and Standard Deviation of the Sample Mean**
  - **Variance of *X*** $Var(X) = \sigma^2$
  - **Standard Deviation**
    - **of *X*** $SD(X) = \sqrt{\sigma^2} = \sigma$
    - **of $\bar{X}$** $SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

Where *n* is the sample size. Also known as the *Standard Error of the Mean*.

LO 7.5

7.2 The Sampling Distribution of the Sample Mean

- **Example: Given that μ = 16 inches and σ = 0.8 inches, determine the following:**
  - **What is the expected value and the standard deviation of the sample mean derived from a random sample of**
    - **2 pizzas** $E\left(\bar{X}\right)=\mu=16$ $SD\left(\bar{X}\right)=\frac{\sigma}{\sqrt{n}}=\frac{0.8}{\sqrt{2}}=0.57$
    - **4 pizzas** $E\left(\bar{X}\right)=\mu=16$ $SD\left(\bar{X}\right)=\frac{\sigma}{\sqrt{n}}=\frac{0.8}{4}=0.40$

LO 7.5

7.2 The Sampling Distribution of the Sample Mean

Sampling from a Normal Distribution

- For any sample size *n*, the sampling distribution of $\bar{X}$ is *normal* if the population X from which the sample is drawn is normally distributed.
- If *X* is normal, then we can transform it into the *standard normal random variable* as:

For a sampling distribution.

$$Z = \frac{\bar{X} - E(\bar{X})}{SD(\bar{X})} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

For a distribution of the values of *X*.

$$Z = \frac{X - E(X)}{SD(X)} = \frac{X - \mu}{\sigma}$$

LO 7.5

7.2 The Sampling Distribution of the Sample Mean

Note that each value $\bar{x}$ on $\bar{X}$ has a corresponding value *z* on *Z* given by the transformation formula shown here as indicated by the arrows.

	Random Variable *X-bar*	Standard Normal *Z*	
$\bar{x}_1$	3	-2.39	$= z_1 = \dfrac{\bar{x}_1 - \mu}{\sigma/\sqrt{n}}$
$\bar{x}_2$	9	4.30	
⋮	4	-1.28	⋮
	2	-3.51	
	10	5.42	
	5	-0.16	
	9	4.30	
	4	-1.28	
	9	4.30	
	2	-3.51	
	3	-2.39	
	8	3.19	
	4	-1.28	
$\bar{x}_{13}$	0	-5.74	$= z_{13} = \dfrac{\bar{x}_{13} - \mu}{\sigma/\sqrt{n}}$
Means	5.14	0.00	
Standard Error	0.90	1.00	

LO 7.5

7.2 The Sampling Distribution of the Sample Mean

- **Example: Given that μ = 16 inches and σ = 0.8 inches, determine the following:**
  - **What is the probability that a randomly selected pizza is less than 15.5 inches?**
    - $$Z = \frac{x - \mu}{\sigma} = \frac{15.5 - 16}{0.8} = -0.63$$
      $$P(X < 15.5) = P(Z < -0.63) = 0.2643 \text{ or } 26.43\%$$
  - **What is the probability that 2 randomly selected pizzas average less than 15.5 inches?**
    - $$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{15.5 - 16}{0.8/\sqrt{2}} = -0.88$$
      $$P(\bar{X} < 15.5) = P(Z < -0.88) = 0.1894 \text{ or } 18.94\%$$

7.2 The Sampling Distribution of the Sample Mean

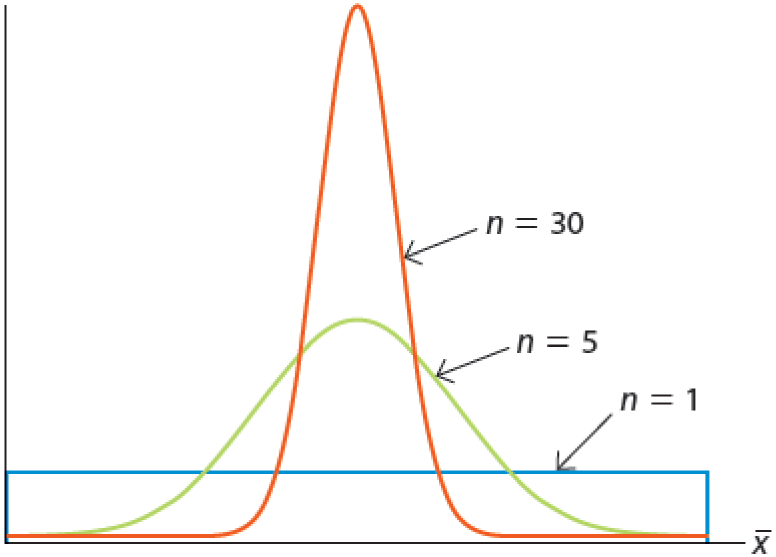
LO 7.6 Explain the importance of the central limit theorem.

- **The Central Limit Theorem**
  - **For any population *X* with expected value μ and standard deviation σ, the sampling distribution of $\bar{X}$ will be approximately normal if the sample size *n* is sufficiently large.**
  - **As a general guideline, the normal distribution approximation is justified when $n \geq 30$.**
  - **As before, if $\bar{X}$ is approximately normal, then we can transform it to** $Z = \dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}}$

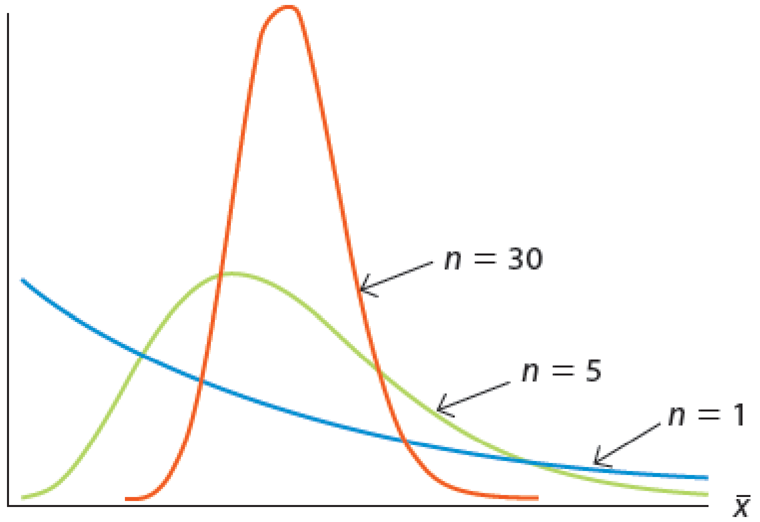
LO 7.6

7.2 The Sampling Distribution of the Sample Mean

- **The Central Limit Theorem**



Sampling distribution of $\bar{X}$ when the population has a uniform distribution.



Sampling distribution of $\bar{X}$ when the population has an exponential distribution.

LO 7.6

7.2 The Sampling Distribution of the Sample Mean

- **Example: From the introductory case, Anne wants to determine if the marketing campaign has had a lingering effect on the amount of money customers spend on iced coffee.**
  - **Before the campaign, μ = \$4.18 and σ = \$0.84. Based on 50 customers sampled after the campaign, $\bar{x}$ = \$4.26.**
  - **Let's find $P(\bar{X} \geq 4.26)$. Since $n \geq 30$, the central limit theorem states that $\bar{X}$ is approximately normal. So,**

$$P(\bar{X} \geq 4.26) = P\left(Z \geq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right) = P\left(Z \geq \frac{4.26 - 4.18}{0.84/\sqrt{50}}\right)$$

$$= P(Z \geq 0.67) = 1 - 0.7486 = 0.2514$$

7.3 The Sampling Distribution of the Sample Proportion

LO 7.7 Describe the properties of the sampling distribution of the sample proportion.

- **Estimator**
  - **Sample proportion $\bar{P}$ is used to estimate the population parameter *p*.**
- **Estimate**
  - **A particular value of the estimator $\bar{p}$.**

LO 7.7

7.3 The Sampling Distribution of the Sample Proportion

- **The Expected Value and Standard Deviation of the Sample Proportion**
  - **Expected Value**
    - **The expected value of $\bar{P}$,**

$$E\left(\bar{P}\right) = p$$

    - **The standard deviation of $\bar{P}$,**

$$SD\left(\bar{P}\right) = \sqrt{\frac{p\left(1-p\right)}{n}}$$

LO 7.7

7.3 The Sampling Distribution of the Sample Proportion

- **The Central Limit Theorem for the Sample Proportion**
  - **For any population proportion *p*, the sampling distribution of $\bar{P}$ is approximately normal if the sample size *n* is sufficiently large .**
  - **As a general guideline, the normal distribution approximation is justified when $np \geq 5$ *and* $n(1 - p) \geq 5$.**

LO 7.7

7.3 The Sampling Distribution of the Sample Proportion

- **The Central Limit Theorem for the Sample Proportion**
  - **If $\bar{P}$ is normal, we can transform it into the standard normal random variable as**

$$Z = \frac{\bar{P} - E(\bar{P})}{SD(\bar{P})} = \frac{\bar{P} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

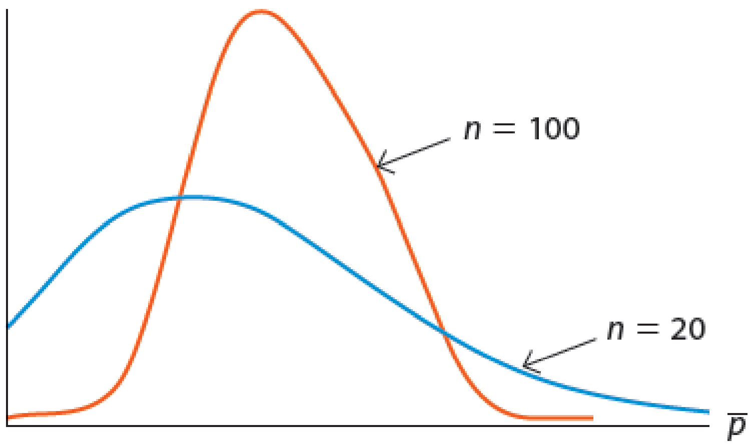
  - **Therefore any value $\bar{p}$ on $\bar{P}$ has a corresponding value z on Z given by**

$$Z = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

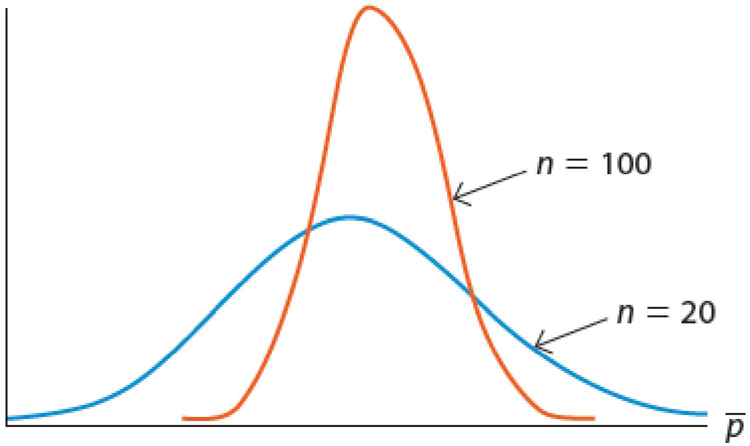
7.3 The Sampling Distribution of the Sample Proportion

LO 7.7

- **The Central Limit Theorem for the Sample Proportion**



Sampling distribution of $\bar{P}$ when the population proportion is p = 0.10.



Sampling distribution of $\bar{P}$ when the population proportion is p = 0.30.

LO 7.7

7.3 The Sampling Distribution of the Sample Proportion

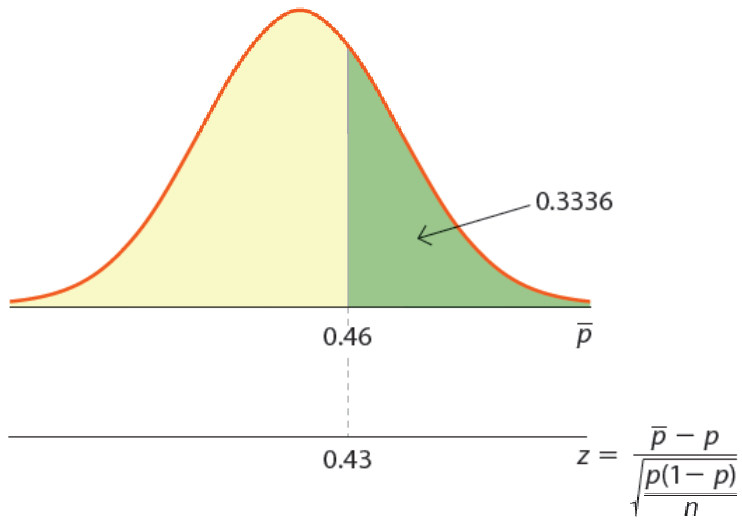
- **Example: From the introductory case, Anne wants to determine if the marketing campaign has had a lingering effect on the proportion of customers who are women and teenage girls.**
  - **Before the campaign, p = 0.43 for women and p = 0.21 for teenage girls. Based on 50 customers sampled after the campaign, p = 0.46 and p = 0.34, respectively.**
  - **Let's find $P\left(\bar{P} \geq 0.46\right)$. Since $n \geq 30$, the central limit theorem states that $\bar{P}$ is approximately normal.**

LO 7.7

7.3 The Sampling Distribution of the Sample Proportion

$$P\left(\bar{P} \geq 0.46\right) = P\left(Z \geq \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \right) = P\left(Z \geq \frac{0.46 - 0.43}{\sqrt{\frac{0.43(1-0.43)}{50}}} \right)$$

$$= P(Z \geq 0.43) = 1 - 0.6664 = 0.3336$$



7.4 The Finite Population Correction Factor

LO 7.8 Use a finite population correction factor.

- **The Finite Population Correction Factor**
  - **Used to reduce the sampling variation of $\bar{X}$.**
  - **The resulting standard deviation is**

$$SD\left(\bar{X}\right) = \frac{\sigma}{\sqrt{n}}\left(\sqrt{\frac{N-n}{N-1}}\right)$$

  - **The transformation of $\bar{X}$ to Z is made accordingly.**

LO 7.8

7.4 The Finite Population Correction Factor

- **The Finite Population Correction Factor for the Sample Proportion**
  - **Used to reduce the sampling variation of the sample proportion $\bar{P}$.**
  - **The resulting standard deviation is**

$$SD\left(\bar{P}\right)=\sqrt{\frac{p\left(1-p\right)}{n}}\left(\sqrt{\frac{N-n}{N-1}}\right)$$

  - **The transformation of $\bar{P}$ to Z is made accordingly.**

LO 7.8

7.4 The Finite Population Correction Factor

- **Example: A large introductory marketing class with 340 students has been divided up into 10 groups. Connie is in a group of 34 students that averaged 72 on the midterm. The class average was 73 with a standard deviation of 10.**
  - **The population parameters are: μ = 73 and σ = 10.**
  - $E\left(\bar{X}\right) = \mu = 73$ **but since *n* = 34 is more than 5% of the population size *N* = 340, we need to use the finite population correction factor.**

$$SD\left(\bar{X}\right) = \frac{\sigma}{\sqrt{n}}\left(\sqrt{\frac{N-n}{N-1}}\right) = \frac{10}{\sqrt{34}}\left(\sqrt{\frac{340-34}{340-1}}\right) = 1.63$$

7.5 Statistical Quality Control

LO 7.9 Construct and interpret control charts for quantitative and qualitative data.

- **Statistical Quality Control**
  - **Involves statistical techniques used to develop and maintain a firm's ability to produce high-quality goods and services.**
  - **Two Approaches for Statistical Quality Control**
    - **Acceptance Sampling**
    - **Detection Approach**

LO 7.9

7.5 Statistical Quality Control

- **Acceptance Sampling**
  - **Used at the completion of a production process or service.**
  - **If a particular product does not conform to certain specifications, then it is either discarded or repaired.**
  - **Disadvantages**
    - **It is costly to discard or repair a product.**
    - **The detection of all defective products is not guaranteed.**

LO 7.9

7.5 Statistical Quality Control

- **Detection Approach**
  - **Inspection occurs during the production process in order to detect any nonconformance to specifications.**
  - **Goal is to determine whether the production process should be continued or adjusted before producing a large number of defects.**
  - **Types of variation:**
    - **Chance variation.**
    - **Assignable variation.**

LO 7.9

7.5 Statistical Quality Control

- **Types of Variation**
  - **Chance variation (common variation) is:**
    - **Caused by a number of randomly occurring events that are part of the production process.**
    - **Not controllable by the individual worker or machine.**
    - **Expected, so not a source of alarm as long as its magnitude is tolerable and the end product meets specifications.**
  - **Assignable variation (special cause variation) is:**
    - **Caused by specific events or factors that can usually be identified and eliminated.**
    - **Identified and corrected or removed.**

LO 7.9

7.5 Statistical Quality Control

- **Control Charts**
  - **Developed by Walter A. Shewhart.**
  - **A plot of calculated statistics of the production process over time.**
  - **Production process is "in control" if the calculated statistics fall in an expected range.**
  - **Production process is "out of control" if calculated statistics reveal an undesirable trend.**
    - **For quantitative data— $\bar{x}$ chart.**
    - **For qualitative data— $\bar{p}$ chart.**

LO 7.9

7.5 Statistical Quality Control

- **Control Charts for Quantitative Data**
  - $\overline{X}$ **Control Charts**
    - **Centerline—the mean when the process is under control.**
    - **Upper control limit—set at +3σ from the mean.**
      - **Points falling above the upper control limit are considered to be *out of control*.**
    - **Lower control limit—set at –3σ from the mean.**
      - **Points falling below the lower control limit are considered to be *out of control*.**

LO 7.9

7.5 Statistical Quality Control

Control Charts for Quantitative Data

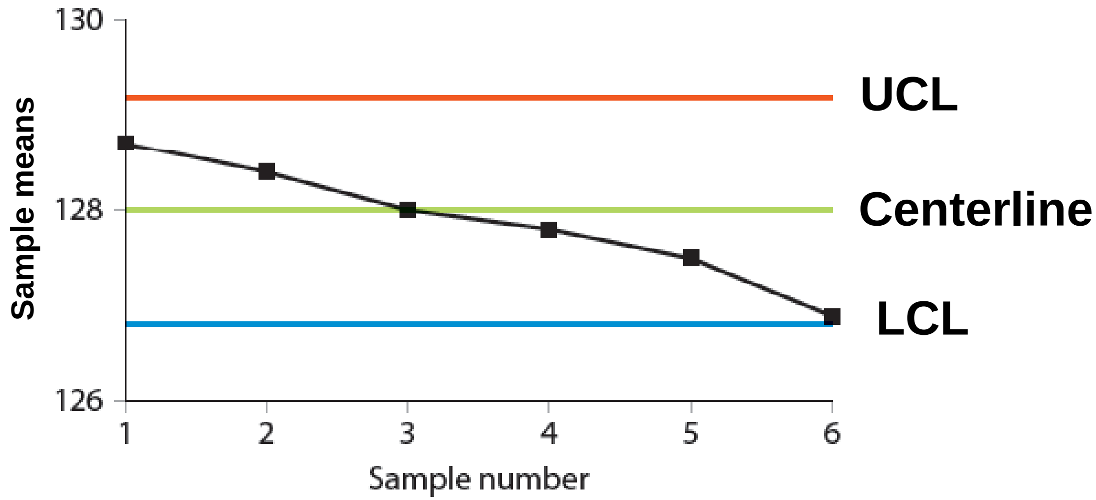
$\bar{X}$ Control Charts

- **Upper control limit (UCL):**

$$\mu + 3\frac{\sigma}{\sqrt{n}}$$

- **Lower control limit (LCL):**

$$\mu - 3\frac{\sigma}{\sqrt{n}}$$



Process is in control—all points fall within the control limits.

LO 7.9

7.5 Statistical Quality Control

- **Control Charts for Qualitative Data**
  - $\overline{p}$ **chart (fraction defective or percent defective chart).**
  - **Tracks proportion of defects in a production process.**
  - **Relies on central limit theorem for normal approximation for the sampling distribution of the sample proportion.**
  - **Centerline—the mean when the process is under control.**
  - **Upper control limit—set at +3σ from the centerline.**
    - **Points falling above the upper control limit are considered to be *out of control*.**
  - **Lower control limit—set at –3σ from the centerline.**
    - **Points falling below the lower control limit are considered to be *out of control*.**

LO 7.9

7.5 Statistical Quality Control

Control Charts for Qualitative Data

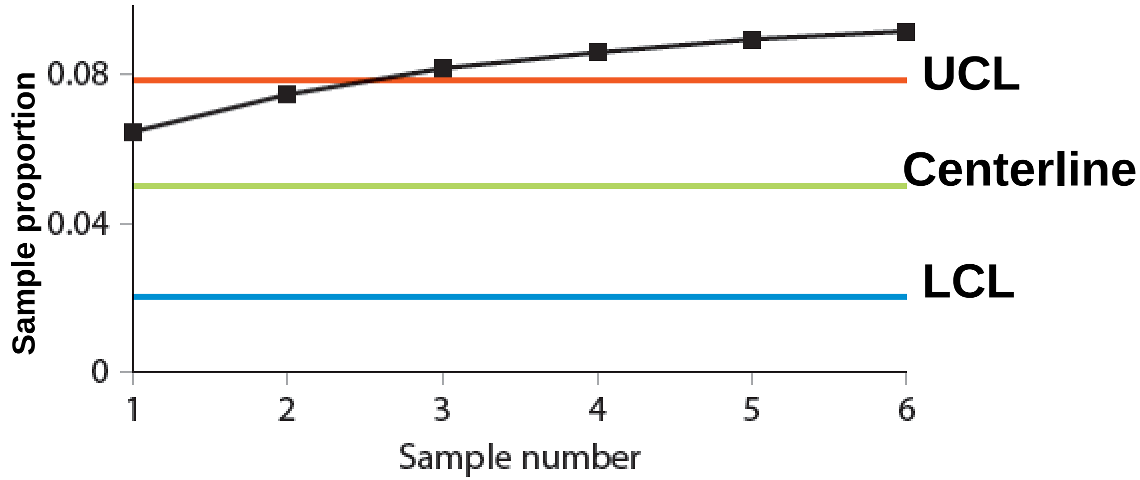
$\bar{p}$ Control Charts

- Upper control limit (UCL):

$$\bar{p}+3\sqrt{\frac{\bar{p}\left(1-\bar{p}\right)}{n}}$$

- Lower control limit (LCL):

$$\bar{p}-3\sqrt{\frac{\bar{p}\left(1-\bar{p}\right)}{n}}$$



Process is out of control—some points fall above the UCL.

End of Chapter