Marmara University, **2021**

Probability and Statistics

Subject 7

Sampling Distributions

Mujdat Soyturk, Ph.D.
Associate Professor

Contents

Most parts of the slides are derived from the textbook: "Mendenhall, Beaver, Beaver, *Introduction to Probability and Statistics*, 14th Ed., Brooks/Cole, Cengage Learning, 2013"

Introduction

- Parameters are numerical descriptive measures for populations.
  - For the normal distribution, the location and shape are described by μ and σ.
  - For a binomial distribution consisting of n trials, the location and shape are determined by p.
- Often the values of parameters that specify the exact form of a distribution are unknown.
- You must rely on the sample to learn about these parameters.

Sampling

Examples:

- A pollster is sure that the responses to his "agree/disagree" question will follow a binomial distribution, but p, the proportion of those who "agree" in the population, is unknown.
- An agronomist believes that the yield per acre of a variety of wheat is approximately normally distributed, but the mean μ and the standard deviation σ of the yields are unknown.
- Rely on the sample to learn about these parameters.

✓ If you want **the sample** to provide *reliable information about the population*, you must select your sample in a certain way!

Sampling Plan

- The **sampling plan** or **experimental design** determines the amount of information you can extract, and often allows you to measure the **reliability of your inference**.
- Sampling Plans
  - **Simple random sampling**.
  - Stratified random sampling
  - Cluster sampling
  - Systematic sampling

IMPORTANT: All sampling plans used for making inferences must involve randomization.

Simple Random Sampling

- **Simple random sampling** is a method of sampling that allows each possible sample of size n an equal probability of being selected.
- Either use random tables given or random number generators.

Example

There are 93 students in a statistics class. The instructor wants to choose 5 students to form a project group. How should he proceed?

1. Give each student a number from 01 to 93.
2. Choose 5 pairs of random digits from the random number table.
3. If a number between 94 and 00 is chosen, choose another number.
4. The five students with those numbers form the group.

06907	11008	42751	27756	53498
72905	56420	69994	98872	31016
91977	05463	07972	18876	20922
14342	63661	10281	17453	18103
36857	53342	53988	53060	59533
69578	88231	33276	70997	79936
40961	48235	03427	49626	69445
93969	52636	92737	88974	33488
61129	87529	85689	48237	52267
97336	71048	08178	77233	13916

Types of Samples

Sampling can occur in two types of practical situations

1. **Observational studies:** The data existed before you decided to study it. Watch out for
   - ✓ **Nonresponse:** Are the responses biased because only opinionated people responded?
   - ✓ **Undercoverage:** Are certain segments of the population systematically excluded?
   - ✓ **Wording bias:** The question may be too complicated or poorly worded.

Observational study: the data already existed before you decided to *observe* or *describe their characteristics*.

Types of Samples

Sampling can occur in two types of practical situations

2. **Experimentation:** The data are generated by imposing an experimental condition or treatment on the experimental units.
   - ✓ **Hypothetical populations** can make random sampling difficult if not impossible.
   - ✓ Samples must sometimes be chosen so that the experimenter believes they are **representative** of the whole population.
   - ✓ Samples must **behave like random samples!**

Other Sampling Plans

There are several other sampling plans that still involve **randomization**:

1. **Stratified random sample:** Divide the population into subpopulations or **strata** and select a simple random sample from each strata.
2. **Cluster sample:** Divide the population into subgroups called **clusters;** select a simple random sample of clusters and take a census of every element in the cluster.
3. **1-in-k systematic sample**: Randomly select one of the first k elements in an ordered population, and then select every k-th element thereafter.

Examples

- Divide California into counties and take a simple random sample within each county. **Stratified**
- Divide California into counties and take a simple random sample of 10 counties. **Cluster**
- Divide a city into city blocks, choose a simple random sample of 10 city blocks, and interview all who live there. **Cluster**
- Choose an entry at random from the phone book, and select every 50th number thereafter. **1-in-50 Systematic**

Non-Random Sampling Plans

There are several other sampling plans that do not involve **randomization**. They should **NOT** be used for statistical inference!

1. **Convenience sample:** A sample that can be taken easily without random selection.
   - People walking by on the street
2. **Judgment sample:** The sampler decides who will and won't be included in the sample.
3. **Quota sample**: The makeup of the sample must reflect the makeup of the population on some selected characteristic.
   - Race, ethnic origin, gender, etc.

Sampling Distributions

- Numerical descriptive measures calculated from the sample are called **statistics**.
- Statistics vary from sample to sample and hence are random variables.
- The probability distributions for statistics are called **sampling distributions**.
- In repeated sampling, they tell us:
  - what values of the statistics can occur and,
  - how often each value occurs.

Sampling Distributions

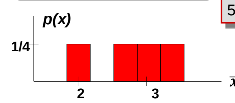
Definition: The **sampling distribution of a statistic** is the probability distribution for the possible values of the statistic that results when random samples of size *n* are repeatedly drawn from the population.

Population: 3, 5, 2, 1

Draw samples of size $n = 3$ without replacement

Possible samples	$\bar{x}$
3, 5, 2	$10/3 = 3.33$
3, 5, 1	$9/3 = 3$
3, 2, 1	$6/3 = 2$
5, 2, 1	$8/3 = 2.67$

Each value of x-bar is equally likely, with probability 1/4



Sampling Distributions

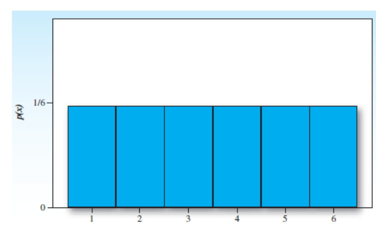
Three ways to find the sampling distributions for a statistics

- Approximated with simulation techniques
- Derived using mathematical theorems
- Use statistical theorems to derive ***exact*** or ***approximate*** sampling distributions. → *The Central Limit Theorem* is one such theorem.

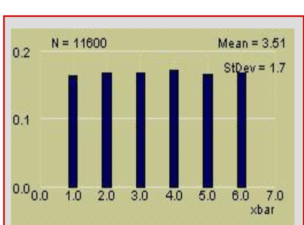
- *Central Limit Theorem* states that, under rather general conditions, sums and means of random samples of measurements drawn from a population tend to have an approximate normal distribution.

Example

Toss a fair dice $n = 1$ time. The distribution of x the number on the upper face is flat or **uniform.**



Example



$$\mu = \sum xp(x)$$

$$= 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + \ldots + 6\left(\frac{1}{6}\right) = 3.5$$

$$\sigma = \sqrt{\sum (x-\mu)^2 p(x)} = 1.71$$

Sample Size (n)	Mean ($\bar{X}$)	Std.Dev. (s)
1	3.5	1.71

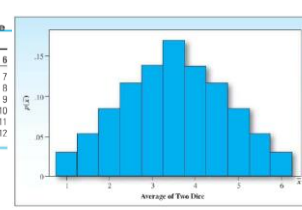
MY APPLET

Example

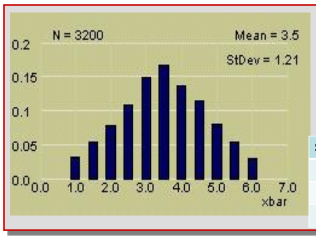
Toss a fair dice $n = 2$ times. The distribution of $\bar{X}$, the average number on the two upper faces, is **mound-shaped.**

Sums of the Upper Faces of Two Dice

Second Die	First Die 1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12



Example



Mean : $\mu = 3.5$

Std Dev :

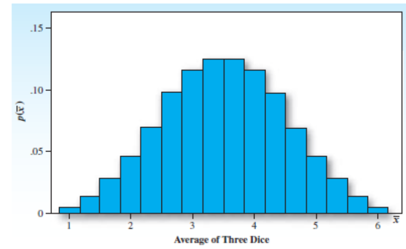
$\sigma/\sqrt{2} = 1.71/\sqrt{2} = 1.21$

Sample Size (n)	Mean ( )	Std.Dev. (s)
1	3.5	1.71
2	3.5	1.21

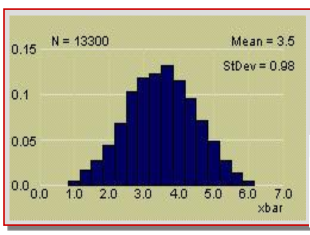
MY APPLET

Example

Toss a fair dice n = 3 times. The distribution of $\bar{X}$, the average number on the three upper faces is approximately normal.



Example



Mean : $\mu = 3.5$

Std Dev :

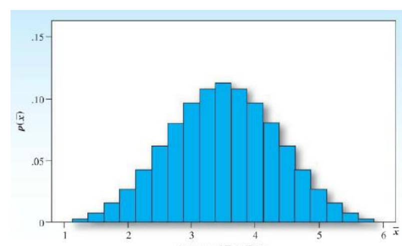
$\sigma/\sqrt{3} = 1.71/\sqrt{3} = .987$

Sample Size (n)	Mean ( )	Std.Dev. (s)
1	3.5	1.71
2	3.5	1.21
3	3.5	0.987

MY APPLET

Example

Toss a fair dice n = 4 times. The distribution of $\bar{X}$, the average number on the four upper faces is **approximately normal.**



Results

As the sample size n increases,

- spread of the distribution is slowly decreasing.
- shape of the normal probability distribution of $\bar{X}$ didn't change (still centered at μ=3.5).
- shape of the normal probability distribution of $\bar{X}$ turns out to be mound-shaped.
- distribution of $\bar{X}$ is approximately normally distributed.

Central Limit Theorem

Central Limit Theorem: If random samples of n observations are drawn from a nonnormal population with finite μ and standard deviation σ, then, when n is large, the sampling distribution of the sample mean $\bar{x}$ is approximately normally distributed, with mean μ and standard deviation $\sigma/\sqrt{n}$. The approximation becomes more accurate as n becomes large.

The sampling distribution of $\bar{x}$ always has a mean μ and standard deviation $\sigma/\sqrt{n}$. The CLT helps describe its shape.

Regardless of its shape, the sampling distribution of $\bar{x}$ always has a mean identical to the mean of the sampled population, and a standard deviation equal to the population standard deviation σ divided by $\sqrt{n}$. *Consequently, the spread of the distribution of sample means is considerably less than the spread of the sampled population.*

Why is this Important?

- The **Central Limit Theorem** also implies that the sum of n measurements is approximately normal with mean $n\mu$ and standard deviation $\sigma\sqrt{n}$.
- Contribution is in *statistical inference*. Many statistics that are used for *statistical inference* are **sums** or **averages** of sample measurements.
- When n is large, these statistics will have approximately **normal** distributions.
- This will allow us to describe their behavior and evaluate the **reliability** of our inferences.

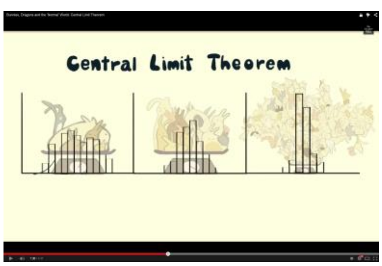
How Large is Large?

When the sample size is large enough to use the CLT:

- If the sampled population is **normal**, then the sampling distribution of $\bar{x}$ will also be normal, no matter what the sample size.
- When the sampled population is approximately **symmetric**, the sampling distribution becomes approximately normal for relatively small values of n.
- When the sampled population is **skewed**, the sample size must be **at least 30** before the sampling distribution of $\bar{x}$ becomes approximately normal.

Central Limit Theorem by a Video

Bunnies, Dragons and the 'Normal' World: *Central Limit Theorem*



The Sampling Distribution of the Sample Mean

- If a random sample of n measurements is selected from a population with mean μ and standard deviation σ, the sampling distribution of the sample mean $\bar{x}$ will have mean μ and standard deviation $\sigma / \sqrt{n}$.
- If the population has a **normal** distribution, the sampling distribution of $\bar{x}$ will be ***exactly*** normally distributed, ***regardless of the sample size, n***.
- If the population distribution is **nonnormal,** the sampling distribution of $\bar{x}$ will be ***approximately*** normally distributed when sample size n is large.

The standard deviation of *x*-bar is sometimes called the STANDARD ERROR (SE).

Standard Error

Definition The standard deviation of a statistic used as an estimator of a population parameter is also called the **standard error of the estimator** (abbreviated **SE**) because it refers to the precision of the estimator. Therefore, the standard deviation of $\bar{x}$—given by $\sigma/\sqrt{n}$—is referred to as the **standard error of the mean** (abbreviated as SE($\bar{x}$), SEM, or sometimes just SE).

Finding Probabilities for the Sample Mean

✓If the sampling distribution of $\bar{x}$ is normal or approximately normal, *standardize or rescale* the interval of interest in terms of

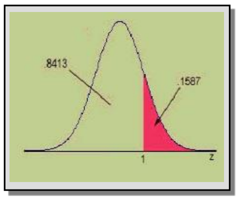
$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

✓Find the appropriate area using Table 3.

Finding Probabilities for the Sample Mean

Example: A random sample of size n = 16 from a normal distribution with μ = 10 and σ = 8.

$$P(\bar{x} > 12) = P(z > \frac{12-10}{8/\sqrt{16}})$$
$$= P(z > 1) = 1 - .8413 = .1587$$

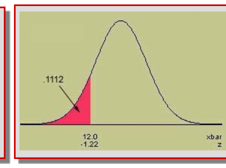


Example

A soda filling machine is supposed to fill cans of soda with 12 fluid ounces. Suppose that the fills are actually normally distributed with a mean of 12.1 oz and a standard deviation of .2 oz. What is the probability that the average fill for a 6-pack of soda is less than 12 oz?

MY APPLET

$$P(\bar{x} < 12) =$$
$$P(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} < \frac{12-12.1}{.2/\sqrt{6}}) =$$
$$P(z < -1.22) = .1112$$



The Sampling Distribution of the Sample Proportion

- We use a *random sample of* ***n*** *people* to estimate the proportion ***p*** of people in the <u>population</u> who have a *specified characteristic (e.g. preference in polls)*.
- If ***x*** of the sampled people have this characteristic, then the sample proportion $\hat{p}$ is;

$$\hat{p} = \frac{x}{n}$$

can be used to <u>*estimate the population proportion* ***p***</u>.

- $\hat{p}$ is simply a *rescaling* of the binomial random variable *x*, dividing it by *n*.

The Sampling Distribution of the Sample Proportion

- The **Central Limit Theorem** can be used to conclude that the binomial random variable *x* is approximately normal when *n* is large, with mean ***np*** and standard deviation $\sqrt{npq}$.
- The sample proportion, $\hat{p} = \frac{x}{n}$ is simply a *rescaling* of the binomial random variable *x*, dividing it by *n*.
- From the Central Limit Theorem, the sampling distribution of $\hat{p} = \frac{x}{n}$ will also be **approximately normal,** with a *rescaled* mean and standard deviation.

The Sampling Distribution of the Sample Proportion

- A random sample of size ***n*** is selected from a binomial population with parameter ***p***.
- The sampling distribution of the sample proportion,

$$\hat{p} = \frac{x}{n}$$

- will have <u>mean ***p***</u> and standard deviation $SE(\hat{p}) = \sqrt{\frac{pq}{n}}$
- If *n* is large, and *p* is not too close to zero or one, the sampling distribution of $\hat{p}$ will be **approximately normal.**

The standard deviation of *p*-hat is sometimes called the STANDARD ERROR (SE) of *p*-hat.

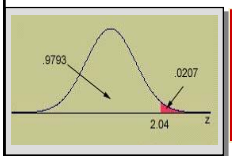
Finding Probabilities for the Sample Proportion

✓If the sampling distribution of $\hat{p}$ is normal or approximately normal, *standardize or rescale* the interval of interest in terms of $z = \frac{\hat{p}-p}{\sqrt{\frac{pq}{n}}}$

✓Find the appropriate area using Table 3.

Finding Probabilities for the Sample Proportion

Example: A random sample of size n = 100 from a binomial population with p = .4.



$$P(\hat{p} > .5) = P(z > \frac{.5 - .4}{\sqrt{\frac{.4(.6)}{100}}})$$
$$= P(z > 2.04) = 1 - .9793 = .0207$$

Example

The soda bottler in the previous example claims that only 5% of the soda cans are underfilled. A quality control technician randomly samples 200 cans of soda. What is the probability that more than 10% of the cans are underfilled?

***n* = 200**
S: underfilled can
***p* = P(S) = .05**
***q* = .95**
***np* = 10 *nq* = 190**

OK to use the normal approximation

$$P(\hat{p} > .10)$$
$$= P(z > \frac{.10 - .05}{\sqrt{\frac{.05(.95)}{200}}}) = P(z > 3.24)$$
$$= 1 - .9994 = .0006$$

This would be very unusual, if indeed p = .05!

Statistical Process Control

- Product quality is usually monitored using statistical control charts. Measurements on a process variable change or vary over time.
- The cause of a change in the variable is said to be **assignable** if it can be found and corrected.
- Other variation that is not controlled is regarded as **random variation.**
- If the variation in a process variable is solely random, the process is said to be **in control**.
- If out of control, we must reduce the variation and get the measurements on the process variable within ***specification limits***.

The $\bar{x}$ Chart for Process Means

- At various times during production, we take a sample of size n and calculate the sample mean $\bar{x}$.
- According to the CLT, the sampling distribution of $\bar{x}$ should be approximately normal; almost all of the values of $\bar{x}$ should fall into the interval

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

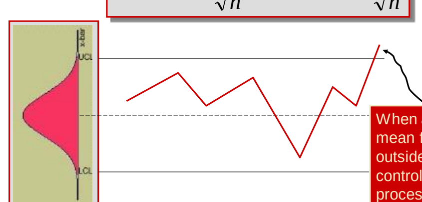
- If a value of $\bar{x}$ falls outside of this interval, the process may be out of control.

The $\bar{x}$ Chart

- To create a control chart, collect data on k samples of size n. Use the sample data to estimate μ and σ.
- The mean μ is estimated with $\bar{\bar{x}}$, the grand average of all the sample statistics calculated for the nk measurements on the process variable.
- The standard deviation σ is estimated by ***s***, the standard deviation of the nk measurements.
- Create the control chart, using a **centerline** and **control limits.** (centerline is estimated mean $\bar{\bar{x}}$, and control limits are placed three standard deviations above and below the centerline.)

The $\bar{x}$ Chart

$$\text{Centerline}: \bar{\bar{x}}$$
$$\text{LCL}: \bar{\bar{x}} - 3\frac{s}{\sqrt{n}} \qquad \text{UCL}: \bar{\bar{x}} + 3\frac{s}{\sqrt{n}}$$



When a sample mean falls outside the control limits, the process may be out of control.

Example

A statistical process control monitoring system samples the inside diameters of $n = 4$ bearings each hour. Table 7.6 provides the data for $k = 25$ hourly samples. Construct an $\bar{x}$ chart for monitoring the process mean.

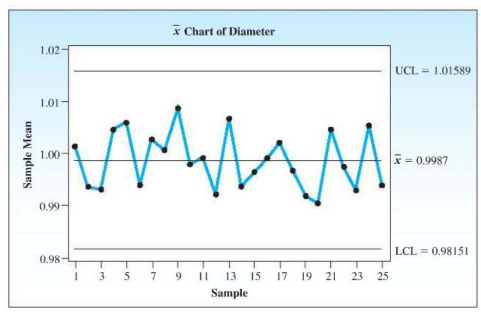
Example

Sample		Sample Measurements			Sample Me $\bar{x}$
1	.992	1.007	1.016	.991	1.00150
2	1.015	.984	.976	1.000	.99375
3	.988	.993	1.011	.981	.99325
4	.996	1.020	1.004	.999	1.00475
5	1.015	1.006	1.002	1.001	1.00600
6	1.000	.982	1.005	.989	.99400
7	.989	1.009	1.019	.994	1.00275
8	.994	1.010	1.009	.990	1.00075
9	1.018	1.016	.990	1.011	1.00875
10	.997	1.005	.989	1.001	.99800
11	1.020	.986	1.002	.989	.99925
12	1.007	.986	.981	.995	.99225
13	1.016	1.002	1.010	.999	1.00675
14	.982	.995	1.011	.987	.99375
15	1.001	1.000	.983	1.002	.99650
16	.992	1.008	1.001	.996	.99925
17	1.020	.988	1.015	.986	1.00225
18	.993	.987	1.006	1.001	.99675
19	.978	1.006	1.002	.982	.99200
20	.984	1.009	.983	.986	.99050
21	.990	1.012	1.010	1.007	1.00475
22	1.015	.983	1.003	.989	.99750
23	.983	.990	.997	1.002	.99300
24	1.011	1.012	.991	1.008	1.00550
25	.987	.987	1.007	.995	.99400

Example

- $\bar{x}_1 = \frac{.992 + 1.007 + 1.016 + .991}{4} = 1.0015$ and find all x
- Centerline, $\bar{\bar{x}} = 24.9675/25 = .9987$
- s, for all observations (n.k)=4 x 25 = 100 is .011458
- Estimated standard error of mean for n=4 observations, $\frac{s}{\sqrt{n}} = \frac{0.011458}{\sqrt{4}} = 0.005729$
- $UCL = \bar{\bar{x}} + 3\frac{s}{\sqrt{n}} = 0.9987 + 3(.005729) = 1.015887$
- $LCL = \bar{\bar{x}} - 3\frac{s}{\sqrt{n}} = 0.9987 - 3(.005729) = .981513$

Example



The *p* Chart for Proportion Defective

- At various times during production, we take a sample of size *n* and calculate the proportion of defective items $\hat{p} = x/n$
- According to the CLT, the sampling distribution of $\hat{p}$ should be approximately normal; almost all of the values of $\hat{p}$ should fall into the interval $\boxed{p \pm 3\sqrt{\frac{pq}{n}}}$
- If a value of $\hat{p}$ falls outside of this interval, the process may be out of control.

The *p* Chart

- To create a control chart, collect data on *k* samples of size *n*. Use the sample data to estimate *p*.
- The population proportion defective *p* is estimated with

$$\bar{p} = \frac{\sum \hat{p}_i}{k}$$

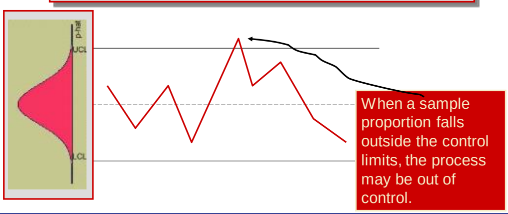
the grand average of all the sample proportions calculated for the *k* samples.

- Create the control chart, using a **centerline** and **control limits.**

The p Chart

$$\text{Centerline}: \bar{p}$$

$$\text{LCL}: \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \qquad \text{UCL}: \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$



Example

A manufacturer of ballpoint pens randomly samples 400 pens per day and tests each to see whether the ink flow is acceptable. The proportions of pens judged defective each day over a 40-day period are listed in Table 7.7. Construct a control chart for the proportion $\hat{p}$ defective in samples of $n = 400$ pens selected from the process.

Example

Proportions of Defectives in Samples of n = 400 Pens

Day	Proportion	Day	Proportion	Day	Proportion	Day	Proportion
1	.0200	11	.0100	21	.0300	31	.0225
2	.0125	12	.0175	22	.0200	32	.0175
3	.0225	13	.0250	23	.0125	33	.0225
4	.0100	14	.0175	24	.0175	34	.0100
5	.0150	15	.0275	25	.0225	35	.0125
6	.0200	16	.0200	26	.0150	36	.0300
7	.0275	17	.0225	27	.0200	37	.0200
8	.0175	18	.0100	28	.0250	38	.0150
9	.0200	19	.0175	29	.0150	39	.0150
10	.0250	20	.0200	30	.0175	40	.0225

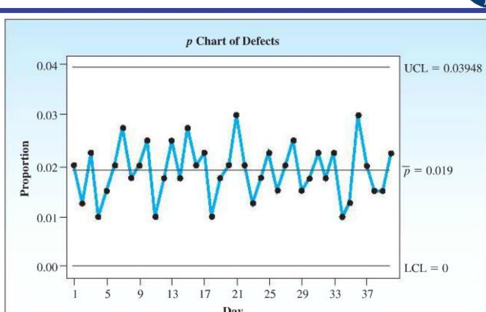
Example

- Centerline $\bar{p} = \frac{\sum \hat{p}_i}{k} = \frac{.0200+.0125+\cdots+.0225}{40} = \frac{.7600}{40} = 0.19$
- Estimate of SE of sample proportion,

$$\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = \sqrt{\frac{.019\,(.981)}{400}} = .00683$$

- $UCL = \bar{p} + 3SE = .0190 + .0205 = .0395$
- $LCL = \bar{p} - 3SE = .0190 - .0205 = -.0015$

Example



Key Concepts

I. Sampling Plans and Experimental Designs

1. Simple random sampling
   a. Each possible sample is equally likely to occur.
   b. Use a computer or a table of random numbers.
   c. Problems are nonresponse, undercoverage, and wording bias.
2. Other sampling plans involving randomization
   a. Stratified random sampling
   b. Cluster sampling
   c. Systematic 1-in-*k* sampling

Key Concepts

3. Nonrandom sampling
   a. Convenience sampling
   b. Judgment sampling
   c. Quota sampling

II. Statistics and Sampling Distributions

1. Sampling distributions describe the possible values of a statistic and how often they occur in repeated sampling.
2. Sampling distributions can be derived mathematically, approximated empirically, or found using statistical theorems.
3. The **Central Limit Theorem** states that sums and averages of measurements from a nonnormal population with finite mean μ and standard deviation σ have approximately normal distributions for large samples of size *n*.

Key Concepts

III. Sampling Distribution of the Sample Mean

1. When samples of size n are drawn from a normal population with mean μ and variance σ^2, the sample mean $\overline{x}$ has a normal distribution with mean μ and variance σ^2/n.
2. When samples of size n are drawn from a nonnormal population with mean μ and variance σ^2, the Central Limit Theorem ensures that the sample mean $\overline{x}$ will have an approximately normal distribution with mean μ and variance σ^2/n when n is large ($n \geq 30$).
3. Probabilities involving the sample mean μ can be calculated by standardizing the value of $\overline{x}$ using $\boxed{z = \dfrac{\overline{x}-\mu}{\sigma/\sqrt{n}}}$

Key Concepts

IV. Sampling Distribution of the Sample Proportion

1. When samples of size n are drawn from a binomial population with parameter p, the sample proportion $\hat{p}$ will have an approximately normal distribution with mean p and variance pq/n as long as $np > 5$ and $nq > 5$.
2. Probabilities involving the sample proportion can be calculated by standardizing the value $\hat{p}$ using $\boxed{z = \dfrac{\hat{p}-p}{\sqrt{\dfrac{pq}{n}}}}$

Key Concepts

V. Statistical Process Control

1. To monitor a quantitative process, use an $\overline{x}$ chart. Select k samples of size n and calculate the overall mean $\overline{\overline{x}}$ and the standard deviation s of all nk measurements. Create upper and lower control limits as

$$\boxed{\mathrm{LCL}: \overline{\overline{x}} - 3\frac{s}{\sqrt{n}} \quad \mathrm{UCL}: \overline{\overline{x}} + 3\frac{s}{\sqrt{n}}}$$

If a sample mean exceeds these limits, the process is out of control.

2. To monitor a binomial process, use a p chart. Select k samples of size n and calculate the average of the sample proportions as $\boxed{\overline{p} = \dfrac{\Sigma \hat{p}_i}{k}}$ Create upper and lower control limits as

$$\boxed{\mathrm{LCL}: \overline{p} - 3\sqrt{\frac{\overline{p}(1-\overline{p})}{n}} \quad \mathrm{UCL}: \overline{p} + 3\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}}$$

If a sample proportion exceeds these limits, the process is out of control.