Lecture 6
Sampling Distributions

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.

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

Simple Random Sampling

- 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**.
- **Simple random sampling** is a method of sampling that allows each possible sample of size *n* an equal probability of being selected.

Example

- There are 89 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 89.
2. Choose 5 pairs of random digits from the random number table.
3. If a number between 90 and 00 is chosen, choose another number.
4. The five students with those numbers form the group.

06907	11008	42751	27756	53498
[illegible]	420	69994	98872	31016
[illegible]	463	07972	18876	20922
[illegible]	661	10281	17453	18103
[illegible]	342	53988	53060	59533
[illegible]	231	33276	70997	79936
[illegible]	235	03427	49626	69445
[illegible]	636	92737	88974	33488
[illegible]	529	85689	48237	52267
[illegible]	048	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.

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 50^{th} 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

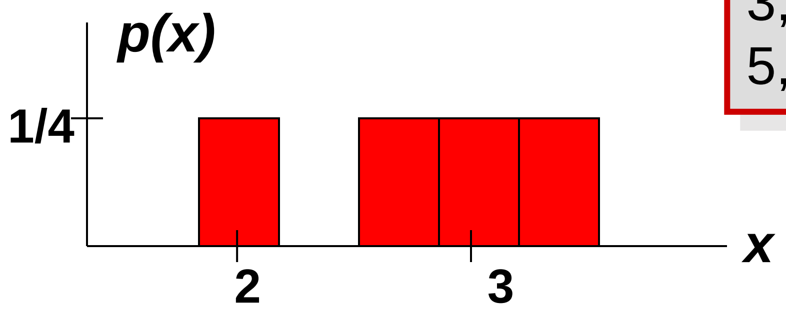
Definition: The sampling distribution of a statistic is the probability distribution for the possible values of the statistic that results when random samples of size *n* are repeatedly drawn from the population.

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

Sampling distributions for statistics can be

✓Approximated with simulation techniques

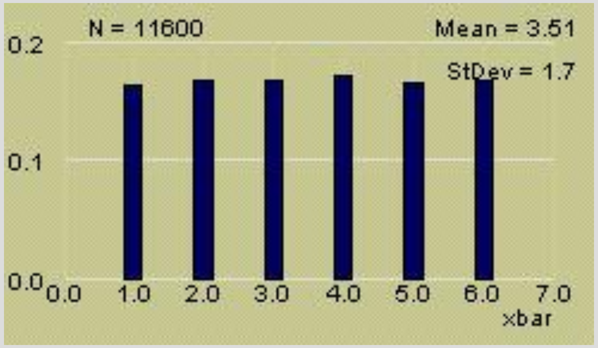
✓Derived using mathematical theorems

✓The Central Limit Theorem is one such 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.

Example

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



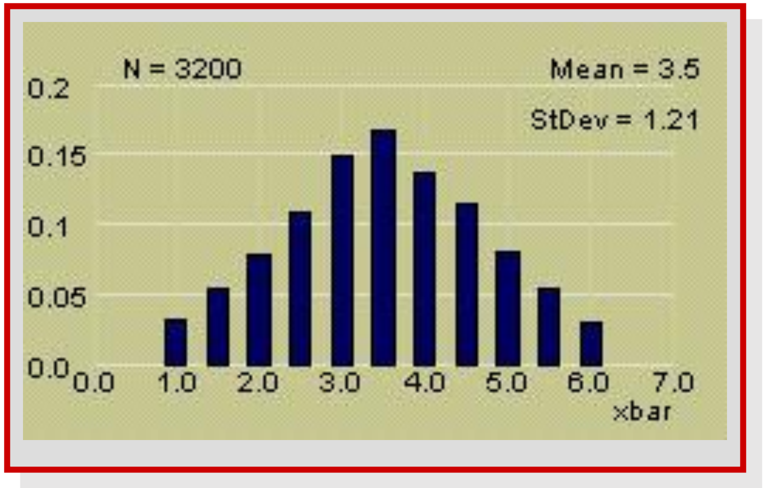
$$\mu = \sum xp(x)$$

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

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

Example

Roll a fair die n = 2 times. The distribution of x the average number on the two upper faces is **mound-shaped.**



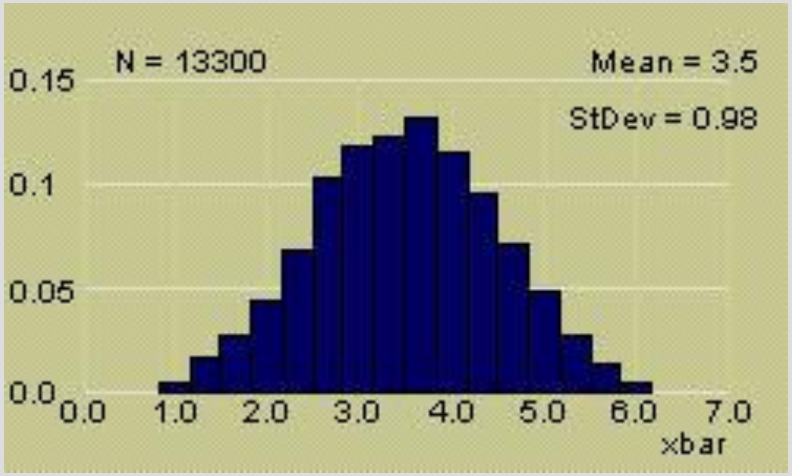
$$\text{Mean}: \mu = 3.5$$

$$\text{Std Dev}:$$

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

Example

Roll a die coin n = 3 times. The distribution of x the average number on the two upper faces is approximately normal.

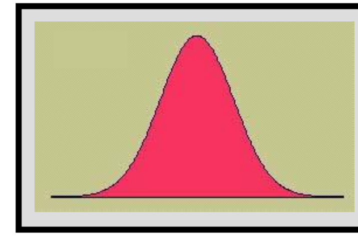


$$\text{Mean}: \mu = 3.5$$

$$\text{Std Dev}:$$

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

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 $\sigma\sqrt{n}$.

✓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?

If the sample is normal, then the sampling distribution of $\overline{x}$ will also be normal, no matter what the sample size.

When the sample population is approximately symmetric, the distribution becomes approximately normal for relatively small values of *n*.

When the sample population is skewed, the sample size must be at least 30 before the sampling distribution of $\overline{x}$ becomes approximately normal.

The Sampling Distribution of the Sample Mean

✓A random sample of size n is selected from a population with mean μ and standard deviation σ.

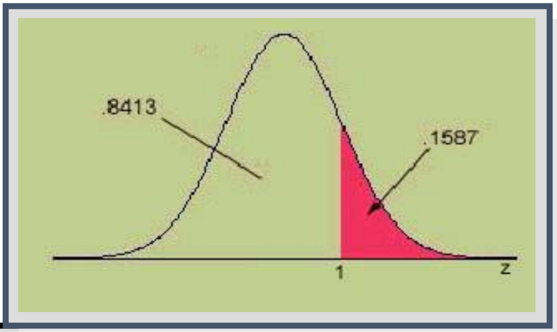
✓The sampling distribution of the sample mean $\bar{x}$ will have mean μ and standard deviation $\sigma / \sqrt{n}$.

✓If the original population is normal, the sampling distribution will be normal for any sample size.

✓If the original population is nonnormal, the sampling distribution will be normal when n is large.

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

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

$$P(\bar{x} > 12) = P\left(z > \frac{12 - 10}{8/\sqrt{16}}\right)$$

$$= P(z > 1) = 1 - .8413 = .1587$$

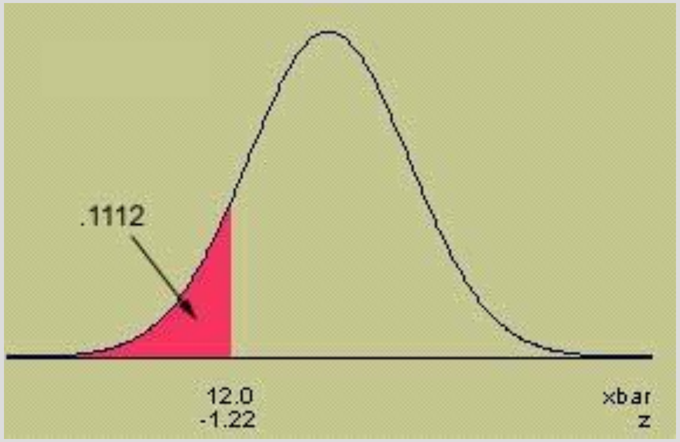
Example

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?

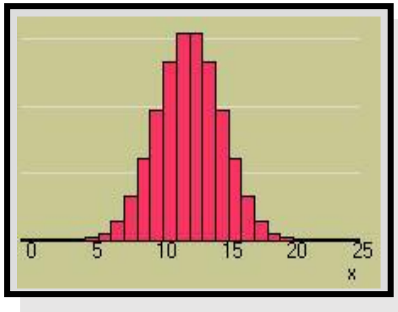
$$P(\bar{x} < 12) =$$

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

$$P(z < -1.22) = .1112$$

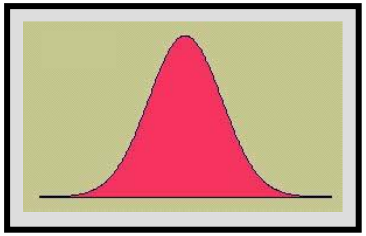


The Sampling Distribution of the Sample Proportion



- ✓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 .
- ✓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 $\hat{p}$ 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***.

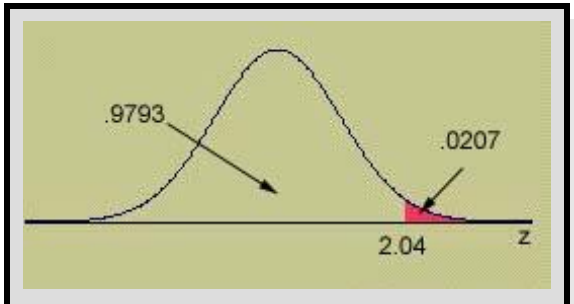
✓The sampling distribution of the sample proportion, $\hat{p}$ will have mean ***p*** and standard deviation

✓If *n* is large (greater than or equal to 30), and *p* is not too close to zero or one *(np and nq greater than or equal to 10),* the sampling distribution of the sample proportion $\hat{p}$ will be **approximately normal.**

$$\sqrt{\frac{pq}{n}}$$

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.

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

$$P(\hat{p} > .5) = P\left(z > \frac{.5 - .4}{\sqrt{\frac{.4(.6)}{100}}}\right)$$

$$= P(z > 2.04) = 1 - .9793 = .0207$$

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

$$P(\hat{p} > .10)$$

$$= P\left(z > \frac{.10 - .05}{\sqrt{\frac{.05(.95)}{200}}}\right) = P(z > 3.24)$$

$$= 1 - .9994 = .0006$$

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

Example

A rowing team consists of four rowers who weigh 152 , 156 , 160 , and 164 pounds.
Find all possible random samples with replacement of size two and compute the sample mean for each one.
Use them to find the probability distribution, the mean, and the standard deviation of the sample mean $\bar{X}$.

Sample	Mean
152, 152	152
152, 156	154
152, 160	156
152, 164	158

Sample	Mean
156, 152	154
156, 156	156
156, 160	158
156, 164	160

Sample	Mean
160, 152	156
160, 156	158
160, 160	160
160, 164	162

Sample	Mean
164, 152	158
164, 156	160
164, 160	162
164, 164	164

Example

A rowing team consists of four rowers who weigh 152 , 156 , 160 , and 164 pounds.
Find all possible random samples with replacement of size two and compute the sample mean for each one.
Use them to find the probability distribution, the mean, and the standard deviation of the sample mean $\bar{X}$.

The table shows that there are seven possible values of the sample mean $\bar{X}$. The value $\bar{x}$ =152 happens only one way (the rower weighing 152 pounds must be selected both times), as does the value $\bar{x}$ =164, but the other values happen more than one way, hence are more likely to be observed than 152 and 164 are. Since the 16 samples are equally likely, we obtain the probability distribution of the sample mean just by counting:

$\bar{x}$	152	154	156	158	160	162	164
$P(\bar{x})$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

For $\mu_{\bar{X}}$, we obtain.

$$\mu_{\bar{X}} = \sum \bar{x} P(\bar{x})$$

$$= 152(116) + 154(216) + 156(316) + 158(416) + 160(316) + 162(216) + 164(116) = 158$$

Example

A rowing team consists of four rowers who weigh 152 , 156 , 160 , and 164 pounds.
Find all possible random samples with replacement of size two and compute the sample mean for each one.
Use them to find the probability distribution, the mean, and the standard deviation of the sample mean $\bar{X}$.

$\bar{x}$	152	154	156	158	160	162	164
$P(\bar{x})$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

For $\sigma_{\bar{X}}$, we first compute $\sum \bar{x}^2 P(\bar{x})$:

$$\sum \bar{x}^2 P(\bar{x}) = 152^2\left(\frac{1}{16}\right) + 154^2\left(\frac{2}{16}\right) + 156^2\left(\frac{3}{16}\right) + 158^2\left(\frac{4}{16}\right) + 160^2\left(\frac{3}{16}\right) + 162^2\left(\frac{2}{16}\right) + 164^2\left(\frac{1}{16}\right)$$

which is $24,974$, so that

$$\begin{aligned} \sigma_{\bar{x}} &= \sqrt{\sum \bar{x}^2 P(\bar{x}) - \mu_{\bar{x}}^2} \\ &= \sqrt{24,974 - 158^2} = \sqrt{10} \end{aligned}$$

The mean and standard deviation of the population {152,156,160,164}
in the example are μ=158 and √σ=20. The mean of the sample mean $\bar{X}$ that we have just computed is exactly the mean of the population.

Example

The mean and standard deviation of the tax value of all vehicles registered in a certain state are μ=$13,525 and σ=$4,180. Suppose random samples of size 100 are drawn from the population of vehicles. What are the mean $\mu_{\bar{x}}$ and standard deviation $\sigma_{\bar{x}}$ of the sample mean $\bar{X}$?

Since $n = 100$, the formulas yield

$$\mu_{\bar{X}} = \mu = \$13,525$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\$4,180}{\sqrt{100}} = \$418$$

Example

Let $\overline{X}$ be the mean of a random sample of size 50 drawn from a population with mean 112 and standard deviation 40.

1. Find the mean and standard deviation of $\overline{X}$.
2. Find the probability that $\overline{X}$ assumes a value between 110 and 114.
3. Find the probability that $\overline{X}$ assumes a value greater than 113.

$$\mu_{\overline{X}} = \mu = 112 \text{ and } \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{40}{\sqrt{50}} = 5.65685$$

Since the sample size is at least 30, the Central Limit Theorem applies: $\overline{X}$ is approximately normally distributed.

just being careful to use $\sigma_{\overline{X}}$ and not σ when we standardize:

$$\begin{aligned} P(110 < \overline{X} < 114) &= P\left(\frac{110 - \mu_{\overline{X}}}{\sigma_{\overline{X}}} < Z < \frac{114 - \mu_{\overline{X}}}{\sigma_{\overline{X}}}\right) \\ &= P\left(\frac{110 - 112}{5.65685} < Z < \frac{114 - 112}{5.65685}\right) \\ &= P(-0.35 < Z < 0.35) \\ &= 0.6368 - 0.3632 \\ &= 0.2736 \end{aligned}$$

Example

Let $\overline{X}$ be the mean of a random sample of size 50 drawn from a population with mean 112 and standard deviation 40.

1. Find the mean and standard deviation of $\overline{X}$.
2. Find the probability that $\overline{X}$ assumes a value between 110 and 114.
3. Find the probability that $\overline{X}$ assumes a value greater than 113.

$$
\begin{aligned}
P(\overline{X} > 113) &= P\left(Z > \frac{113 - \mu_{\overline{X}}}{\sigma_{\overline{X}}}\right) \\
&= P\left(Z > \frac{113 - 112}{5.65685}\right) \\
&= P(Z > 0.18) \\
&= 1 - P(Z < 0.18) \\
&= 1 - 0.5714 \\
&= 0.4286
\end{aligned}
$$

Example

The numerical population of grade point averages at a college has mean 2.61 and standard deviation 0.5. If a random sample of size 100 is taken from the population, what is the probability that the sample mean will be between 2.51 and 2.71?

The sample mean $\overline{X}$ has mean $\mu_{\overline{X}} = \mu = 2.61$ and standard deviation $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{0.5}{10} = 0.05$, so

$$
\begin{aligned}
P(2.51 < \overline{X} < 2.71) &= P\left(\frac{2.51 - \mu_{\overline{X}}}{\sigma_{\overline{X}}} < Z < \frac{2.71 - \mu_{\overline{X}}}{\sigma_{\overline{X}}}\right) \\
&= P\left(\frac{2.51 - 2.61}{0.05} < Z < \frac{2.71 - 2.61}{0.05}\right) \\
&= P(-2 < Z < 2) \\
&= P(Z < 2) - P(Z < -2) \\
&= 0.9772 - 0.0228 \\
&= 0.9544
\end{aligned}
$$

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 $\bar{x}$ 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 $\bar{x}$ 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 $\bar{x}$ using

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

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 $np > 10$ and $nq > 10$.

2. Probabilities involving the sample proportion can be calculated by standardizing the value $\hat{p}$ using

$$\boxed{z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}}$$

Key Concepts

V. Statistical Process Control

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

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

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

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

Create upper and lower control limits as

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

If a sample proportion exceed t of control.