Introduction to Probability and Statistics Twelfth Edition

Robert J. Beaver • Barbara M. Beaver • William Mendenhall

Presentation designed and written by:
Barbara M. Beaver

Edited by: Dr. Worapan Kusakunniran and Dr. Rob Egrot

Introduction to Probability and Statistics Twelfth Edition

Chapter 7

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**.
  - E.g. $\mu = 0$ and $\sigma = 1$ for the **standard** normal distribution
- We must rely on the **sample** to learn about these parameters.

Sampling

- A **Sampling** is a process of taking a sample.

  Examples:
  - 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**.
  - 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**.
- If we want the sample to provide **reliable** information about the population, we must select our sample **carefully**.

Sampling Plans

- The **sampling plan** or **experimental design** determines the amount of information we can extract, and often allows us to measure the **reliability of our inference**.

- It involves either:
  1. Non-randomization
  2. Randomization

Non-randomized Sampling Plans

Non-randomized sampling plans should not be used for statistical inference! They are probably biased in some way.

1.1 Convenience sampling: A sample that can be taken easily without random selection.

- People walking by on the street

1.2 Judgment sampling: The sampler decides who will and won't be included in the sample.

Methods of (Randomized) Sampling

1. Simple random sampling
2. Stratified random sampling
3. Cluster sampling
4. 1-in-k systematic sampling

Simple Random Sampling

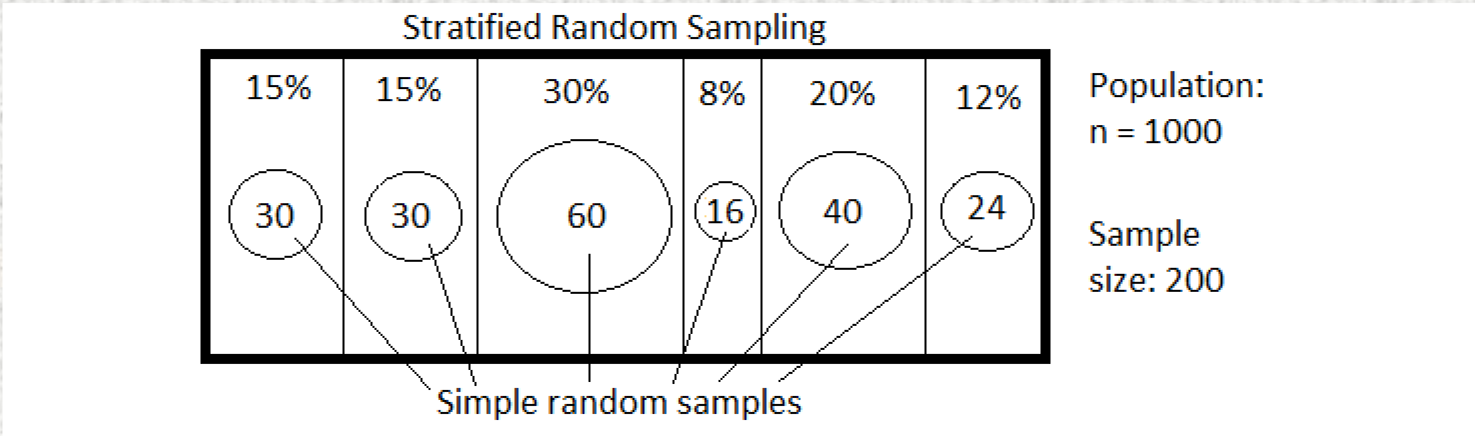
- **Simple random sampling** is a method of sampling that allows each possible sample of size n an **equal probability** of being selected.

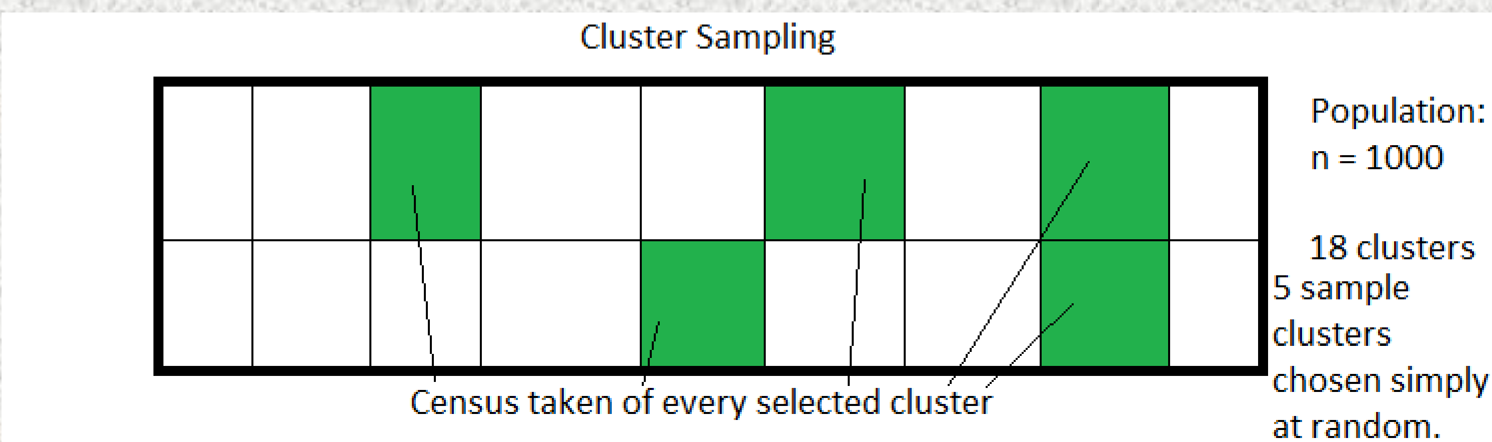
- **Example:** There are 89 students in a statistics class. The instructor wants to choose 5 students to form a project group. How should he proceed?

Methods of (Randomized) Sampling

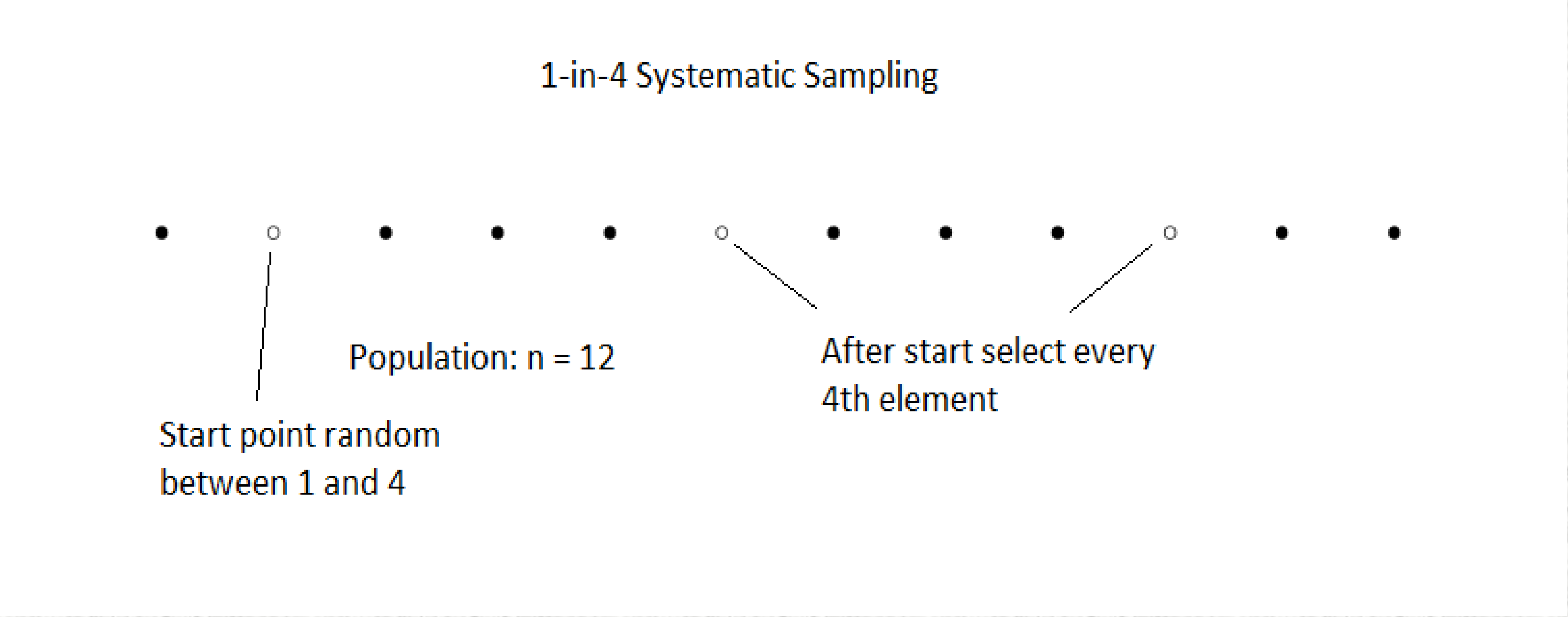
1. **Simple random sampling:**
2. **Stratified random sampling**: Divide the population into subpopulations or **strata** and select a simple random sample from each strata.
3. **Cluster sampling**: Divide the population into subgroups called **clusters**; select a simple random sample of clusters and take a census of every element in the cluster.
4. **1-in-k systematic sampling**: Randomly select one of the first k elements in an ordered population, and then select every k-th element thereafter.

Stratified vs Cluster





Systematic


1-in-4 Systematic Sampling
Population: n = 12
After start select every
4th element
Start point random
between 1 and 4

Examples

- Divide Bangkok into districts and take a simple random sample within each districts. **Stratified**
- Divide Bangkok into districts, take a simple random sample of 10 districts. **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

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.

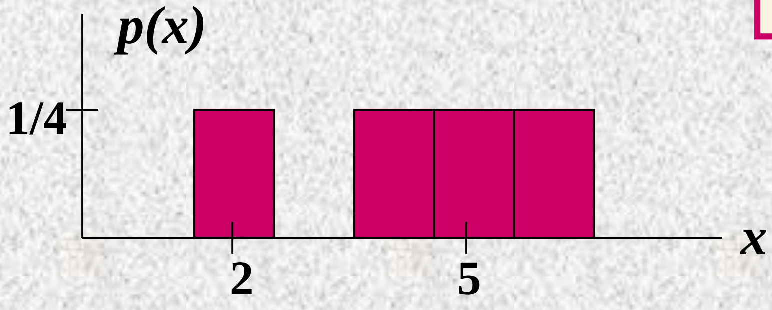
Sampling Distributions

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: 2, 4, 5, 6

Draw samples of size $n = 3$ without replacement

Possible samples	$\bar{x}$	$p(\bar{x})$
2, 4, 5	11/3 = 3.67	1/4
2, 4, 6	12/3 = 4	1/4
2, 5, 6	13/3 = 4.33	1/4
4, 5, 6	15/3 = 5	1/4



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

A Sampling Distribution of $\bar{x}$

$\bar{x}$ is a statistic.

Its distribution is called the **sampling distribution of** $\bar{x}$**.**

$\bar{x}$	$p(\bar{x})$
3.67	**1/4**
4	**1/4**
4.33	**1/4**
5	**1/4**

A Sampling Distribution of $\bar{x}$

$\bar{x}$	$p(\bar{x})$
3.67	**1/4**
4	**1/4**
4.33	**1/4**
5	**1/4**

One can find

$P(\bar{x} < 4.33)$,

the mean of $\bar{x}$,

the standard deviation of $\bar{x}$,

etc.

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

Sampling Distributions

How to find a **sampling distribution**?

Sampling distributions for statistics can be

✓Approximated with simulation techniques.

✓Derived using mathematical theorems

✓The Central Limit Theorem is one such theorem.

The Sampling Distributions of the Sample Mean

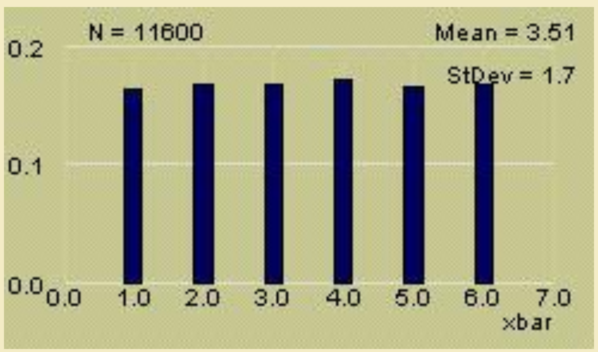
Central Limit Theorem:

If random samples of n observations are drawn from a **non-normal population** with finite mean 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

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



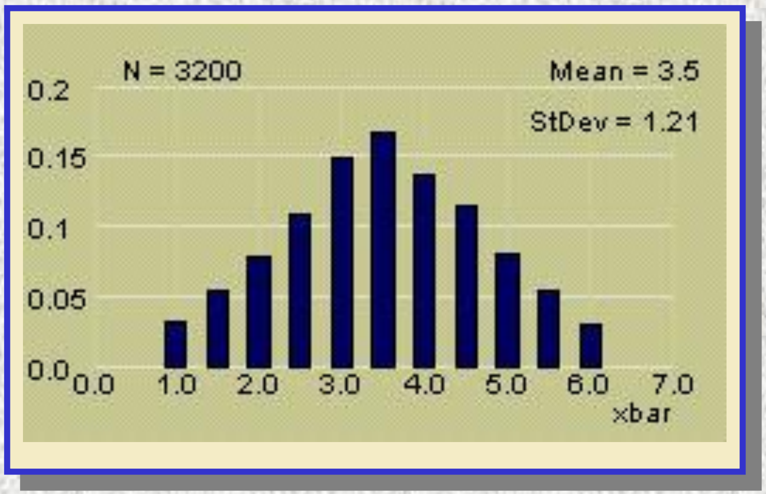
$$\mu = \sum xp(x)$$

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

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

Example

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



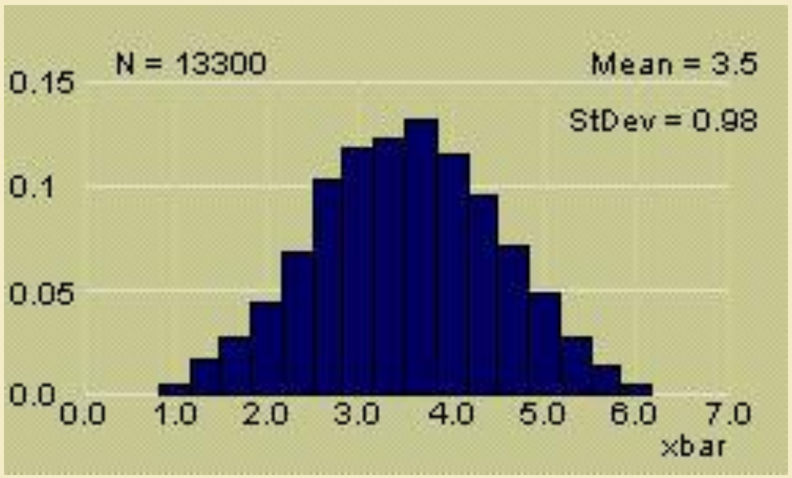
$$\text{Mean}: \mu = 3.5$$

$$\text{Std Dev}:$$

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

Example

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

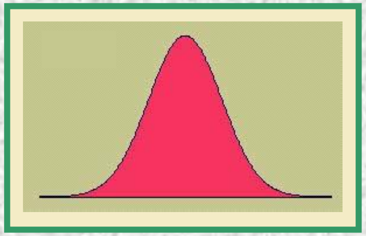


$$\text{Mean}: \mu = 3.5$$

$$\text{Std Dev}:$$

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

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.

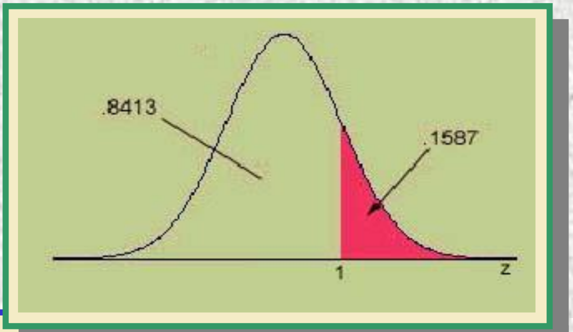
How Large is Large?

If the population is **normal**, then the sampling distribution of $\bar{x}$ 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 $\bar{x}$ becomes approximately normal.

Finding Probabilities for the Sample Mean



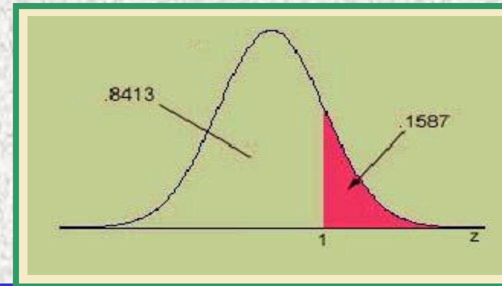
✓If the sampling distribution of $\bar{x}$ is normal or approximately normal, we can find probabilities just like we would with any other normal distribution.

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

✓Find the appropriate area using the standard normal distribution table.

Example: A random sample of size n = 36 from an unknown distribution with μ = 10 and σ = 8. What is the probability that the sample mean will be more than 11?

Finding Probabilities for the Sample Mean



✓If the sampling distribution of $\bar{x}$ is normal or approximately normal, we can find probabilities just like we would with any other normal distribution.

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

✓Find the appropriate area using the standard normal distribution table.

Example: A random sample of size n = 36 from an unknown distribution with μ = 10 and σ = 8. What is the probability that the sample mean will be more than 11?

P(x > 11)=1- P(x<11)=1-P(z<(11-10)/(8/sqrt(36)))=1-P(z<0.75)

=1- 0.7734 = 0.2266

Answer: 0.2266

Example

A soda filling machine is supposed to fill cans of soda with 12 fluid ounces. Suppose that the distribution of the fills is unknown but its mean is 12.05 oz. and standard deviation is 0.2 oz. What is the probability that the average fill for 49 soda cans is more than 12 oz.?

Example

A soda filling machine is supposed to fill cans of soda with 12 fluid ounces. Suppose that the distribution of the fills is unknown but its mean is 12.05 oz. and standard deviation is 0.2 oz. What is the probability that the average fill for 49 soda cans is more than 12 oz.?

$$P(\bar{X} > 12) = P\left(Z > \frac{12 - 12.05}{0.2/7}\right)$$

$$= P(Z > -1.75)$$

$$= 1 - P(Z < -1.75)$$

$$= 1 – 0.0401$$

$$= 0.9599$$

Example

A soda filling machine is supposed to fill cans of soda with 12 fluid ounces. Suppose that the distribution of the fills is unknown but its mean is 12.05 oz. and standard deviation is 0.2 oz. What is the probability that the average fill for 49 soda cans is more than 12 oz.?

$$P(\bar{X} > 12) = P\left(Z > \frac{12 - 12.05}{0.2/7}\right)$$
$$= P(Z > -1.75)$$
$$= P(Z < 1.75)$$
$$= 0.9599$$

Class Activity/ Homework 9

1. Random sample of size n were selected from populations with the means and variances given here. Find the mean and standard deviation of the sampling distribution of the sample mean in each case:
   a) $n = 36$, $\mu = 10$, $\sigma^2 = 9$
   b) $n = 100$, $\mu = 5$, $\sigma^2 = 4$
   c) $n = 16$, $\mu = 120$, $\sigma^2 = 1$

Class Activity/ Homework 9

2. A random sample of size n = 49 is selected from a population with mean $\mu = 53$ and standard deviation $\sigma = 21$.
   a) What will be the approximate shape of the sampling distribution of $\bar{x}$?
   b) What will be the mean and standard deviation of the sampling distribution of $\bar{x}$?

Class Activity/ Homework 9

3. Suppose a random sample of $n = 25$ observations is selected from a population that is normally distributed with mean equal to 106 and standard deviation equal to 12.
   a) Give the mean and the standard deviation of the sampling distribution of the sample mean $\bar{x}$.
   b) Find the probability that $\bar{x}$ exceeds 110.
   c) Find the probability that the sample mean deviates from the population mean $\mu = 106$ by no more than 4.

Class Activity/ Homework 9

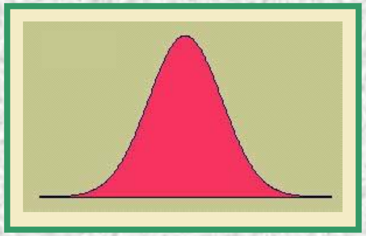
4. **Potassium Levels.** The amount of potassium in food varies, but banana are often associated with high potassium, with approximately 422 mg in a medium-sized banana. Suppose the distribution of potassium in a banana is normally distributed, with mean equal to 422 mg and standard deviation equal to 13 mg per banana. You eat 3 bananas per day, and T is the **total** number of milligrams of potassium you receive from them.
   a) Find the mean and standard deviation of T.
   b) Find the probability that your total daily intake of potassium from three bananas will exceed 1300 mg.

The Normal approximation to the binomial distribution

- We do n independent trials, each with probability p of success. What is the probability we get at least k successes?
- Think about it like this: We take a random sample of n trials. Each success has value 1, each failure has value 0. The number of successes is the sum of this sample.
- If our sample contains x successes, then the mean of the sample is $\frac{x}{n}$. We call this the **sample proportion** $\hat{p} = \frac{x}{n}$.
- So the probability we are interested in is the probability that the mean of our sample is at least $\frac{k}{n}$.

- The mean value of a single trial is p, and the standard deviation is $\sqrt{pq}$ (we can see this by setting $n = 1$).
- So the Central Limit Theorem says that when n is large enough, $\hat{p} = \frac{x}{n}$ is approximately normally distributed with mean p and standard deviation $\frac{\sqrt{pq}}{\sqrt{n}}$.
- So, we want to find $P(\hat{p} \geq \frac{k}{n})$, in a normal distribution with mean p and standard deviation $\frac{\sqrt{pq}}{\sqrt{n}}$.
- This is the same as finding $P(x \geq k)$ in a normal distribution with mean np and standard deviation $\sqrt{npq}$ (we just multiplied everything by n).
- This is the normal approximation to the binomial distribution.
- Note: For accuracy we usually want to do a continuity correction (see week 07 slides).

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,

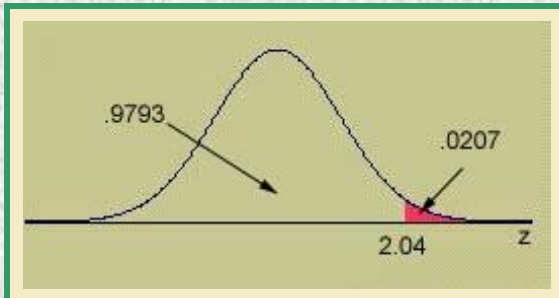
$$\hat{p} = \frac{x}{n}$$

will have mean ***p*** and standard deviation $\sqrt{\frac{pq}{n}}$

✓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, *we can find probabilities for* $\hat{p}$ *using the standard normal distribution table:*

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

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**

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!

Class Activity/ Homework 9

5. Random samples of size n were selected from binomial populations with population parameters p given here. Find the mean and the standard deviation of the sampling distribution of the sample proportion $\hat{p}$ in each case:

a) n = 100, p =0.3
b) n = 400, p =0.1
c) n = 250, p =0.6

Class Activity/ Homework 9

6. Random samples of size n = 75 were selected from a binomial population with p = 0.4. Use the normal distribution to approximate the following probabilities:
   a) $P(\hat{p} \le 0.43)$
   b) $P(0.35 \le \hat{p} \le 0.43)$

Class Activity/ Homework 9

7. **M&M'S.** An advertiser claims that the average percentage of brown M&M'S candies in a package of milk chocolate M&M'S is 13%. Suppose you randomly select a package of milk chocolate M&M'S that contains 55 candies and determine the proportion of brown candies in the package.
   a) What is the approximate distribution of the sample proportion of brown candies in a package that contains 55 candies?
   b) What is the probability that the sample percentage of brown candies is less than 20%?
   c) What is the probability that the sample percentage exceeds 35%?
   d) Within what range would you expect the sample proportion to lie about 95% of the time?

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

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 ($n >= 30$).

3. Probabilities involving the sample mean μ can be calculated by standardizing the value of $\bar{x}$ using $\boxed{z = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}}}$

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 $np > 5$ and $nq > 5$.

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

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

Some Useful Videos

- Sampling methods
  www.youtube.com/watch?v=be9e-Q-jC-0
- The central limit theorem
  https://www.youtube.com/watch?v=Pujol1yC1_A
- The binomial distribution
  https://www.youtube.com/watch?v=qIzC1-9PwQo
- Galton's board
  https://www.youtube.com/watch?v=03tx4v0i7MA