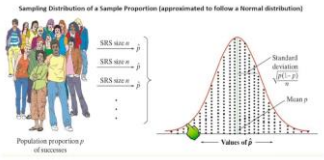
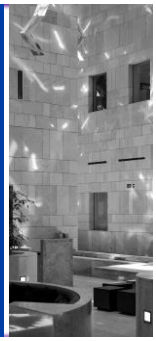
Chapter 9 - Sampling Distributions





Introduction

Probability Distribution
In Chapters 7 and 8, we used knowledge of the population and the parameter(s) of a probability distribution to make probability statements about individual members of the population.

Population & Parameters ⇒ *Probability Distribution* ⇒*Individual*

Sampling Distribution
In this chapter, we are going to develop the sampling distribution, wherein knowledge of the parameter(s) and some information about the distribution allow us to make probability statements about a sample statistic.

Population & Parameters ⇒ *Sampling Distribution* ⇒*Statistic*

Statistical Inference
Starting in Chapter 10, we will assume that most population parameters are unknown. A sample drawn from the population will provide the required statistic, and the sampling distribution of such statistic will enable us to draw inferences about the parameter.

Statistic ⇒ *Sampling Distribution* ⇒*Parameter*

2

Sampling Distribution of the Mean

A **sampling distribution** is created by sampling a population.

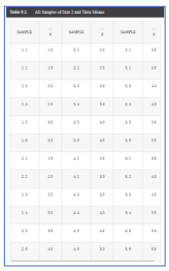
There are two methods to create a sampling distribution:

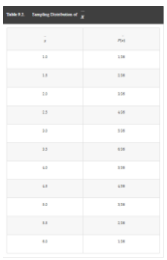
1. *Theoretical method*: apply the rules of probability and the laws of expected value and variance to derive the sampling distribution from sample statistics.
2. *Empirical method*: draw several samples of the same size from a population, calculate the statistic of interest, and then use descriptive techniques to build the sampling distribution.

In this chapter we describe the former method and then briefly demonstrate the latter one.

3

Constructing the sampling distribution of $\bar{x}$





4

Chapter 9 - Sampling Distributions

Distribution of the Mean of Two Dice

Population Distribution of One Die

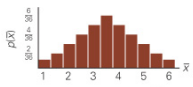
(a) Distribution of X

$\mu = \sum xp(x) = 1\frac{1}{6} + 2\frac{1}{6} + \cdots 6\frac{1}{6} = 3.5$

$\sigma^2 = \sum(x-\mu)^2 p(x) = (1-3.5)^2\frac{1}{6} + (2-3.5)^2\frac{1}{6} + \cdots (6-3.5)^2\frac{1}{6} = 2.92$

$\sigma = \sqrt{\sigma^2} = \sqrt{2.92} = 1.71$

Sampling Distribution of the Mean of Two Dice



(b) Sampling distribution of $\overline{X}$

$\mu_{\bar{x}} = \sum \bar{x}p(\bar{x}) = 1.0\frac{1}{36} + 1.5\frac{2}{36} + \cdots 3.5\frac{6}{36} + \cdots 6.0\frac{1}{36} = 3.5$

$\sigma_{\bar{x}}^2 = \sum(\bar{x}-\mu_{\bar{x}})^2 p(\bar{x}) = (1.0-3.5)^2\frac{1}{36} + (1.5-3.5)^2\frac{1}{36} + \cdots (3.5-3.5)^2\frac{6}{36} + \cdots (6.0-3.5)^2\frac{1}{36} = 1.46$

$\sigma_{\bar{x}} = \sqrt{\sigma_{\bar{x}}^2} = \sqrt{1.46} = 1.21$

5

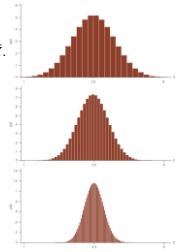
Central Limit Theorem (CLT)

From the sampling distribution of the mean of two dice, we learned:

- The distribution of X is different from the distribution of $\bar{X}$.
- The mean of X is the same as the mean of $\bar{X}$:
  $\mu_{\bar{x}} = \mu$
- The variance of $\bar{X}$ is half the variance of X:
  $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{2}$

It can be further demonstrated that, for a sample of size n:

- as the sample size increases, the distribution of $\bar{X}$ narrows.
- the relation between the variances becomes:
  $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$



6

Central Limit Theorem

When the population from which we are selecting a random sample does not have a normal distribution, the central limit theorem is helpful in identifying the shape of the sampling distribution of $\bar{x}$.

CENTRAL LIMIT THEOREM

In selecting random samples of size n from a population, the sampling distribution of the sample mean $\bar{x}$ can be approximated by a normal distribution as the sample size becomes large.

7

Sampling Distribution of the Sample Mean

The sampling distribution of the sample mean for large populations can be summarized as:

1. $\mu_{\bar{x}} = \mu$
2. If the population is infinite:
   $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
3. If the population is finite, then the **finite population correction factor** is applied:
   $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$ when $N/n < 20$
4. If X is normal, then $\bar{X}$ is normal. If X is nonnormal, then $\bar{X}$ is approximately normal for sufficiently large sample sizes. The definition of "sufficiently large" depends on the extent of nonnormality of X.
   → In many practical situations, a sample size of 30 may be sufficiently large to allow us to use the normal distribution as an approximation for the sampling distribution of $\bar{X}$.

8

Creating the Sampling Distribution Empirically

We can create the sampling distribution of the mean of two dice empirically by:

1. Tossing two fair dice repeatedly.
2. Calculating the sample mean for each sample.
3. Counting the number of times each value of *X* occurs.
4. Computing the relative frequencies to estimate theoretical probabilities.

Obviously, such empirical approach is not practical because of the excessive amount of time required to toss the dice enough times to make the relative frequencies good approximations for the theoretical probabilities.

9

Example – Content of a 32-Ounce Bottle

The amount of soda in each 32-ounce bottle is a normally distributed random variable, with mean of 32.2 oz and standard deviation of 0.3 oz.

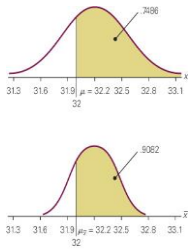
What is the probability that:

1. A single bottle contains more than 32 oz?
2. The mean amount of four bottles is greater than 32 oz?

We define:

X = amount of soda per bottle (normal)

μ = 32.2 oz, σ = 0.3 oz



10

Example – Solution

Solution

1. We need to find $P(X > 32)$:

$$P(X > 32) = P\left(\frac{X-\mu}{\sigma} > \frac{32-32.2}{0.3}\right)$$
$$= P(Z > -0.67) = 1 - P(-0.67)$$
$$= 1 - .2514 = .7486$$

2. We need to find $P(\bar{X} > 32)$ with $n = 4$:

$$\mu_{\bar{x}} = \mu = 32.2 \text{ and } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.3}{\sqrt{4}} = 0.15$$
$$P(\bar{X} > 32) = P\left(\frac{\bar{X}-\mu_{\bar{x}}}{\sigma_{\bar{x}}} > \frac{32-32.2}{0.15}\right)$$
$$= P(Z > -1.33) = 1 - P(-1.33)$$
$$= 1 - .0918 = .9082$$

11

Example: Salaries of Business School Graduates

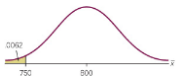
In the advertisements for a large university, the dean of the School of Business claims that the average salary of the school's graduates one year after graduation is $800 per week with a standard deviation of $100.

A student surveys 25 graduates and calculates a mean salary of $750 per week.

Solution: To verify the dean's claim, we need to calculate $P(\bar{X} < 750)$. The distribution of X, the weekly income, is likely to be positively skewed, but not sufficiently so to make the distribution of $\bar{X}$ nonnormal. Thus, we may assume that X is normal with:

$$\mu_{\bar{x}} = \mu = 800, \text{ and } \sigma_{\bar{x}} = \sigma/\sqrt{n} = 100/\sqrt{25} = 20$$
$$P(\bar{X} < 750) = P\left(\frac{\bar{X}-\mu_{\bar{x}}}{\sigma_{\bar{x}}} < \frac{750-800}{20}\right) = P(Z < -2.5)$$
$$= .0062$$

The probability is very small. The dean's claim is not justified.



12

Using the Sampling Distribution for Inference

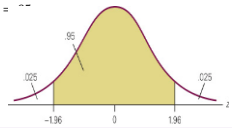
In preparation for the next chapter on interval estimation, we present **another way** of expressing the probability associated with the sampling distribution.

We define $z_{.025}$ to be the value of Z such that the area to the right of $z_{.025}$ under the standard normal curve is equal to .025. We calculated $z_{.025} = 1.96$.

Because the standard normal distribution is symmetric about 0, the area to the left of −1.96 is also .025.

Thus, the area between −1.96 and 1.96 is .95, that is:

$P(-1.96 < Z < 1.96) =$



13

Using the Sampling Distribution for Inference

Because the sampling distribution is normally distributed, we can substitute the equation for Z into the previous probability statement:

$$P\left(-1.96 < \frac{\bar{X} - \mu_{\bar{X}}}{\sigma/\sqrt{n}} < 1.96\right) = .95$$

Using algebra, we produce:

$$P(\mu_{\bar{X}} - 1.96\,\sigma/\sqrt{n} < \bar{X} < \mu_{\bar{X}} + 1.96\,\sigma/\sqrt{n}) = .95$$

If we now plug in mean and standard deviation:

$$P\left(800 - 1.96\frac{100}{\sqrt{25}} < \bar{X} < 800 + 1.96\frac{100}{\sqrt{25}}\right) = .95$$

$$P(760.8 < \bar{X} < 839.2) = .95$$

We can conclude that there is a 95% probability that a sample mean of size 25 will fall between \$760.8 and \$839.2, which does not include the sample mean of \$750. Thus, the dean's claim is not supported.

14

Sampling Distribution of a Proportion

Before introduced the binomial distribution, described by the parameter p, the probability of success in any trial.

When p is unknown, we estimate it with the sample proportion (read as *p hat)*:

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

Where X is the count of successes and n the sample size. The random variable X is binomially distributed.

However, a binomial random variable is discrete, and it does not lend well to the calculations needed for inference. Next, we show how to approximate a discrete binomial random variable with a continuous normal random variable.

15

Binomial Distribution Example

Consider a binomial distribution with $n = 20$, and $p = .5$.

The probability that X = 10, can be written as:

$$P(X = 10) = \frac{n!}{n!\,(n-X)!}p^X(1-p)^{(n-X)}$$

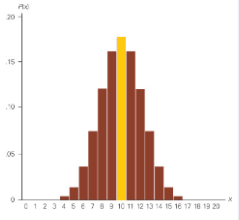
$$= \frac{20!}{10!\,(20-10)!}p^{10}(1-p)^{(20-10)} = .1762$$

From Section 7-4, we also learned that the binomial distribution parameters are:

$$\mu = np = 20(.5) = 10$$

$$\sigma = \sqrt{np(1-p} = \sqrt{20(.5)(.5)} = \sqrt{5}$$

$$= 2.236$$



16

Chapter 9 - Sampling Distributions

Normal Approximation to the Binomial Distribution

- To approximate the probability that $X = 10$ using the normal distribution requires that we find the area under the normal curve between 9.5 and 10.5 (0.5 called correction factor).

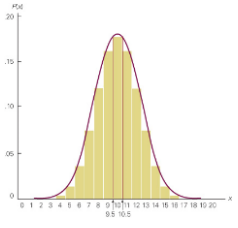
$$P(X = 10) \approx P(9.5 < Y < 10.5)$$

where Y is a normal random variable that approximates the binomial random variable X.

- We standardize Y to find:

$$P(9.5 < Y < 10.5)$$
$$= P\left(\frac{9.5 - 10}{2.236} < \frac{Y - \mu}{\sigma} < \frac{10.5 - 10}{2.236}\right)$$
$$= P(-0.2236 < Z < 0.2236)$$
$$= P(Z < 0.2236) - P(Z < -0.2236)$$
$$= .5885 - .4115 = .1770$$

Which is an excellent approximation.



17

Omitting the Correction Factor for Continuity

When calculating the probability of individual values of X as we did when we computed the probability that X equals 10 earlier, the **continuity correction factor** must be used.

When computing the probability of a *range* of values of X, we can omit the correction factor.

However, the omission of the correction factor will considerably decrease the accuracy of the approximation when the value of *X* is near the center of the distribution, but it can be tolerated for large sample sizes or when the calculations involve the tails of the distribution.

Consider finding $P(X \le 8)$ from the previous example.

Binomial distribution: $P(X \le 8) = .2517$

Normal approximation

with the correction factor: $P(Y < 8.5) = P(Z < -0.894) = .1857$

without the correction factor: $P(Y < 8) = P(Z < -0.671) = .2511$

18

Approximate Sampling Distribution of a Sample Proportion

Using the laws of expected value and variance, we can determine the mean, variance, and standard deviation of $\hat{P}$:

Sampling Distribution of a Sample Proportion

$\hat{P}$ is approximately normally distributed provided that $np \ge 5$ and $n(1-p) \ge 5$

The expected value: $\mathrm{E}(\hat{P}) = p$

The variance: $\mathrm{V}(\hat{P}) = \sigma_{\hat{P}}^2 = \frac{p(1-p)}{n}$

The standard deviation: $\sigma_{\hat{P}} = \sqrt{\frac{p(1-p)}{n}}$

The standard deviation of $\hat{P}$ is called the **standard error of the proportion**.

19

Example – Political Survey

In the last election, a state representative received 52% of the votes cast. One year after the election, the representative organized a survey that asked a random sample of 300 people how they would vote in the next election. If we assume that the representative's popularity has not changed, what is the probability that more than half of the sample would vote to re-elect?

Solution:

The number of respondents who would vote for the representative is a random binomial variable with $n = 300$, and $p = .52$.

We want to find: $P(\hat{P} > .50)$.

$\hat{P}$ is approximately normal with $\mathrm{E}(\hat{P}) = p = .52$ and $\sigma_{\hat{P}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(.52)(.48)}{300}} = .0288$

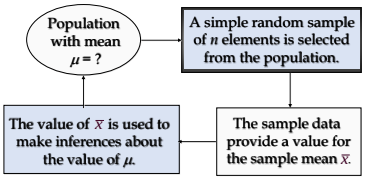
$$P(\hat{P} > .50) = P\left(\frac{\hat{P} - p}{\sigma_{\hat{P}}} > \frac{.50 - .52}{.0288}\right) = P(Z > -0.69) = 1 - P(Z -< 0.69) = 1 - .2451$$
$$= .7549$$

If the level of support has not changed, the probability of a favorable sample exceeds 75%.

20

Sampling Distribution of $\bar{x}$

Process of Statistical Inference



21

Sampling Distribution of $\bar{x}$

When the population has a normal distribution, the sampling distribution of $\bar{x}$ is normally distributed for any sample size.

In most applications, the sampling distribution of $\bar{x}$ can be approximated by a normal distribution whenever the sample is size 30 or more.

In cases where the population is highly skewed or outliers are present, samples of size 50 may be needed.

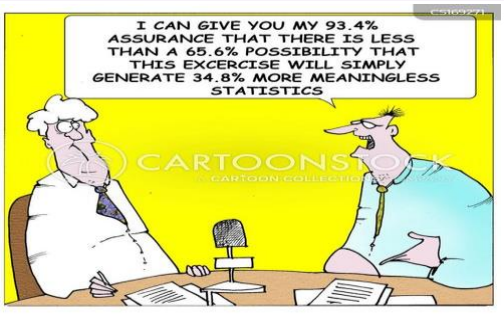
The sampling distribution of $\bar{x}$ can be used to provide probability information about how close the sample mean $\bar{x}$ is to the population mean μ.

22

Sample Mean Vs Population Mean

Traits	Sample Mean www.facebook.com/thesishelper01	Population Mean
Sampling Error	Errors occur when the sample doesn't represent the population or when samples are collected unevenly.	No errors—the entire population is considered for mean computation.
Bias	The sample mean may be biased due to sampling method used or characteristics of the sample.	The population mean is unbiased as it represents the true average of the entire population.
Accuracy	Since the sample mean considers random sample values taken from the whole population, it may not provide accurate results.	The average calculated for the entire population is accurate and reliable.
Level of Difficulty	For sample mean, data collection and computation are easy.	The collection of population data and computation is a very tedious process.
Data Source	The sample data is acquired from the population.	The population data is collected from records, censuses, etc.
Calculation	The sample mean is calculated by summing the values in the sample and dividing by sample size.	The population mean is calculated by summing all the values in the population and dividing by the population size.

23



24