

CHAPTER 5 – Sampling Distributions

Sections: 5.1 & 5.2

Introduction

In this chapter we focus our attention on the sampling distribution of the sample mean and the sampling distribution of the sample proportion. These sampling distributions will give us the tools that we need to derive statistical methods that will allow us to estimate the value for the following population parameters:

1. Population mean (μ)
2. Population proportion (p)

The foundation of our work in all of the chapters that follow is based primarily on the following fact: Information that we seek about a particular population is also embedded in any random sample that we draw from this population.

The Binomial Distribution

Assumptions:

A random experiment with only two possible outcomes, success or failure, is repeated for a given number of times, n .

Let p be the probability that the outcome of the experiment is a success (this implies that the probability of a failure is $1 - p$) and that the outcomes of successive trials are independent of each other.

Let X be the number of successes when the experiment is repeated n times (n trials).

X is a discrete random variable with possible values:

$$S = \{0, 1, 2, 3, 4, \dots, n\}.$$

The probability distribution that the random variable X follows is called the **Binomial Distribution**. The probability that X will take a value k (one of the numbers $0, 1, 2, 3, 4, \dots, n$) can be computed by using the following formula.

$$P[X = k] = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

Tables with binomial probabilities are also available for values of n for up to 15 or 20 and for the most common values of p . In the textbook for this class, this table is found in the Tables section at the end of the book (Table C).

Mean and Standard Deviation for a Binomial Distribution

The mean μ (mu) and the standard deviation σ (sigma) of any binomial random variable can be found by using the following formulas:

$$\mu = np$$

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

Example:

A basketball player takes 12 shots at the basket. On each throw he makes a basket with probability 0.7 and he misses the basket with probability 0.3. Let X be the number of times that the basketball player misses the basket.

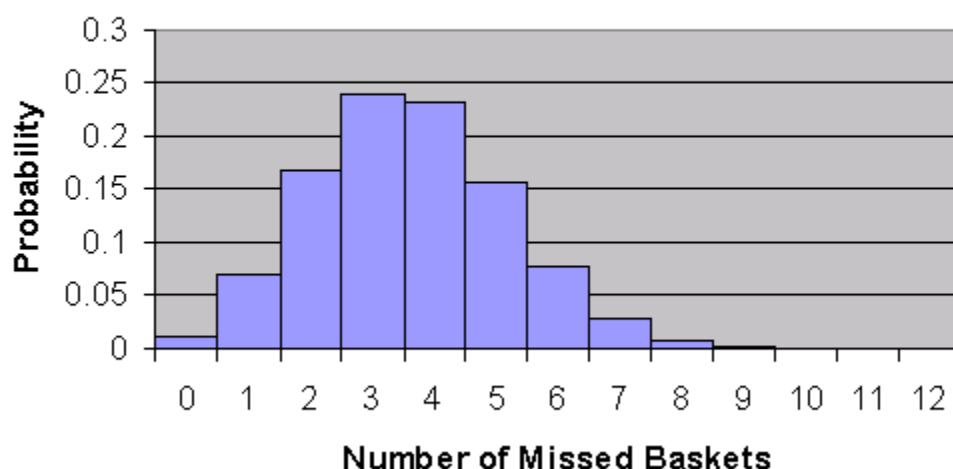
1. Draw the probability distribution of X . Is this a symmetric distribution or is it a skewed distribution?
2. What is the mean and what is the standard deviation of the number of baskets that the player misses?
3. What is the probability that the basketball player will make more than 8 baskets?

Solution:

1. By using Table C, we construct the following table, which describes the probability distribution of X , the number of baskets the players misses in 12 attempts.

k	1	2	3	4	5	6	7	8	9	10	11	12
P[X = k]	0.0712	0.1678	0.2397	0.2311	0.1585	0.0792	0.0291	0.0078	0.0015	0.0002	0.0000	0.0000
P[X ≤ k]	0.0850	0.2528	0.4925	0.7237	0.8822	0.9614	0.9905	0.9983	0.9998	1.0000	1.0000	1.0000

$$P[X=k]$$



$$2. \mu = np = 12 \times .3 = 3.6, \quad \sigma = \sqrt{12(.3)(1-.3)} = 1.59$$

3. $P\{\text{Player makes more than 8 baskets}\} = P\{\text{player misses less than 4}\}$
 $P[X < 4] = P[X \leq 3] = 0.4925$

Normal Approximation to Binomial Probabilities

If X is a binomial distribution with parameters n and p , then it can be shown that

$$P[X=k] = P\left\{\frac{(k - 1/2) - np}{\sqrt{np(1-p)}} < Z < \frac{(k + 1/2) - np}{\sqrt{np(1-p)}}\right\}, \text{ approximately.}$$

This formula gives satisfactory approximations for large values of n , and values of p not near 0 or 1 (say, $0.05 < p < 0.95$).

In general the approximations are good if both np and $n(1-p)$ are at least 10. The addition and subtraction of $1/2$ is called the **continuity correction**.

EXERCISES

1. A new vaccine was tested on 100 persons to determine its effectiveness. The drug company claims that the vaccine is 80% effective, find the probabilities that:
 - a. less than 74 people will develop immunity
 - b. between 74 and 85 people, inclusive, will develop immunity.

Answers:

- a. 0.0668
- b. 0.8276

Note: These answers were derived without using the continuity correction.

2. When a certain seed is planted, the probability that it will sprout is 0.1. If 1000 seeds are planted, find the approximate probability that:
 - a. more than 130 seeds will sprout
 - b. between ninety and ninety-five seeds, inclusive, will sprout.

Answers:

- a. 0.0008
- b. 0.1512

Sampling Distributions

Parameter

A parameter is number that can be used to describes a population as a whole.

Statistic

A statistic is a number derived from a sample drawn from a specific population.

In statistical practice, the value of a population parameter is not known. A statistic is used to estimate a parameter.

Example:

A telemarketing firm in L.A. uses a device that dials residential telephone numbers in that city at random. Of the first 100 numbers dialed, 48% are unlisted. This is not surprising because 52% of all L.A. residential phones are unlisted. Here 48% is a statistic. 52% is a parameter.

Sampling Distribution (of a statistic)

The distribution of the values taken by a statistic in all possible samples of the same size from the same population is called the sampling distribution of the statistic.

Unbiased statistics

A statistic that is used to estimate a parameter is said to be unbiased if the mean of its sampling distribution equals the value of the population parameter being estimated.

Two of the population parameters that we are interested in are:

- The population proportion (p)
- The population mean (μ)

The sample proportion $\hat{p} = \frac{X}{n}$ and sample mean $\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$ are unbiased statistics for the population proportion p and population mean μ , respectively.

Sampling Distribution of the Sample Proportion

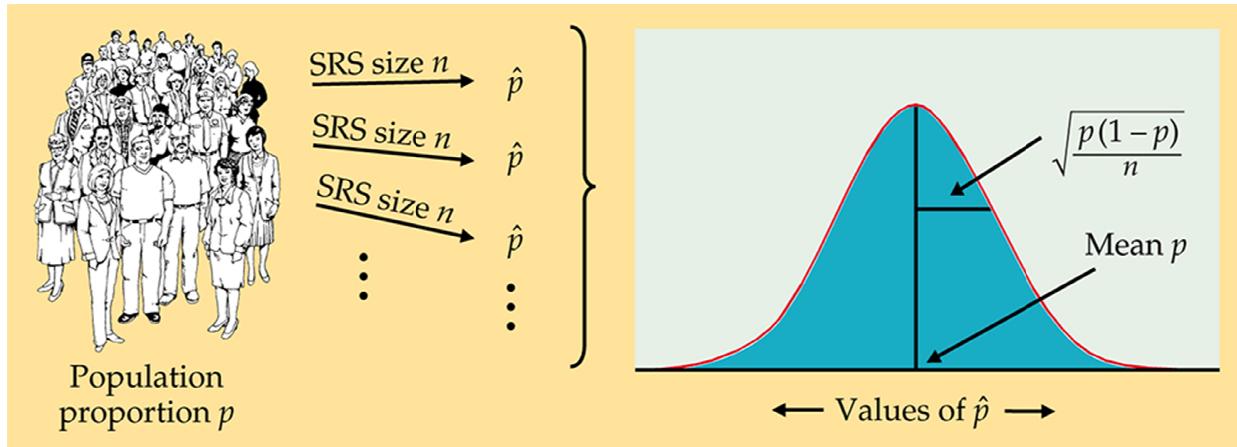


Figure 5.4
Introduction to the Practice of Statistics, Sixth Edition
© 2009 W.H. Freeman and Company

Example

The probability that a person is left-handed is approximately 0.10. Take a SRS of size $n=200$ and let \hat{p} (p-hat) be the proportion of left-handed people in the sample (the sample proportion). The value of p-hat varies in different SRSs of size $n = 200$.

Because $\hat{p} = \frac{X}{n}$ (for this example $p\text{-hat} = X/200$), where X is the random variable whose value equals the number of left-handed people in the sample (we call X the count), the mean and standard deviation of p-hat can be found using the rules for means and variances.

The count X follows the binomial distribution with parameters n & p with mean $\mu_X = np$ and standard deviation $\sigma_X = \sqrt{np(1-p)}$.

Using the rules for means and variances from section 4.4 we get:

$$\mu_{\hat{p}} = \frac{\mu_X}{n} = \frac{np}{n} = p$$

$$\sigma_{\hat{p}}^2 = \left(\frac{1}{n}\right)^2 \sigma_X^2 = \frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n}$$

Or

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

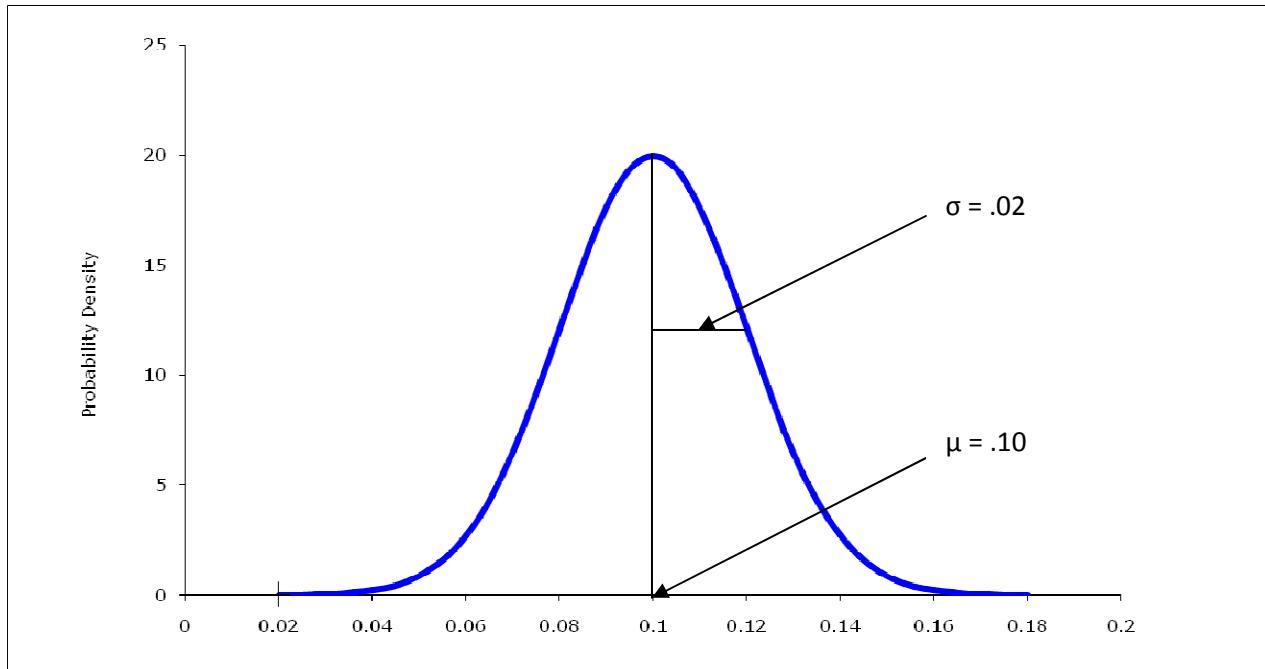
In our case, the mean and standard deviation for the sample proportion of left-handed people in a sample of 200 are:

$$\mu_{\hat{p}} = .10 \text{ (or } 10\%)$$

$$\sigma_{\hat{p}} = \sqrt{\frac{(.10)(.90)}{200}} = 0.021213 \text{ (or approx } 2\%)$$

Because the distribution of the count X is approximately $N(np, \sqrt{np(1-p)})$, the distribution of \hat{p} is also approximately $N(p, \sqrt{\frac{p(1-p)}{n}})$

The approximate distribution of the sample proportion of left-handed people in a SRS of 200 is shown below.



Sampling Distribution of the Sample Mean

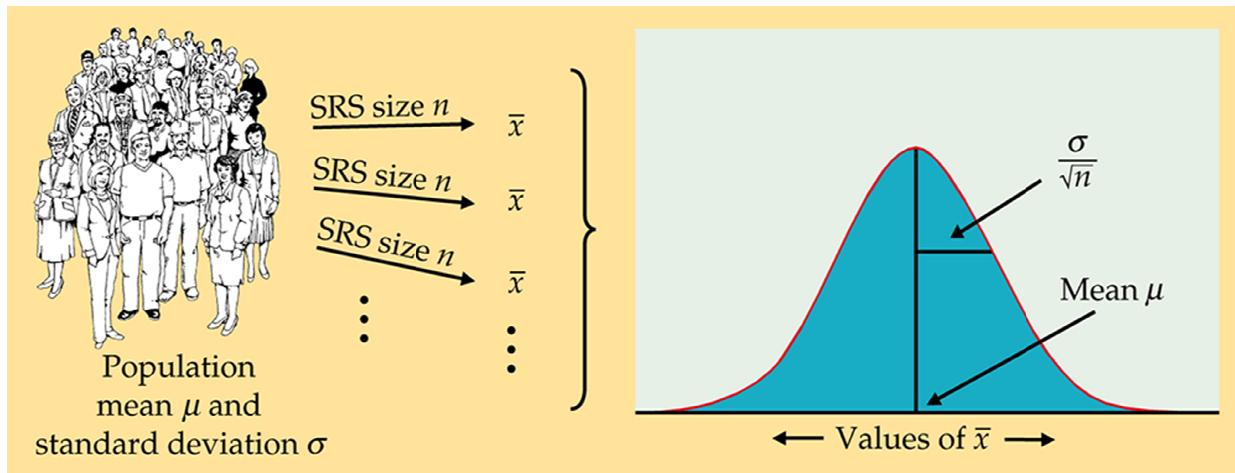
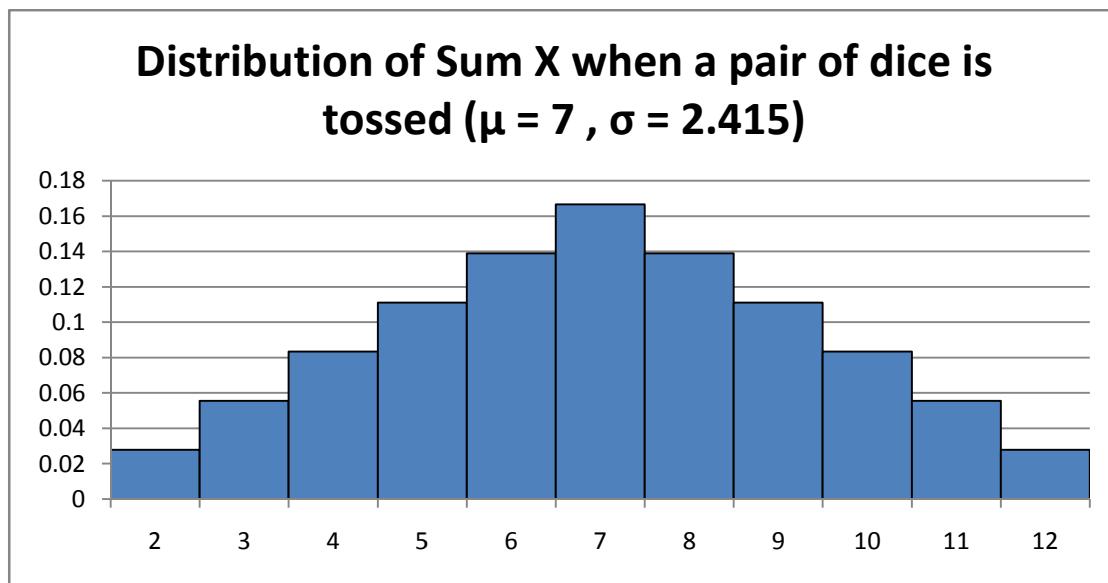


Figure 5.12
Introduction to the Practice of Statistics, Sixth Edition
©2009 W.H. Freeman and Company

Example

The distribution of the sum X that shows up when a pair of fair dice is tossed is shown below.



Take a SRS of size $n = 9$ from this population and compute its sample mean \bar{x} . You can do this by tossing a pair of dice 9 times and record the sum that shows up each time. In Excel you can use Random Number Generation from Data Analysis to simulate the toss of a pair of dice.

The table below shows 10 samples of size $n = 9$ generated using Excel. The last row is the average for the 9 tosses of the dice for each sample.

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	Sample 9	Sample 10
X ₁	5	9	6	2	11	6	7	4	9	5
X ₂	8	2	3	9	4	8	5	9	5	5
X ₃	7	10	8	4	10	10	6	7	9	7
X ₄	5	6	10	4	9	7	8	9	9	7
X ₅	11	9	11	3	7	8	8	9	6	5
X ₆	4	5	10	3	8	3	9	7	7	8
X ₇	10	3	7	12	5	7	6	5	2	6
X ₈	5	10	10	7	9	8	10	6	11	8
X ₉	9	9	8	7	4	6	7	4	9	11
X-bar	7.1	7.0	8.1	5.7	7.4	7.0	7.3	6.7	7.4	6.9

Notice that the value of the sample mean varies from sample to sample.

In general, the sample mean \bar{X} in a SRS of size n from the same population is a random variable.

THE MEAN $\mu_{\bar{X}}$ AND STANDARD DEVIATION $\sigma_{\bar{X}}$ OF THE RANDOM VARIABLE \bar{X}

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{X_1 + X_2 + X_3 + \dots + X_n}{n} = \frac{1}{n} (X_1 + X_2 + X_3 + \dots + X_n)$$

Applying the rules for means and variances from section 4.4 we get:

$$\mu_{\bar{X}} = \frac{1}{n} (n * \mu_X) = \mu_X$$

$$\sigma_{\bar{X}}^2 = \left(\frac{1}{n}\right)^2 * (n \sigma_X^2) = \frac{1}{n} \sigma_X^2$$

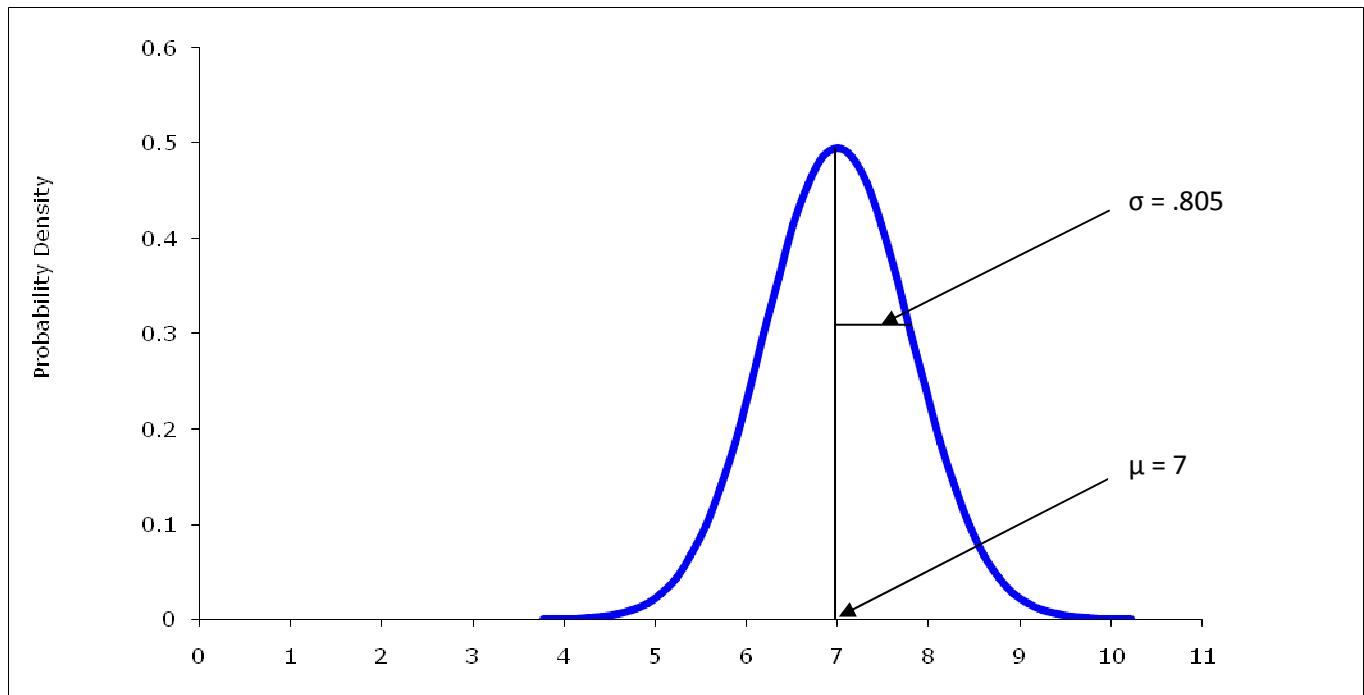
Or

$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$$

SHAPE OF THE DISTRIBUTION OF THE SAMPLE MEAN \bar{X}

Based on a well known result in Statistics called the Central Limit Theorem, the shape of the distribution of \bar{X} is approximately normal.

According to these results, the distribution of the sample mean \bar{X} in 9 tosses of a pair of dice is approximately normal and has mean $\mu_{\bar{X}} = 7$ and standard deviation $\sigma_{\bar{X}} = \frac{2.415}{\sqrt{9}} = \frac{2.415}{3} = 0.805$



EXAMPLES:

1. In the long run, annual real returns on common stocks have varied with mean 9% and standard deviation 28%. You plan to retire in 45 years and you are considering investing in stocks. What is the approximate probability (assuming market conditions do not change dramatically in the next 45 years) that the mean annual return on your investment over the next 45 years will:
 - (a) exceed 15%
 - (b) be lower than 5%

Answers:

- a. 0.0753
- b. 0.1690

2. According to government data, 21% of American children under the age of six live in households with incomes less than the official poverty level. A study of learning in early childhood chooses a random sample of 300 children. Find the approximate probability that at least 80 of the children in the sample selected come from households with incomes less than the official poverty level.

Answer: 0.0080