

## Sampling Distributions :

### Example 1:

Consider a sequence of  $n$  independent, identical experiments of flipping a fair coin. For each experiment, the sample space  $\Omega$  contains two possible outcomes: head and tail. Define  $X_1 = 1$  if observe a head for the first experiment, otherwise  $X_1 = 0$  for the first experiment;  $X_2 = 1$  if observe a head for the second experiment, otherwise  $X_2 = 0$  for the second experiment; ....;  $X_n = 1$  if observe a head for the  $n$ th experiment, otherwise  $X_n = 0$  for the  $n$ th experiment. In this way, we get a sequence of  $n$  random variables. They are independent and each corresponds to the identical experiment.

### Definition 17.1

Let  $X_1, \dots, X_n$  be random variables corresponding to the outcomes of  $n$  independent, identical experiments, respectively. Then,  $X_1, \dots, X_n$  is called a random sample drawing from the population of the experiment. A statistic is a map or function of a random sample without any unknown parameters. Since a statistic is r.v., it has a distribution. Its probability distribution is called a sampling distribution.

### Questions:

Why do we need statistics? Why do we need the distribution of a statistic?

### Example 2:

Consider a quality control problem that a shipment

of 100,000 bulbs. That the mean lifetime of the bulbs is over 3,000 hours is bottom line of the qualification. The testing is destructive. We can't check one by one and we only allow to check limited number of bulbs. In this case, we must draw finite sample observations of their lifetimes  $X_1, \dots, X_n$ . Find their mean lifetime  $\bar{X} = (X_1 + \dots + X_n)/n$ . If we know the distribution of  $\bar{X}$ , then we can make a judgment whether the population mean is over 3,000 hours. This is called hypothesis testing.

Back to Example 1:

For a given integer  $n$ , define  $Y = X_1 + X_2 + \dots + X_n$ .

- (1) Is  $Y$  a statistic?
- (2) What is the distribution for each  $X_i$ ? What is the sampling distribution of  $Y$ ?

For a given integer  $n$ , define  $Z = (X_1 + X_2 + \dots + X_n)/n$

- (1) Is  $Z$  a statistic?
  - (2) What is the sampling distribution of  $Z$ ?
- (See pictures,  $n \rightarrow \infty$ , the distribution of  $Z$  converges to Normal distribution.) More generally, we have

Central Limit Theorem Let  $X_1, \dots, X_n$  be a random sample from  $n$  independent, identical experiments. Let  $\bar{X} = (X_1 + \dots + X_n)/n$  be the sample mean.

- (A) If each  $X_i$  has mean  $\mu$  and standard deviation  $\sigma$ , then  $\bar{X}$  has mean  $\mu$  and standard deviation  $\sigma/\sqrt{n}$ .
- (B) If each  $X_i$  has a normal distribution  $N(\mu, \sigma)$ , then  $\bar{X}$  has exactly normal distribution  $N(\mu, \sigma/\sqrt{n})$ .

(C) If each  $X_i$  has a non-normal distribution, then  $\bar{X}$  has an approximate normal distribution  $N(\mu, \sigma/\sqrt{n})$  when  $n$  is large ( $n \geq 25$ ).

**Example 3:**

Consider  $n$  independent, identical experiments of observing the height of an American person with age between 20 and 60 year's old. Then, the sample space  $\Omega$  is all the heights of American people with age between 20 and 60. Let  $X_1$  denote the observation of an American's height in the first random experiment; Let  $X_2$  denote the observation of an American's height in the second random experiment; ...; Let  $X_n$  denote the observation of an American's height in the  $n^{th}$  random experiment. The random experiments are all with replacement. Define  $\bar{X} = (X_1 + \dots + X_n)/n$ . Each  $X_i$  has mean  $\mu = 1.7m$  and variance  $\sigma^2 = 0.2$ . If  $n = 200$ , what is the mean and variance of  $\bar{X}$ ? Find the approximate probability of  $(\bar{X} \geq 1.6)$ .

**Example :**

Suppose that a random sample  $\{X_1, \dots, X_6\}$  of  $n = 6$  observations is selected from a population that is normally distributed, with mean equal to 1 and standard deviation equal to 0.3.

(1) Find the mean and the standard deviation of  $\bar{X} = (1/6) \sum_{i=1}^6 X_i$ .

**(2) Find the probability  $\mathbb{P}(\bar{X} \geq 1.03)$ .**

**Solution:**

**(1) According to the central limit theorem, we have that the mean of  $\bar{X}$  is :**

$$\mu = 1$$

**the standard deviation of  $\bar{X}$  is :**

$$\sigma/\sqrt{n} = 0.3/\sqrt{6} = 0.1224$$

**(2)**

$$\begin{aligned}\mathbb{P}(\bar{X} \geq 1.03) &= \mathbb{P}\left(\frac{\bar{X}-1}{0.1224} \geq \frac{1.03-1}{0.1224}\right) \\ &= \mathbb{P}\left(\frac{\bar{X}-1}{0.1224} \geq 0.245\right) \\ &= 1 - \mathbb{P}\left(\frac{\bar{X}-1}{0.1224} \leq 0.245\right) = 0.404\end{aligned}$$

**Example :**

Suppose that a random sample  $\{X_1, \dots, X_{30}\}$  of  $n = 30$  observations is selected from a population that is not normally distributed, with mean equal to 1 and standard deviation equal to 0.3.

(1) Find the mean and the standard deviation of  $\bar{X} = (1/30) \sum_{i=1}^{30} X_i$ .

(2) Find the probability  $\mathbb{P}(\bar{X} \geq 1.01)$ .

**Solution:**

(1) According to the central limit theorem, we have that the mean of  $\bar{X}$  is :

$$\mu = 1$$

the standard deviation of  $\bar{X}$  is :

$$\sigma/\sqrt{n} = 0.3/\sqrt{30} = 0.01$$

(2)

$$\begin{aligned} \mathbb{P}(\bar{X} \geq 1.01) &= \mathbb{P}\left(\frac{\bar{X}-1}{0.01} \geq \frac{1.01-1}{0.01}\right) \\ &= \mathbb{P}\left(\frac{\bar{X}-1}{0.1224} \geq 1\right) \\ &= 1 - \mathbb{P}\left(\frac{\bar{X}-1}{0.1224} \leq 1\right) = 0.1587 \end{aligned}$$

**Example :**

The number of wiring packages that can be assembled by a company's employees has a normal distribution, with a mean equal to 16.4 per hour and a standard deviation of 1.3 per hour.

(1) What are the mean and standard deviation  $X/8$

where  $X$  is the number of packages produced per worker in an 8 hour day?

(2) What is the exactly distribution of  $X/8$ ?

(3) Find the probability  $\mathbb{P}(X \geq 135)$ .

**Solution:**

(1) Let  $X_i$  be the number of packages produced per worker in the  $i$ th hour. Then  $X_1, \dots, X_8$  are the outcomes of independent, identical 8 one-hour experiments. According to CLT,  $X/8$  has mean 16.4 and standard deviation  $1.3/\sqrt{8} = 0.4596$ .

(2) According to CLT,  $X/8$  has normal distribution  $N(16.4, 0.4596)$  exactly.

(3)

$$\begin{aligned} \mathbb{P}(X \geq 135) &= \mathbb{P}(X/8 \geq 16.875) \\ &= \mathbb{P}\left(\frac{X/8-16.4}{0.4596} \geq \frac{16.875-16.4}{0.4596}\right) \\ &= \mathbb{P}\left(\frac{X/8-16.4}{0.4596} \geq 1.0335\right) \\ &= 1 - \mathbb{P}\left(\frac{X/8-16.4}{0.4596} \leq 1.0335\right) = 0.1515 \end{aligned}$$

**Example :**

The fracture strengths of a certain type of glass average 14 (in thousands of pounds per square inch) and have a standard deviation of 2.

(1) What is the probability of that the average fracture strength for 100 pieces of this glass exceeds 14.5?

(2) Let  $\bar{X}$  be the average fracture strength for 100 pieces

of this glass. Find a positive  $x_0$  such that

$$\mathbb{P}(-x_0 \leq \bar{X} - 14 \leq x_0) = 0.95$$

**Solution:**

**(1) The mean of  $\bar{X}$  is 14 and its standard deviation is  $2/\sqrt{100} = 0.2$ . Therefore,**

$$\begin{aligned} \mathbb{P}(\bar{X} \geq 14.5) &= \mathbb{P}\left(\frac{\bar{X}-14}{0.2} \geq \frac{14.5-14}{0.2}\right) \\ &= \mathbb{P}\left(\frac{\bar{X}-14}{0.2} \geq 2.5\right) \\ &= 1 - \mathbb{P}\left(\frac{\bar{X}-14}{0.2} \leq 2.5\right) \\ &= 0.0062 \end{aligned}$$

**(2)**

$$\mathbb{P}(-x_0 \leq \bar{X} - 14 \leq x_0) = 0.95$$

$\Leftrightarrow$

$$\mathbb{P}\left(\frac{-x_0}{0.2} \leq \frac{\bar{X} - 14}{0.2} \leq \frac{x_0}{0.2}\right) = 0.95$$

$\Leftrightarrow$

$$\mathbb{P}(-1.96 \leq \frac{\bar{X} - 14}{0.2} \leq 1.96) = 0.95$$

**Therefore,**

$$\begin{aligned} \frac{x_0}{0.2} &= 1.96 \\ x_0 &= 0.392 \end{aligned}$$

**Sample proportion :**

**Consider  $n$  independent, identical experiments with**

two possible outcomes called success and failure. Let

$$X_i = \begin{cases} 1 & \text{if success with probability } p \\ 0 & \text{if failure with probability } q = 1 - p. \end{cases}$$

be the r.v. of  $i$ th experiment. Define

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i$$

Then,  $\hat{p}$  is called the sample proportion of the sample  $\{X_1, \dots, X_n\}$ . According to normal approximation to binomial, we have following sample proportion procedure:

- (1) Check whether  $np > 5$  and  $nq > 5$ ;
- (2) The mean of  $\hat{p}$  is equal  $p$ , the standard deviation of  $\hat{p}$  is equal to  $\sqrt{\frac{pq}{n}}$ ;
- (3) According to normal approximation to binomial,

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

has approximate standard normal distribution  $N(0, 1)$ .

Remark:

If we denote  $Y = \sum_{i=1}^n X_i$ , then  $Y$  has binomial distribution with mean  $np$  and standard deviation  $\sqrt{npq}$ .

$$\begin{aligned} \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} &= \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \cdot \frac{n}{n} \\ &= \frac{Y - np}{\sqrt{npq}} \end{aligned}$$

which is an approximate standard normal r.v. by normal approximation to binomial.

**Example :**

Suppose that  $\{X_1, \dots, X_6\}$  are  $n = 6$  independent, identical observations of flipping a biased coin. For each experiment, a head appears with probability 0.3 and a tail appears with probability 0.7.

(1) Find the mean and the standard deviation of  $Y = \sum_{i=1}^6 X_i$ .

(2) Find the probability  $\mathbb{P}(Y \leq 3)$ .

**Solution:**

(1)  $Y$  has a binomial distribution with mean  $6(0.3) = 1.8$  and standard deviation  $\sqrt{6(0.3)(0.7)} = 1.122$

(2) According to table( $n = 6, p = 0.3$ ) on page 710, we have

$$\mathbb{P}(Y \leq 3) = 0.93$$

**Example :**

Suppose that  $\{X_1, \dots, X_{60}\}$  are  $n = 60$  independent, identical observations of flipping a biased coin. For each experiment, a head appears with probability  $1/3$  and a tail appears with probability  $2/3$ .

(1) Find the mean and the standard deviation of  $Y/(60)$  where  $Y = \sum_{i=1}^{60} X_i$ .

(2) Find the probability  $\mathbb{P}(Y \leq 18)$ .

**Solution:**

(1) According the sample proportion procedure,  $Y/(60)$

has mean  $1/3$  and standard deviation  $\sqrt{[(1/3)(2/3)]/(60)} = 0.06$

(2) Since no binomial table for  $n = 60$  and  $np = (60)(1/3) > 5$ ,  $nq = (60)(2/3) > 5$ , we can use sample proportion procedure to get

$$\begin{aligned}
 \mathbb{P}(Y \leq 18) &= \mathbb{P}\left(\frac{Y}{60} \leq \frac{18}{60}\right) \\
 &= \mathbb{P}\left(\frac{Y/(60)-1/3}{0.06} \leq \frac{0.3-(1/3)}{0.06}\right) \\
 &= \mathbb{P}\left(\frac{Y/(60)-1/3}{0.06} \leq -0.55\right) \\
 &= \mathbb{P}\left(\frac{Y/(60)-1/3}{0.06} \geq 0.55\right) \\
 &= 0.5 - \mathbb{P}\left(0 \leq \frac{Y/(60)-1/3}{0.06} \leq 0.55\right) \\
 &= 0.5 - 0.2088 = 0.2912
 \end{aligned}$$

Example :

According to *Chance* magazine, the average percentage of brown M&M candies in a package of plain M&Ms is 30%. Suppose that you randomly select a package of plain M&Ms that contains 55 candies and determine the proportion of brown candies in the package.

(1) What is the approximate distribution of the sample proportion of brown candies in a package that contains 55 candies?

(2) What is the probability that the sample proportion of brown candies is less than 20%?

Solution:

(1) The random variable  $\hat{p}$ , the sample proportion of

**brown M&Ms in a package of  $n = 55$ , has a binomial distribution with  $n = 55$ ,  $p = 0.3$ . Since  $np = (55)(0.3) = 16.5 > 5$  and  $nq = 38.5 > 5$ ,  $\hat{p}$  has approximate normal distribution with mean  $p = 0.3$  and standard deviation  $\sqrt{(0.3)(0.7)/(55)} = 0.06179$ .**

**(2)**

$$\begin{aligned}\mathbb{P}(\hat{p} \leq 0.2) &= \mathbb{P}\left(\frac{\hat{p}-0.3}{0.06179} \leq \frac{0.2-0.3}{0.06179}\right) \\ &= \mathbb{P}\left(\frac{\hat{p}-0.3}{0.06179} \leq -1.62\right) \\ &= \mathbb{P}\left(\frac{\hat{p}-0.3}{0.06179} \geq 1.62\right) \\ &= 0.5 - \mathbb{P}\left(0 \leq \frac{\hat{p}-0.3}{0.06179} \leq 1.62\right) \\ &= 0.5 - 0.4474 = 0.0526\end{aligned}$$

### Large Sample Estimation:

Before this chapter, we have studied basic probability theory which provide basic models for applications. Starting from this chapter, we are concerned with the question that if we know that a population has certain distribution, how to find its parameters. The methods for making inferences about population parameters fall into one of two categories.

#### Estimation:

estimating the value of the parameter by an estimator or a statistic. Estimation consists of point estimation, which use a single number calculated based on sample data to estimate the population parameter (e.g. sample mean estimates the population mean) and interval estimation, which use two numbers calculated based on sample data to form an interval to cover the population parameter (e.g. using minimum sample and maximum sample to cover the population mean). The resulting value of an estimator is called an estimate.

#### Hypothesis Testing:

making a decision about the value of a parameter based on an assumption about what its value might be.

Example: In a box, we have 100 balls which are identical except their colors. We know that there are only white and black colors and only one ball has different color from others. Find out the proportion  $p$  of white ball(s).

**Definition:** An estimator is said to be unbiased if its mean is equal to the corresponding parameter of the population. (see picture)

The second measure to judge the goodness of an estimator is the spread or variance of the sampling distribution.

**Definition:** The distance between an estimate and the estimated parameter is called the error of estimation. Especially the standard deviation of the estimator is called standard error of the estimator. For any point estimator with normal distribution, the Empirical Rule states that 95% of all the point estimates will fall into two (or more exactly, 1.96) standard deviations of the mean of that distribution. Therefore, we called the  $(1.96) \times$  standard error of the estimator margin of error. For the convenience, we have following procedure for the estimation of the population mean:

(1) Since in this chapter the sample size is always  $n \geq 25$ , the point estimator  $\bar{X} = (1/n) \sum_{i=1}^n X_i$  for the population mean is an unbiased with standard error

$$\frac{\sigma}{\sqrt{n}}$$

and margin of error

$$\pm 1.96 \left( \frac{\sigma}{\sqrt{n}} \right)$$

(2) If the population standard deviation  $\sigma$  is unknown,

**the sample standard deviation**

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

can be used to approximate the  $\sigma$ .

For the binomial population which consists of  $n$  independent, identical experiments with two possible outcomes called success and failure. Let

$$X_i = \begin{cases} 1 & \text{if success with probability } p \\ 0 & \text{if failure with probability } q = 1 - p. \end{cases}$$

be the r.v. of  $i$ th experiment. Here  $p$  is called the population proportion. Define

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i$$

Then,  $\hat{p}$  is the sample proportion of the sample  $\{X_1, \dots, X_n\}$ .

For the convenience, we have following procedure for the estimation of the population proportion:

If  $n\hat{p} > 5$  and  $n\hat{q} > 5$  (where  $\hat{q} = 1 - \hat{p}$ ), then  $\hat{p}$  is an unbiased estimator of the population proportion with standard error

$$\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

and margin of error

$$\pm 1.96 \left( \sqrt{\frac{\hat{p}\hat{q}}{n}} \right)$$

Example 21.1: Estimate of Americans' heights is important to help governments, their different departments, and different companies to make decisions on policies and product designs. A sample of 10,000 Americans are randomly chosen, the sample mean is 1.7m and

**the standard deviation is 0.26m. What is an unbiased estimator of the Americans' average height? Find the margin of error of your estimate.**

**Solution:** The sample mean is an unbiased estimator of the Americans' average height. The margin of error of the sample mean is

$$\pm 1.96 \left( \frac{s}{\sqrt{n}} \right) = \pm 1.96 (0.26 / \sqrt{10,000}) = \pm 0.005096$$

**Example 21.2:** Although most school districts do not specifically recruit men to be elementary school teachers, those men who do choose a career in elementary education are highly valued and find the career very rewarding. If there were 40 men in a random sample of 250 elementary school teachers, find the unbiased estimator of the male proportion in the entire population and estimate the proportion of male elementary school teachers in the entire population. Give the margin of error for your estimate.

**Solution:** The sample proportion is an unbiased point estimator of number of the male elementary school teachers in the entire population. The estimate is

$$\hat{p} = \frac{x}{n} = \frac{40}{250} = 0.16$$

The margin of error of the sample proportion is

$$\pm 1.96 \left( \sqrt{\frac{\hat{p}\hat{q}}{n}} \right) = \pm 1.96 \left( \sqrt{\frac{0.16(0.84)}{250}} \right) = \pm 0.045$$