Chapter 4

Sampling distributions

Statistics is the science of answering questions on the basis of empirical data. Often, the question is such that it is impossible to obtain all of the relevant observations (the population) and a conclusion must be drawn on the basis of a sample of observations.

Sampling refers to the process of selecting a sample from a population. From the sample, we calculate statistics (numerical summaries of the sample) which used as estimates of the corresponding parameters (numerical summaries of the population).

The nature of Sampling Distributions

Sampling involves the selection of a sample ~~sample~~ from which statistics are calculated to estimate the corresponding parameters of the population. Once an estimate have been obtained, it is reasonable to ask how close the estimate is to the true value. What is the probability that estimate differs from the unknown parameter by less than so & so many units.

Def^{n}: A sampling distribution of a statistic is a description of the probabilities with which various values or set of values of the statistic occur.

4.3

9.2 ▪ Sampling Distribution of the Sample Mean

It is important to distinguish between the distribution of the population of values X and the sampling distribution of the sample mean $\bar{X}$. The following example utilizes sampling without replacement and illustrates how to obtain the sampling distribution of $\bar{X}$.

Example 9.3

The Sampling Distribution of $\bar{X}$: The data in Table 9.1 show 200 random digits. Consider these values to be a population. The population mean for these values is $\mu = 4.62$. Table 9.2 shows the frequency and relative frequency distributions for the data in Table 9.1, and Figure 9.3 shows the relative frequency histogram.

TABLE 9.1
200 random digits

1	2	6	5	1	6	1	6	4	6	1	1	7	6	9	7	5	1	0	9
8	6	9	9	6	9	7	6	6	9	2	5	7	5	7	3	2	5	3	5
0	7	1	2	2	7	6	7	6	3	8	1	7	6	9	7	4	4	3	6
0	2	6	3	0	7	2	3	1	0	4	5	0	4	9	1	8	0	2	9
0	7	4	6	9	4	2	3	4	1	9	8	1	7	3	7	9	2	6	0
3	6	7	3	7	9	8	8	6	3	7	7	2	4	0	7	6	2	5	1
0	0	6	5	4	6	4	6	8	8	0	9	3	4	3	7	0	2	7	8
6	7	3	3	1	9	8	7	2	9	8	2	8	6	1	5	4	3	7	1
7	6	6	1	0	9	4	9	3	4	7	2	7	4	8	4	4	1	2	4
0	5	6	1	0	5	3	7	5	0	9	5	9	3	8	0	1	4	8	5

TABLE 9.2 **Distribution of data in Table 9.1**

X	0	1	2	3	4	5	6	7	8	9
Frequency	19	20	17	19	20	15	27	28	15	20
Relative frequency	.095	.100	.085	.085	.100	.075	.135	.140	.075	.100

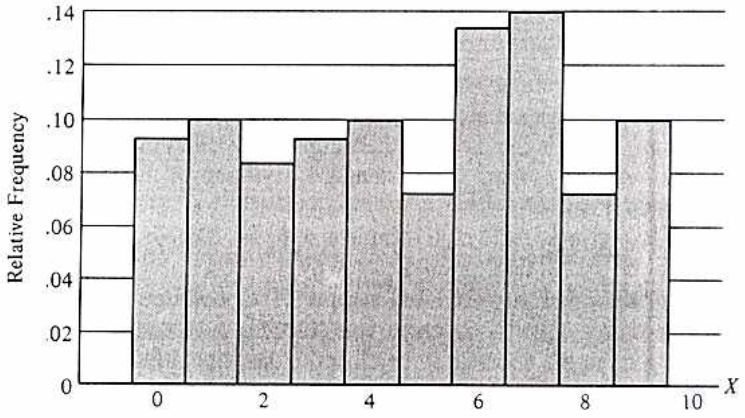


FIGURE 9.3
Relative frequency histogram of data in Table 9.1.

In Table 9.1 the values $x_1, x_2, \ldots, x_{20}$ are in the first row; $x_{21}, x_{22}, \ldots, x_{40}$ are in the second row; and so forth. To get information about the sampling distribution of the sample mean, we took 80 different samples of size $n = 10$ and calculated the sample mean $\bar{x}$ for each sample. The first sample, which contained the first 10 observations in the first row of Table 9.1, had a sample mean of 3.8. The second sample contained the last 10 observations in the first row of Table 9.1 and had a sample mean of 4.6. An additional 78 samples of size 10 were obtained, and the sample mean was calculated for each sample. The 80 values of $\bar{X}$ obtained in this fashion are shown in Table 9.3. Note that if we sample without replacement, ${}_{200}C_{10}$ different samples of size 10 can be selected from Table 9.1. We have chosen 80 of these possible samples to illustrate that $\bar{X}$ is a random variable with a value that varies from sample to sample. Table 9.4 shows the relative frequency distribution of the 80 sample means, and Figure 9.4 shows the relative frequency histogram of these 80 sample means.

Table 9.3
Eighty sample means from samples of size 10 shown in Table 9.1

3.8	4.6	7.5	4.4	4.1	5.5	2.4	4.2	4.0	5.2
6.0	4.1	4.7	4.3	5.5	4.5	4.9	4.3	3.2	5.2
2.5	4.8	5.4	3.8	3.0	7.1	4.5	6.2	4.5	4.3
5.5	4.5	5.1	4.9	5.7	4.8	4.3	2.4	4.3	4.8
3.3	5.1	5.1	4.3	4.0	4.4	5.0	6.2	4.0	4.3
4.2	4.1	5.0	6.4	4.8	5.6	5.0	4.3	2.7	4.8
3.5	5.7	4.6	3.2	4.7	5.1	3.5	4.8	4.8	4.9
4.9	6.2	5.0	3.4	4.5	5.0	5.5	5.2	4.4	3.5

Table 9.4
Distribution of 80 sample means in Table 9.3

$\bar{X}$	Frequency	Relative Frequency
2.0 to under 2.5	2	.0250
2.5 to under 3.0	2	.0250
3.0 to under 3.5	5	.0650
3.5 to under 4.0	5	.0650
4.0 to under 4.5	19	.2375
4.5 to under 5.0	20	.2500
5.0 to under 5.5	13	.1625
5.5 to under 6.0	7	.0875
6.0 to under 6.5	5	.0625
6.5 to under 7.0	0	.0000
7.0 to under 7.5	2	.0250
Total	80	1.0000

Figure 9.3 shows that the relative frequency distribution of the population of data is approximately a uniform distribution. This result is reasonable because if the digits are truly random, each digit should be equally likely with a relative frequency of .1.

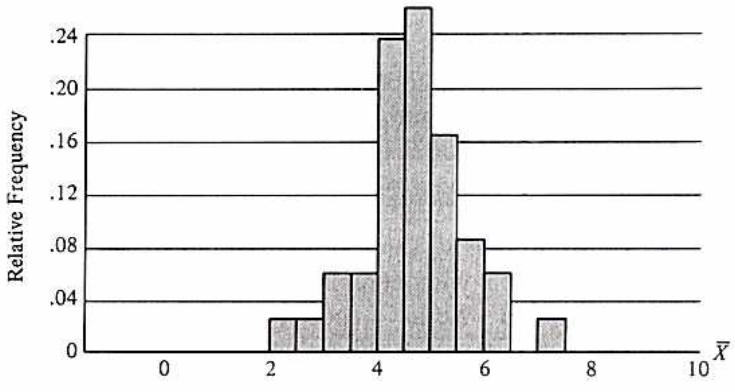


FIGURE 9.4
Relative frequency histogram of 80 sample means in Table 9.3.

Figure 9.4 shows the distribution of $\bar{X}$ values. This distribution is approximately bell shaped, and most of the values of $\bar{X}$ are clustered around 4.62, which is the mean of the original population of data. Even though the population is approximately uniformly distributed, the sampling distribution of $\bar{X}$ is approximately normal. It can be shown that this approximation to the normal distribution improves as the sample size increases. For example, if we had taken samples of size $n = 30$ rather than $n = 10$, the approximation to normality would have been more exact. ■

Mean and Variance of the Random Variable $\bar{X}$

The mean and variance of the sampling distribution of $\bar{X}$ depend on the mean and variance of the population and on the sample size n. The formulas for the mean and variance of the sampling distribution of $\bar{X}$ are given in the accompanying box.

Formulas for the Mean and Variance of $\bar{X}$

Let $x_1, x_2, \ldots, x_n$ denote a random sample of size n selected from a population having mean μ_x (or μ) and variance σ_x^2 (or σ^2). Let $\bar{X}$ denote the sample mean.

1. The mean value, or expected value, of $\bar{X}$ is
$$E(\bar{X}) = \mu_{\bar{X}} = \mu$$
2. If the population is infinitely large or if sampling is done with replacement, then the variance of $\bar{X}$ is
$$\text{Var}(\bar{X}) = \sigma_{\bar{X}}^2/n$$
and the standard deviation of $\bar{X}$ is given by $\sigma_{\bar{X}}/\sqrt{n}$.
3. If sampling is done *without replacement from a finite population* containing N elements, then the variance of $\bar{X}$ is
$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

The value $[(N - n)/(N - 1)]$ in the formula for calculating the variance when sampling is done without replacement is called the *finite population correction factor*

where N is the population size and n is the sample size. If the population size N is much larger than the sample size n, this correction factor is close to 1; in such cases it makes little difference whether we sample with or without replacement. Unless explicitly noted, we will assume henceforth that sampling is done from an extremely large population so that this correction factor can be ignored.

According to the theorem, the standard deviation of the $\bar{X}$ values is $\sigma_{\bar{X}} = \sigma/\sqrt{n}$. This shows that $\sigma_{\bar{X}}$ will be small if σ is small or if n is large. If the population standard deviation σ is small, there is little variation in the population and we would expect sample means obtained from different samples to be highly concentrated. If the sample size is large, we would expect the different estimates $\bar{X}$ obtained from different samples to be highly concentrated about the mean μ.

Effect of a Small Population Standard Deviation: An employment agency wants to estimate the mean hourly wage earned by high school students who work as cashiers at supermarkets. The agency hires several pollsters and asks each one to obtain a random sample of $n = 25$ students. Because most students earn between \$3.35 and \$5.00 per hour, the population standard deviation is relatively small, say, $\sigma = \$1.00$. Because the students' wages are highly concentrated and do not exhibit much variation, the estimates obtained by the different pollsters should not exhibit much variation either. That is, we would expect the values of $\bar{X}$ obtained by each pollster to be quite close to one another. Thus, if the population has a small standard deviation, the sampling distribution of sample means will also be highly concentrated. ■

Effect of a Large Population Standard Deviation: The employment agency in Example 9.4 is also interested in estimating the mean annual income earned by lawyers. Again pollsters are hired, and each is told to obtain a random sample of $n = 25$ lawyers. The population mean is, say, \$70,000 per year. However, the standard deviation of this population is very large, say $\sigma = \$20{,}000$, because the annual incomes are widely dispersed about the mean. As a result, the sample mean obtained by one pollster could be quite different from the sample mean obtained by another. We would not necessarily expect the values of $\bar{X}$ obtained by the pollsters to be quite so close to one another. If the population has a large standard deviation, the sampling distribution of the sample mean can also have a large amount of dispersion. ■

The following example illustrates the point that the variance of the sampling distribution of $\bar{X}$ depends on the sample size n. If the sample size is large, then the sampling distribution of $\bar{X}$ will be highly concentrated about the population mean.

Effect of Sample Size on the Variance of $\bar{X}$: Several economists are interested in estimating the mean annual income of accountants. Suppose the unknown population mean is $\mu = \$50{,}000$ and the population standard deviation is $\sigma = \$10{,}000$. Now suppose one economist takes a random sample of $n = 4$ employees. Because this sample size is very small, the estimate obtained by the economist could be quite different from the true population mean. For example, the standard deviation of the sampling distribution of $\bar{X}$ is $\sigma/\sqrt{n} = 10{,}000/\sqrt{4} = \$5000$. From the Empirical Rule, approximately 68% of the data lie within 1 standard deviation of the mean if the distribution is approximately bell shaped, so only about 68% of the potential estimates of the population mean will be within \$5000 of the true mean. Since the remaining 32% of the estimates will be

more than \$5000 from the mean, the probability is approximately .32 that the sampling error will be more than \$5000. An estimate of the mean income that is in error by more than \$5000 could be useless in many economic studies. Because the sample size is very small, the sampling distribution of $\bar{X}$ is widely dispersed about the population mean.

Now suppose another economist takes a random sample of $n = 1600$ employees. Because this sample size is quite large, we would expect to obtain a very good estimate of the population mean. In this case, the standard deviation of the sampling distribution of $\bar{X}$ is $\sigma/\sqrt{n} = \$10{,}000/\sqrt{1600} = \$250$. From the Empirical Rule, the probability is approximately .95 that the sampling error $|\bar{X} - \mu|$ will be less than 2 standard deviations, namely \$500, if the distribution is approximately bell shaped. Thus, in approximately 95% of the potential samples of size 1600, the estimated sample mean will be within \$500 of the true population mean. An estimate of the mean income that is in error by less than \$500 would probably be quite useful in most economic studies. When the sample size is large, the sampling distribution of the sample mean is highly concentrated about the population mean. ■

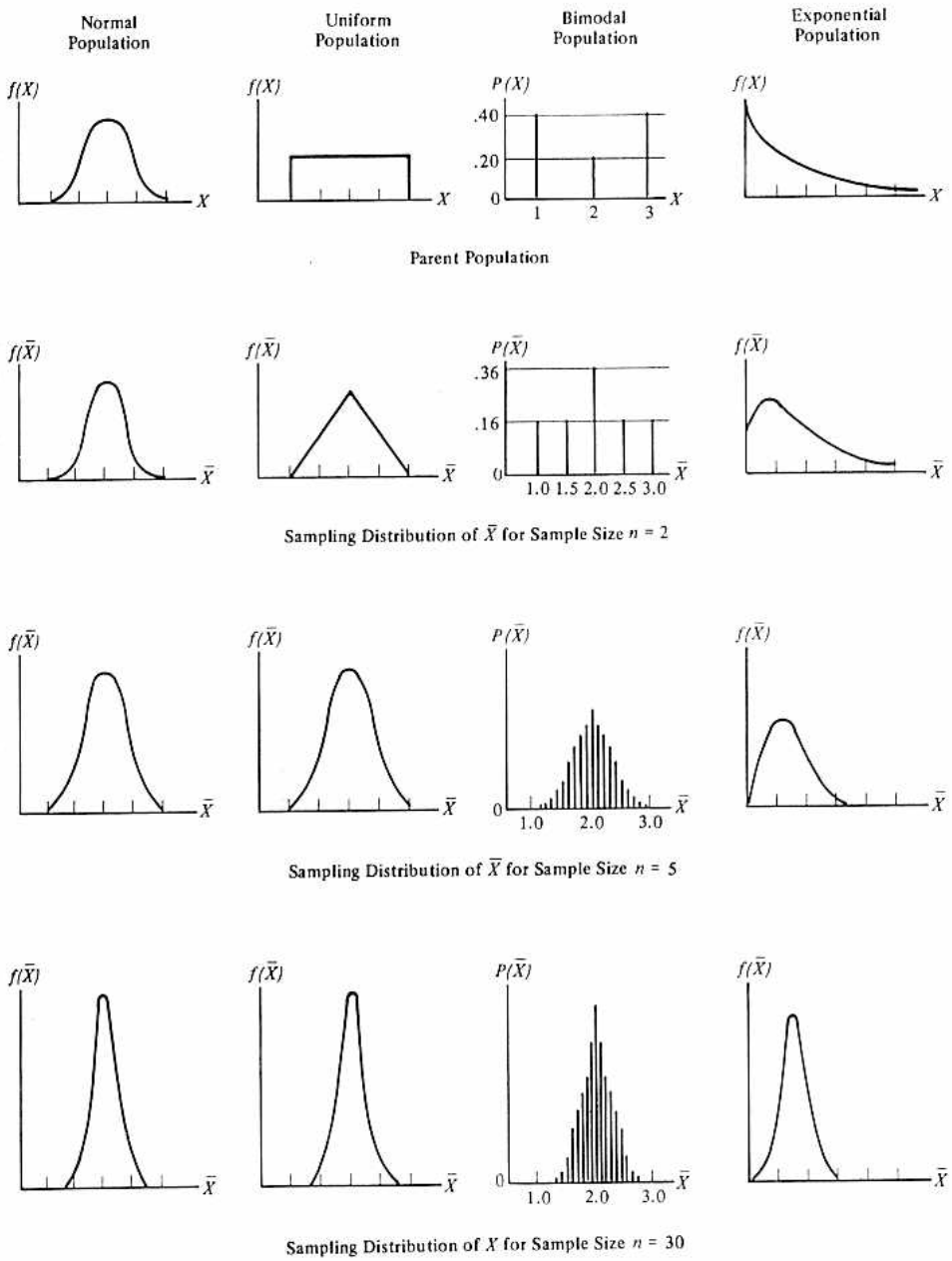
Graphical Representation of the Distribution of $\bar{X}$

The graphs in Figure 9.5 show that the sampling distribution of $\bar{X}$ is centered at the population mean and becomes more concentrated as the sample size increases. Also note that as the sample size gets larger, the distribution of $\bar{X}$ begins to approach normality regardless of the shape of the original distribution.

Exercises for Section 9.2

1. Suppose the average IQ of all college seniors is 125 and the variance is 144. You take a random sample of 100 students and calculate the sample mean IQ ($\bar{X}$). The sample mean $\bar{X}$ is a random variable having a probability distribution. What is the mean, variance, and standard deviation of this distribution?
2. A city council is studying the tourist trade. During a certain year, the mean expenditure per tourist was $\mu = \$500$ with variance 8100. The city takes a random sample of 10 tourists and calculates their average expenditure $\bar{X}$. What is the mean, variance, and standard deviation of this distribution?
3. In Exercise 2 the city takes a random sample of 1000 tourists and calculates the mean expenditure. Now what is the mean, variance, and standard deviation of the distribution of $\bar{X}$? In this case would you be very confident that your value of $\bar{X}$ is very close to μ?
4. You take a random sample of size $n = 10$ and obtain $x_1 = 8$, $x_2 = 6$, $x_3 = 5$, $x_4 = 5$, $x_5 = 8$, $x_6 = 7$, $x_7 = 6$, $x_8 = 2$, $x_9 = 4$, and $x_{10} = 6$. Calculate the sample mean and sample standard deviation.
5. A person takes a random sample of 10 college seniors to obtain their IQs and gets the values $x_1 = 110$, $x_2 = 145$, $x_3 = 150$, $x_4 = 130$, $x_5 = 120$, $x_6 = 105$, $x_7 = 130$, $x_8 = 105$, $x_9 = 125$, and $x_{10} = 120$. Use the formulas in Chapter 4 to calculate the sample mean and sample variance. Use the formula S^2/n to estimate the variance of $\bar{X}$.

FIGURE 9.3 **Sampling distribution of $\bar{X}$ for $n = 2$, 5, and 30.**



6. Five randomly selected tourists are asked how much they spent on their vacation. We obtain $x_1 = \$600$, $x_2 = \$800$, $x_3 = \$400$, $x_4 = \$1000$, and $x_5 = \$500$. Calculate the sample mean and sample variance. Estimate the variance of $\bar{X}$ using the estimator S^2/n.
7. If the population is approximately normal, the variance of the sample median equals $(\pi/2)(\sigma^2/n)$. If a sample median is to have a standard deviation equal to that of a sample mean (from the same population), how much bigger must the sample size be?

The Central Limit Theorem (C.L.T.)

The Central Limit Theorem, one of the most important theorems in statistics, is the reason that the normal distribution is the most important of all probability distribution.

C. L. T. :

Let $x_1, x_2, \ldots x_n$ denote a random sample selected from a population having mean μ and variance σ^2. Let $\bar{X}$ denote the sample mean. If n is large, then $\bar{X}$ has approximately a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$

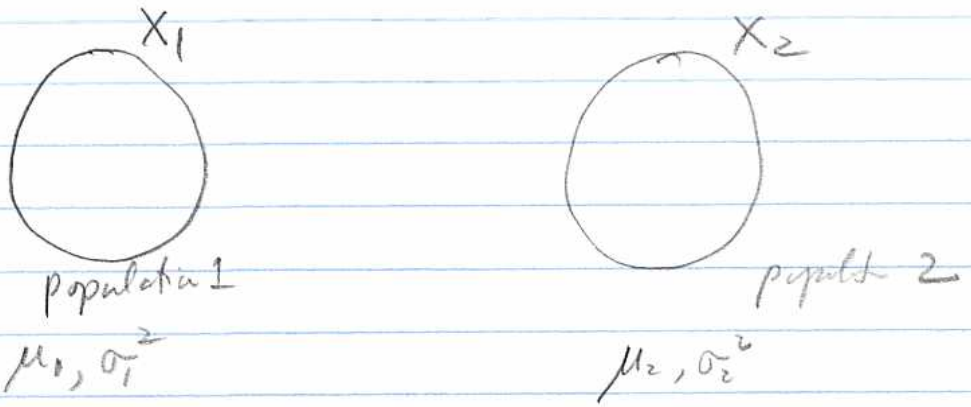
The C.L.T. would be of little practical value if the approximation was applicable only if n were extremely large, but in fact it works in most cases, even when n is fairly small. A common rule of thumb is that the theorem applies as n gets up to 20 or 30.

If the original population follows the normal distribution, then the distribution of $\bar{X}$ is exactly a normal distribution, regardless of the sample size. The C.L.T. is important because it states that if the original population is NOT Normal, the distribution of $\bar{X}$ will be approximately normal (provided $n \geq 30$)

e.g. Application of the C.L.T. : Suppose the Bureau of Labour Statistics wants to estimate the mean starting salary of newly graduated

Sampling distribution of the difference between two sample means

Frequently we are interested in determining if the mean of one population is equal to the mean of another.
e.g. (a) Is the mean height of male the same as the mean height of female.
(b) Is the mean number of cavities for children who use Brand A toothpaste the same as mean # of cavities for children who use Brand B.



To make comparisons, we take <u>independent</u> random sample of n_1 observations from group 1 & n_2 obs. from population 2 and then calculate $\bar{x}_1, \bar{x}_2$.

$\bar{X}_1, \bar{X}_2$ are random variables with values that vary from sample to sample. $\bar{X}_1 - \bar{X}_2$ is also a random variable with values vary from sample to sample.
Now, we have 5 different prob. distribution
$X_1, X_2, \bar{X}_1, \bar{X}_2, \bar{X}_1 - \bar{X}_2$

If n_1, n_2 is large enough ($n_1, n_2 \geq 30$), we

have

$$\bar{X}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right)$$

$$\bar{X}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

and $\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$

e.g A financial loan officer claims that the mean monthly payment for credit cards is \$80 with a variance of 1400 for single females and \$80 with a variance of 1320 for single males. We take a random sample of 100 single females & 120 single males.

Single females: $\mu_1 = 80$, $\sigma_1^2 = 1400$, $n_1 = 100$

Single males: $\mu_1 = 80$, $\sigma_2^2 = 1320$, $n_2 = 120$

What is the probability that the sample mean for females will be at least \$5 higher than the sample mean for males?

Soln: by CLT

$$\bar{X}_1 - \bar{X}_2 \sim N\left(0, \frac{1400}{100} + \frac{1320}{120}\right)$$

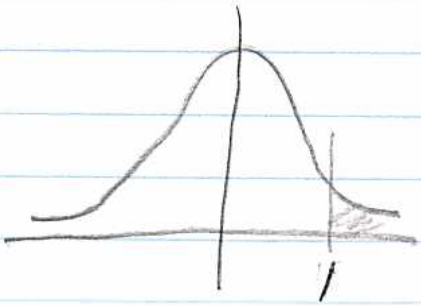
$$= N(0, 25)$$

We seek the prob that $\bar{X}_1$ is at least \$5 more than $\bar{X}_2$.

i.e. we seek

$$P\{\bar{X}_1 - \bar{X}_2 \geq 5\} = P\left\{\frac{\bar{X}_1 - \bar{X}_2 - 0}{5} \geq \frac{5-0}{5}\right\}$$

$$= P\{Z > 1\} = 0.1587$$



Sampling Distribution of the sample proportion

There are numerous problems in real life where we want to know the proportion of items in a population that posses a certain characteristic

e.g. (1) TV producers wants to know what proportion of viewers of a certain program have incomes exceeding $40k per year

(2) proportion of lung cancer patient out of a smoking population.

In each case we have to estimate the parameter p, the proportion of the population having the characteristic of interest.

Let x denote the # of observations in the sample that possess the characteristic of interest. We use the sample proportion $\hat{p} = \frac{x}{n}$ to estimate the population proportion p.

If $np \geq 5$, $nq \geq 5$, $X \sim N(np, npq)$

under the same constraints, the sample distribution of $\hat{p}$ is approximately normal

i.e. $\hat{p} = \frac{x}{n} \sim N\left(p, \frac{pq}{n}\right)$

i.e. $\text{mean}(\hat{p}) = p$, $\sigma^2(\hat{p}) = \frac{pq}{n}$

e.g. TV producer claims that 30% of all homes in a city have cable TV. To test this claim, a random sample of 400 homes is examined.

(a) What is the probability that the sample proportion is less than 0.25?

(b) What is the probability that the sample proportion is bigger than 0.20?

(c) Suppose the sample proportion is 0.18 Would this make you doubt the producer's claim?

Solution:

$$\sigma = \sqrt{\frac{0.3 \times 0.7}{400}} = 0.02291$$

(a) $$P\{\hat{p} < 0.25\} = P\left\{\frac{\hat{p} - 0.3}{0.02291} \leq \frac{0.25 - 0.3}{0.02291}\right\}$$

$$= P\{Z \leq -2.18\} = 0.0146$$

(b)
$$P\{\hat{p} > 0.20\} = P\left\{\frac{\hat{p}-0.3}{0.02291} > \frac{0.20-0.3}{.02291}\right\}$$
$$= P(z > -4.36) = 1$$

(c) Yes, there is virtually NO chance of it being below 0.20.

e.g. A medical journal states that the probability of surviving open heart surgery is .9. To check this claim, you examine the records of a sample of 200 patients who had open heart surgery. In the sample, 77% survived.

(a) If the journal's claim is correct, find the probability that the proportion of survivors is 77% or less

(b) Does the result in part (a) make you doubt the journal's claim? Explain.

Soln: (a)
$$\sigma = \sqrt{\frac{(.9)(.1)}{200}} = 0.02121$$

(a)
$$P\{\hat{p} \le 0.77\} = P\left\{\frac{\hat{p}-0.9}{0.02121} \le \frac{0.77-0.9}{0.02121}\right\}$$
$$= P\{z \le -6.128\} = 0$$

(b) Yes, there is NO chance of it being below 0.77.