Chapter I

Sampling Distribution

D. 1. 1 (*Population*)
Any well defined set of objects about which a statistical enquiry is being made is called a *population* or *universe.*
The total number of objects (individuals or members) in a population is known as the *size of the population*.

R. 1. 1.
The size of the population can be finite or infinite. A population is finite if it contains a finite number of individuals. It is infinite if it contains an infinite number of individuals.

Ex. 1. 1.
1. The ages of 10000 students of a university form a finite population.
2. The pressures at various points in the atmosphere are an example of an infinite population.

R. 1. 2.
It is often impossible or at least expensive or time consuming to study the whole population. Therefore, we confine ourselves to a small selection from the population.

D. 1. 2 (*Sample*)
A finite set of objects drawn from the population with an aim is called *sample*.

D. 1. 3. (*Random Sampling*)
Random sampling consists of the selection of individuals from the population in such a way that each individual of the population has an equal chance of being selected (i.e. a sample so selected must be a *true representative* of the population.)

R. 1. 3.
The aim of the theory of sampling is to get as much information as possible, ideally all the information about the population from which the sample has been drawn. From the parent population, in particular, we would like to estimate the parameters of the population or specify the limits or ranges within which the population parameters are expected to lie with a specified degree of confidence.
The logic of the sampling theory is the logic of induction, that is, we go from particular (i.e. sample) to general (i.e. population). That is why we speak of "Inductive Statistics".

R. 1. 4.
In the theory of sampling, we often come across two words, '*parameter*' and '*statistic*'.
The word 'parameter' is associated with population'. It is understood as the measure of the characteristics of the population, such as means and standard deviation etc.
The word 'statistic' is used for a random sample and it is understood as the measure of the characteristics of the random sample, such as means and standard deviation etc.
Different symbols are usually used to denote parameters and statistics:

Symbols for Corresponding Statistics and Parameters

Measure	Symbol for Statistic (Sample)	Symbol for Parameter (Population)
Mean	$\bar{x}$	μ
Standard deviation	s	σ
Number of Items	n	N
Proportion	p	P

Inferential or inductive statistics helps us making an educated guess about the population based on the statistics of its random sample.

D. 1. 4. (*Sampling Distribution*)
A distribution of the statistic obtained from the samples is called the *sampling distribution of the statistic*.

Ex. 1. 2.
If the sample size is two and the population size is three (items A, B, C), it is possible to draw three samples (AB, BC, AC) from the population. We may compute the mean for each sample. Thus, we have three sample means. The three sample means form a distribution. The distribution of the means is called the *distribution of sample means* or the *sampling distribution of the mean*.
Likewise, the distribution of the proportions (or percent rates) obtained from all possible samples is called the *sampling distribution of the proportion*.

D. 1. 5. (*Standard Error*)
The standard deviation of a sampling distribution of a statistic is called the *standard error of the statistic*.

R. 1. 5.
The difference between the terms "standard deviation" and "standard error" is that the former concerns original values, whereas the latter concerns computed values.
Standard error measures the precision of the statistic as an estimate of a population or model parameter.

D. 1. 6. (*Sampling Error*)
The absolute difference between the result obtained from a sample (a statistic) and the result which would have been obtained from the population (the corresponding parameter) is called the *sampling error*:

$$\text{Sampling error} := |\bar{x} - \mu| .$$

R. 1. 6.
A sampling error usually occurs when the complete survey of the population is not carried out, but a sample is taken for estimating the characteristic of the population. The sampling error is measured by the standard error of the statistic in terms of probability under the normal curve. The result of the measurement indicates the *precision* of the estimate of the population based on the sample study. The smaller the sampling error, the greater is the precision of the estimate. For example, the difference between the mean of a sample and the mean of the population, if it were obtained, is a type of sampling error of the mean.
It should be noted that the error made in a sample survey, such as answers being inconsistent, incomplete, or not determinable, are not considered sampling errors. The nonsampling errors may also occur in a complete survey of the population.

R. 1. 7. (*Methods of Selecting Samples*)
Depending on how a sample is drawn, it may be a *random sample* or a *non-random sample*.

A *random sample* is a sample drawn in such a way that each member of the population has some chance of being selected in the sample.

In a *non-random sample*, some members of the population may not have any chance of being selected in the sample

There are many ways to select a random sample. We discuss four of them:

- Simple Random Sampling
- Systematic Random Sampling
- Stratified Random Sampling
- Cluster Sampling

A sampling technique under which each sample of the same size has the same possibility of being selected is called *simple random sampling*.

(*Example:*
Given a list of 3000 e-mail addresses of possible survey respondents, a researcher would like to survey a sample of 200. Then a simple random sample is one in which every possible group of 200 e-mail addresses out of the 3000 are equally likely.)

In *systematic random sampling*, we first randomly select one member from the first k units. Then every k th member, starting with the first selected member, is included in the sample.

(*Example:*
For example, suppose we want to sample 200 e-mail addresses from the sampling frame that contains 3000 e-mail addresses. Then we have that $3000/200=15$, so every 15th e-mail address in the list is chosen after beginning at a random point between 1 and 15. If the random starting point is 5, then the e-mail addresses selected are $5,\ 20,\ 35,\ 60,...,2975,\ 3000$.
Now, if there were 3050 e-mail addresses, $3050/200=15.25$, so we take every 15th address. (If we were to take every 16th address, $200\cdot 16=3200$, so that the last few e-mail addresses 'chosen' will not exist and the resulting sample size will be smaller than desired. That is, we would only select 190 e-mail addresses.) However, note that $200\cdot 15=3000$, which means that if we use a random starting point between 1 and 15 the last 50 e-mail address will never have a chance of being selected. Hence, the random starting point should now be chosen to be between 1 and 65 to ensure that every e-mail address has some chance of being selected.)

In a *stratified random sampling*, we first divide the population into subpopulations, which are called *strata*. Then one sample is selected from each of these strata.

(*Example:*
Perhaps the 3000 e-mail addresses just described could be grouped into computer power users ($N_1 = 100$), average computer users ($N_2 = 1800$), and computer novices ($N_3 = 1100$). A sample drawn from such a population using stratified random sampling will very likely result in the selection of only a few power users. (For example, a stratified random sample of size $n = 200$ will likely yield about six or seven power users.) If it is important to obtain a larger sub-sample of power users in the overall sample, one way to do this is to draw a stratified random sample from each of the user strata according to pre-defined sub-sample sizes, $n_1, n_2,$ and n_3 .)

In *cluster sampling*, the whole population is first divided into (geographical) groups called clusters. Each cluster is representative of the population. Then a random sample of clusters is selected. Finally, a random sample of elements from each of the selected clusters is selected.

(*Examples:*
To survey individuals it is sometimes useful to use household as the sampling unit. The Household is the cluster and in this scheme a sample of households are first selected, and then individuals are randomly selected from within each household.
In surveys of Internet users, it is sometimes useful or convenient to first sample by discussion groups or Internet domains and then to sample individual users.)

T. 1. 1.
The mean of the sampling distribution of means, denoted by $\mu_{\bar{x}}$, is given by

$$\mu_{\bar{x}} = \mu \tag{1. 1.}$$

where μ is the mean of the population.

R. 1. 8.
Theorem T. 1. 1. states that the expected value of the sample mean is the population mean.

T. 1. 2.
If a population is infinite and the sampling is random or if the population is finite and the sampling is with replacement, then the variance of the sampling distribution of means, denoted by $\sigma_{\bar{x}}^2$, is given by

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \tag{1. 2.}$$

where σ^2 is the variance of the population.

T. 1. 3.
If the population is of size N , if sampling is without replacement, and if the sample size is $n \le N$, then (1. 2.) is replaced by

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \cdot \left(\frac{N-n}{N-1} \right) \tag{1. 3.}$$

where $\mu_{\bar{X}}$ is still given by (1. 1.)

R. 1. 9.
Note that (1. 3.) reduces to (1. 2.) as $N \to \infty$.

Ex. 1. 3.
A population consists of the five numbers 2, 3, 6, 8, 11. Consider all possible samples of size two which can be drawn with replacement from this population. Find

1. the mean of the population,
2. the standard deviation of the population,
3. the mean of the sampling distribution,
4. the standard deviation of the sampling distribution of means, i.e., the standard error of means.

Solution:
1.

$$\mu = \frac{2+3+6+8+11}{5} = 6.0 .$$

2.

$$\sigma^2 = \frac{(2-6)^2 + (3-6)^2 + (6-6)^2 + (8-6)^2 + (11-6)^2}{5} = 10.8 ,$$

$$\sigma = 3.29.$$

3.
There are 25 samples of size two which can be drawn with replacement:

(2,2)	(2,3)	(2,6)	(2,8)	(2,11)
(3,2)	(3,3)	(3,6)	(3,8)	(3,11)
(6,2)	(6,3)	(6,6)	(6,8)	(6,11)
(8,2)	(8,3)	(8,6)	(8,8)	(8,11)
(11,2)	(11,3)	(11,6)	(11,8)	(11,11)

The corresponding sample means are

2.0	2.5	4.0	5.0	6.5
2.5	3.0	4.5	5.5	7.0
4.0	4.5	6.0	7.0	8.5
5.0	5.5	7.0	8.0	9.5
6.5	7.0	8.5	9.5	11.0

And the mean of distribution is

$$\mu_{\bar{x}} = \frac{\text{sum of all sample means above}}{25} = \frac{150}{25} = 6.0$$

illustrating the fact that $\mu_{\bar{x}} = \mu$.

4.

The variance $\sigma^2_{\bar{x}}$ of the sampling distribution means is obtained by subtracting the mean 6 from each number, squaring the result, adding all 25 numbers obtained, and dividing by 25. The final result is

$$\sigma^2_{\bar{x}} = \frac{135}{25} = 5.40$$

so that

$$\sigma_{\bar{x}} = \sqrt{5.40} = 2.32.$$

This illustrates the fact that for finite populations involving sampling with replacement (or infinite populations), $\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$ since the right-hand side $\frac{10.8}{2} = 5.40,$ agreeing with the above value.

Ex. 1. 4.

Solve Exercise 1. 3. in case sampling is without replacement.

Solution:

1. – 2.

As in 1. and 2. of Exercise 1. 3., $\mu = 6$ and $\sigma^2 = 10.8,$ $\sigma = 3.29.$

3.

There are 10 samples of size two which can be drawn without replacement from the population, namely

$$(2,3),\ (2,6),\ (2,8),\ (2,11),\ (3,6),\ (3,8),\ (3,11),\ (6,8),\ (6,11),\ (8,11).$$

The corresponding sample means are

$$2.5,\ \ 4.0,\ \ 5.0,\ \ 6.5,\ \ 4.5,\ 5.5,\ 7.0,\ 7.0,\ 8.5,\ 9.5$$

and the mean of sampling distribution of means is

$$\mu_{\bar{x}} = \frac{2.5 + 4.0 + 5.0 + 6.5 + 4.5 + 5.5 + 7.0 + 7.0 + 8.5 + 9.5}{10} = 6.0$$

illustrating the fact that $\mu_{\bar{x}} = \mu$.

4.

The variance of the sampling distribution of means is

$$\sigma^2_{\bar{x}} = \frac{(2.5-6.0)^2 + (4.0-6.0)^2 + (5.0-6.0)^2 + \ldots + (9.5-6.0)^2}{10} = 4.05$$

and $\sigma_{\bar{x}} = 2.01$.

This illustrates $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \cdot \left(\frac{N-n}{N-1} \right)$, since the right side equals $\frac{10.8}{2} \cdot \left(\frac{5-2}{5-1} \right) = 4.05$,
as obtained above.

T. 1. 4.
If the population from which samples are taken is normally distributed with mean μ and variance σ^2, then the sample mean is normally distributed with mean μ and variance σ^2 / n.

Ex. 1. 5.
Assume that the heights of 3000 male students at a British university are normally distributed with mean 68.0 inches and standard deviation 3.0 inches. If 80 samples consisting of 25 students each are obtained, what would be the mean and standard deviation of the resulting sample of means if sampling were done

a) with replacement,

b) without replacement?

Solution:
The number of samples of size 25 that could be obtained theoretically from a group of 3000 students with and without replacement are 3000^{25} and $\binom{3000}{25}$ which are much larger than 80. Hence, we do not get a true sampling distribution of means but only an *experimental* sampling distribution. Nevertheless, since the number of samples is large, there should be close agreement between the two sampling distributions. Hence, the mean and standard deviation of the 80 sample means would be close to those of the theoretical distribution. Therefore, we have:

a)

$$\mu_{\bar{x}} = \mu = 68.0 \text{ inches}, \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{25}} = 0.6 \text{ inches}.$$

b)

$$\mu_{\bar{x}} = \mu = 68.0 \text{ inches}, \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = \frac{3}{\sqrt{25}} \sqrt{\frac{3000-25}{3000-1}} = 0.5975943772 \text{ inches}$$

which is only very slightly less than 0.6 inches and can for all practical purposes be considered the same as in sampling with replacement.
Thus we could expect the experimental sampling to be approximately normally distributed with mean 68.0 inches and standard deviation 0.6 inches.

T. 1. 5.

Suppose that the population from which samples are taken has a probability distribution with mean μ and variance σ^2 that is not necessarily a normal distribution. Then the standardised variable associated with $\bar{x}$, given by

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \qquad (1.\ 4.)$$

is asymptotically normal, i. e.,

$$\lim_{n\to\infty} P(Z \le z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/dt}. \qquad (1.\ 5.)$$

R. 1. 10.

Theorem T. 1. 5. is a consequence of the central limit theorem[1]. It is assumed here that the population is infinite or that sampling is with replacement.. Otherwise, the above is correct if we replace $\sigma/\sqrt{n}$ in (1. 4.) by $\sigma_{\bar{x}}$ as given by (1. 3.).

R. 1. 11. (***Shape of the Sampling Distribution of the Mean***)

To summarise:

1.

If a sample selected from a *normally distributed population* then the distribution of the sample mean will also be *normal* with

$$\mu_{\bar{x}} = \mu, \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

2.

If a "large" sample ($n \ge 30$) is selected from "*any*" *population*, then the distribution of the mean will be *approximately normal* with

$$\mu_{\bar{x}} = \mu, \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

Ex. 1. 6.

The portfolio of wealthy people over the age of 50 produce yearly retirement incomes which are normally distributed with mean equal to \$125000 and standard deviation equal to \$25000.

---

[1] Let $X_1, X_2, ..., X_n$ be independent random variables that are identically distributed (i.e., all have the *same* probability function in the discrete case or density function in the continuous case) and have finite mean μ and variance σ. Then if $S_n = X_1, X_2, ..., X_n,\ n = 1, 2, ...,$

$$\lim_{n\to\infty} P\left(a \le \frac{S_n - n\mu}{\sigma\sqrt{n}} < b \right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2}\, dt$$

that is, the random variable $(S_n - n\mu)/\sigma\sqrt{n}$, which is the standardised variable corresponding to S_n, is asymptotically normal.

1.
Describe the distribution of the mean of samples of size 16 from the population.

2.
Find the probability that the mean retirement income of the 16 wealthy individuals exceeds \$135000.

Solution:
1.
Since the population of yearly retirement incomes is normally distributed, the distribution of the sample mean is also normal with

$$\mu_{\bar{x}} = \$125000\,, \qquad \sigma_{\bar{x}} = \frac{25000}{\sqrt{16}} = \$6250\,.$$

2.

$$P(\bar{x} > 135000) = 1 - P(\bar{x} \le 135000) \approx 1 - P(\bar{x} < 135000)$$

$$= 1 - F(135000) = 1 - \Phi\left(\frac{135000 - 125000}{6250}\right) = 1 - \Phi(1.6)$$

$$= 1 - 0.945201 = 0.054799$$

Ex . 1. 7.
The mean age of non-residential buildings is 30 years and their standard deviation 5 years. The distribution of the ages is not normal.

1.
Describe the distribution of the mean of the sample means for samples of sizes

i) 10
ii) 50.

2.
Determine the probability that a random sample of 50 non-residential buildings will have a mean age of

i) 27.5 years or less
ii) between 28 and 33 years.

Solution:
1.
The distribution of ages is not normal.

i) The distribution of the mean of the sample means of size 10 is not known ($n = 16 \le 30$).

ii) The distribution of the mean of the sample means of size 50 is approximately normal with ($n = 16 \le 30$) with

$$\mu = 30, \quad \sigma_{\bar{x}} = \frac{5}{\sqrt{50}} \approx 0.707 .$$

2.

i)

$$P(\bar{x} \le 27.5) \approx P(\bar{x} < 27.5) = F(27.5) = \Phi\left(\frac{27.5-30}{0.707}\right)$$

$$= \Phi(-3.54) = 1 - \Phi(3.54) = 1 - 0.9998 = 0.0002 .$$

ii)

$$P(28 \le \bar{x} < 33) = F(33) - F(28) = \Phi\left(\frac{33-30}{0.707}\right) - \Phi\left(\frac{28-30}{0.707}\right)$$

$$= \Phi(4.24) - \Phi(-2.82) = \Phi(4.24) - (1 - \Phi(2.82))$$

$$= 1 - 1 + \Phi(2.82) = 0.997599$$

D. 1. 6. (***Sampling Distribution of Proportions***)

The probability distribution of the sample distribution, $\bar{p}$, is called its *sampling distribution*.

T. 1. 6.

$$\mu_{\bar{P}} = P ,$$

$$\sigma_{\bar{p}} = \begin{cases} \sqrt{\dfrac{P(1-P)}{n}} & \text{,when} \quad \dfrac{n}{N} \le 0.05 \\ \sqrt{\dfrac{P(1-P)}{n}} \cdot \sqrt{\dfrac{N-n}{N-1}} & \text{,when} \quad \dfrac{n}{N} > 0.05 \end{cases}$$

Ex. 1. 8.

The following table describes a population consisting of five states and indicates whether or not there is at least one woman on death row for that sate:

State	At least one woman on death row
A: Alabama	yes
B: California	yes
C: Colorado	no
D: Kentucky	no
E: Nebraska	no

For this population 40% of the states have at least one woman on death row.

1. List all samples of size 2 from this population.

2. Find the proportion having at least one woman on death row for each sample.
3. Use the above listing to derive the sampling distribution for the sample proportion.
4. Find the mean and standard deviation of the sample proportion and verify the formulas in the script.

Solution:
1.- 2.

Sample	At least one woman on death row	Sample proportion	Probability
A, B	y, y	1.0	0.1
A, C	y, n	0.5	0.1
A, D	y, n	0.5	0.1
A, E	y, n	0.5	0.1
B, C	y, n	0.5	0.1
B, D	y, n	0.5	0.1
B, E	y, n	0.5	0.1
C, D	n, n	0.0	0.1
C, E	n, n	0.0	0.1
D, E	n, n	0.0	0.1

3.

$\bar{p}$	0.0	0.5	1.0
$P(\bar{p})$	0.3	0.6	0.1

4.

$$P = 0.4, \qquad \mu_{\bar{p}} = 0 \cdot 0.3 + 0.5 \cdot 0.6 + 1 \cdot 0.1 = 0.4 \Rightarrow P = \mu_{\bar{p}},$$

$$\sqrt{\frac{P(1-P)}{n}} \cdot \sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{0.4(1-0.4)}{2}} \cdot \sqrt{\frac{5-2}{5-1}} = 0.3 \quad (\text{since } \frac{n}{N} = \frac{2}{5} > 0.05)$$

$$\sigma_{\bar{p}} = \sqrt{0^2 \cdot 0.3 + 0.5^2 \cdot 0.6 + 1^2 \cdot 0.1 - 0.4^2} = 0.3$$

$$\Rightarrow \sqrt{\frac{P(1-P)}{n}} \cdot \sqrt{\frac{N-n}{N-1}} = \sigma_{\bar{p}}.$$

<u>R. 1. 12.</u> (*Shape of the Sampling Distribution of the Sample Proportion*)

The sampling distribution of the sample proportion, $\bar{p}$, is approximately normally distributed for a sufficiently large sample size.

Rule of thumb 1: The sample size is considered to be sufficiently large if

$$n \cdot P > 5 \ \wedge \ n \cdot (1-P) > 5.$$

Rule of thumb 2:

$$n \geq 30.$$

The $z-$value for a value of $\bar{p}$ is calculated as

$$z = \frac{\bar{p} - \mu_{\bar{p}}}{\sigma_{\bar{p}}}.$$

Ex. 1. 9.
According to a survey, 50% of Americans were in 2005 satisfied with their job. Assume that the result is true for the current proportion of Americans.

1. Find the mean and standard deviation of the proportion for a sample of 1000.
2. Describe the shape of its sampling distribution.

Solution:
1.

$$\mu_{\bar{p}} = P = 0.50; \qquad \sigma_{\bar{p}} = \sqrt{\frac{P(1-P)}{n}} = \sqrt{\frac{0.50 \cdot 0.50}{1000}} \approx 0.0158.$$

2. Because of $n \cdot P = n \cdot (1-P) = 500 > 5$, we can apply the central limit theorem to make an inference about the shape of the sampling distribution of $\bar{p}$. Therefore, the sampling distribution of $\bar{p}$ is approximately normal with the mean of 0.50 and the standard deviation of 0.0158.

Ex. 1. 10.
In a poll conducted during August 16-18, 2004, 38% of adult Americans said that they were very satisfied with the way things were going in their lives at that time.
Suppose this result is true for the current population of adult Americans. Find the probability that in a sample of 1000 the proportion of the adult Americans that are very satisfied lies between 0.40 and 0.42.

Solution:

$$\mu_{\bar{p}} = P = 0.38, \qquad \sigma_{\bar{p}} = \sqrt{\frac{P(1-P)}{n}} = \sqrt{\frac{0.38(1-0.38)}{1000}} = 0.01534927.$$

Because of $n \cdot P = 1000 \cdot 0.38 = 380 > 5, \quad n \cdot (1-P) = 1000 \cdot 0.62 = 620 > 5,$ we can apply the central limit theorem to make an inference about the shape of the sampling distribution of $\bar{p}$.

Therefore, the sampling distribution of $\bar{p}$ is approximately normal with the mean of 0.38 and the standard deviation of 0.01534927.

$$P\left(0.40 \le \bar{p} < 0.42\right) = F\left(0.42\right) - F\left(0.40\right)$$

$$= \Phi\left(\frac{0.42-0.38}{0.01534927}\right) - \Phi\left(\frac{0.40-0.38}{0.01534927}\right) = \Phi\left(2.61\right) - \Phi\left(1.30\right)$$

$$= 0.9955 - 0.9032 = 0.0923.$$

Ex. 1. 11.

A woman, who is running for mayor in a large city, claims that she is favoured by 53% of all eligible voters of that city. Assume that this claim is true.
What is the probability that in a random sample of 400 registered voters taken from this city, less than 49% will favour her?

Solution:

$$\mu_{\bar{p}} = P = 0.53, \qquad \sigma_{\bar{p}} = \sqrt{\frac{P\left(1-P\right)}{n}} = \sqrt{\frac{0.53\left(1-0.53\right)}{400}} = 0.02495496\,.$$

Because of $n \cdot P = 400 \cdot 0.53 = 212 > 5, \quad n \cdot \left(1-P\right) = 400 \cdot 0.47 = 188 > 5,$ we can apply the central limit theorem to make an inference about the shape of the sampling distribution of $\bar{p}$.

Therefore, the sampling distribution of $\bar{p}$ is approximately normal with the mean of 0.38 and the standard deviation of 0.02495496.

$$P\left(\bar{p} < 0.49\right) = F(0.49) = \Phi\left(\frac{0.49-0.53}{0.02495496}\right)$$

$$= \Phi\left(-1.60\right) = 1 - \Phi\left(1.60\right) = 0.0548.$$