Chapter 6 - Part 2

Sampling Distributions

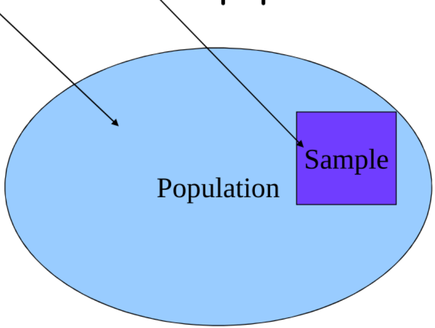
Sampling

Recall that:

- A population is a set of all units (usually people, objects, transactions, or events) that we are interested in studying.
- A sample is a subset of the units of a population.
- A parameter is a numerical descriptive measure of a population. Because it is based on the observations in the population, its value is almost always unknown.
- A sample statistic is a numerical descriptive measure of a sample. It is calculated from the observations in the sample.

Sampling

- A **statistic** can be computed from the sample and used to estimate a **parameter** of the population.



Sampling

	Population Parameter	**Sample Statistic**
Mean	μ	$\bar{x}$
Standard deviation	σ	s
Proportion	p	$\hat{p}$

Sampling

Randomly select one individual from the population.

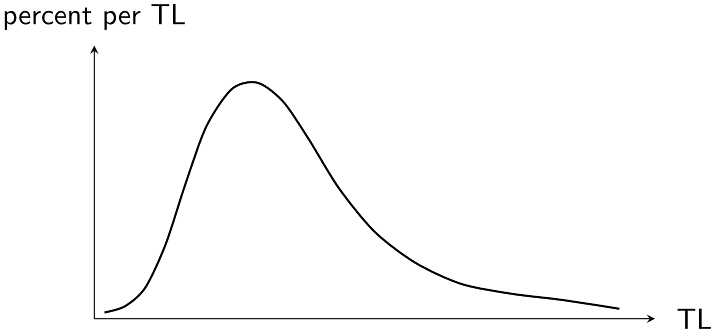
Record the variable of interest, and denote it by X_1.

For example, it may represent height, income, or number of siblings.

Then X_1 is a random variable with the same distribution as the variable in the population.

Example

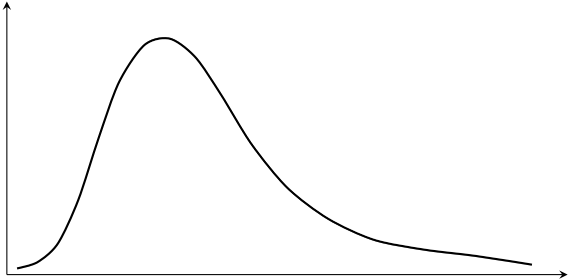
Consider the following histogram for yearly income in the population.



Randomly select one person from this population. Let X_1 be the random variable representing that person's yearly income.

Example (cont'd)

Then the PDF of X_1 has the same graph as this histogram.



Example

Suppose a population consists of 5.1 million women and 4.9 million men.

Randomly select one person from this population.

Define

$$X_1 = \begin{cases} 1, & \text{if the selected person is a woman,} \\ 0, & \text{if the selected person is a man.} \end{cases}$$

Then, the PDF of X_1 is given by

x	0	1
$P(X_1 = x)$	0.49	0.51

Sampling

Continuing the previous discussion, select a second person, then a third, and so on, until n people have been selected at random from the population.

This gives a sample of size n.

Let X_i denote the value of the variable of interest for the ith selected person, for $i = 1, 2, \ldots, n$.

Then $X_1, X_2, \ldots, X_n$ all have the same distribution as the population variable.

Sampling

Sometimes we are interested in the sample sum:

$$\Sigma X = \sum_{i=1}^{n} X_i = X_1 + X_2 + \ldots + X_n.$$

More often, we are interested in the sample mean:

$$\overline{X} = \frac{1}{n}\Sigma X$$

For example, if X_i represents yearly income, then $\overline{X}$ is the average yearly income in the sample.

To apply the Central Limit Theorem, we assume that $X_1, X_2, \ldots, X_n$ are independent.

A Note on Independence

In practice, when we select a sample from a population, the selection is without replacement.

So, strictly speaking, $X_1, X_2, \ldots, X_n$ are not independent.

However, if the sample size is very small compared with the population size (which is usually the case), then they can be treated as independent.

Example

Suppose the population consists of 5.1 million women and 4.9 million men.

Then

$$P(X_1 = 1) = 0.51 \qquad \text{and} \qquad P(X_2 = 1) = 0.51.$$

But

$$P(X_1 = 1,\ X_2 = 1) = \frac{5.1 \text{ million}}{10 \text{ million}} \cdot \frac{5.1 \text{ million} - 1}{10 \text{ million} - 1},$$

while

$$P(X_1 = 1)\, P(X_2 = 1) = \frac{5.1 \text{ million}}{10 \text{ million}} \cdot \frac{5.1 \text{ million}}{10 \text{ million}}.$$

These two quantities are very close, so in practice we treat X_1 and X_2 as independent.

A Note on Independence

The X_i are "approximately independent" if the sample size is very small compared with the population size.

Otherwise, we need to use a "correction factor". (We will return to this at the end of the topic.)

Sampling Distribution

A sampling distribution is the probability distribution of a sample statistic.

It describes how the statistic varies from sample to sample when we repeatedly take random samples of the same size n from a population.

For example, the probability distribution of

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

is the sampling distribution of the sample mean.

Theorem: Sampling Distribution of the Sample Mean for Normal Population

By the result on page 8 of 6.1, we obtain the following theorem.

If a random sample of n observations is selected from a population that has a normal distribution with mean μ and standard deviation σ, then the sampling distribution of $\overline{X}$ is normal with mean μ and standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

In other words, if the population distribution is $N(\mu, \sigma)$, then

$$\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

Example

The exam scores of students in a course follow a normal distribution with mean 48 and standard deviation 16. What is the probability that the mean score of a random sample exceeds 52 if the sample contains

(i) 16 students?

(ii) 64 students?

Solution: Let $\overline{X}$ be the sample mean.
Since the population is normally distributed with mean $\mu = 48$ and standard deviation $\sigma = 16$, the sampling distribution of $\overline{X}$ is normal with

$$E(\overline{X}) = 48 \qquad \text{and} \qquad \sigma_{\bar{X}} = \frac{16}{\sqrt{n}}.$$

We want to find

$$P(\overline{X} > 52).$$

Example (cont'd)

$$\overline{X} \sim N\left(48, \frac{16}{\sqrt{n}}\right)$$

(i) For $n = 16$,

$$\overline{X} \sim N\left(48, \frac{16}{\sqrt{16}}\right) = N\,(48, 4)$$

So

$$P(\overline{X} > 52) = P\left(Z > \frac{52 - 48}{4}\right) = P(Z > 1) = 0.1587.$$

(ii) For $n = 64$,

$$\overline{X} \sim N\left(48, \frac{16}{\sqrt{64}}\right) = N\,(48, 2)$$

So

$$P(\overline{X} > 52) = P\left(Z > \frac{52 - 48}{2}\right) = P(Z > 2) = 0.0228.$$

Therefore, the answers are

Theorem: Sampling Distribution of the Sample Mean for Large Samples

By the Central Limit Theorem (the result on page 11 of 6.1), we obtain the following theorem.

If a random sample of n observations is selected from a population with mean μ and standard deviation σ, and if n is sufficiently large, then the sampling distribution of $\overline{X}$ is approximately normal with mean μ and standard deviation $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$.

In other words, if the population mean is μ and the population standard deviation is σ, then for large enough n,

$$\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \quad \text{approximately.}$$

In our course, 'large enough' means $n \geq 30$.

Example

The average salary in Los Angeles is known to be \$50 000 per year, with standard deviation \$40 000.

What is the probability that a random sample of 100 people will have an average salary under \$45 000?

Solution: Let $\overline{X}$ be the sample mean salary.

Here,

$$\mu = 50\,000, \qquad \sigma = 40\,000, \qquad n = 100.$$

Therefore,

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{40\,000}{\sqrt{100}} = 4\,000.$$

We want to find

$$P(\overline{X} < 45\,000).$$

Example (cont'd)

Since $\overline{X} \sim N(50\,000, 4\,000)$, we have

$$P(\overline{X} < 45\,000) = P\left(Z < \frac{45\,000 - 50\,000}{4\,000}\right) = P(Z < -1.25).$$

Using the standard normal table,

$$P(Z < -1.25) = 0.5 - 0.3944 = 0.1056.$$

Sampling Distribution of the Sample Proportion

To study a population proportion, we code each observation as

$$X_i = \begin{cases} 1, & \text{if the } i\text{th observation has the characteristic of interest,} \\ 0, & \text{otherwise.} \end{cases}$$

Then the population mean is the population proportion: $\mu = p$.

And the sample mean is the sample proportion: $\overline{X} = \hat{p}$.

Example: If a population has 2 million smokers and 8 million non-smokers, then proportion of smokers is $p = 0.20$. If we code smokers by 1 and non-smokers by 0, then this is also the mean of

$$\underbrace{0, 0, \ldots, 0}_{\text{8 million}}, \underbrace{1, 1, \ldots, 1}_{\text{2 million}}$$

Theorem: Sampling Distribution of the Sample Proportion for Large Samples

By the result on page 18 of 6.1, we obtain the following theorem.

If a random sample of n observations is selected from a population with proportion p, and if np and $n(1-p)$ are sufficiently large, then the sampling distribution of $\hat{p}$ is approximately normal with mean p and standard deviation $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$.

In other words, if the population proportion is p, then for large enough np and $n(1-p)$,

$$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right) \quad \text{approximately.}$$

In our course, 'large enough' means $np \geq 15$ and $n(1-p) \geq 15$.

Example

In Dublin, 20% of people earn more than €50 000 a year.

In a random sample of 400 people, what is the probability that between 18% and 23% of the people in the sample earn more than €50 000 a year?

Solution: Let $\hat{p}$ be the sample proportion of people who earn more than €50 000 a year.

Here,

$$p = 0.20, \qquad n = 400.$$

Therefore,

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.2 \cdot 0.8}{400}} = 0.02$$

and

$$\hat{p} \sim N(0.2, 0.02).$$

We want to find $P(0.18 < \hat{p} < 0.23)$.

Example (cont'd)

Since

$$\hat{p} \sim N(0.2, 0.02).$$

we have

$$P(0.18 < \hat{p} < 0.23) = P\left(\frac{0.18 - 0.20}{0.02} < Z < \frac{0.23 - 0.20}{0.02}\right) = P(-1 < Z < 1.5).$$

Using the standard normal table,

$$P(-1 < z < 1.5) = 0.3413 + 0.4332 = 0.7745.$$

Standard Error

The standard deviation of the sampling distribution

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad \left(\text{or } \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \text{ for proportions} \right)$$

is called the standard error.

As the sample size n increases, the standard error gets smaller.

For $n = 1$, the standard error equals the population standard deviation.

As n becomes very large, the standard error gets very close to 0.

This makes sense because if we could observe the entire population, there would be no sampling variability.

Standard Error vs Standard Deviation

Standard deviation describes the spread of individual observations in the population.

Standard error describes the spread of a sample statistic from sample to sample.

In short:

- Standard deviation = spread of the data
- Standard error = spread of the statistic

A larger sample size gives a smaller standard error.

Example

Suppose that adult heights in a population are normally distributed with mean 175 cm and standard deviation 10 cm.

So for one randomly selected person,

$$X \sim N(175, 10).$$

Now take a random sample of size 100 and let $\overline{X}$ be the sample mean height. Then

$$\overline{X} \sim N\left(175, \frac{10}{\sqrt{100}}\right) = N(175, 1).$$

So:

- the **standard deviation** is 10 cm,
- the **standard error** is 1 cm.

Individual heights vary by about 10 cm, but sample means vary by only about 1 cm.

Correction Factor

$$\begin{matrix}\text{standard error when} \\ \text{sampling without replacement}\end{matrix} = \begin{matrix}\text{correction} \\ \text{factor}\end{matrix} \times \begin{matrix}\text{standard error when} \\ \text{sampling with replacement}\end{matrix}$$

Here

$$\begin{matrix}\text{correction} \\ \text{factor}\end{matrix} = \sqrt{\frac{N-n}{N-1}}$$

where n is the sample size and N is the population size.

N	n	correction factor
5 000	2 500	0.70718
10 000	2 500	0.86607
100 000	2 500	0.98743
500 000	2 500	0.99750
1 500 000	2 500	0.99917
15 000 000	2 500	0.99992

Correction Factor

$$\begin{matrix}\text{correction} \\ \text{factor}\end{matrix} = \sqrt{\frac{N-n}{N-1}}$$

where n is the sample size and N is the population size.

- As N becomes large compared with n, the correction factor gets very close to 1.
- Therefore, in most sampling situations, this correction factor is close to 1 and can be ignored.
- In this course, we will take the correction factor to be 1 and ignore it, unless stated otherwise.