Chapter 4: Statistics II
New Section 4.7A: Sampling

Editor: Priscilla O'Connor

Designer: Liz White

Layout: Compuscript

Illustrations: Compuscript

ISBN: 978-1-78090-250-0
1509

Folens Publishers, Hibernian Industrial Estate, Greenhills Road, Tallaght, Dublin 24, Ireland

Contents

Following clarifications to the Leaving Certificate Higher Level syllabus, a new section on Sampling has been added to Chapter 4.

This booklet contains the new material, including a new Exercise section and extra revision questions.

Answers for all new questions are also included.

4.7A SAMPLING

Sample data must be collected in an appropriate way, such as through the process of random selection.

A **census** is the collection of data from every member of the population.

A **sample** is a subset of members selected from the population.

In a particular school, 85% of the students play sport. The figure of 85% is a **parameter**, as it is based on the entire population of the school.

A **parameter** is a numerical measurement describing some characteristic of a population.

Based on a sample of 400 First Year college students, it is found that 45% of them achieved more than 400 points in their Leaving Certificate. The figure of 45% is a **statistic** because it is based on a sample, not on the entire population of all First Year college students.

A **statistic** is a numerical measurement describing some characteristic of a sample.

When we select a sample from a population and the sample is representative of the population, then we can make inferences about the population from the sample.

The Sampling Distribution of a Statistic

The **sampling distribution** of a statistic is the distribution of all values of that statistic when all possible samples of the same size are taken from the same population.

OR

The **sampling distribution** of a sample statistic calculated from a sample of *n* measurements is the probability distribution of the statistic.

Worked Example 4.19A

A census was carried out and the following measurements were recorded for the population:

1, 1, 2, 2, 2, 3, 3

(i) Show the distribution in a frequency table.

(ii) Is the mean of the data a parameter or a statistic? Explain.

(iii) What is the mean of the data?

(iv) How many different possible samples of size 6 can be taken from the population?

(v) List all the possible samples of size 6.

(vi) Find the mean of each sample listed in part (v).

(vii) Is the mean of each sample an example of a parameter or a statistic? Explain.

(viii) Show the sampling distribution of the sample mean in a frequency table.

(ix) Find the mean of the sampling distribution. What do you notice?

Solution

(i)

Measurement	1	2	3
Frequency	2	3	2

(ii) The mean of the data is a parameter, as it is a measurement describing a characteristic of the population.

(iii) $\mu = \dfrac{1(2) + 2(3) + 3(2)}{2 + 3 + 2}$

$= \dfrac{14}{7}$

$\therefore \mu = 2$

μ = population mean
σ = population standard deviation

(iv) $\binom{7}{6} = 7$

(v) 1, 1, 2, 2, 2, 3

1, 1, 2, 2, 2, 3

1, 1, 2, 2, 3, 3

1, 1, 2, 2, 3, 3

1, 1, 2, 2, 3, 3

1, 2, 2, 2, 3, 3

1, 2, 2, 2, 3, 3

(vi)

Sample number	Sample	Mean
1	1, 1, 2, 2, 2, 3	$1\frac{5}{6}$
2	1, 1, 2, 2, 2, 3	$1\frac{5}{6}$
3	1, 1, 2, 2, 3, 3	2
4	1, 1, 2, 2, 3, 3	2
5	1, 1, 2, 2, 3, 3	2
6	1, 2, 2, 2, 3, 3	$2\frac{1}{6}$
7	1, 2, 2, 2, 3, 3	$2\frac{1}{6}$

(vii) The mean of each sample is an example of a statistic, as it is a measurement describing a characteristic of a sample.

(viii)

Sample mean	$1\frac{5}{6}$	2	$2\frac{1}{6}$
Frequency	2	3	2

(ix) $\bar{x} = \dfrac{1\frac{5}{6}(2) + 2(3) + 2\frac{1}{6}(2)}{2 + 3 + 2}$

$= \dfrac{14}{7}$

$\therefore \bar{x} = 2$

$\bar{x}$ = sample mean
s = sample standard deviation

The mean of the sampling distribution is equal to the mean of the population ($\bar{x} = \mu$).

The Central Limit Theorem

When selecting a simple random sample of size n from a population with mean μ and standard deviation σ, then:

(a) If $n > 30$, the sample means for a given sample size have a distribution that can be approximated by a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. (This guideline is commonly used regardless of the distribution of the underlying population; in other words, the population measurements may or may not be normal.)

(b) If $n \leqslant 30$ and the underlying population is normal, then the sample means for a given sample size have a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

The standard deviation of the sampling distribution of the sample means is often referred to as the standard error of the mean.

In the case of (a) and/or (b),
the corresponding z-score for the sample mean is:

FORMULA

$$z = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$$

Formulae and Tables, pages 34–5

Central Limit Theorem Summary

Mean: $\mu_{\bar{x}} = \mu$

Standard deviation: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Worked Example 4.20A

A car battery manufacturer claims that the distribution of the lengths of life of its premium battery has a mean of 54 months and a standard deviation of 6 months. A consumer magazine tests a random sample of 50 batteries to investigate the company's claim.

(i) Describe the sampling distribution of the mean lifetime of a sample of 50 batteries, assuming the company's claim is true.

(ii) What theorem have you used in answering part (a)?

(iii) Sketch the sampling distribution described in part (a).

(iv) Assuming that the company's claim is true, what is the probability that the magazine's sample has a mean life of 52 months or less?

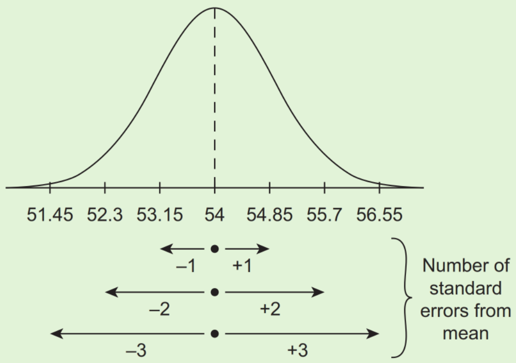
(v) If the magazine's sample had a mean life of 52 months and you were the journalist at the magazine reporting on the company's claim, what would you conclude regarding the claim?

Solution

(i) Approximately normal with mean = 54 months and standard deviation of $\frac{6}{\sqrt{50}} \approx 0.85$ months.

(ii) The Central Limit Theorem, which states that if $n > 30$, then the sampling distribution of the sample mean (sample size n) will be approximately normal with mean μ (population mean) and standard deviation of $\frac{\sigma}{\sqrt{n}}$ (where σ is the population standard deviation).

(iii)



(iv) $z = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$

$= \frac{52 - 54}{0.85}$

≈ -2.35

$P(z \leqslant -2.35) = 1 - P(z < 2.35)$

$= 1 - 0.9906$

$= 0.0094$

$= 0.94\%$

(v) I would conclude that the investigation provides evidence that the company's claim is **not** true. If the company's claim were true, there would be a less than 1% chance of a random sample of size 50 having a sample mean of 52 or less.

OR

As $|z| > 2$, we dismiss the company's claims. A z-score of –2.35 seems to suggest that the true mean is lower than what the company is claiming.

Worked Example 4.21A

In a computer simulation, random samples of size 400 are selected from a population. The mean measurement of each sample is recorded. Five thousand such sample means are recorded.

Describe the expected distribution of the sample means.

Solution

The sample means will have an approximately normal distribution. The mean of the distribution will approximate the mean of the population. The Central Limit Theorem tells us that the mean of the sampling distribution is the mean of the population.

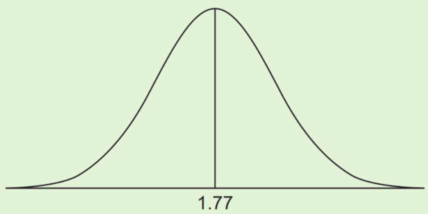
The standard deviation of the distribution will approximate $\frac{\sigma}{\sqrt{400}}$, where σ is the standard deviation of the population.

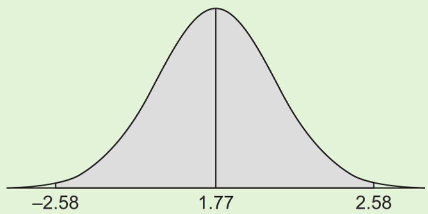
Worked Example 4.22A

A simple random sample of 400 is selected from a population having a mean height of 1.77 m and a standard deviation of 0.0775 m.

Find the probability that the sample mean $\bar{x}$ lies in the range 1.76 m $\leqslant \bar{x} \leqslant$ 1.78 m.

Solution





Since $n > 30$, the sampling distribution of the means is approximately normal, with a mean $\mu_{\bar{x}} = \mu = 1.77$ and a standard deviation of $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.0775}{20} = 0.003875$.

We will now convert 1.76 and 1.78 to standard z-scores using the formula $z = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$.

$$z_1 = \frac{1.76 - 1.77}{0.003875} \approx -2.58$$

$$z_2 = \frac{1.78 - 1.77}{0.003875} \approx 2.58$$

Therefore, the probability that the sample mean $\bar{x}$ lies in the range 1.76 m $\leqslant x \leqslant$ 1.78 m is given by:

$$\begin{aligned} P(-2.58 \leqslant z \leqslant 2.58) &= 2P(z \leqslant 2.58) - 1 \\ &= 2(0.9951) - 1 \\ &= 0.9902 \end{aligned}$$

There is a probability of approximately 99.02% that the sample mean lies in the range 1.76 m $\leqslant \bar{x} \leqslant$ 1.78 m.

Worked Example 4.23A

A population is normally distributed with a mean of 11 and a standard deviation of 3.
Find the sample size such that $P(\bar{x} > 11.5) = 0.05$, where $\bar{x}$ is the sample mean.

Solution

$P(z > z_1) = 0.05$

$P(z \leqslant z_1) = 0.95$

$\therefore z_1 = 1.645$

$$1.645 = \frac{11.5 - 11}{\frac{3}{\sqrt{n}}}$$

$$\frac{0.5\sqrt{n}}{3} = 1.645$$

$$\sqrt{n} = 9.87$$

$$n = 97.4169$$

Sample size = 98 (Round up)

Exercise 4.8A

1. Numerical data for a large population is stored on a database. The data has a mean of μ and a standard deviation of σ. A computer simulation randomly selects samples of size 400 from the database and calculates the mean of each sample. This is done repeatedly until 1,200 such means are recorded.

(i) Describe the expected distribution of the sample means.

(ii) Write in terms of μ and σ the mean and standard deviation of the sampling distribution of the means.

2. A simple random sample of size 49 is chosen from a population that is known to be normal. The mean of the population is 9 with a standard deviation of 2.

Find the probability that the sample mean is greater than 10.

3. Women's heights in Ireland are normally distributed with a mean given by $\mu = 164.4$ cm and a standard deviation given by $\sigma = 6.25$ cm (Source: Jaume Garcia and Climent Quintana-Domeque, 'The evolution of adult height in Europe', *Economics and Human Biology* 5(2) (2007), pp 340–349).

(i) If a woman is selected at random, find the probability that her height is less than 167 cm.

(ii) A simple random sample of 25 women are selected. Find the probability that the mean height of the sample is less than 167 cm.

(iii) Explain why you were able to use the Central Limit Theorem in part (ii), even though the sample size does not exceed 30.

4. Men's heights in Ireland are normally distributed with a mean given by $\mu = 177.5$ cm and a standard deviation given by $\sigma = 6.3$ cm (Source: Garcia and Quintana–Domeque, 'The evolution of adult height in Europe').

(i) If a man is selected at random, find the probability that his height is between 173 cm and 182 cm.

(ii) A simple random sample of 25 men is conducted. Find the probability that they have a mean height of between 173 cm and 182 cm.

5. A ski lift has a maximum capacity of 12 people or 910 kg. The capacity will be exceeded if 12 people have weights with a mean greater than $\frac{910 \text{ kg}}{12} = 75.8$ kg. As men tend to weigh more than women, a 'worst case' scenario involves 12 passengers who are all men. Men's weights are normally distributed with a mean of 78.1 kg and a standard deviation of 13.2 kg.

(i) Find the probability that if an individual man is selected at random, his weight will be greater than 75.8 kg.

(ii) Find the probability that 12 randomly selected men will have a mean weight greater than 75.8 kg.

(iii) Does the ski lift appear to have the correct weight limit?

6. The manager of a hotel finds that guests spend a mean of 12.5 minutes each day in the shower. Assume shower times are normally distributed with a standard deviation of 2.8 minutes.

(i) Find the percentage of guests who shower for more than 13 minutes.

(ii) The hotel has installed a hot water system that can provide enough hot water, provided that the mean shower time for 100 guests is less than 13 minutes. Find the probability that there will not be enough hot water on a morning that the hotel has 100 guests.

7. A normal distribution has a mean of 80 and a standard deviation of 8. A sample of size n is selected at random and the mean of the sample is $\bar{x}$.

Find n if the $P(\bar{x} > 82) = 0.1977$.

8. A sample is chosen at random from a population that is strongly skewed to the left.

(i) Describe the shape of the sampling distribution of the means if the sample size is large.

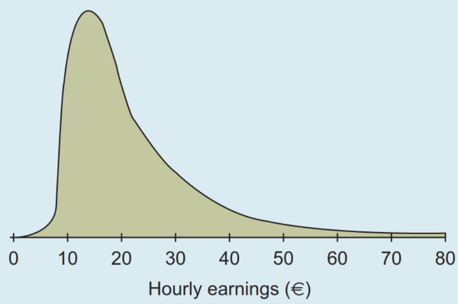
(ii) Describe the mean and standard deviation of the sampling distribution of the means in terms of the underlying population.

9. Records indicate that the value of homes in a large town is positively skewed with a mean of 150,000 and standard deviation of 65,000. To check the accuracy of the records, officials plan to conduct a survey of 100 randomly selected homes.

Draw and label an appropriate sampling model for the mean value of the homes selected.

10. The distribution of the hourly earnings of all employees in Ireland in October 2009 is shown in the diagram. It can be seen that the distribution is positively skewed.

- The mean is €22.05.
- The median is €17.82.
- The standard deviation is €10.64.
- The lower quartile is €12.80.
- The upper quartile is €26.05.



(Source: Adapted from: CSO. *National Employment Survey 2008 and 2009*)

(i) If six employees are selected at random from this population, what is the probability that exactly four of them had hourly earnings of more than €12.80?

In a computer simulation, random samples of size 200 are repeatedly selected from this population and the mean of each sample is recorded. A thousand such sample means are recorded.

(ii) Describe the expected distribution of these sample means. Your description should refer to the shape of the distribution and to its mean and standard deviation.

(iii) How many of the sample means would you expect to be greater than €23?

SEC Project Maths Sample Paper 2, Leaving Certificate Higher Level, 2012

STATISTICS II

11. Perfluoro-octanoic acid (PFOA) is a chemical used in Teflon-coated cookware. The US Environmental Protection Agency (EPA) is investigating if PFOA causes cancer. It is known that the blood concentration of PFOA in the general population has a mean of 6 ppb (parts per billion) and a standard deviation of 10 ppb. Tests for PFOA exposure were conducted on a sample of 326 people who live near DuPont's Teflon-making plant in West Virginia.

(i) What is the probability that the average blood concentration of PFOA in the sample is greater than 7.5 ppb?

(ii) What assumption did you make in answering part (i)?

(iii) If the actual study resulted in a sample mean of 7 ppb, what inference would you make about the true mean PFOA concentration for the population that lives near the DuPont Teflon facility?

12. On 1 May 1994, the triple World Champion, Ayrton Senna, was killed at the San Marino F1 Grand Prix at Imola, Italy, following a mechanical failure in his car. Later research revealed that the time x (in hours) to the first mechanical failure in an F1 Grand Prix is distributed with mean of 0.1 and standard deviation of 0.1; i.e. $x \sim (0.1, 0.01)$.

A random sample of 40 F1 Grand Prix is selected by a computer.

(i) Describe the distribution of the sample mean (sample size of 40).

(ii) Sketch this distribution.

(iii) Find the probability that the sample mean time for the first mechanical failure exceeds 0.13 hours.

(iv) A computer generates 1,500 random samples each of 40 F1 Grand Prix. How many of these 1,500 random samples would you expect to have a sample mean exceeding 0.13 hours?

(v) Describe the expected distribution of the 1,500 sample means referred to in part (iv).

Revision Exercises

11. An aircraft strobe light is designed so that the time between flashes is normally distributed, with a mean of 3 seconds and a standard deviation of 0.4 seconds.

(a) Find the probability that an individual time is greater than 4 seconds.

(b) Find the probability that the mean for 60 randomly selected times is greater than 3.10 seconds.

12. According to a National Business Travel Association 2008 survey, the average salary of a travel management professional is \$97,300. Assume that the standard deviation of such salaries is \$30,000. Consider a random sample of 50 travel management professionals.

(a) What is the mean of the sampling distribution of the sample mean with sample size 50?

(b) What is the standard deviation of the same sampling distribution?

(c) Describe the shape of this sampling distribution.

(d) Sketch this sampling distribution.

(e) Find the z-score for a value of \$89,500 for the sample mean.

(f) What is the probability that a random sample of size 50 will have a sample mean of less than \$89,500?

Answers

Exercise 4.8A

1. (i) Approximately normal distribution (ii) Mean $= \mu$; SD $= \frac{\sigma}{20}$ **2.** 0.0011 **3.** (i) 0.6628 (ii) 0.9812 (iii) The underlying population is normally distributed. **4.** (i) 0.5222 (ii) ≈ 1 **5.** (i) 0.5675 (ii) 0.7257 (iii) No **6.** (i) 42.86% (ii) 0.0367 **7.** $n = 12$ **8.** (i) Approximately normal distribution (ii) Mean $= \mu$; SD $= \frac{\sigma}{\sqrt{n}}$ **10.** (i) $\approx$ 29.66% (ii) Approximately normal distribution (iii) ≈ 103 or ≈ 104 **11.** (i) 0.34% (ii) Assumption: Population from which sample was chosen has mean of 6 ppb and SD of 10 ppb. **12.** (i) $\approx$ 0.0158 hours (iii) 2.87% (iv) 43 (v) Expected distribution would be very similar to actual sample distribution of sample means found in part (i)

Revision Exercises

11. (a) 0.0062 (b) 0.0262 **12.** (a) \$97,300 (b) \$4,242.64 (c) Approximately normal distribution (e) –1.838 (f) 3.29%

Notes