UNIT 1.4 SAMPLING DISTRIBUTION OF THE MEAN

Prof. Ryung Kim
ryung.kim@einstein.yu.edu

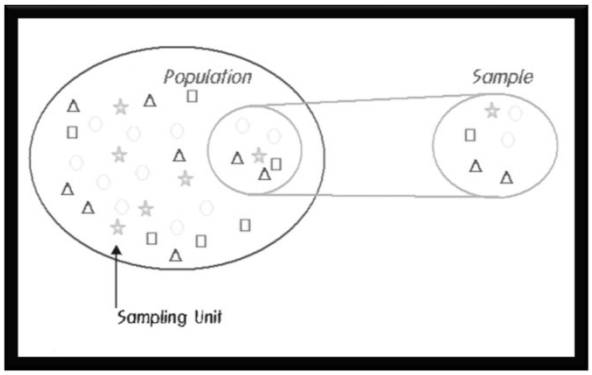
UNIT 1.4 (PnG p.4)

- *This unit investigates the properties of the sample mean (i.e. average) when repeated samples are drawn from a population, thus introducing an important concept known as the central limit theorem. This theorem provides a foundation for quantifying the uncertainty associated with the inferences being made.*

1. SAMPLING DISTRIBUTIONS

Prof. Ryung Kim
ryung.kim@einstein.yu.edu

What is Statistics?



Population and Sample

• Population: the complete collection of all elements to be studied. All subjects to be studied are included.

• Sample: a subcollection of members selected from a population.

Often, it is impossible to investigate characteristics of whole population. So we investigate a part (sample) of population and make inference about population using the sample characteristics.

Image by Lee J, SNU

Example

5

- To investigate the mean daily calorie intake among adults in Worcester, MA, an agency called randomly selected 2000 adults in Worcester who are older than 20 for phone interview.
  - Population
  - Sample

Population and Sample

6

- A parameter is a number that describes the population.
  - e.g. Population mean
  - e.g. True mean daily calorie intake of all adults in Worcester
- A statistic is a number that describes a sample.
  - e.g. Sample mean
  - e.g. Average daily calorie intake of the sample of 2,000 people

Sampling variability and Sampling distribution

7

- The value of a statistic, such as the sample mean $\overline{x}$, depends on the particular values included in the sample, and varies from sample to sample. This variability of a statistic is called *sampling variability*.

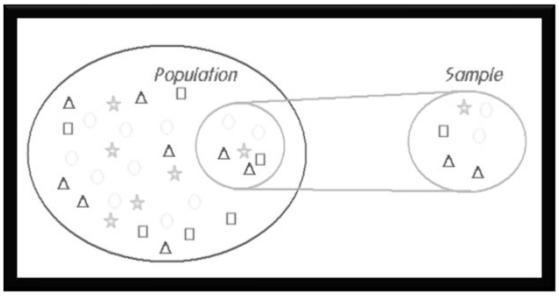


Image by Prof. Lee J, SNU

Sampling variability and Sampling distribution

8

- If we repeatedly choose samples from a population, value of sample statistic will vary.
- The *sampling distribution* of the distribution of sample statistic value (e.g. sample mean or sample proportion) of all possible samples of the same size *n* taken from the same population.
- In other words, sampling distribution is a table describing the sampling variability.

Sampling distribution

9

- **This is one of the most important concept in this course.**

Difference sources of uncertainty:
Variability often called 'Error'

- Sampling error (non-systemic error; random error)
  - Variability due to sampling
  - We can quantify the uncertainty.
- Systematic error
  - Measurement (instrumentation) error or observer Error
  - We cannot quantify the uncertainty.

2. THE CENTRAL LIMIT THEOREM

Prof. Ryung Kim
ryung.kim@einstein.yu.edu

Distribution of a sample mean

If $X_1, X_2, \ldots, X_n$ is an independent random sample with size *n* from a common distribution:

- The mean of the sample:

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

- The standard deviation of sample mean (often called the *standard error*):

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

Distribution of a sample mean

13

1. If the original population is itself normally distributed, then the sample means will be normally distributed for any sample size (*n*).

2. For samples of size (*n*) larger than 30, the distribution of the sample means can be approximated reasonably well by a normal distribution. The approximation gets better as the sample size *n* becomes larger. ("Central Limit Theorem")

Formal Central Limit Theorem

14

Let $X_1, X_2, \ldots, X_n$ be an <u>independent</u> random sample of size *n* from a distribution (which may or may not be normal) with mean μ and standard deviation σ.

1. If sample size (*n*) is large enough, the distribution of sample mean $\bar{X}$ is approximately a normal distribution.

2. And the mean and standard deviation of the sample means are μ and $\sigma/\sqrt{n}$.

3. APPLICATIONS OF THE DISTRIBUTION OF SAMPLE MEANS

Prof. Ryung Kim
ryung.kim@einstein.yu.edu

Example – serum cholesterol levels

- Serum cholesterol levels for all 20~74 years old males in the U.S. have the mean 211 mg/100ml and the standard deviation 46 mg/100ml. (Population distribution is approximately normal.)
- If we select repeated samples of size 25 from the population, what proportion of the samples will have a mean value of 230 mg/ml or above?

  (Distribution of sample mean)

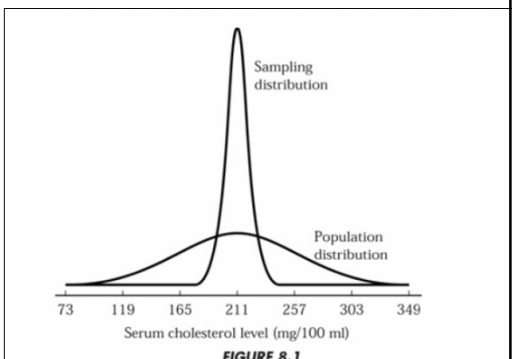


FIGURE 8.1
Distributions of individual values and means of samples of size 25 for the serum cholesterol levels of 20- to 74-year-old males, United States, 1976–1980

PnG

Example – cont.

17

- The sample mean approximately follows normal distribution with mean 230 mg/ml and standard deviation $46/\sqrt{25}$ (mg/ml).

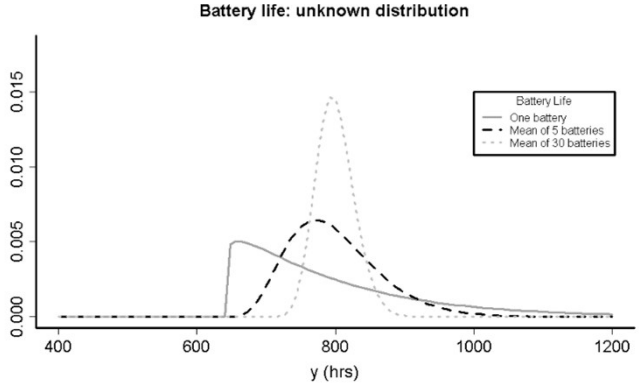
$$z = \frac{230-211}{\frac{46}{\sqrt{25}}} = 2.065$$

P(Z > 2.07)=0.0195

Example – Life of batteries

18

- The life of a certain brand battery has mean 800 hours and a standard deviation of 150 hours. The distribution of the battery life is known to be skewed (hence, not normal). What is the probability that the mean life time of 30 batteries is less than 700 hours? (→ Central Limit Theorem)



$$P(Z < \frac{700-800}{150/\sqrt{30}})$$
$$= P(Z < -3.651)$$
$$= 0.0339$$

'Very unlikely'

- Reference
  - Principles of Biostatistics (Pagano and Gauvreau)
  - *Elementary Statistics* by Triola, 10th edition.
  - *Applied Statistics for Engineers and Scientists* by Petruccelli et al.
- Acknowledgements
  - Prof. Jayson Wilbur, WPI
  - Prof. Balgobin Nandram, WPI
  - Prof. Lee Jaeyong, Seoul National University
  - Some slides provided by Pearson Education, Inc Publishing as Pearson Addison-Weslely