Dec 1, 2022 (updated on Dec 4, 2024)

Engineering Statistics

Week 6: Sampling distributions

Introduction

The concepts of statistical inference

Population

A population is the complete set of all items that interest an investigator.

Sample

A sample is an observed subset of a population.

The concepts of statistical inference

Sample

A sample is an observed subset of a population.

The concepts of statistical inference

Parameter: A numerical feature of a population is called a parameter.

- The true value of a population parameter is an unknown constant.
- It can be correctly determined only by a complete study of the population.
- However, studying on a population is not practical because of time and cost restrictions.
- Therefore, a **proper sample** from the population should ve drawn and inferences about a parameter should be based on this sample.

The concepts of sampling distribution

Sampling from a population

The observations $X_1, X_2, \ldots, X_n$ are a **random sample** of size n from the population distribution if they result from **independent selections and each observation has the same distribution as the population**.

Sampling from a population

Therefore, a random sample is a collection of independent and identically distributed (iid) random variables.

Sampling from a population

A statistic is a numerical valued function of the sample observations.

Sampling from a population

Let $X_1, X_2, \ldots, X_n$ be a sample from a population having a parameter θ. Any function of the sample is called **statistic or point estimator*** of θ, i.e.,

$$\hat{\theta} = T(X_1, X_2, \ldots, X_n)$$

Sampling from a population

- Statistic or point estimator are used in same manner.
- A statistic or point estimator is also a random variable.
- A value that the point estimator takes called as point estimate and denoted by lower case letters, i.e. while T is point estimator of the parameter θ, t is point estimate.

Example: Normal distribution case

Let $X_1, X_2, \ldots, X_n$ be a sample from $N(\mu, \sigma^2)$ distribution,

- The statistic $\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is a point estimator of population mean μ.
- The statistic $\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is a point estimator of population variance σ^2.

Example: Binomial distribution case

Let $X_1, X_2, \ldots, X_n$ be a sample from $Binom(n, p)$ distribution. The statistic,

$$\hat{p} = \frac{\text{number of success in the sample}}{n}$$

is a point estimator of population proportion p.

List of population parameters and their point estimators

	Population parameter	Point estimator
Mean	μ	$\bar{X}$
Variance	σ^2	s^2
Standard deviation	σ	s
Proportion	p	$\hat{p}$

Sampling distribution

- Since any statistic (point estimator) is a random variable, it has its own probability distribution.
- The distribution of a statistic is called its **sampling distribution**.
- The sampling distribution of a statistic is determined from the distribution of the population, and it also depends on the sample size n.

Sampling distribution

- Populations for various statistical studies are modeled as random variables whose probability distributions have a mean and variance, which are generally not known as we conduct our statistical sampling and analysis.

- We will select a sample of observations-realizations of a random variable from our population and compute **sample statistics** that will be used to obtain inferences about the population, such as the population mean and variance.

- To make inferences we need to know the **sampling distribution** of the observations and the computed sample statistics.

Example

In a single toss of a fair coin, let x equal the number of heads observed. Now consider a sample of $n = 2$ tosses. Find the sampling distribution of $\bar{x}$, the sample mean.

Example

Outcome	Probability	$\bar{x}$
HH	1/4	1
HT	1/4	0.5
TH	1/4	0.5
TT	1/4	0

Take H = 1 and T = 0

Example

$\bar{x}$	0	0.5	1
$p(\bar{x})$	0.25	0.50	0.25

Example

During the production process of a metal alloy, each sample is measured using a hardness test. This test classifies whether the metal alloy meets the standards as pass or fail. A pass is represented by a sample meeting the required hardness value (> 200 Vickers), while a fail indicates that the sample's hardness is below the standard. In a single inspection of a metal alloy, let x equal the number of pass observed. Now consider a sample of $n = 3$ inspections.

Find the sampling distribution of $\bar{x}$, the sample mean.

Example

Outcome	Probability	$\bar{x}$
PPP	0.8 x 0.8 x 0.8 = 0.512	1
PPF	0.8 x 0.8 x 0.2 = 0.128	2 / 3
PFF	0.8 x 0.2 x 0.2 = 0.032	1 / 3
FFF	0.2 x 0.2 x 0.2 = 0.008	0
PFP	0.8 x 0.2 x 0.8 = 0.128	2 / 3
FPF	0.2 x 0.8 x 0.2 = 0.032	1 / 3
FFP	0.2 x 0.2 x 0.8 = 0.032	1 / 3
FPP	0.2 x 0.8 x 0.8 = 0.128	2 / 3

Take Pass = 1 and Fail = 0

Example

$\bar{x}$	0	1 / 3	2 / 3	1
$p(\bar{x})$	0,008	0,096	0,384	0,512

Properties of sampling distributions

Unbiasedness

If the expected value of a sample statistic is equal to the population parameter, the statistic is said to be an unbiased estimator of the parameter, i.e.

$$E(T) = \theta \to T \text{ is unbiased estimator of } \theta$$

If the expected value of the sampling distribution is not equal to the parameter, the statistic is said to be a biased estimator of the parameter.

Properties of sampling distributions

Standard error

The standard deviation of a statistic is called standard error (SE) of the statistic.

$$SE(T) = \sqrt{Var(T)}$$

Standard error in practice

- **Standard error** provides crucial information about the reliability of the sample mean.
- Standard error shows how reliable the sample mean is in representing the population mean. A low standard error indicates that the sample mean is close to the true population mean, while a high standard error suggests more variability in the sample mean.
- A low standard error indicates that the sample mean is more reliable and that products are close to the desired quality in production processes.
- A high standard error suggests more variability in the production process, and improvements or better control may be necessary.
- In quality control processes, standard error and confidence intervals are essential to understanding variability and error margins in production.

Example

In a single toss of a fair coin, let x equal the number of heads observed. Now consider a sample of $n = 2$ tosses. **Find the standard errors of the sample mean.**

Example

In a single toss of a fair coin, let x equal the number of heads observed. Now consider a sample of $n = 2$ tosses. **Find the standard errors of the sample mean.**

$$\mu = n \times p = 2 \times 0.5 = 1$$

$$\sigma^2 = n \times p \times (1 - p) = 2 \times (0.5) \times (1 - 0.5) = 0.5$$

$$SE = \frac{\sigma}{\sqrt(n)} = \frac{\sqrt(0.5)}{\sqrt(2)} \sim 1.58$$

Example

During the production process of a metal alloy, each sample is measured using a hardness test. This test classifies whether the metal alloy meets the standards as pass or fail. A pass is represented by a sample meeting the required hardness value (> 200 Vickers), while a fail indicates that the sample's hardness is below the standard. In a single inspection of a metal alloy, let x equal the number of pass observed. Now consider a sample of $n = 3$ inspections.

Find the standard errors of the sample mean.

Example

Find the standard errors of the sample mean.

$$\mu = n \times p = 3 \times 0.8 = 2.4$$

$$\sigma^2 = n \times p \times (1 - p) = 3 \times (0.8) \times (1 - 0.8) = 0.48$$

$$SE = \frac{\sigma}{\sqrt(n)} = \frac{\sqrt(0.48)}{\sqrt(3)} \sim 0.4$$

Sampling distribution of $\bar{X}$

Sampling from a normal distribution

Let the random variables $X_1, X_2, \ldots, X_n$ denote a random sample from a normal distribution with mean μ and variance σ^2. Then, sampling distribution of the $\bar{X}$ is normal with mean μ and variance $\frac{\sigma^2}{n}$, i.e.

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

$$E(\bar{X}) = \mu, \quad Var(\bar{X}) = \frac{\sigma^2}{n}, \quad SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

Sampling from a normal distribution

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \rightarrow Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Example

The weight of a pepperoni and cheese pizza from a local provider is a random variable whose distribution is normal with mean 16 ounces and standard deviation 1 ounce. You intend to purchase four pepperoni and cheese pizzas. What is the probability that:

(a) The average weight of the four pizzas will be greater than 17.1 ounces?

(b) The total weight of the four pizzas will not exceed 61 ounces?

Example

(a) The average weight of the four pizzas will be greater than 17.1 ounces?

$$P(\bar{X} > 17.1) = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > \frac{17.1 - 16}{1/\sqrt{4}}\right) = P(Z > 2.2)$$

Example

(b) The total weight of the four pizzas will not exceed 61 ounces?

$$P(X_1 + X_2 + X_3 + X_4 \leq 61) = P(\bar{X} \leq 15.25)$$

$$P(\bar{X} \leq 15.25) = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq \frac{15.25 - 16}{1/\sqrt{4}}\right) = P(Z \leq -1.5)$$

Example

The weight of a metal alloy component follows a normal distribution. The mean weight is 21.5 kg, and the standard deviation is 0.5 kg. For a random sample of ten components, answer the following questions:

(a) What is the probability that the average weight of the ten components exceeds 20 kg?

(b) What is the probability that the total weight of the ten components does not exceed 220 kg?

Example

X: weight of a metal alloy component

$\bar{X}$: mean weight of the metal alloy components

$$\bar{X} \sim N(21.5, \frac{0.5^2}{10})$$

Example

(a) What is the probability that the average weight of the ten components exceeds 21 kg?

$$P(\bar{X} > 21) = P\left(\frac{\bar{X} - \mu}{SE} > \frac{21 - 21.5}{0.5/\sqrt{10}}\right) = P(Z > -3.16) = 0.9992$$

Example

(b) What is the probability that the total weight of the ten components does not exceed 216 kg?

$$P(X_1 + X_2 + \ldots + X_{10} \leq 216) = P(\bar{X} \leq 21.6)$$

$$P(\bar{X} \leq 21.6) = P\left(\frac{\bar{X} - \mu}{SE} \leq \frac{21.6 - 21.5}{1/\sqrt{10}}\right) = P(Z \leq 0.6235) \sim 0.737$$

Example

A manufacturer of automobile batteries claims that the distribution of the lengths of life of its best battery has a mean of 54 months and a standard deviation of 6 months. Suppose a consumer group decides to check the claim by purchasing a sample of 50 batteries and subjecting them to test that estimate the battery's life.

(a) Assuming that the manufacturer's claim is true, describe the sampling distribution of the mean lifetime of a sample of 50 batteries.

(b) Assuming that the manufacturer's claim is true, what is the probability that the consumer group's sample has a mean life of 52 or fewer months?

Example

(a) Assuming that the manufacturer's claim is true, describe the sampling distribution of the mean lifetime of a sample of 50 batteries.

X: lifetime of a battery

$\bar{X}$: mean lifetime of the batteries

$$\bar{X} \sim N(54, \frac{6^2}{50})$$

Example

(b) Assuming that the manufacturer's claim is true, what is the probability that the consumer group's sample has a mean life of 52 or fewer months?

$$P(\bar{X} \leq 52) = P\left(\frac{\bar{X} - \mu}{SE} \leq \frac{52 - 54}{6/\sqrt{50}}\right) = P(Z \leq -2.357023)$$

Course materials

You can download the notes and codes from:

https://github.com/mcavs/ESTUMatse_2022Fall_EngineeringStatistics

SCAN ME

Contact

Do not hesitate to contact me on:

https://twitter.com/mustafa__cavus

https://www.linkedin.com/in/mustafacavusphd/

mustafacavus@eskisehir.edu.tr