Elementary Statistics Lecture 5

Sampling Distributions

Chong Ma

Department of Statistics
University of South Carolina

Outline

1 Introduction

2 Sampling Distribution of Sample Statistic

3 Examples

Recall

- **Parameter**: A numerical summary of the population, such as a population proportion p for a categorical variable fixed but usually unknown.
- **Statistic**: A numerical summary of a sample taken from the population, such as the sample mean, sample proportion, sample median and so on.

Sampling Distribution

The **sampling distribution** of a statistic is the probability distribution that specifies probabilities for the possible values the statistic can take.

Summary of jargons in terms of distributions

Summary

Population distribution: The distribution from which we take the sample

Data distribution: The distribution of the data obtained from the sample. The larger the sample, the more closely the data distribution resembles the population distribution.

Sampling distribution: The distribution of a statistic such as a sample proportion or a sample mean.

Outline

1. Introduction
2. Sampling Distribution of Sample Statistic
3. Examples

Central Limit Theorem(CLT)

Given certain conditions, the arithmetic mean of a sufficiently large number of independent random variables, each with a well-defined(finite) expected value(μ) and finite variance(σ^2), will be approximately normally distributed, regardless of the underlying distribution. Mathematically, it can be rewritten as follows.

CLT

Suppose $\{X_1, X_2, \ldots, X_n\}$ is a sequence of i.i.d random variables with $E[X_i] = \mu$ and $Var(X_i) = \sigma^2 < \infty$. Then as n approaches infinity, the random variable $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$ converge in distribution to the standard normal distribution $N(0, 1)$.

In other words,

$$\bar{X}_n \overset{aprox}{\sim} N(\mu, \frac{\sigma}{\sqrt{n}})$$

Sampling distribution of sample proportion $\hat{p}$

For a random sample of a size n from a population with proportion p of outcomes in a particular category, the sampling distribution of the sample proportion in that category approximately follows a normal distribution

$$\hat{p} \overset{aprox}{\sim} N(p, \sqrt{\frac{p(1-p)}{n}})$$

In practice, the above statement holds when the assumptions of $np \geq 15, n(1-p) \geq 15$ are satisfied.

Sampling distribution of sample mean $\bar{x}_n$

For a random sample of size n from a population having mean μ and standard deviation σ, then as the sample size n increases, the sampling distribution of the sample mean $\bar{x}_n$ approaches an approximately normal distribution as follows.

$$\bar{x}_n \overset{aprox}{\sim} N(\mu, \frac{\sigma}{\sqrt{n}})$$

In practice, the above statement holds when $n \geq 30$.

Sampling distribution

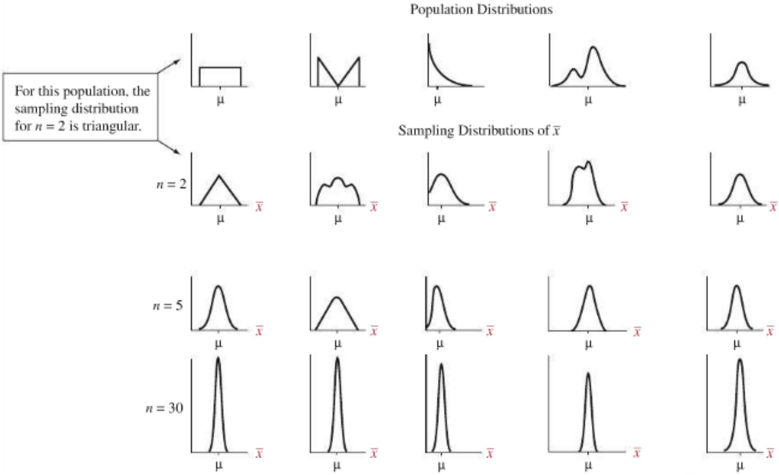

Figure 1: Five population distributions and the corresponding sampling distributions of $\bar{x}_n$. Regardless of the shape of the population distribution, the sampling distribution becomes more bell shaped as the sample size n increases.

Outline

1 Introduction

2 Sampling Distribution of Sample Statistic

3 Examples

Defective Chips

A supplier of electronic chips for tablets claims that only 4% of his chips are defective. A manufacture tests 500 randomly selected chips from a large shipment from the supplier for potential defects.

(a) Find the mean and the standard deviation for the distribution of the sample proportion of defective chips in the sample of 500.

(b) Is it reasonable to assume a normal shape for the sampling distribution? Explain.

(c) The manufacture will return the entire shipment if he finds more than 5% of the 500 sampled chips to be defective. Find the probability that the shipment will be returned.

Defective Chips

A supplier of electronic chips for tablets claims that only 4% of his chips are defective. A manufacture tests 500 randomly selected chips from a large shipment from the supplier for potential defects.

(a) Find the mean and the standard deviation for the distribution of the sample proportion of defective chips in the sample of 500.

Solution

The population of defective chips of the supplier is $p = 0.04$. The sample size is $n = 500$.

$$\begin{aligned} \text{mean} \quad & p = 0.04 \\ \text{standard deviation} \quad & \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.04 \times 0.96}{500}} = 0.0088 \end{aligned}$$

Defective Chips

A supplier of electronic chips for tablets claims that only 4% of his chips are defective. A manufacture tests 500 randomly selected chips from a large shipment from the supplier for potential defects.

(b) Is it reasonable to assume a normal shape for the sampling distribution? Explain.

Solution

Yes.

Since $np = 500 \times 0.04 = 20 \geq 15, n(1-p) = 500 \times 0.96 = 480 \geq 15$, the central limit theorem guarantees the sampling distribution of the sample proportion of defective chips is approximately normal.

Defective Chips

A supplier of electronic chips for tablets claims that only 4% of his chips are defective. A manufacture tests 500 randomly selected chips from a large shipment from the supplier for potential defects.

(c) The manufacture will return the entire shipment if he finds more than 5% of the 500 sampled chips to be defective. Find the probability that the shipment will be returned.

Solution

Note that

$$\hat{p} \overset{aprox}{\sim} N(p, \sqrt{\frac{p(1-p)}{n}}) = N(0.04, 0.0088)$$

Then

$$\begin{aligned} P(\hat{p} \geq 0.05) &= P(Z \geq \frac{0.05 - 0.04}{0.0088}) \\ &= P(Z \geq 1.14) \\ &= 0.127 \end{aligned}$$

Average income

A large corporation employs 27,251 individuals. The average income in 2008 for all employees was \$74,550 with a standard deviation of \$19,872. You are interested in comparing the incomes of today's employee's with those of 2008. A random sample of 100 employees of the corporation yields $\bar{x} = \$75,207$ and $s = \$18,901$.

(a) Describe the center and variability of the population distribution. What shape does it probably have?

(b) Describe the center and variability of the data distribution. What shape does it probably have?

(c) Describe the center and variability of the sampling distribution of the sample mean for $n = 100$. What shape does it have?

(d) Explain why it would not be unusual to observe an individual who earns more than \$100,000, but it would be highly unusual to observe a sample mean income of more than \$100,000 for a random sample size of 100 people?

Average income

A large corporation employs 27,251 individuals. The average income in 2008 for all employees was \$74,550 with a standard deviation of \$19,872. You are interested in comparing the incomes of today's employee's with those of 2008. A random sample of 100 employees of the corporation yields $\bar{x} = \$75,207$ and $s = \$18,901$.

(a) Describe the center and variability of the population distribution. What shape does it probably have?

Solution

The mean and standard deviation for the population is

$$\begin{aligned} \text{mean} \quad \mu &= 74,550 \\ \text{standard deviation} \quad \sigma &= 19,872 \end{aligned}$$

The shape of the population distribution of employee's income is probably highly right-skewed.

Average income

A large corporation employs 27,251 individuals. The average income in 2008 for all employees was \$74,550 with a standard deviation of \$19,872. You are interested in comparing the incomes of today's employee's with those of 2008. A random sample of 100 employees of the corporation yields $\bar{x} = \$75,207$ and $s = \$18,901$.

(b) Describe the center and variability of the data distribution. What shape does it probably have?

Solution

The mean and standard deviation for the data population is

$$\begin{aligned} \text{mean} \quad & \bar{x} = 75,207 \\ \text{standard deviation} \quad & s = 18901 \end{aligned}$$

Because the data distribution resembles the population distribution, thus the shape of the data distribution is probably right-skewed as well.

Average income

A large corporation employs 27,251 individuals. The average income in 2008 for all employees was \$74,550 with a standard deviation of \$19,872. You are interested in comparing the incomes of today's employee's with those of 2008. A random sample of 100 employees of the corporation yields $\bar{x} = \$75,207$ and $s = \$18,901$.

(c) Describe the center and variability of the sampling distribution of the sample mean for $n = 100$. What shape does it have?

Solution

The mean and standard deviation for the data population is

$$\begin{aligned} \text{mean} \quad & \mu_{\bar{x}_n} = \mu = 74,550 \\ \text{standard deviation} \quad & \sigma_{\bar{x}_n} = \tfrac{\sigma}{\sqrt{100}} = 1,987 \end{aligned}$$

The central limit theorem guarantees that the sampling distribution of the sample mean of employee's income for $n = 100$ is approximately normal since $n = 100 \geq 30$.

Average income

A large corporation employs 27,251 individuals. The average income in 2008 for all employees was \$74,550 with a standard deviation of \$19,872. You are interested in comparing the incomes of today's employee's with those of 2008. A random sample of 100 employees of the corporation yields $\bar{x} = \$75,207$ and $s = \$18,901$.

(d) Explain why it would not be unusual to observe an individual who earns more than \$100,000, but it would be highly unusual to observe a sample mean income of more than \$100,000 for a random sample size of 100 people?

Solution

Note that

$$X \sim N(\mu, \sigma) = N(74,550, 19,872)$$

$$\bar{X}_n \overset{aprox}{\sim} N(\mu, \frac{\sigma}{\sqrt{n}}) = N(74,550, 1,987)$$

Average income

A large corporation employs 27,251 individuals. The average income in 2008 for all employees was \$74,550 with a standard deviation of \$19,872. You are interested in comparing the incomes of today's employee's with those of 2008. A random sample of 100 employees of the corporation yields $\bar{x} = \$75,207$ and $s = \$18,901$.

(d) Explain why it would not be unusual to observe an individual who earns more than \$100,000, but it would be highly unusual to observe a sample mean income of more than \$100,000 for a random sample size of 100 people?

Solution

Note that

$$P(X \geq 100,000) = P(X \geq \frac{100,000 - 74,550}{19,872}) = P(Z \geq 1.28) = 0.1$$

$$P(\bar{X}_n \geq 100,000) = P(\bar{X}_n \geq \frac{100,000 - 74,550}{1,987}) = P(Z \geq 12.8) = 0$$

Coin-toss distribution

For a single coin toss of a balanced coin, let $x = 1$ for a head and $x = 0$ for a tail. Say a coin is flipped 30 times. Let Y denote the number of heads occurring in the 30 flips.

(a) Find the sampling distribution of the sample proportion of head.

(b) Find the probability of observing more than 10 heads for the 30 flips of a balanced coin.

Coin-toss distribution

For a single coin toss of a balanced coin, let $x = 1$ for a head and $x = 0$ for a tail. Say a coin is flipped 30 times. Let Y denote the number of heads occurring in the 30 flips.

(a) Find the sampling distribution of the sample proportion of head.

Solution

Note $p = 0.5, n = 30$, then

$$\hat{p} \overset{aprox}{\sim} N(p, \sqrt{p(1-p)/n}) = N(0.5, 0.09)$$

The CLT guarantees the sampling distribution of $\hat{p}$ is approximately normal since $np = 15, n(1-p) = 15$.

Coin-toss distribution

For a single coin toss of a balanced coin, let $x = 1$ for a head and $x = 0$ for a tail. Say a coin is flipped 30 times. Let Y denote the number of heads occurring in the 30 flips.

(b) Find the probability of observing more than 10 heads for the 30 flips of a balanced coin.

Solution

Note that $Y \sim \mathit{Binomial}(n, p) = \mathit{Binomial}(30, 0.5)$, then

$$\begin{aligned} P(Y > 10) &= 1 - P(Y \leq 10) \\ &= 1 - \{P(Y = 0) + P(Y = 1) + \cdots + P(Y = 10)\} \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ &= 0.95 \end{aligned}$$

It's tedious for using the binomial distribution to calculate this probability. An easier way is to use the sampling distribution of the sample proportion $\hat{p}$.

Coin-toss distribution

For a single coin toss of a balanced coin, let $x = 1$ for a head and $x = 0$ for a tail. Say a coin is flipped 30 times. Let Y denote the number of heads occurring in the 30 flips.

(b) Find the probability of observing more than 10 heads for the 30 flips of a balanced coin.

Solution

The question is equivalent to finding the probability of sample proportion more than 0.3.

Note

$$\hat{p} \overset{aprox}{\sim} N(p, \sqrt{p(1-p)/n}) = N(0.5, 0.09)$$

Thus

$$\begin{aligned} P(\hat{p} > 0.3) &= P(\hat{p} > \frac{0.3 - 0.5}{0.09}) \\ &= P(Z > -2.22) \\ &= 0.986 \end{aligned}$$