ECO 391 Economics and Business Statistics

Lecture 3: Sampling and Sampling Distribution

Xiaozhou Ding

February 4, 2019

Overview

1. Introduction
2. Several Concepts
   - Basic Definition and Sampling Bias
   - Sampling Methods
3. The Sampling Distribution of the Sample Mean
4. The Central Limit Theorem

Introduction

Outline

- Explain common sample bias.
- Describe various sampling methods.
- Describe the sampling distribution of the sample mean.
- Explain the importance of the central limit theorem.

Several Concepts

Sampling

- **Population** consists of all items of interest in a statistical problem.
  - **Population Parameter is unknown**.
- **Sample** is a subset of the population.
  - **Sample statistic** is calculated from sample and used to make inferences about the population.
- **Bias** is the tendency of a sample statistic to systematically over- or underestimate a population parameter.

Classic Case of a "Bad" Sample

- During the1936 presidential election, the Literary Digest predicted a landslide victory for Alf Landon over Franklin D. Roosevelt (FDR) with only a 1% margin of error.
- They were wrong! FDR won in a landslide election.
- The Literary Digest had committed selection bias by randomly sampling from their own subscriber/membership lists, etc.
- In addition, with only a 24% response rate, the Literary Digest had a great deal of non-response bias.

Bias

- Selection bias: a systematic exclusion of certain groups from consideration for the sample.
  - The Literary Digest committed selection bias by excluding a large portion of the population (e.g., lower income voters).
- Nonresponse bias: a systematic difference in preferences between respondents and non-respondents to a survey or a poll.
  - The Literary Digest had only a 24% response rate. This indicates that only those who cared a great deal about the election took the time to respond to the survey. These respondents may be atypical of the population as a whole.

Sampling Methods

Definition

Simple random sample is a sample of n observations which has the *same probability of being selected* from the population as any other sample of n observations.

- Most statistical methods presume simple random samples.
- However, in some situations other sampling methods have an advantage over simple random samples.

Sampling Methods

Definition

Stratified Random Sampling divide the population into mutually exclusive and collectively exhaustive groups, called strata. Then it randomly select observations from each stratum, which are proportional to the stratum's size.

- Guarantees that the each population subdivision is represented in the sample.
- Parameter estimates have greater precision than those estimated from simple random sampling.

Sampling Methods

Definition

Cluster Sampling divide the population into mutually exclusive and collectively exhaustive groups, called clusters. A cluster sample includes observations from randomly selected clusters.

- Less expensive than other sampling methods.
- Less precision than simple random sampling or stratified sampling.
- Useful when clusters occur naturally in the population.

Sampling Methods

Compare stratified and cluster sampling.

- Stratified sampling
  - Sample consists of elements from each group.
  - Preferred when the objective is to increase precision.
- Cluster Sampling
  - Sample consists of elements from the selected groups.
  - Preferred when the objective is to reduce costs.

The Sampling Distribution of the Sample Mean

Housekeeping

- Population is described by **parameters**.
  - A parameter is a constant, whose value may be unknown.
  - Only one population.
- Sample is described by **statistics**.
  - A statistic is a **random variable** whose value depends on the chosen random sample.
  - Statistics are used to make *inferences* about the population parameters.
  - Can draw multiple random samples of size n.

The Sampling Distribution of the Sample Mean

Definition

An **estimator** is a statistic that is used to estimate a population parameter.

Example

$\overline{X}$, the mean of the sample, is an estimator of μ, the mean of the population.

Definition

An estimate is a particular value of the estimator.

Example

The mean of the sample $\bar{x}$ is an estimate of μ, the mean of the population.

The Sampling Distribution of the Sample Mean

- Each random sample of size n drawn from the population provides an estimate of μ-the sample mean $\bar{x}$.
- Drawing many samples of size n results in many different sample means, one for each sample.
- The sampling distribution of the mean is the frequency or probability distribution of these sample means.

Example

	Random Variable				
	X_1	X_2	X_3	X_4	
	6	10	8	4	
	5	10	4	3	
	1	8	4	3	
	4	1	6	2	
	6	6	8	4	
	7	7	8	6	
	1	5	10	5	
	5	5	9	1	
	4	6	4	2	
	7	4	9	5	
	8	5	8	6	
	9	2	7	7	
	9	1	2	3	
	6	10	2	6	
Means	5.57	5.71	6.36	4.07	5.43

The Expected Value and the Standard Error of the Sample Mean

- The expected value of X,
$$\mathrm{E}(X) = \mu.$$
- The expected value, or estimator, of the mean,
$$\mathrm{E}(\overline{X}) = \mathrm{E}(X) = \mu.$$

The Sampling Distribution of the Sample Mean

- Variance of X:
$$\text{Var}(X) = \sigma^2.$$
- Variance of $\overline{X}$:
$$\text{Var}(\overline{X}) = \frac{\sigma^2}{n}.$$
- Standard deviation of X:
$$\text{SD}(X) = \sqrt{\sigma^2} = \sigma.$$
- Standard error of $\overline{X}$:
$$se(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

The standard error is the standard deviation of the sampling distribution for all samples of size n.

Example

Given that $\mu = 16$ inches and $\sigma = 0.8$ inches, determine the following:

- The expected value
- The standard error of the sample mean derived from the sample of
  - 2 pizzas
  - 4 pizzas

- 2 pizzas: $\mathrm{E}(\overline{X}) = \mu = 16$, $se(\overline{X}) = \frac{\sigma}{\sqrt{n}} = \frac{0.8}{\sqrt{2}} = 0.57$.
- 4 pizzas: $\mathrm{E}(\overline{X}) = \mu = 16$, $se(\overline{X}) = \frac{\sigma}{\sqrt{n}} = \frac{0.8}{\sqrt{4}} = 0.40$.

Sampling from a Normal Distribution

- For any sample size n, the sampling distribution of $\overline{X}$ is normal if the population X from which the sample is drawn is normally distributed.
- If X is normal, then we can transform it into the standard normal random variable as:
  - For a sampling distribution

$$Z = \frac{\overline{X} - \mathrm{E}(\overline{X})}{se(\overline{X})} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

  - For a distribution of the values of X

$$Z = \frac{x - \mathrm{E}(X)}{SD(X)} = \frac{x - \mu}{\sigma}$$

Any value $\bar{x}$ has a corresponding value z given by $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$.

Sampling from a Normal Distribution

	Random Variable $\overline{X}$	Standard Normal Z	
	3	-2.39	$z_1 = \frac{\overline{X}_1 - \mu}{\sigma/\sqrt{n}}$
	9	4.3	
	4	-1.28	
	2	-3.51	
	10	5.42	
	5	-0.16	
	9	4.3	
	4	-1.28	
	9	4.3	
	2	-3.51	
	3	-2.39	
	8	3.19	
	4	-1.28	
	0	-5.74	$z_{13} = \frac{\overline{X}_{13} - \mu}{\sigma/\sqrt{n}}$
Means	5.14	0	
Standard Error	0.9	1	

Example

Example: Given that $\mu = 16$ inches and $\sigma = 0.8$ inches, determine the following:

- What is the probability that a randomly selected pizza is less than 15.5 inches?
- What is the probability that *2 randomly selected pizzas average* less than 15.5 inches?

- $$Z = \frac{X - \mu}{\sigma} = \frac{15.5 - 16}{0.8} = -0.63$$
$$P(X < 15.5) = P(Z < -0.63) = 0.2643$$

- $$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{15.5 - 16}{0.8/\sqrt{2}} = -0.88$$
$$P(X < 15.5) = P(Z < -0.88) = 0.1894$$

The Central Limit Theorem

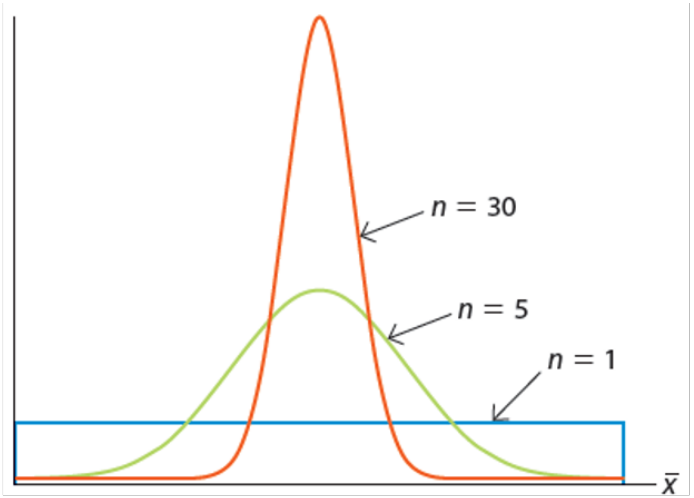
The Central Limit Theorem

Definition

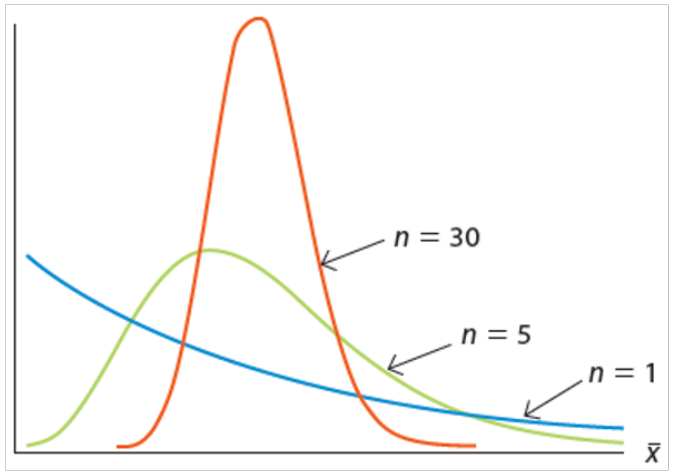
For any population X with expected value μ and standard deviation σ, the sampling distribution of X will be approximately normal if the sample size n is sufficiently large.

- As a general guideline, the normal distribution approximation is justified when $n \geq 30$.
- As before, if $\overline{X}$ is approximately normal, then we can transform it to $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$.

The Central Limit Theorem



Sampling distribution of $\bar{X}$ when the population has a uniform distribution.



Sampling distribution of $\bar{X}$ when the population has an exponential distribution.

Example

In order to capitalize on the iced coffee trend, Starbucks offered for a limited time half-priced Frappuccino beverages between 3 pm and 5 pm. Anne Jones, manager at a local Starbucks, determines the following from past historical data:

- 43% of iced-coffee customers were women.
- 21% were teenage girls.
- Customers spent an average of \$4.18 on iced coffee with a standard deviation of \$0.84.

One month after the marketing period ends, Anne surveys 50 of her iced-coffee customers and finds:

- 46% were women.
- 34% were teenage girls.
- They spent an average of \$4.26 on the drink.

Anne wants to use this survey information to calculate the probability that: Customers spend an average of \$4.26 or more on iced coffee.

From the introductory case, Anne wants to determine if the marketing campaign has had a lingering effect on the amount of money customers spend on iced coffee.

- Before the survey, $\mu = 4.18$ and $\sigma = 0.84$. Based on 50 customers sampled after the survey, $\bar{x} = 4.26$.
- Let's find $P(\overline{X} \geq 4.26)$. Since $n > 30$, the central limit theorem states that $\overline{X}$ is approximately normal. So,

$$\begin{aligned} P(\overline{X} \geq 4.26) =& P\left(Z \geq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right) \\ =& P\left(Z \geq \frac{4.26 - 4.18}{0.84/\sqrt{50}}\right) \\ =& P(Z \geq 0.67) \\ =& 1 - P(Z < 0.67) \\ =& 0.2514 \end{aligned}$$