CSCI 3022: Intro to Data Science

Lecture 13: The Central Limit Theorem

Rachel Cox

Department of Computer Science

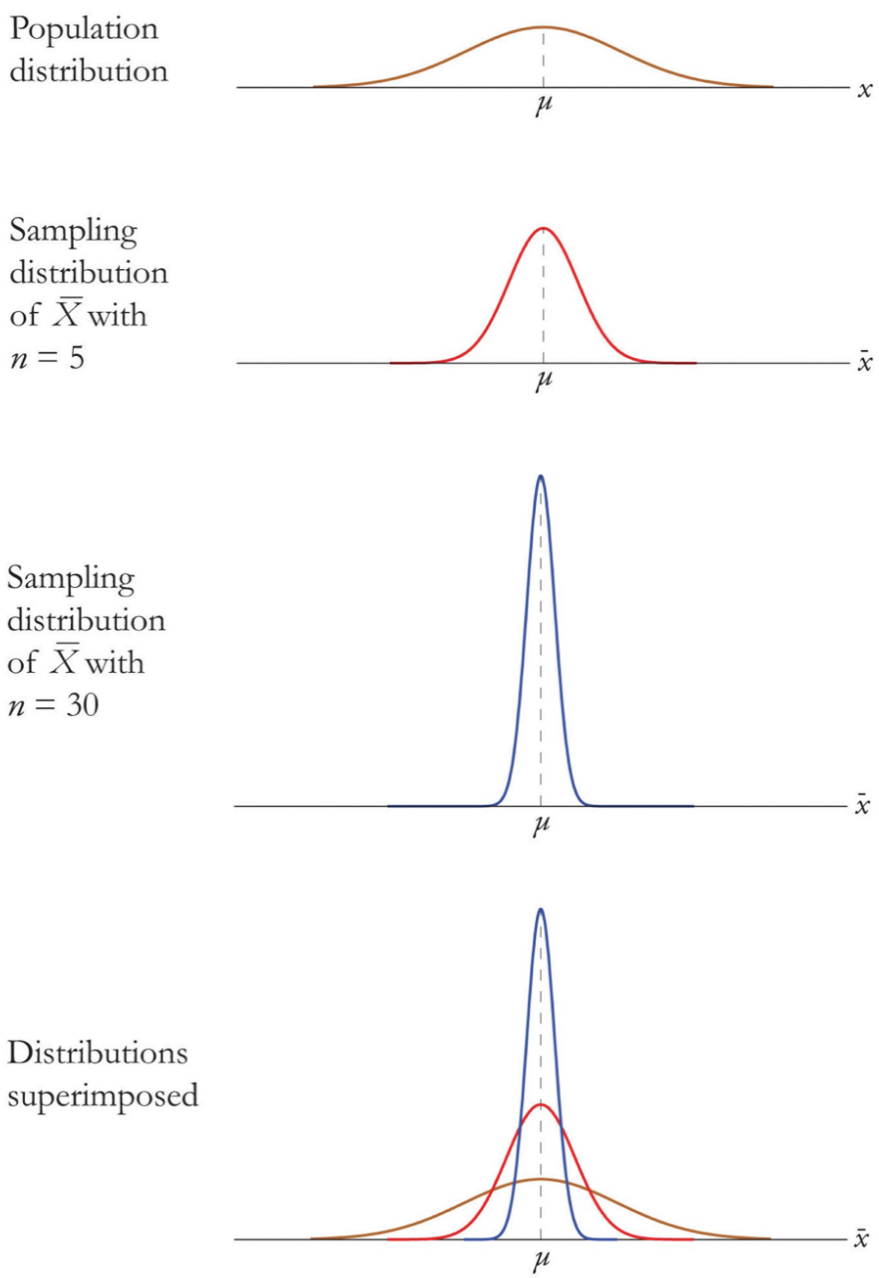


***Figure* 6.2.4**: *Distribution of Sample Means for a Normal Population*

Announcements & Reminders

- ☐ Quizlet Due tonight
- ☐ Homework 3 posted - Due March 13.

What will we learn today?

- ❑ Central Limit Theorem
- ❑ iid samples
- ❑ Distribution of samples vs. distribution of sample means

- ❑ *A Modern Introduction to Probability and Statistics, Chapter 14*

Review from Last Time

A continuous random variable X has a normal (or Gaussian) distribution with parameters μ and σ^2 if its pdf is given by the following. We say that $X \sim N(\mu, \sigma^2)$.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Proposition: If X is a normally distributed random variable with mean μ and standard deviation σ, then Z follows a standard normal distribution if we define:

$$Z = \frac{X-\mu}{\sigma} \quad \text{and} \quad X = \sigma Z + \mu$$

If Z is a standard normal random variable, then we can compute probabilities using the standard normal cdf

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} f(x)\, dx$$

Random Samples

The random variables $X_1, X_2, \ldots, X_n$ are said to form a random sample of size n if:

1. All $X_k's$ are independent.

2. All $X_k's$ come from the same distribution. "identically distributed"

independent and identically distributed
iid

We use estimators to summarize our iid sample.

μ, σ^2 "true" statistics

$\bar{X}$ is the sample mean estimator of the population mean μ

$\bar{x}, s^2$ observed quantities of sample.

$\hat{p}$ is the sample proportion (# in the sample satisfying some characteristic of interest/total #)

s^2 is the sample estimator for σ^2

Estimators and their distributions

Any estimator, including the sample mean, is a random variable (since it is based on a random sample.)

This means that $\bar{X}$ has a distribution of its own, which is referred to as the **sampling distribution of the sample mean**.

The sampling distribution depends on:

1) Population distribution

2) Sample size n

3) Method of sampling

Distribution of the Sample Mean

Proposition: Let $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$.

- As sample size increases, the variance decreases

Then for any n, $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

$$E[\bar{x}] = E\left[\frac{1}{n}\sum_{k=1}^{n} x_k\right]$$

$$= \frac{1}{n} E\left[\sum_{k=1}^{n} x_k\right]$$

$$= \frac{1}{n} \sum_{k=1}^{n} E[x_k]$$

$$= \frac{1}{n} \sum_{k=1}^{n} \mu$$

$$= \frac{n\mu}{n}$$

$$= \mu$$

$$Var(\bar{x}) = Var\left(\frac{1}{n}\sum_{k=1}^{n} x_k\right)$$

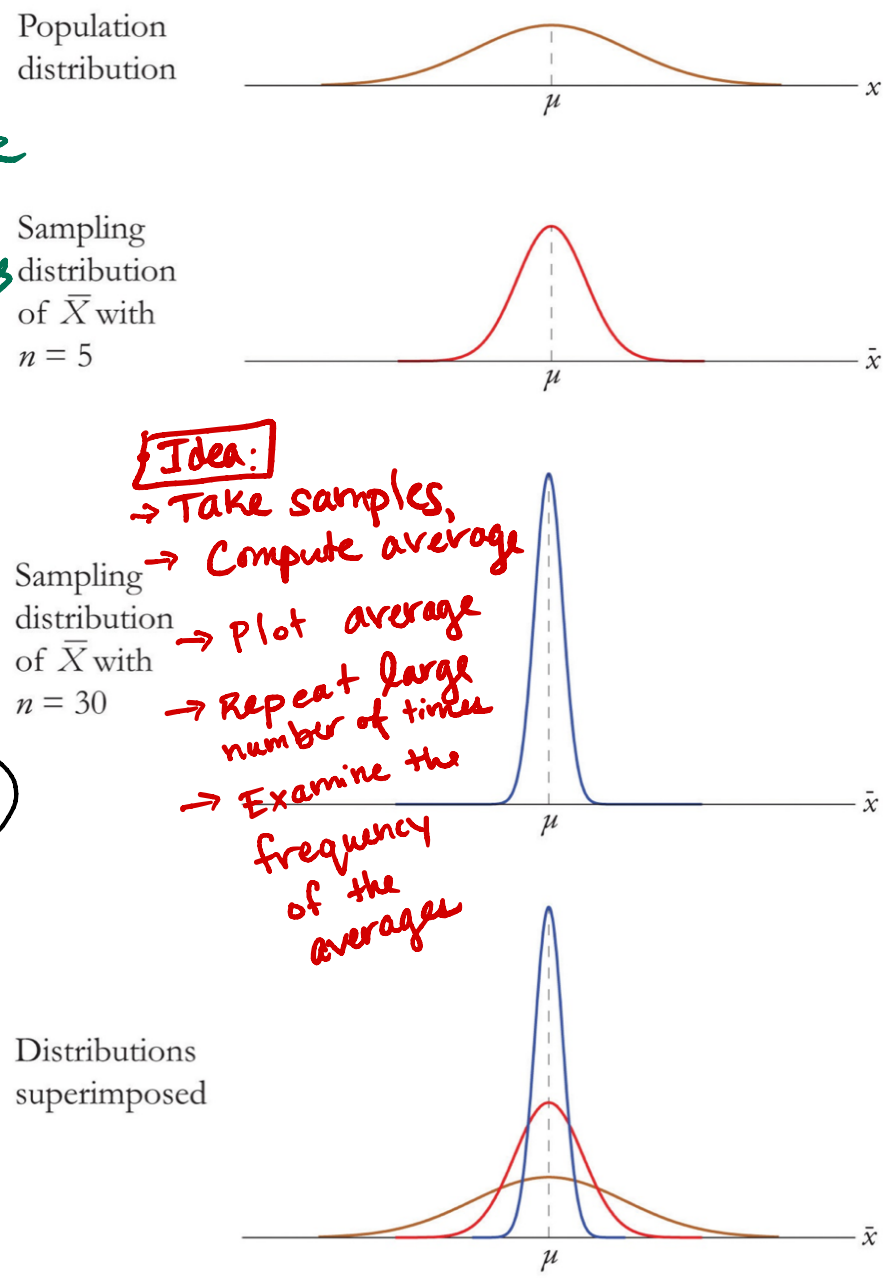
$$= \frac{1}{n^2} Var\left(\sum_{k=1}^{n} x_k\right)$$

$$= \frac{1}{n^2} \sum_{k=1}^{n} Var(x_k)$$

$$= \frac{1}{n^2} \sum_{k=1}^{n} \sigma^2$$

$$= \frac{n\sigma^2}{n^2}$$

$$= \frac{\sigma^2}{n}$$



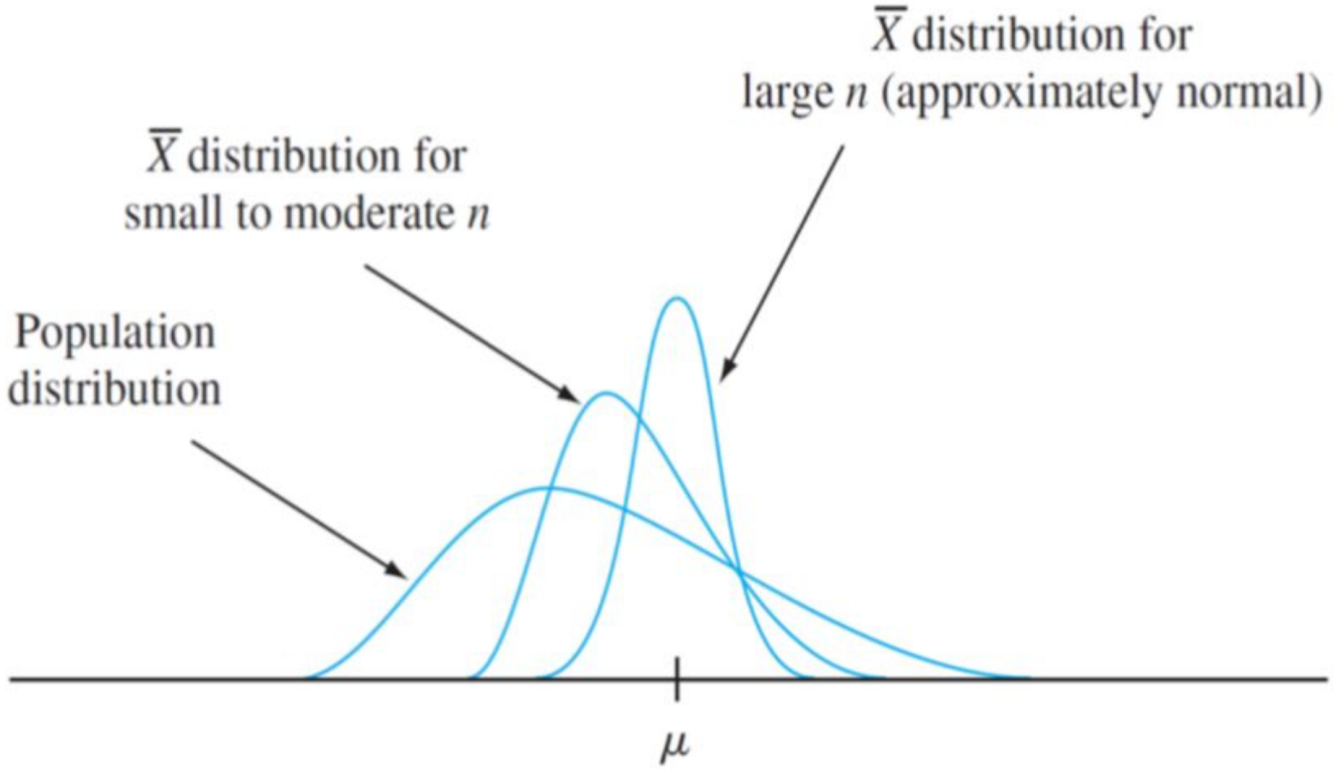
Idea:
→ Take samples,
→ Compute average
→ Plot average
→ Repeat large number of times
→ Examine the frequency of the averages

***Figure* 6.2.4**: *Distribution of Sample Means for a Normal Population*

Distribution of the Sample Mean

What if the population distribution is not normally distributed?

- When the population distribution is non-normal, averaging produces a distribution more normal (bell-shaped) than the one being sampled.



Visualization from onlinestatbook.com

The Central Limit Theorem

The Central Limit Theorem: Let $X_1, X_2, \ldots, X_n$ be iid draws from some distribution. Then, as n becomes large

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

From the book: Let $X_1, X_2, \ldots$ be any sequence of independent identically distributed random variables with finite positive variance. Let μ be the expected value and σ^2 the variance of each X_i. For $n \geq 1$, let Z_n be defined by

$$Z_n = \frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}}$$

Then for any number a, $\lim_{n\to\infty} F_{Z_n}(a) = \Phi(a)$, where Φ is the distribution function of the N(0, 1) distribution.

- The distribution function of Z_n converges to the distribution function Φ of the standard normal distribution.

The Central Limit Theorem

Example: A hardware store receives a shipment of bolts that are supposed to be 12 cm long. The mean is indeed 12 cm, and the standard deviation is 0.2 cm. For quality control, the hardware store chooses 100 bolts at random to measure.

They will declare the shipment defective and return it to the manufacturer if the average length of 100 bolts is less that 11.97 cm or greater than 12.04 cm. Find the probability that the shipment is found satisfactory.

$\mu = 12$ cm

$\sigma = 0.2$ cm

$n = 100$

We want $P(11.97 \le \bar{X} \le 12.04)$

Using a transformation to move towards a standard normal.

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}, \quad \bar{X} \sim N\left(\mu = 12, \text{var} = \frac{0.2^2}{100}\right)$$

$$\Rightarrow \sigma_{\bar{x}} = 0.2/10$$

So $$P\left(\frac{11.97-12}{0.2/10} \le Z \le \frac{12.04-12}{0.2/10}\right) = \Phi\left(\frac{.04}{.02}\right) - \Phi\left(-\frac{.03}{.02}\right)$$

$$= \Phi(2) - \Phi(-1.5)$$

$$= \text{stats.norm.cdf}(2) - \text{stats.norm.cdf}(-1.5)$$

$$= 0.977250 - 0.066807 = 0.910443$$

The Central Limit Theorem

Example: (continued)

Next Time:

- Notebook Day