Introduction to Data Science With Probability and Statistics

Lecture 16: The Central Limit Theorem

CSCI 3022 - Summer 2020
Sourav Chakraborty
Dept. of Computer Science
University of Colorado Boulder

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"

We use estimators to summarize our iid sample.

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

$\hat{p}$ 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.)

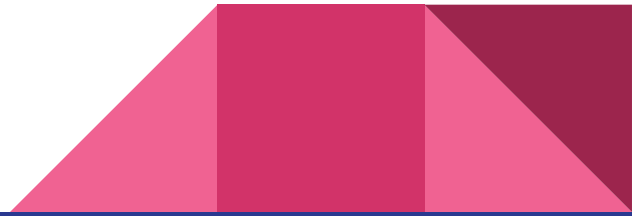
This means that $\bar{X}$ has a distribution of its own, which is referred to as the **sampling distribution of the sample mean**.

The sampling distribution depends on:

1) Population distribution

2) Sample size n

3) Method of sampling



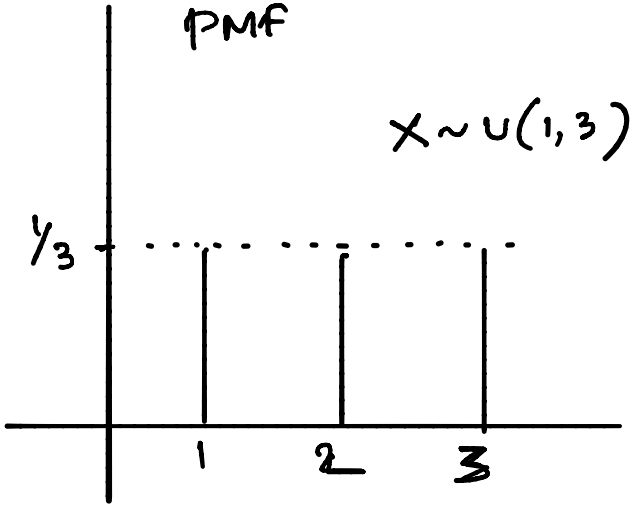
Random Sampling Example

Sample 1: n = 2, [2, 3], mean = 2.5, x_1

Sample 2: n = 2, [2, 1], mean = 1.5, x_2

Sample 3: n = 2, [3, 3], mean = 3, x_3

Sample 4, n = 2, [1, 3], mean = 2, x_4



Random Sampling Example

$\bar{X}$ is the random variable which represents the sample mean.

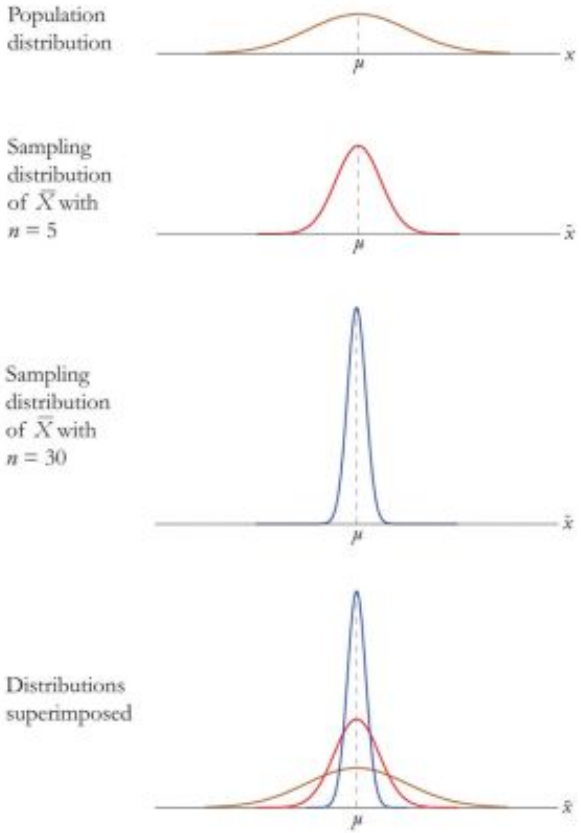
$$\therefore \quad \bar{x} = \frac{x_1 + x_2 + x_3 + \cdots - x_n}{n}$$

So, for the current example;

$$\bar{x} = \{2{,}5,\ 1{\cdot}5,\ 3,\ 2 \cdots\}$$

We are concerned about it's distribution.

Distribution of the Sample Mean



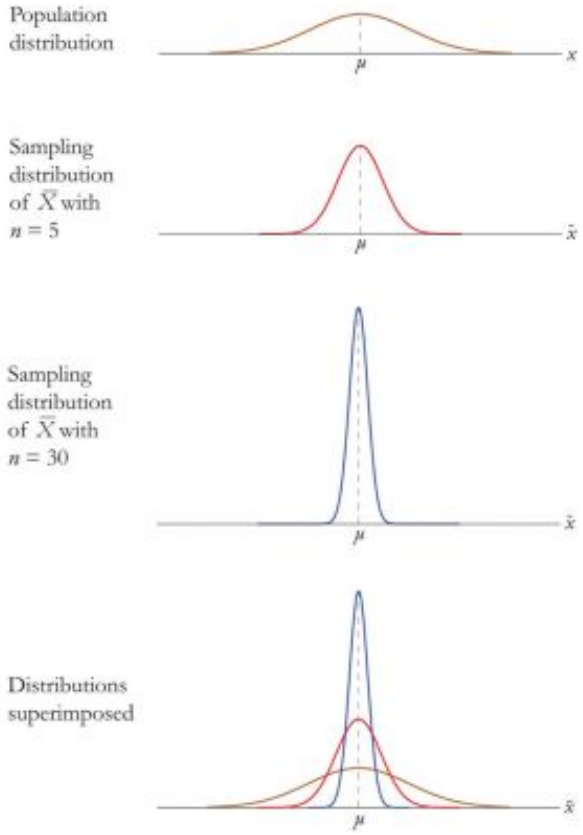
Proposition: Let $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$.

Then for any n, $\bar{X} \sim$ N(?, ?)

$E[\bar{X}]$?

$Var(\bar{X})$?

Distribution of the Sample Mean



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

Then for any n, $\bar{X} \sim N(\mu, ?)$

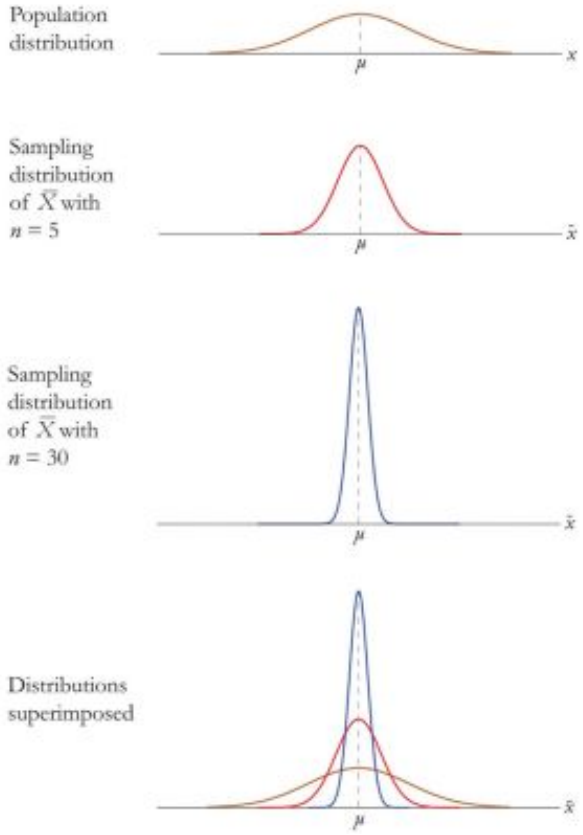
$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

$$E[\bar{x}] = E\left[\frac{x_1}{n} + \frac{x_2}{n} + \cdots + \frac{x_n}{n}\right]$$

$$= \frac{1}{n} \cdot \sum_{i}^{n} E[x_i] = \frac{1}{n} \cdot n \cdot \mu$$

$$\boxed{E[\bar{x}] = \mu}$$

Distribution of the Sample Mean



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

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

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{X_1}{n} + \frac{X_2}{n} + \cdots + \frac{X_n}{n}\right)$$

$$= \frac{1}{n^2} \sum_{i}^{n} \text{Var}(X_i)$$

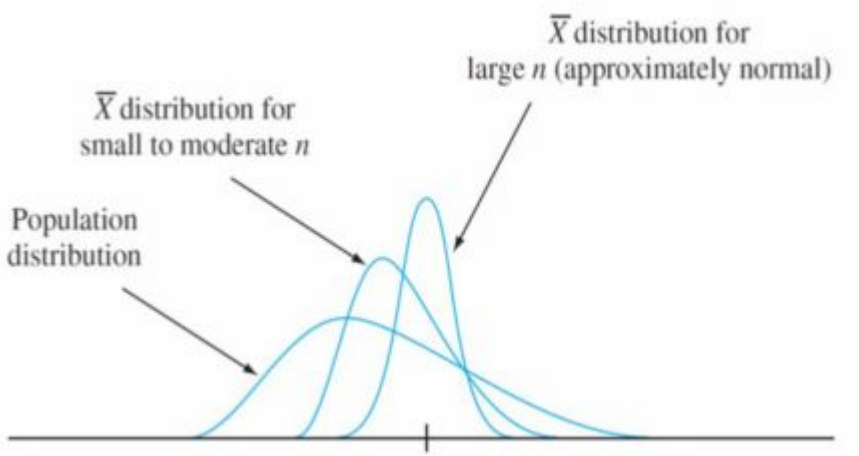
$$= \frac{1}{n^2} \cdot n \cdot \sigma^2$$

$$\boxed{\text{Var}(\bar{X}) = \sigma^2/n}$$

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.

$\mu = 12\,cm$

$\sigma = 0.2\,cm$

$n = 100$

} population characteristics.

The Central Limit Theorem

Example: (continued)

Let $\bar{x}$ be random variable which represents the avg. length of 100 bolts. So, we want $P(11.97 \le \bar{x} \le 12.04)$

from CLT, we know $\bar{x} \sim N(\mu = 12, \text{var} = (0.2)^2/100)$

Let transform $\bar{x}$ to Z.

$$P(11.97 \le \bar{x} \le 12.04)$$

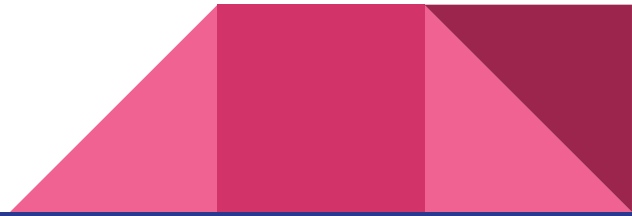
$$= P\left(\frac{11.97 - 12}{0.2/10} \le Z \le \frac{12.04 - 12}{0.2/10}\right)$$

$$\because \quad Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$$= P(-1.5 \le z \le 2) = \boxed{\phi(2) - \phi(-1.5) = 0.91}$$

Heights revisited

Bob told Alice that the heights of the people in the city of Boulder follow some probability distribution that he does not know, but he knows that the mean height is 5 ft. and the standard deviation is 2.2 ft. Now Alice wonders what is the probability that the average height of a person in her building of 100 people will be more than 6 ft?



Heights revisited

Bob told Alice that the heights of the people in the city of Boulder follow some probability distribution that he does not know, but he knows that the mean height is 5 ft. and the standard deviation is 2.2 ft. Now Alice wonders what is the probability that the average height of a person in her building of 100 people will be more than 6 ft?

Population stats.

$\mu = 5$

$\sigma = 2.2$

$n = 100$

So, we want.

$P(\bar{x} > 6)$?

$\bar{X} \sim N\left(5, (2.2)^2/100\right)$

Heights revisited

Bob told Alice that the heights of the people in the city of Boulder follow some probability distribution that he does not know, but he knows that the mean height is 5 ft. and the standard deviation is 2.2 ft. Now Alice wonders what is the probability that the average height of a person in her building of 100 people will be more than 6 ft?

$$P(\bar{x} > 6) = P\left(Z > \frac{6-5}{2.2/10}\right)$$

$$= P(Z > 4.54)$$

$$= 1 - \phi(4.54) \approx 1 - 0.99$$

$$\approx 0.01$$

$$\bar{X} \sim N\left(5, \frac{(2.2)^2}{100}\right)$$

Next Time:

- Notebook Day !