The Central Limit Theorem

Administrivia

- **Homework 3** due Friday.
- **Midterm** coming up **in-class** on Wednesday **October 18th**
  - Mix of Multiple Choice and Free-Response Questions
  - Allowed one 8.5 x 11in sheet of handwritten notes (no magnifying glasses)
  - Allowed a calculator that can't connect to internet or store large large data
- We'll do a Q&A style **Midterm Review** in class on **Monday October 16th**

Previously on CSCI 3022

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

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

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

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

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

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

Motivating Example

Soon we'll be talking about statistical inference wherein we'll try to infer things about the true mean of a population using sample datasets.

Examples:

- MEAN GPA OF ALL CS STUDENTS : SAMPLE OF 30 STUDENTS
- ZEBRAS
- POLITICAL POLLING

$\bar{x} \overset{?}{\Leftrightarrow} \mu$

$s \Leftrightarrow \sigma^2$

Random Samples

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

1. ALL X_k'S ARE INDEPENDENT
2. ALL X_k'S COME FROM SAME DISTRIBUTION

We say these X_k's are:

iid = INDEPENDENT & IDENTICALLY DISTRIBUTED

Estimators and Their Distributions

We use **estimators** to summarize our i.i.d. sample

Examples:

1): $\bar{X}$ IS THE SAMPLE MEAN ESTIMATOR OF POP MEAN μ

2) $\hat{p}$ IS THE SAMPLE PROPORTION

3) s^2 IS THE ESTIMATOR FOR σ^2

Estimators and Their Distributions

We use **estimators** to summarize our i.i.d. sample

Any estimator, including the **sample mean**, $\bar{X}$, is a random variable (since it's based on a random sample)

This means that $\bar{X}$ has a distribution of it's 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

$$\mathrm{VAR}\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) = \frac{1}{n^2}\mathrm{VAR}\left(\sum_{k=1}^{n} X_k\right) = \frac{1}{n^2}\sum \mathrm{VAR}(X_k) = \frac{1}{n^2}\sum \sigma^2 = \frac{\sigma^2}{n}$$

But what does this sampling distribution actually look like?

Proposition: Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$. Then for any n

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

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

$$\mathrm{VAR}(\bar{X}) = \frac{\sigma^2}{n}$$

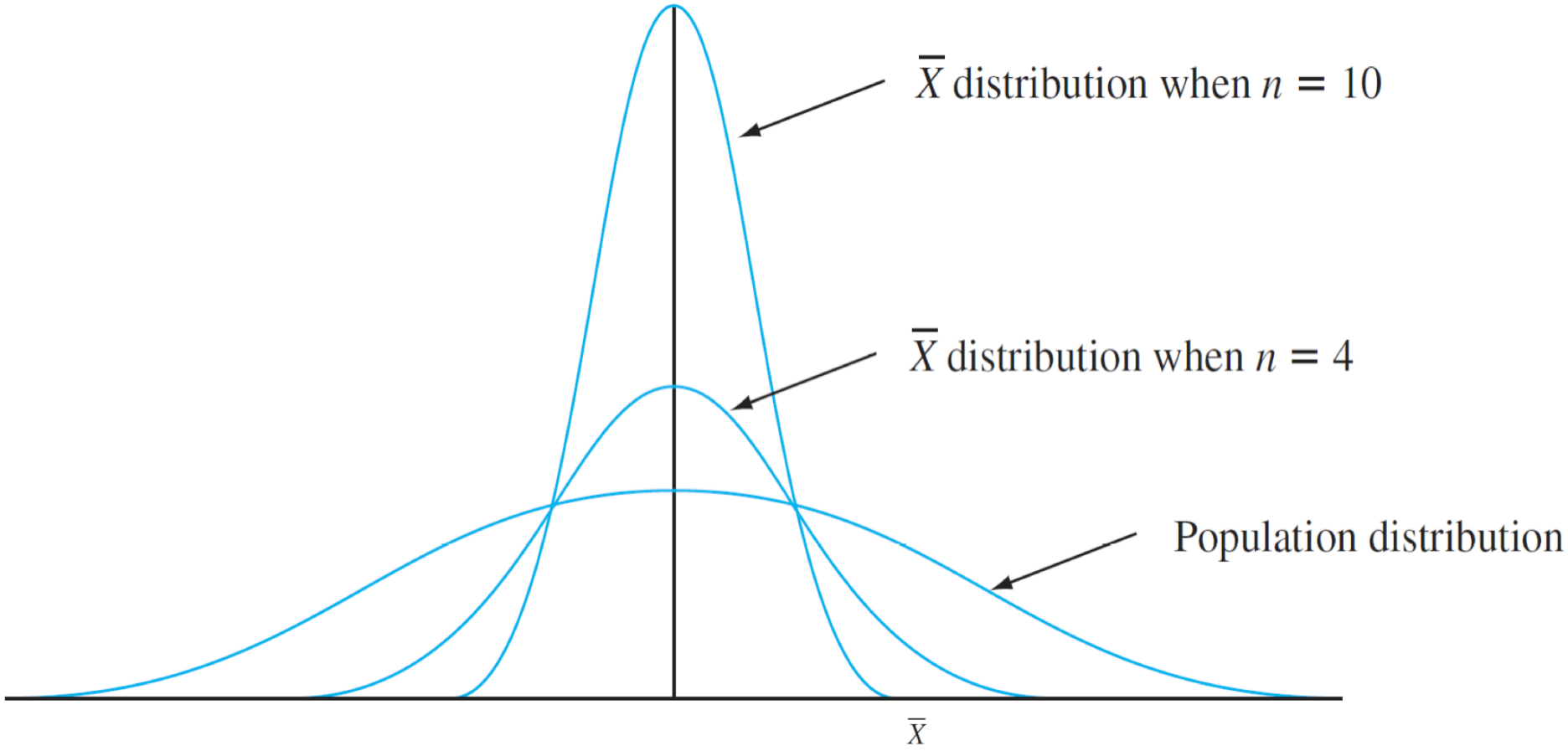
We know everything there is to know about the distribution of the sample mean when the population distribution is normal.

$$E[\bar{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_k\right] = \frac{1}{n}\sum E[X_k] = \frac{1}{n}\sum_{k=1}^{n} \mu$$

$$E[\bar{X}] = \mu$$

Distribution of the Sample Mean

If the population is normally distributed:



Distribution of the Sample Mean

But what if the population distribution is **NOT** normally distributed?!

The Central Limit Theorem

But what if the population distribution is **NOT** normally distributed?!

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

A reasonable assumption is that if n is **large**, a suitable normal curve will well-approximate the actual distribution of the sample mean.

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

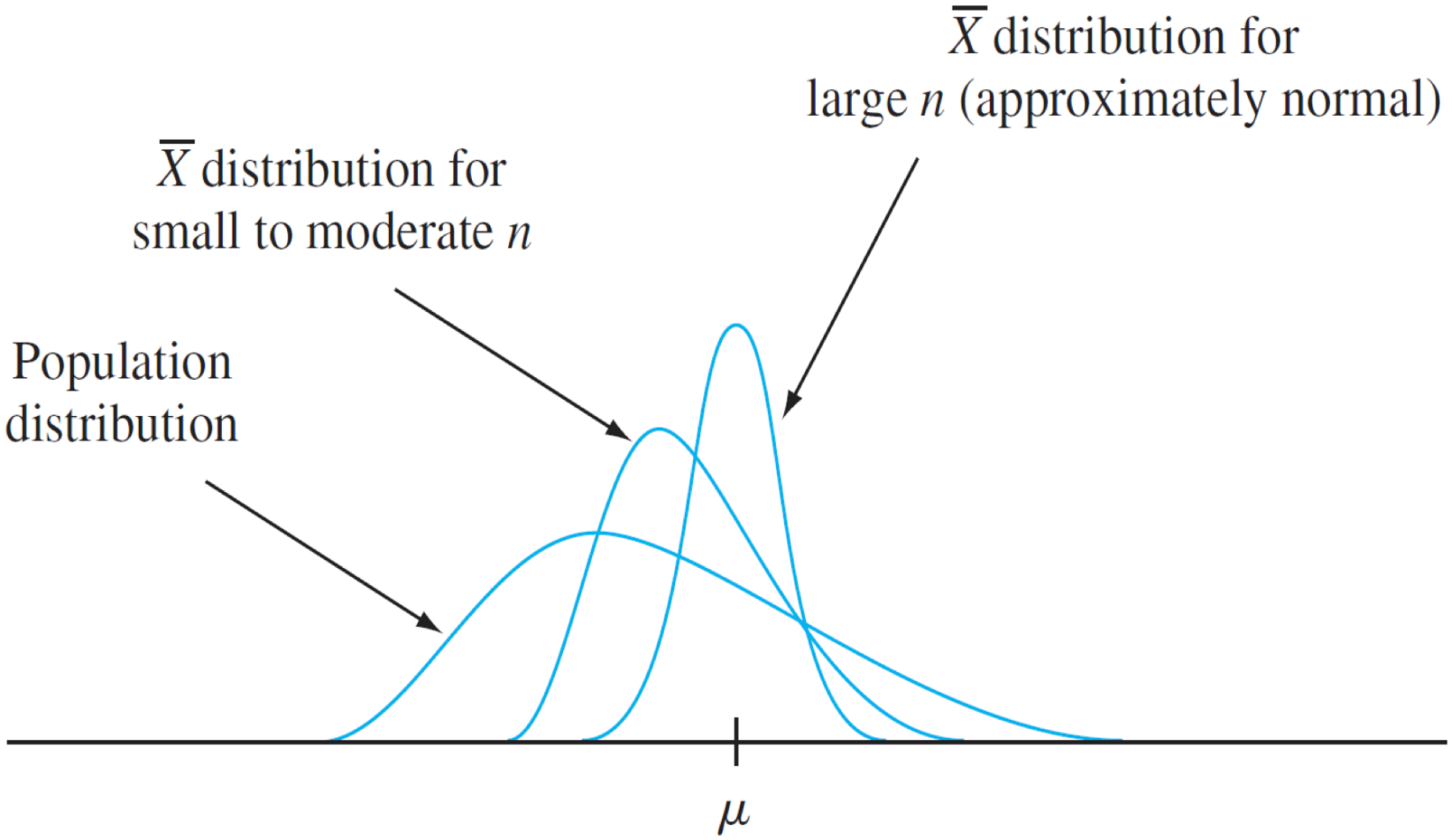
$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\text{VAR}(\bar{X}) = \frac{\sigma^2}{n}$$

Rule of Thumb: $n \geq 30$

Distribution of the Sample Mean

If the population is **NOT** normally distributed:



Examples

Example: A hardware store receives a shipment of bolts that are supposed to be 12cm long. The mean is indeed 12cm, and the standard deviation is 0.2cm. For quality control, the hardware store chooses 100 bolts at random to measure. They will call the shipment defective and return it to the manufacturer if the average length of the 100 bolts is less than 11.97cm or greater than 12.04cm. Find the probability that the shipment is found satisfactory.

$\mu = 12\text{cm} \quad \sigma = 0.2\text{cm}$

$P(\bar{X} \le 11.97 \text{ or } \bar{X} \ge 12.04)$

$= 1 - P(11.97 \le \bar{X} \le 12.04)$

$n = 100$

$\bar{X} \sim N\left(12, \frac{(.2)^2}{100}\right)$

$11.97 \rightarrow \frac{11.97-12}{.2/10} = \frac{-.03}{.02} \rightarrow -1.5$

$12.04 \rightarrow \frac{12.04-12}{.2/10} = \frac{.04}{.02} = 2$

$P(11.97 \le \bar{X} \le 12.04) = P(-1.5 \le Z \le 2)$

$P(\bar{X} \leq 11.97 \text{ or } \bar{X} \geq 12.04)$

Examples

Example: A hardware store receives a shipment of bolts that are supposed to be 12cm long. The mean is indeed 12cm, and the standard deviation is 0.2cm. For quality control, the hardware store chooses 100 bolts at random to measure. They will call the shipment defective and return it to the manufacturer if the average length of the 100 bolts is less than 11.97cm or greater than 12.04cm. Find the probability that the shipment is found satisfactory.

$$1 - P(11.97 \leq \bar{X} \leq 12.04) = 1 - P(-1.5 \leq Z \leq 2)$$

$$= 1 - (\Phi(2) - \Phi(-1.5))$$

$$= 1 - (.977250 - .066811)$$

$$= 1 - (0.910439) = 0.0896$$

Examples

Example: Suppose you have a jar of lemon and banana jelly beans where it is known that the true proportion of lemon jelly beans is 0.5. You try to estimate the proportion of lemon beans by reaching in and drawing 50 jelly beans and testing them (by eating them). What is the probability that your sample is 75% or more lemon jelly beans?

The CLT and Monte Carlo Simulation

Rolling 6-SIDED DIE

$X_1 = 5 \quad X_2 = 6, \quad X_3 = 1, \quad X_4 = 3, \quad X_5 = 3$

$$\bar{X} = \frac{1}{n}\left[X_1 + X_2 + \cdots + X_n\right]$$

CODE SPITS THIS

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

$\frac{\sigma}{\sqrt{n}}$ = STANDARD ERROR

The CLT and Monte Carlo Simulation

The CLT and Monte Carlo Simulation

OK! Let's Go to Work!

Get in groups, get out laptop, and open the Lecture 13 In-Class Notebook

Let's:

- See the Central Limit Theorem in Action!
- Experiment with estimating the mean of Bin(n,p) with various parameters
- See how we can use the CLT to estimate the accuracy of our simulations