Lecture 13: The Central Limit Theorem

Announcements and reminders

- Practicum 1 due today by 11:59 PM
- HW 3 to be posted later!

  And due Monday 18 March (2 weeks)

Previously, on CSCI 3022…

Definition: 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 $X \sim N(\mu, \sigma^2)$.

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

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:

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

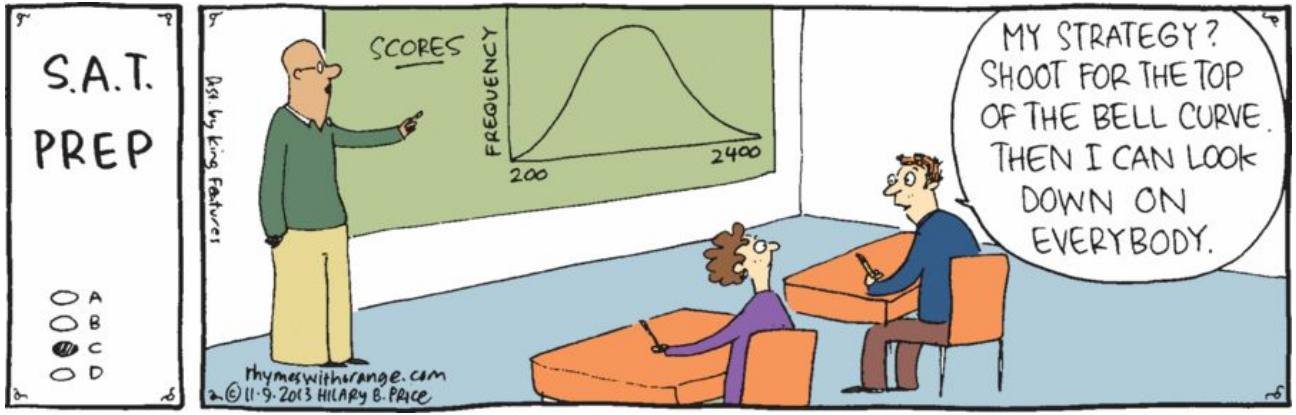
Fact: If Z is a standard normal random variable, then we can compute probabilities using the standard normal cdf

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

Motivating Example

Soon, we'll talk about **statistical inference**, wherein we'll infer things about the true mean of a population using sample data sets.

Examples:

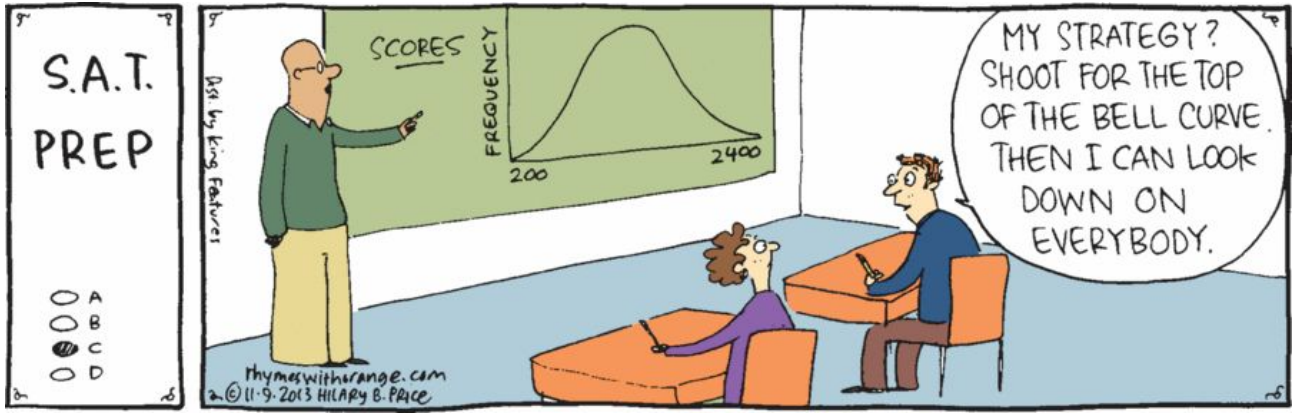


Motivating Example

Soon, we'll talk about **statistical inference**, wherein we'll infer things about the true mean of a population using sample data sets.

Examples:

- Mean GPA of all CS students -- sample of 30 students
- Mean weight of all puppies -- sample of a bunch of puppies at a shelter?
- Political polling



Random samples

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

1. _____

2. _____

We say these X_k's are _____

Random samples

Definition: 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 the **same distribution**

We say these X_k's are **independent & identically distributed** → "**iid**"

Estimators and their distributions

We use **estimators** to summarize our iid sample

Examples:

1.

2.

3.

Estimators and their distributions

We use **estimators** to summarize our iid sample

Examples:

1. $\bar{x}$ is the sample mean estimator of the population mean μ

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

3. s^2 is the sample estimator for σ^2

Estimators and their distributions

We use **estimators** to summarize our iid sample

Fun fact: Any estimator, including the **sample mean** $\bar{X}$, 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)

2)

3)

Estimators and their distributions

We use **estimators** to summarize our iid sample

Fun fact: Any estimator, including the **sample mean** $\bar{X}$, 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

But what does this sampling distribution actually look like?

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

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

Distribution of the sample mean

But what does this sampling distribution actually look like?

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

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

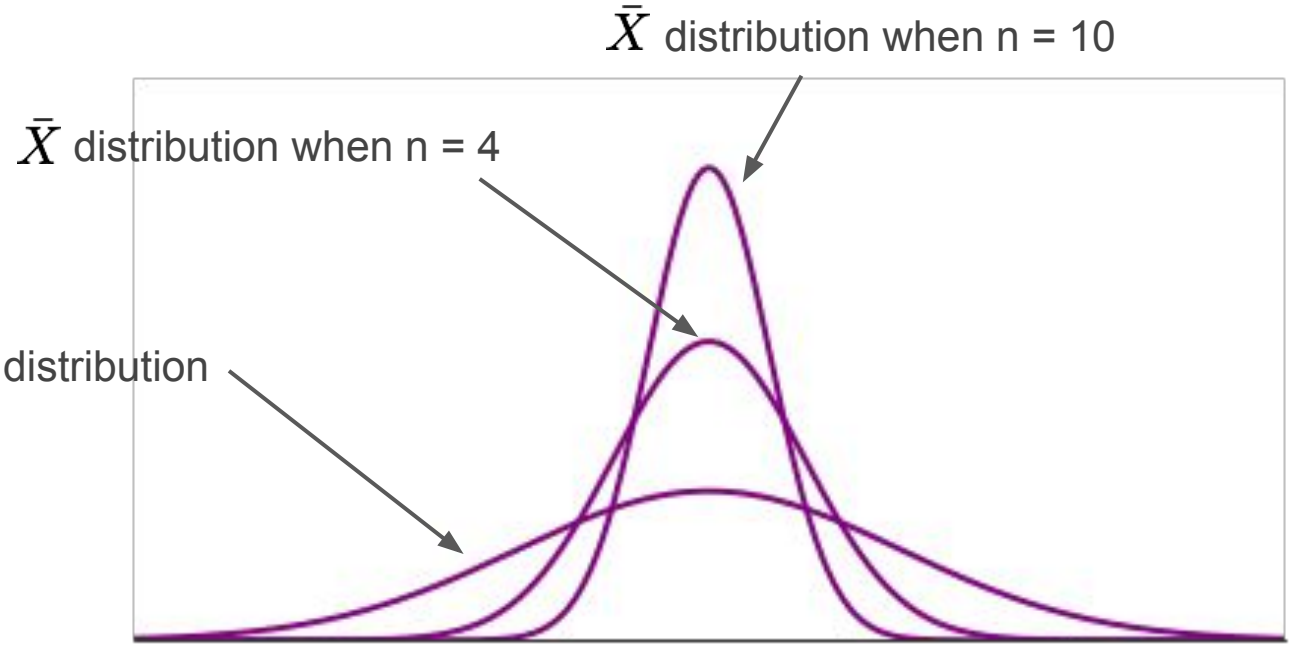
We know everything there is to know about the distribution of the sample mean when the population distribution is normal.

$$E[\overline{X}] = E[\frac{1}{n}\sum_{k=1}^{n} X_k] = \frac{1}{n}\sum_{k=1}^{n} E[X_k] = \frac{1}{n}\sum_{k=1}^{n} \mu = \mu$$

$$Var(\overline{X}) = \frac{1}{n^2} Var(\sum_{k=1}^{n} X_k) = \frac{1}{n^2}\sum_{k=1}^{n} Var(X_k) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

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?!

Distribution of the sample mean

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

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

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

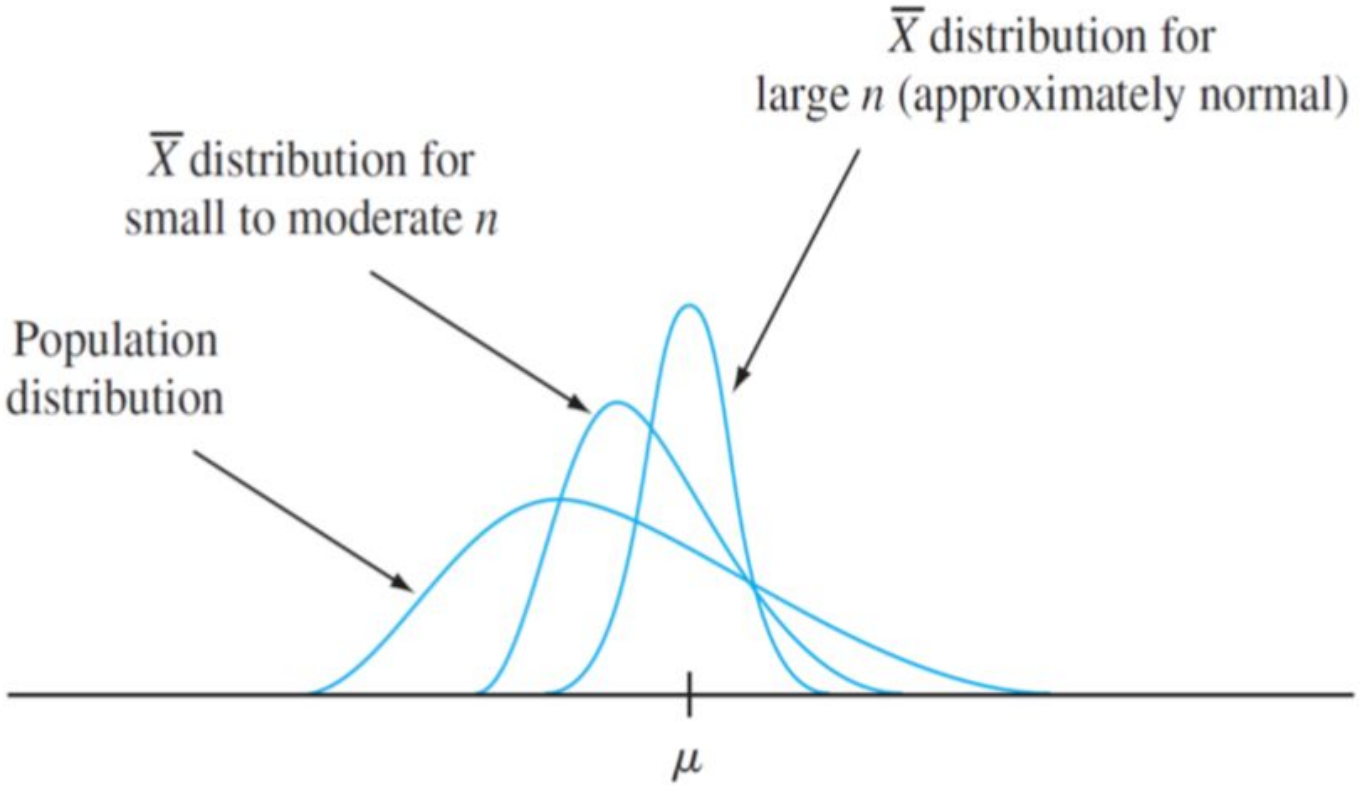
The Central Limit Theorem: Let $X_1, X_2, \ldots, X_n$ be iid draws from some distribution. Then, as n becomes large

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

Rule of Thumb: $n \geq 30$

Distribution of the sample mean

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



Examples

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 the 100 bolts is less than 11.97 cm or greater than 12.04 cm. Find the probability that the shipment is found satisfactory.

Examples

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 the 100 bolts is less than 11.97 cm or greater than 12.04 cm. Find the probability that the shipment is found satisfactory.

$$\text{Want } P(\bar{X} \leq 11.97 \text{ or } \bar{X} \geq 12.04) = 1 - P(11.97 \leq \bar{X} \leq 12.04)$$

$$\bar{X} \sim N\left(12, \frac{0.2^2}{100}\right)$$

$$X = 11.97 \longrightarrow Z = \frac{11.97 - 12}{0.2/10} = \frac{-0.3}{0.2} = -1.5$$

$$X = 12.04 \longrightarrow Z = \frac{12.04 - 12}{0.2/10} = \frac{0.4}{0.2} = 2$$

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

Examples

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 the 100 bolts is less than 11.97 cm or greater than 12.04 cm. Find the probability that the shipment is found satisfactory.

$$\begin{aligned} 1 - P(11.97 \leq \bar{X} \leq 12.04) &= 1 - P(-1.5 \leq Z \leq 2) \\ &= 1 - (\Phi(2) - \Phi(-1.5)) \\ &= 1 - (0.977250 - 0.066807) \\ &= 1 - 0.910443 \\ &= 0.089557 \end{aligned}$$

Examples

Example: S'pose you have a jar of lemon and banana jelly beans, and 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?

Examples

Example: S'pose you have a jar of lemon and banana jelly beans, and 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?

Population: p = 0.5
Sample: n = 50, and $\hat{p} = \ ?\ $ = (# successes) / n

And # successes out of n is… Binomial!! Bin(n, p)

Mean: $\bar{X} = \hat{p} = \dfrac{Bin(n,p)}{n}$

Variance: $Var(\hat{p}) = Var\left(\dfrac{Bin(n,p)}{n}\right) = \dfrac{1}{n^2}Var(Bin(n,p))$

$$= \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$$

The CLT and Monte Carlo Simulation

S'pose X is a random variable, in {0, 1, 2, …, 9}, but we do not know the mean and probabilities associated with each number. But! We have a sampler that spits out samples for X.

(Real life example: slot machines)

Sample: $X_1 = 5$, $X_2 = 1$, $X_3 = 8$, $X_4 = 4$, $X_5 = 5$, $X_6 = 2$, $X_7 = 9$, … Python: $X = [X_1, X_2, X_3, \ldots X_n]$

The CLT and Monte Carlo Simulation

S'pose X is a random variable, in {0, 1, 2, …, 9}, but we do not know the mean and probabilities associated with each number. But! We have a sampler that spits out samples for X.

(Real life example: slot machines)

Sample: $X_1 = 5$, $X_2 = 1$, $X_3 = 8$, $X_4 = 4$, $X_5 = 5$, $X_6 = 2$, $X_7 = 9$, … Python: $X = [X_1, X_2, X_3, \ldots X_n]$

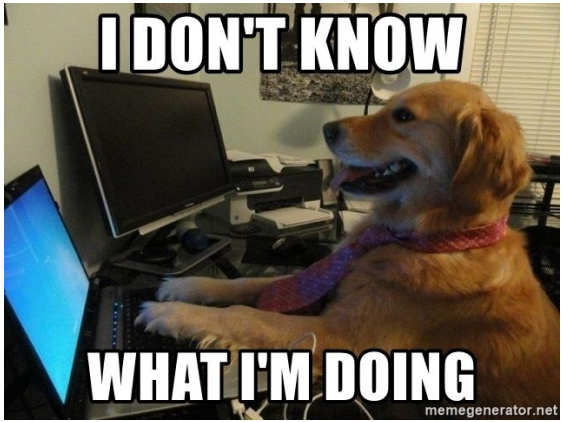
$$\bar{X} = \frac{1}{n}[x_1 + x_2 + \ldots + x_n]$$ ← Python: mean(X)

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

Standard error = $\frac{\sigma}{\sqrt{n}}$

Problem-solving hints

- **First,** identify the population and identify the sample.
- **Second,** is the problem about means or proportions?
- **Finally,** we're off to the races using the CLT, Box-Muller transform, and the handy-dandy standard normal distribution!



What just happened?

- “iid” samples
- Distribution of **samples** vs distribution of **sample means**
- **Central Limit Theorem**