Stat 104: Quantitative Methods

Class 19: The Central Limit Theorem for Means

Stat 104 - Roadmap

Data

Intro
- Population vs. sample
- Parameters vs. statistics

Graphs
- Dotplots and Histograms

Descriptive Stats
- Central tendency
- Variability
- Relative standing
- 2 variable stats

Basic Probability
- Marginal prob.
- Conditional probability
- Laws of probability
- Probability tables

Inference

Discrete Distributions
- Random variable
- E[X], Var[X] Binomial
- Joint distributions

Continuous Distributions
- Density functions
- Uniform distribution
- Normal distribution

Central Limit Theorem
- $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$
- $\hat{P} \sim N\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right)$

Estimation
- Point estimators
- Confidence intervals
- Levels of confidence: $1-\alpha$

Analysis

Hypothesis Testing
- Testing parameters: mu and pi
- Left tailed, right tailed, two-tailed
- Rejection regions approach
- P-values
- Levels of significance
- Type I and Type II errors

Two Sample Tests
- Compare mean across two populations
- Compare two proportions

Linear Regression
- Least squares lines: scatterplot
- Multivariate regression
- Dummy variables: 0/1
- Regression Diagnostics

(preview) Example CLT Problem

- Suppose a population has mean μ = 8 and standard deviation σ = 3. Suppose a random sample of size n = 36 is selected.
- What is the probability that the sample mean is between 7.8 and 8.2?

CLT let’s you answer questions about X’ i.e. sample mean

Wait! Lets ask a different question

- What is the probability that a single observation is between 7.8 and 8.2?

Now this is asking about X, not X’

Do we know if it’s a normal distribution?

Do we know if it’s a binomial distribution?

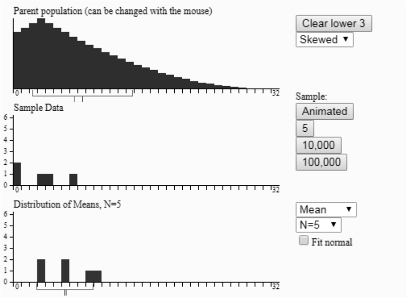
With the information we are given, we can only answer questions about the **sample mean** (because of the clt).

Population to Sample to Population

- We usually never have enough time and/or money to study the entire population that interests us.
- Fortunately we can learn a lot about a population by taking a random sample.
- Starting with this section, we being the process of relating sample statistics to the summary values of the population that produced our data.

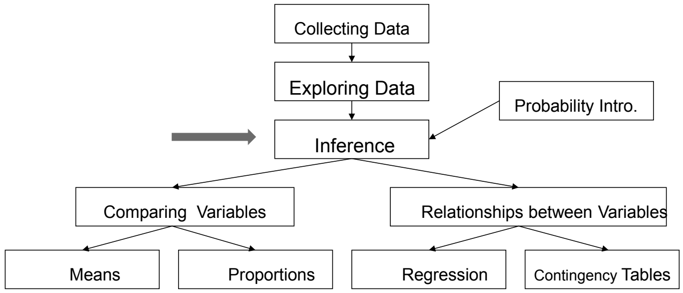
Visual Demo

- We find that many students find the clt hard to grasp so we try several demos to try to make it clearer.



http://onlinestatbook.com/stat_sim/sampling_dist/

Class Roadmap

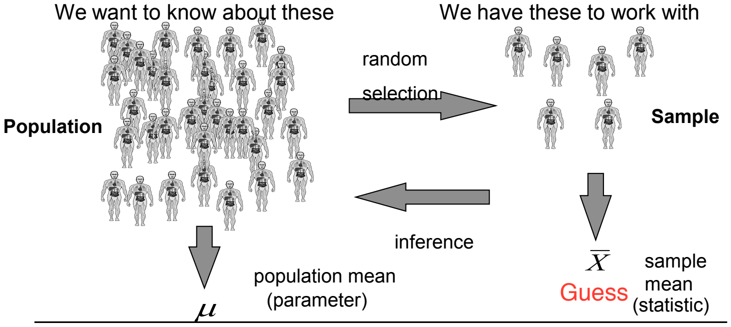


Population Parameters and Sample Statistics

Population parameter	*Value*	*Sample statistic used to estimate*
p *proportion of population with a certain characteristic*	*Unknown*	$\hat{p}$
μ *mean value of a population variable*	*Unknown*	$\overline{x}$

- The value of a population parameter is a **fixed** number, it is NOT random; its value is **not known.**
- The value of a sample statistic is calculated from sample data
- The value of a sample statistic will vary from sample to sample (sampling distributions)

The Inference Setup:



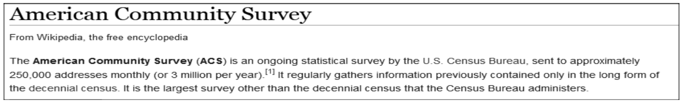
True but unknown

Introduction

- Suppose the government wanted to determine the mean income of all U.S. households.
- One approach the government could take is to literally survey each U.S. household to determine the population mean, μ. This would be a very expensive and time-consuming survey!
- A second approach the government could (and does) take is to survey a random sample of U.S. households and use the results to estimate the mean household income.
- This is done through the American Community Survey, which is administered to approximately 250,000 randomly selected households each month.

American Community Survey

AMERICAN COMMUNITY SURVEY
U.S. CENSUS BUREAU



American Community Survey

From Wikipedia, the free encyclopedia

The **American Community Survey** (**ACS**) is an ongoing statistical survey by the U.S. Census Bureau, sent to approximately 250,000 addresses monthly (or 3 million per year).[1] It regularly gathers information previously contained only in the long form of the decennial census. It is the largest survey other than the decennial census that the Census Bureau administers.

Question Corner
for Survey Respondents

Q Can I respond to the survey online?

Yes. Most people can respond to the American Community Survey online. If you received a letter or postcard inviting you to complete the ACS online, you will need the materials to begin. More.

☎ **Call Us**

General 1-800-923-8282

What is the American Community Survey?

The American Community Survey (ACS) is a mandatory, ongoing statistical survey that samples a small percentage of the population every year -- giving communities the information they need to plan investments and services.

How do I respond to the survey?

Learn ways to respond to the ACS or get help with the survey. Learn more about how we protect your privacy; why you were selected; why it's important to participate; why we ask specific questions and more in About the Survey.

How do I get started using ACS data?

We release new data every year — get the latest on American FactFinder, or get advice on choosing the right tool or data table for your needs. Learn more about our annual data releases or browse the supporting documentation.

American Community Survey

- For example, in 2009 the mean annual household income in the United States was estimated to be $\overline{x}$ = *67,976.*
- *The government might* infer from this survey that the mean annual household income of *all U.S. households in* 2009 was μ = 67,976.
- But they might be wrong……

X' is random and depends on the sample

Variability of Estimates

- The households in the American Community Survey were determined by chance (random sampling).
- A second random sample of households would likely lead to a different sample mean, such as $\bar{x}$ = *67,731,*
- *and a third random sample of households* would likely lead to a third sample mean, such as $\bar{x}$ = *67,978.*

Crucial Point

- *Because the households* selected will vary from sample to sample, the sample mean of household income will also vary from sample to sample.
- For this reason, the sample mean $\bar{x}$ *is a random variable, so* it has a probability distribution.
- Our goal in this section is to describe the distribution of the sample mean.
- Remember, when we describe a distribution, we do so in terms of its shape, center, and spread.

The Main Idea Summary

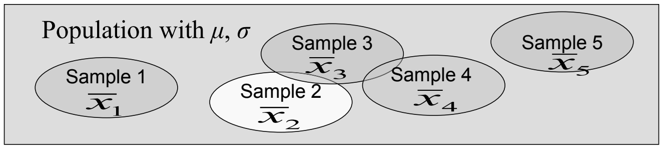
- Statistics such as $\bar{X}$ are random variables since their value varies from sample to sample.
- As such, they have probability distributions associated with them. In this section we focus on the shape, center and spread of $\bar{X}$**.**

Bunnies and Dragons

- Lets watch a short film

Bunnies, Dragons and the 'Normal' World: Central Limit Theorem | The New York Times

Sampling Distribution of Sample Means



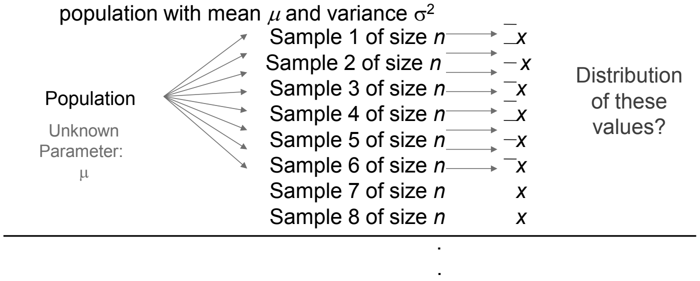
The sampling distribution consists of the values of the sample means, $\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4, \bar{x}_5, ...$

The Sampling Distribution

- A sampling distribution is a distribution of all of the possible values of a statistic for a given size sample selected from a population.
- Once we know the sampling distribution of $\bar{x}$, we will be able to tell how far our guess $\bar{x}$, is from μ, without even knowing μ. Freaky!

Sampling Distribution of Sample Mean

- Distribution of values taken by statistic in all possible samples of size *n* from the same population
- Model assumption: our observations x_i are sampled from a population with mean μ and variance σ^2



Population

Unknown Parameter: μ

Sample 1 of size *n* → $\bar{x}$

Sample 2 of size *n* → $\bar{x}$

Sample 3 of size *n* → $\bar{x}$

Sample 4 of size *n* → $\bar{x}$

Sample 5 of size *n* → $\bar{x}$

Sample 6 of size *n* → $\bar{x}$

Sample 7 of size *n* x

Sample 8 of size *n* x

.

.

Distribution of these values?

Mean of Sample Mean

- First, we examine the **center** of the sampling distribution of the sample mean.
- Center of the sampling distribution of the sample mean is the unknown population mean:

The expcted value of X' is u

mean($\bar{X}$) = μ

- Over repeated samples, the sample mean will, *on average*, be equal to the population mean
  - ❑ ***no guarantees for any one sample!***

Variance of Sample Mean

- Next, we examine the **spread** of the sampling distribution of the sample mean
- The variance of the sampling distribution of the sample mean is

variance($\bar{X}$) = σ^2/n

- As sample size increases, variance of the sample mean decreases!
  - Averaging over many observations is more accurate than just looking at one or two observations

Law of Large Numbers

- The Law of Large Numbers:
  - If one draws independent samples from a population with mean μ, then as the number of observations increases, the sample mean x gets closer and closer to the population mean μ
- This is easy to see since we know that

mean($\bar{X}$) = μ

variance($\bar{X}$) = σ^2/n $\longrightarrow$ 0 as n gets large

The Central Limit Theorem

The CLT states that if random samples of size n are repeatedly drawn from **any** population with mean μ and variance σ^2, then **when n is large**, the distribution of the sample means will be approximately normal :

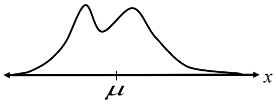
$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

Is distributed as

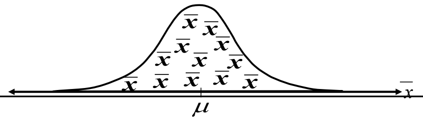
If the population is normal this is true for any sample size.

The Central Limit Theorem

1. If samples of size $n \geq 30$ are drawn from **any** population with mean = μ and standard deviation = σ,

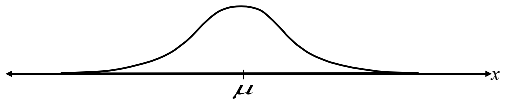


then the sampling distribution of sample means approximates a normal distribution. The greater the sample size, the better the approximation.

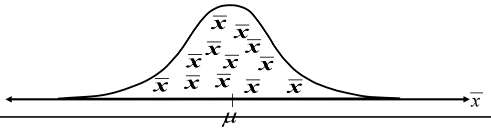


The Central Limit Theorem

2. If the population itself is normally distributed,



then the sampling distribution of sample means is normally distribution for ***any*** sample size *n*.



The Central Limit Theorem

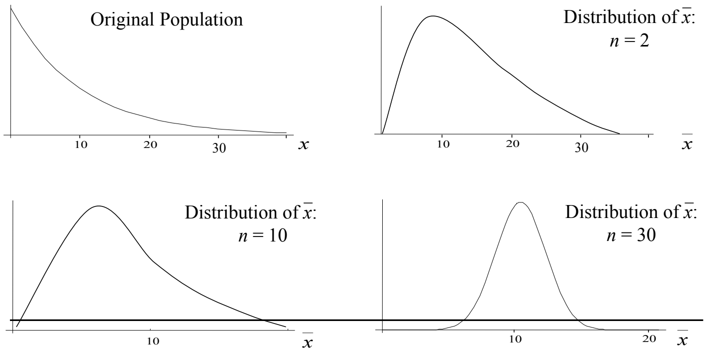
- In either case, the sampling distribution of sample means has a mean equal to the population mean.

$\mu_{\bar{x}} = \mu$ Mean

- The sampling distribution of sample means has a standard deviation equal to the population standard deviation divided by the square root of *n.*

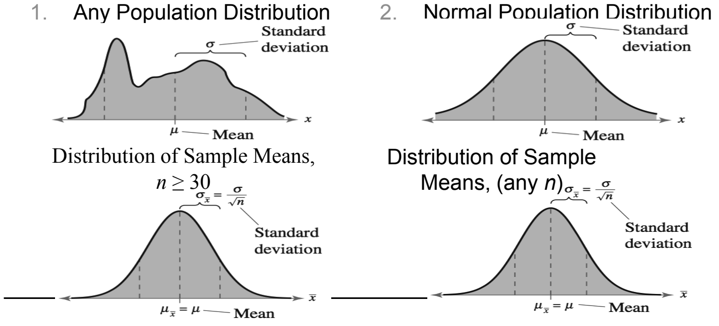
$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ Standard deviation (**standard error of the mean)**

The Central Limit Theorem



The Central Limit Theorem

1. Any Population Distribution
2. Normal Population Distribution



Distribution of Sample Means, $n \geq 30$

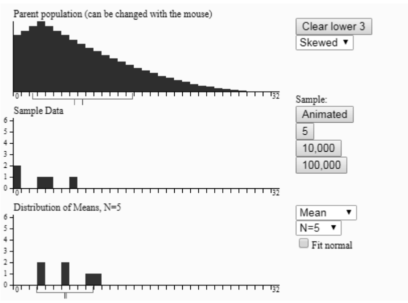
Distribution of Sample Means, (any *n*)

This is a really cool and powerful result. It says that **no matter what** our initial data looks like, when we take **averages** we end up with the normal distribution.

This result is so useful because we usually don't know where our data comes from-i.e. what the shape of the underlying true distribution looks like. The CLT theorem says that as long as we work with <u>averages</u>, it doesn't matter.

Visual Demo

- We find that many students find the clt hard to grasp so we try several demos to try to make it clearer.



http://onlinestatbook.com/stat_sim/sampling_dist/

Simulation

- One reason the CLT is a difficult concept to grasp is it deals with the theoretical idea of repeated sampling.
- That is, what does the distribution of $\bar{X}$ look like if we took many samples, calculating the mean of each sample.

Simulation

- Simulation makes it easier to see this repeated sampling concept.
- In simulation we generate random data that mimics the idea of repeated sampling.

Example: Tossing Dice

- Throw a fair die
- Random variable X=# spots that appear

x	1	2	3	4	5	6
P(x)	1/6	1/6	1/6	1/6	1/6	1/6

- One can show that E(X)=3.5 and Var(X)=2.92. StdDev(X) = 1.71.

One Dice Roll in R

- We use the `sample` command in R

```
> sample(1:6,1,replace=TRUE)
[1] 6
> sample(1:6,1,replace=TRUE)
[1] 4
> sample(1:6,1,replace=TRUE)
[1] 3
> sample(1:6,1,replace=TRUE)
[1] 1
> sample(1:6,1,replace=TRUE)
[1] 1
> sample(1:6,1,replace=TRUE)
[1] 3
```

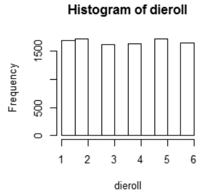
Roll one die 1000 times

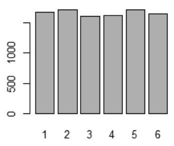
```
dieroll=sample(1:6,10000,replace=TRUE)

> dieroll[1:100]
  [1] 6 6 4 4 1 4 2 1 4 1 4 6 5 5 6 3 3 4 3 4 6 3 1 2 5 6 3 1 6 6 6 6 2 5 2 2 6
 [38] 6 2 5 6 2 4 1 4 6 3 4 2 4 5 3 4 2 1 5 5 3 2 5 6 5 3 1 4 2 2 2 2 3 2 3 6 2
 [75] 4 1 5 4 5 1 4 4 1 1 3 4 3 5 6 1 2 4 4 6 4 1 6 3 2 2

> mean(dieroll)
[1] 3.4923
> sd(dieroll)
[1] 1.71261
```

What does theory say these values should be?

The Histogram

- Histogram sucks (breaks are weird) so we use barplot.





```
> hist(dieroll)
> barplot(table(dieroll))
```

Now 2 dice

- Throw two dice and take the average

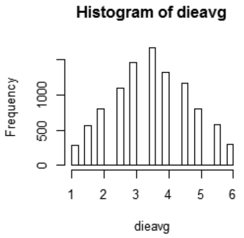
```
> die1=sample(1:6,10000,replace=TRUE)
> die2=sample(1:6,10000,replace=TRUE)
> dieavg=(die1+die2)/2
> mean(dieavg)
[1] 3.49555
> sd(dieavg)
[1] 1.200791
```

Note that (so what??)

```
> 1.71/sqrt(2)
[1] 1.209153
```

Histogram of two dice average

- What does it look like? Where is it centered?

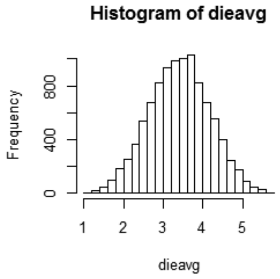


Now do average of 5 dice

- Roll 5 dice and take the average

```
> die1=sample(1:6,10000,replace=TRUE)
> die2=sample(1:6,10000,replace=TRUE)
> die3=sample(1:6,10000,replace=TRUE)
> die4=sample(1:6,10000,replace=TRUE)
> die5=sample(1:6,10000,replace=TRUE)
> dieavg=(die1+die2+die3+die4+die5)/5
> mean(dieavg)
[1] 3.5027
> sd(dieavg)
[1] 0.7610484
> 1.71/sqrt(5)
[1] 0.7647352
```

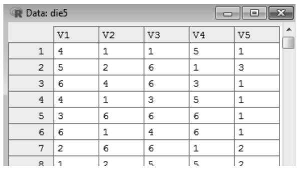
The 5 dice histogram



Some R Tricks

- We don’t have to simulate the dice individually. Examine the following code:

```
> die5=matrix(nrow=10000,ncol=5,sample(1:6,50000,replace=TRUE))
> View(die5)
```



	V1	V2	V3	V4	V5
1	4	1	1	5	1
2	5	2	6	1	3
3	6	4	6	3	1
4	4	1	3	5	1
5	3	6	6	6	1
6	6	1	4	6	1
7	2	6	6	1	2
8	1	2	5	5	2

Each row
represents 5
rolls of a die

Finding the average of the rolls

- The `apply` command finds the mean of each row of our matrix

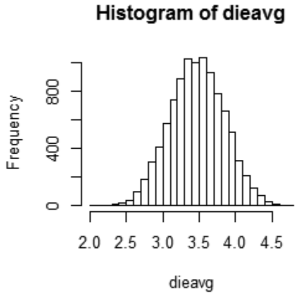
```
> die5=matrix(nrow=10000,ncol=5,sample(1:6,50000,replace=TRUE))
> dieavg=apply(die5,1,mean)

> mean(dieavg)
[1] 3.50614
> sd(dieavg)
[1] 0.7645579
> 1.71/sqrt(5)
[1] 0.7647352
```

Now do twenty dice

- Using the `apply` command makes it easier

```
> die20=matrix(nrow=10000,ncol=20,sample(1:6,200000,replace=TRUE))
> dieavg=apply(die20,1,mean)
> mean(dieavg)
[1] 3.494325
> sd(dieavg)
[1] 0.376201
> 1.71/sqrt(20)
[1] 0.3823676
```

The Histogram for 20 Dice



Recap: The Central Limit Theorem

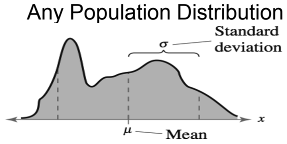
The CLT states that if random samples of size n are repeatedly drawn from **<u>any</u>** population with mean μ and variance σ^2, then **when n is large**, the distribution of the sample means will be approximately normal :

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

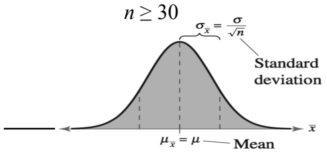
If the population is normal this is true for any sample size.

The Central Limit Theorem

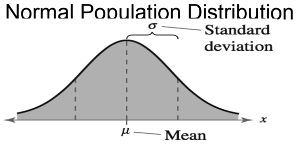
1. Any Population Distribution



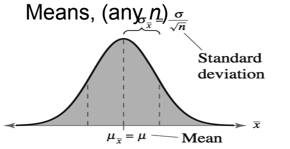
Distribution of Sample Means, $n \geq 30$



2. Normal Population Distribution



Distribution of Sample Means, (any n)



Note: Standardizing Sample Mean

- The sample mean can be standardized to the standard normal distribution using the following formulation:

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

Example of Using the CLT

- Suppose a population has mean μ = 8 and standard deviation σ = 3. Suppose a random sample of size n = 36 is selected.

- What is the probability that the sample mean is between 7.8 and 8.2?

Wait! Lets ask a different question

- What is the probability that a single observation is between 7.8 and 8.2?

Do we know if it's a normal distribution?

Do we know if it's a binomial distribution?

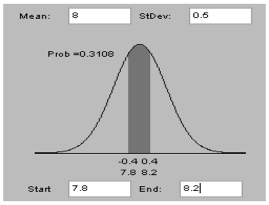
With the information we are given, we can only answer questions about the **sample mean** (because of the clt).

So: What is the probability that the **sample mean** is between 7.8 and 8.2?

- Even if the population is not normally distributed, the central limit theorem can be used (n > 30)
- … so the sampling distribution of $\bar{x}$ is approximately normal
- … with mean $\mu_{\bar{x}} = 8$
- and standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{36}} = 0.5$

Solution (cont)

$$P(7.8 < \bar{X} < 8.2) = P\left(\frac{7.8\text{-}8}{3/\sqrt{36}} < \frac{\bar{X}\text{-}\mu}{\sigma/\sqrt{n}} < \frac{8.2\text{-}8}{3/\sqrt{36}} \right)$$
$$= P(-0.4 < Z < 0.4) = 0.3108$$



Mean: 8 StDev: 0.5

Prob =0.3108

-0.4 0.4
7.8 8.2

Start 7.8 End: 8.2

Recap

- Unless you are explicitly told the distribution of a random variable X, there is no way to evaluate P(a<X<b).
- However, without needing to know the underlying distribution, if *n* is sufficiently large, the CLT allows one to evaluate $P(a < \bar{X} < b)$
- This is a powerful result.

Example

- Suppose that the mean time for an oil change at a “10-minute oil change joint” is 11.4 minutes with a standard deviation of 3.2 minutes.
- If a random sample of *n* = 35 oil changes is selected, what is the probability the mean oil change time is less than 11 minutes?

Example

- We want to find $P(\bar{X}<11)$.
- By the Central Limit Theorem

$$\bar{X} \sim N(11.4, (3.2)^2 / 35)$$

- So the answer is

```
> pnorm(11,11.4,3.2/sqrt(35))
        [1] 0.2297987
```

Another CLT Example

- We want to determine the true population mean tread life of a brand of tires.
- We will sample 100 tires.
- What is the probability that the sample mean will be within 300 miles of the population mean?
- So we want to know how good our guess is in a fashion.

CLT Example

- Assume we know σ=2000 miles.
- We want to find

$$\mathrm{P}(\mu-300<\ \bar{\mathrm{X}}\ <\mu+300\ )$$

- From the CLT we know that

$$\bar{\mathrm{X}}\sim\mathrm{N}(\mu,40000).$$

CLT Example

- So

$$\mathrm{P}(\mu-300<\ \bar{\mathrm{X}}\ <\mu+300\ )$$

$$=\mathrm{P}\left(\frac{\mu-300-\mu}{40000}<\frac{\bar{\mathrm{X}}\text{-}\mu}{40000}<\frac{\mu+300-\mu}{40000}\right)$$

$$=P(-1.5<Z<1.5)$$

$$=0.866$$

Wait, what did we just do??

- Study what we just did at home.
- We are able to determine how far our guess is from the true value, without needing to even know what the true μ is.
- Pretty tricky and powerful.
- We will use this in practice when we cover confidence intervals.

Example of using the CLT

- The service times for customers coming through a checkout counter in a retail store are independent random variables with a mean of 1.5 minutes and a variance of 1.0.
- Approximate the probability that 88 customers can be serviced in less than 2 hours of total service time by this one checkout counter.

Example

- We wish to find

$$P(\sum_{i=1}^{88} X_i < 120)$$

- If we divide both sides by 88 we obtain

$$P(\bar{X}<1.36)$$

- From the Central Limit Theorem we know

$$\bar{X}\sim N(1.5,\frac{1}{88})$$

Example

- Then

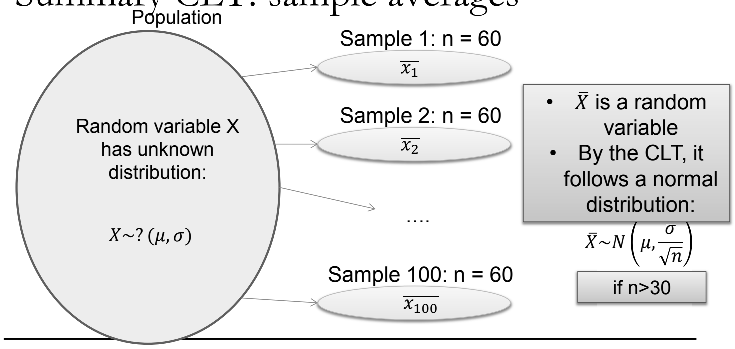
$$\begin{aligned} P(\bar{X} < 1.36) &= P\left(\frac{\bar{X} - 1.5}{.1066} < \frac{1.36 - 1.5}{.1066}\right) \\ &= P(Z < -1.313) \\ &= 0.095 \end{aligned}$$

- So there is about a 10% chance 88 customers can be serviced within 2 hours.

How Large is Large Enough?

- For most distributions, n > 30 will give a sampling distribution that is nearly normal
- For fairly symmetric distributions, n > 15
- For normal population distributions, the sampling distribution of the mean is always normally distributed

Summary CLT: sample averages



Population

Random variable X has unknown distribution:

$X \sim ?\,(\mu, \sigma)$

Sample 1: n = 60

$\overline{x_1}$

Sample 2: n = 60

$\overline{x_2}$

….

Sample 100: n = 60

$\overline{x_{100}}$

- $\bar{X}$ is a random variable
- By the CLT, it follows a normal distribution:

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

if n>30

Things you should know

❑Sampling distribution of the sample mean