

## Stat 104: Quantitative Methods

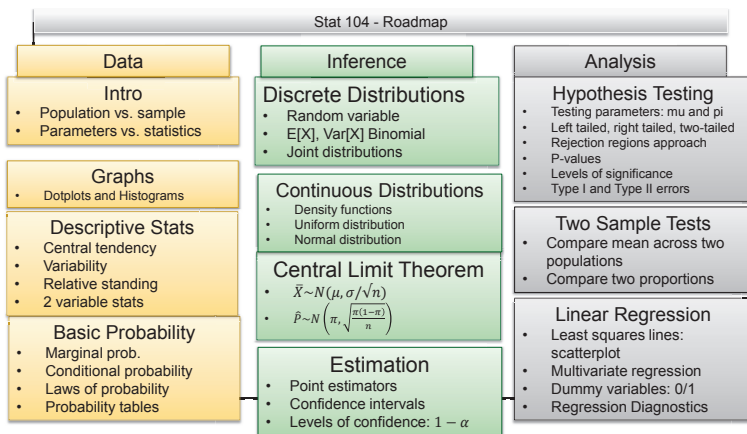
Class 20: More on the Central Limit Theorem

## Recap: The Central Limit Theorem

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

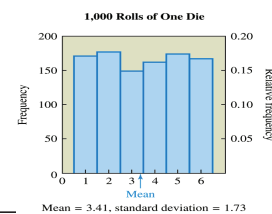
$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

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



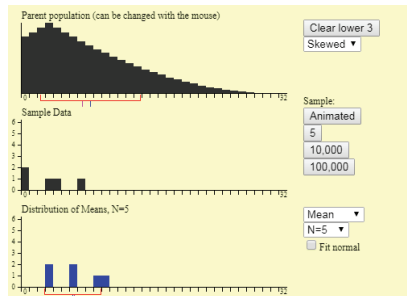
## Visualizing the Central Limit Theorem Using Dice

Suppose we roll *one* die 1,000 times and record the outcome of each roll, which can be the number 1, 2, 3, 4, 5, or 6.



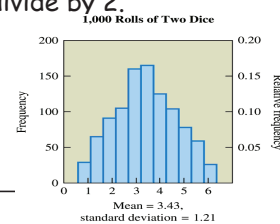
## Visual Demo

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



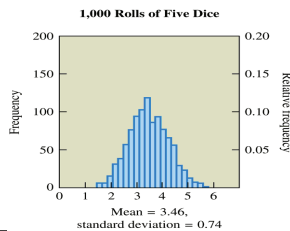
## Visualizing the Central Limit Theorem Using Dice

Now suppose we roll *two* dice 1,000 times and record the *mean* of the two numbers that appear on each roll. To find the mean for a single roll, we add the two numbers and divide by 2.



## Visualizing the Central Limit Theorem Using Dice

Suppose we roll *five* dice 1,000 times and record the mean of the five numbers on each roll.

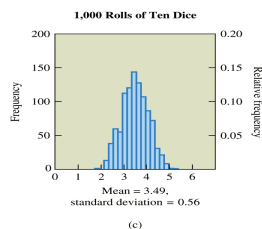


## The Central Limit Theorem

1. The distribution of means will be approximately a normal distribution for larger sample sizes
2. The mean of the distribution of means approaches the population mean,  $\mu$ , for large sample sizes
3. The **standard deviation of the distribution of means** approaches  $\sigma/\sqrt{n}$  for large sample sizes, where  $\sigma$  is the standard deviation of the population and  $n$  is the sample size

## Visualizing the Central Limit Theorem Using Dice

Now we will further increase the number of dice to *ten* on each of 1,000 rolls.



## The Central Limit Theorem Side Notes

1. For practical purposes, the distribution of means will be nearly normal if the sample size is larger than 30
2. If the original population is normally distributed, then the sample means will remain normally distributed for *any* sample size  $n$ , and it will become narrower
3. The original variable can have any distribution, it does not have to be a normal distribution

## Visualizing the Central Limit Theorem Using Dice

Number of dice rolled each time	Mean of the distribution of means	Standard deviation of the distribution of means
1	3.41	1.73
2	3.43	1.21
5	3.46	0.74
10	3.49	0.56

What do you notice about the shape of the distribution as the sample size increases?

**It approximates a normal distribution**

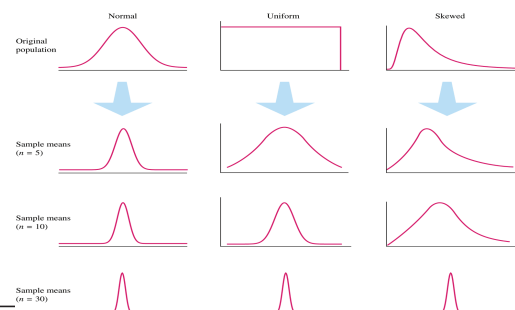
What do you notice about the mean of the distribution of sample means as the sample size increases in comparison to the true mean of the population (3.5)?

**It approaches the population mean**

What do you notice about the standard deviation of the distribution of means as the sample size increases?

**It gets smaller representing a lower variation**

## Shapes of Distributions as Sample Size Increases



## Example: Predicting Test Scores

You are a middle school principal and your 100 eighth-graders are about to take a national standardized test. The test is designed so that the mean score is  $\mu = 400$  with a standard deviation of  $\sigma = 70$ . Assume the scores are normally distributed.

- a. What is the likelihood that one of your eighth-graders, selected at random, will score below 375 on the exam?

```
> pnorm(375,400,70)
[1] 0.3604924
```

There is a 36% chance that a randomly selected student will score below 375

## Example: Predicting Test Scores

You are a middle school principal and your 100 eighth-graders are about to take a national standardized test. The test is designed so that the mean score is  $\mu = 400$  with a standard deviation of  $\sigma = 70$ . Assume the scores are normally distributed.

- b. Your performance as a principal depends on how well your entire group of eighth-graders scores on the exam. What is the likelihood that your group of 100 eighth-graders will have a mean score below 375?

According to the C.L.T. if we take random groups of say 100 students and study their means, then the means distribution will approach normal. Hence, the  $\mu = 400$  and its standard deviation is  $\sigma/\sqrt{n} = 70/\sqrt{100} = 70/10 = 7$  according to the C.L.T.

```
> pnorm(375,400,7)
[1] 0.0001775197
```

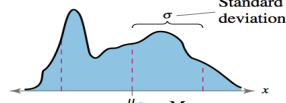
## Example

A soda filling machine is supposed to fill cans of soda with 12 fluid ounces. Suppose that the fills are actually normally distributed with a mean of 12.1 oz and a standard deviation of .2 oz. What is the probability that the average fill for a 6-pack of soda is less than 12 oz?

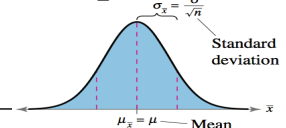
$$P(\bar{x} < 12) = P\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < \frac{12 - 12.1}{.2/\sqrt{6}}\right) = P(z < -1.22) = .1112$$

## The Central Limit Theorem

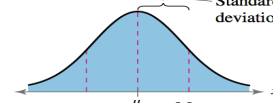
1. Any Population Distribution



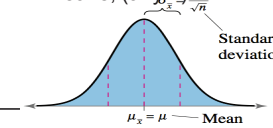
Distribution of Sample Means,  $n \geq 30$



2. Normal Population Distribution



Distribution of Sample Means, (any  $n$ )



## Example of Using the CLT

- Suppose a population has mean  $\mu = 8$  and standard deviation  $\sigma = 3$ . 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  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{36}} = 0.5$

19

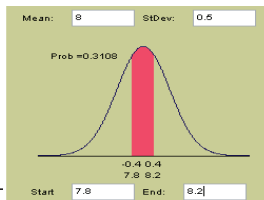
## Example

- 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?

22

## Solution (cont)

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



20

## 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
```

23

## 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

$$P(a < \bar{X} < b)$$

21

## 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.

24

- This is a powerful result.

## CLT Example

25

- Assume we know  $\sigma=2000$  miles.
- We want to find
$$P(\mu - 300 < \bar{X} < \mu + 300)$$
- From the CLT we know that
$$\bar{X} \sim N(\mu, 40000).$$

## Example of using the CLT

28

- 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.

## CLT Example

26

- So
$$P(\mu - 300 < \bar{X} < \mu + 300)$$
$$= P\left(\frac{\mu - 300 - \mu}{40000} < \frac{\bar{X} - \mu}{40000} < \frac{\mu + 300 - \mu}{40000}\right)$$
$$= P(-1.5 < Z < 1.5)$$
$$= 0.866$$

## Example

29

- We wish to find
$$P\left(\sum_{i=1}^{88} X_i < 120\right)$$
- If we divide both sides by 88 we obtain
$$P(\bar{X} < 1.36)$$
- From the Central Limit Theorem we know
$$\bar{X} \sim N\left(1.5, \frac{1}{88}\right)$$

## Wait, what did we just do??

27

- 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  $\mu$  is.
- Pretty tricky and powerful.
- We will use this in practice when we cover confidence intervals.

## Example

30

- Then
$$P(\bar{X} < 1.36) = P\left(\frac{\bar{X} - 1.5}{.1066} < \frac{1.36 - 1.5}{.1066}\right)$$
$$= P(Z < -1.313)$$
$$= 0.095$$
- So there is about a 10% chance 88 customers can be serviced within 2 hours.

## How Large is Large Enough?

31

- 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

## The CLT for Proportions

34

- If  $n$  is large, we obtain the following result:

$$\hat{P} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

Technical requirement  
 $n * p \geq 5$     $n * (1 - p) \geq 5$

## We need a CLT for Proportions

32

- The CLT we saw so far works for the sample mean of numerical questions.
- We also want a CLT that will work for the sample proportion of category questions

## Example

35

- The CLT is a thought experiment.
- In practice we only have one sample!
- The CLT tells us how variable our one sample could be.
- This will be very useful when we construct confidence intervals to know how close our guess is to the true population value.

33

A sample proportion can be thought of as a sample mean.

$$\hat{p} = \frac{4}{6}$$

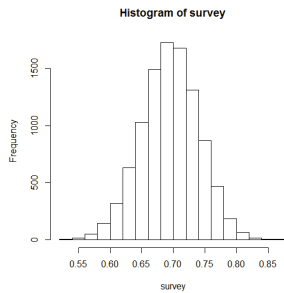
## Example

36

- Suppose the true population proportion of people who like Rocky Road Ice Cream is 70%.
- We sample 100 people and find 65 like rocky road
- Sample another 100 now 81 like
- Sample another 100 and now its 71
- And so on

## Histogram of Many Samples

37



We don't know exactly where the estimate from our **one** survey will end up, but the CLT theorem lets us determine how far we are from the true value.

## Example: Sample Proportions

40

- Suppose the proportion of voters that support the Democratic party in the US is 52% (this is  $p$ ).

What is the probability of selecting a sample of 100 voters, in which the proportion of Democrats is:

- Equal to 55%
- Less than 50%
- Between 40% and 52%

## The CLT Result

38

Instead of saying:

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

When dealing with proportions we say

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

## Example

41

- From the CLT we know that

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

$$\hat{p} \sim N\left(0.52, \frac{0.52(0.48)}{100}\right) = N(0.52, 0.002496)$$

## Example

A soda bottler claims that only 5% of the soda cans are under filled. A quality control technician randomly samples 200 cans of soda. What is the probability that more than 10% of the cans are under filled?

$$\begin{aligned} P(\hat{p} > .10) \\ &= P\left(z > \frac{.10 - .05}{\sqrt{\frac{.05(.95)}{200}}}\right) = P(z > 3.24) \\ &= 1 - .9994 = .0006 \end{aligned}$$

## P(sample proportion=55%)

42

- We have

$$\hat{p} \sim N(0.52, 0.002496)$$

- So  $P(\hat{p} = 0.55) = 0$  [why?]

P(sample proportion < 50%)

■ We have

$$\hat{p} \sim N(0.52, 0.002496)$$

■ So  $P(\hat{p} < 0.50) = P\left(Z < \frac{.5 - .52}{\sqrt{.002496}}\right) = 0.344$



Example: binge drinking

$$\hat{p} \sim N(.44, .032)$$

$$\begin{aligned} \text{So } P(\hat{p} \leq .36) &= P\left(\frac{\hat{p} - .44}{.032} \leq \frac{.36 - .44}{.032}\right) \\ &= P(z \leq -2.5) = .0062 \end{aligned}$$

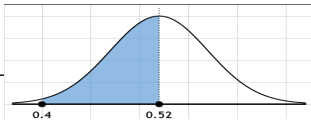
P(40% < sample proportion < 52%)

■ We have

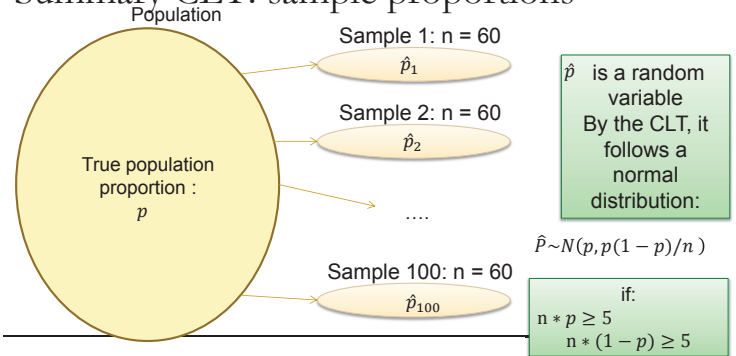
$$\hat{p} \sim N(0.52, 0.002496)$$

■ So

$$P(.4 < \hat{p} < 0.52) = P\left(\frac{.4 - .52}{\sqrt{.002496}} < Z < 0\right) = 0.4918$$



Summary CLT: sample proportions



Example: binge drinking

- Study by Harvard School of Public Health: 44% of college students binge drink.
- Assume the value 0.44 given in the study is the proportion  $p$  of college students that binge drink; that is 0.44 is the population proportion  $p$
- Compute the probability that in a sample of 244 students, 36% or less have engaged in binge drinking.

Recap-What is our goal?

- To get on the Owl's invite list?
- Explore new space frontiers?
- Get a job?
- Make new friends?



- No! our goal is to make **statistical inference**

We want to draw conclusions from **sample** data about the larger **populations** from which the samples are drawn.



## Recap-Terminology

- A **parameter** is a characteristic of a population. A **statistic** is a characteristic of a sample
- Inferential statistics enables you to make an educated guess about a population parameter based on a statistic computed from a random sample drawn from that population.

	Sample	Population
	Statistic	Parameter
Mean	$\bar{X}$	$\mu$
Proportion	$\hat{p}$	$\pi$
Variance	$s^2$	$\sigma^2$
Correlation	$r$	$\rho$

greek letters

49

## What kind of estimators do we want?

- Statisticians spend lots of time trying to develop estimators of parameters.
- In particular, there are two properties we want our estimators to have:
  - ☐ They should be unbiased
  - ☐ They should have minimum variance
- We'll discuss these concepts in turn.

52

## Examples

- The sample mean is a statistic used to estimate  $\mu$ :
 
$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$
- There could be many possible estimators!
- For example, the sample median is another statistic that could be used to estimate  $\mu$ .

50

## Unbiased Estimates

- What kind of estimator do we want ? Probably one that always gives a good approximation to the truth.
- Can we be right all the time ? Probably not. What about being correct **on average** ?
- Generically, let  $\hat{\theta}$  denote the estimate of some parameter  $\theta$ .
- The bias of  $\hat{\theta}$  is defined to be

$$\text{bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

- As estimator is called **unbiased** if its bias=0.

53


## Examples

- The sample variance is a statistic used to estimate  $\sigma^2$ .
 
$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$
- But one could also use the range, IQR or even the above divided by n to estimate  $\sigma^2$ .

$$s_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

51

## Example

- The sample mean is an unbiased estimate of the population mean.
- That is  $E(\bar{X}) = \mu$  
- But there are many other unbiased estimates of the population mean.
- We also want our estimators to have minimum variance.

54

## The Sample Variance

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

55

- Why do we divide by  $n-1$ ? So it is unbiased!
- It is a bit of work but it can be shown that

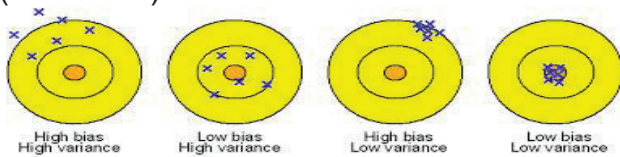
$$E(s^2) = \sigma^2.$$

- So in short, we divide by “ $n-1$ ” to make the sample variance an unbiased estimator.

## Bias-Variance Trade-off

56

- Given a choice, we want estimators with low (or no bias) and low variance.



### Things you should know

57

- ☐ Sampling distribution of the sample proportion
- ☐ Understand what it means to be unbiased
- ☐ Understand that in comparing unbiased estimators, we want the one with smaller variance.