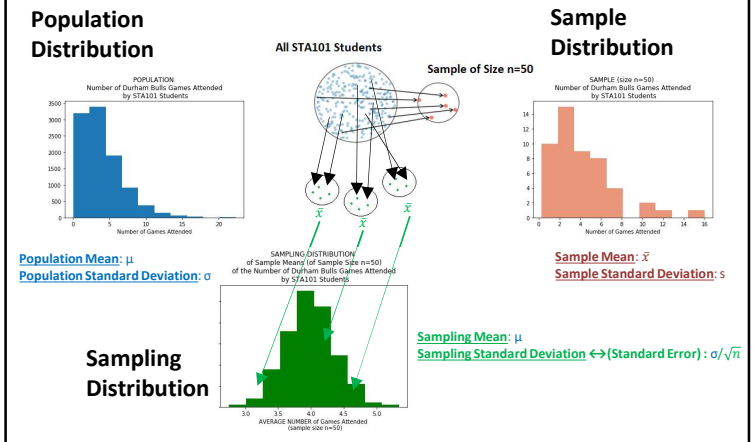
Unit 3: Foundations for inference
1. Variability in estimates and CLT

Sta 101 – Spring 2019

Duke University, Department of Statistical Science

Dr. Ellison

Slides posted at
https://www2.stat.duke.edu/courses/Spring19/sta101.001/



Outline

1. Housekeeping

2. Main ideas

   Problem: We need a probability distribution of values that a sample statistic can take on to assess how "extreme" (or likely) a sample statistic is (given an assumption about the population).

   One Solution: If you can, use a theoretical probability distribution of the sample statistic, using the Central Limit Theorem.

   1. Observation: ✋ 🎧 💻 Sample statistics vary from sample to sample
   2. Important Theorem: CLT describes the shape, center, and spread of sampling distributions
   3. Necessary Conditions: CLT only applies when independence and sample size/skew conditions are met

3. Summary

Coming up…

- Lab Assignment 3 is due **Thursday just before your lab section time.**
- Problem Set 2 due **Friday 2/8 11:55 pm**
- Performance Assessment 2 due **Sunday 2/10 11:55 pm** (opens today)
- Midterm 1 Review **Wednesday 2/13**
- Midterm 1 (Unit 1-Unit 3.2) **Monday 2/18**

Outline

1. Housekeeping

2. Main ideas

   Problem: We need a probability distribution of values that a sample statistic can take on to assess how "extreme" (or likely) a sample statistic is (given an assumption about the population).

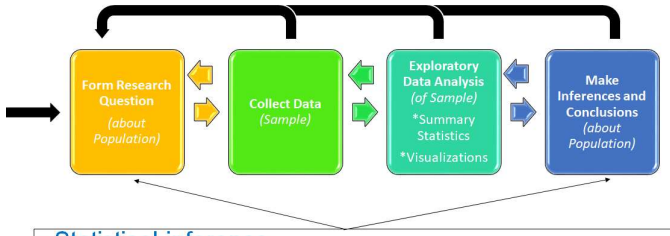
   One Solution: If you can, use a theoretical probability distribution of the sample statistic, using the Central Limit Theorem.

   1. Observation: Sample statistics vary from sample to sample.
   2. Important Theorem: CLT describes the shape, center, and spread of sampling distributions
   3. Necessary Conditions: CLT only applies when independence and sample size/skew conditions are met

3. Summary

Outline

Course Overview



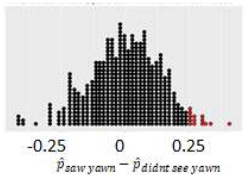
- Statistical inference
  - *–Unit 3 - Framework for inference:*
    - –Central Limit Theorem
    - –Sampling distributions
    - –Introduction to Theoretical Inference.

Making an Inference

Outline

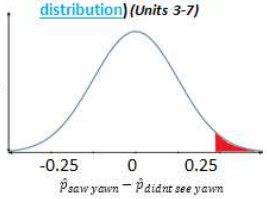
Problem: We need a probability distribution of values that a sample statistic can take on (that assumes Ho is true) to assess how "extreme" (or likely) an observed sample statistic is (under this assumption).

Randomization Distribution *(Unit 1)*



Vs.

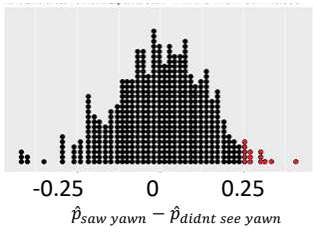
Sampling Distribution
(Special kind of normal distribution) *(Units 3-7)*



Making an Inference

One Solution: Generate a randomization distribution.

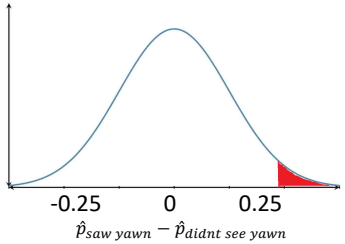
Randomization Distribution *(Unit 1)*



Making an Inference

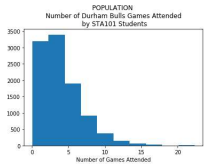
Another Solution: *If you can (see conditions)*, use a theoretical probability distribution *(probabilities come from an equation)* of the sample statistics, using the **Central Limit Theorem**.

Sampling Distribution
(Units 3-7)



Goal: We are often interested in *population parameters*.

Population Distribution



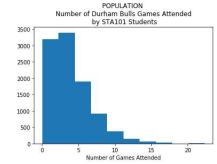
All STA101 Students

Population Mean: μ

Goal: We are often interested in *population parameters*.

Issue: Complete populations are difficult (or impossible) to collect data on.

Population Distribution



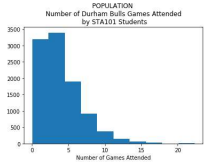
All STA101 Students

Population Mean: μ

Goal: We are often interested in *population parameters*.

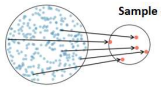
Issue: Complete populations are difficult (or impossible) to collect data on.
Solution: We use *sample statistics* as *point estimates* for the unknown population parameters of interest.

Population Distribution

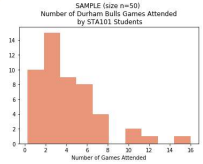


All STA101 Students

Sample of Size n=50



Sample Distribution

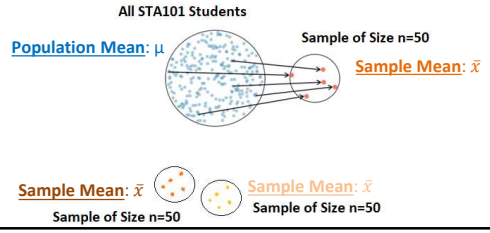


Population Mean: μ

Sample Mean: $\bar{x}$

Goal: We use *sample statistics* as *point estimates* for the unknown population parameters of interest.

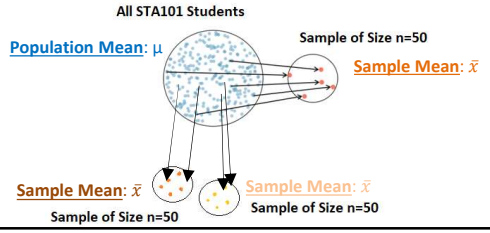
Issues:
- Sample statistic rarely = *population parameter*
- Sample statistics vary from sample to sample.



Goal: We use *sample statistics* as *point estimates* for the unknown population parameters of interest.

Issues:
- *Sample statistic* rarely = *population parameter*
- Sample statistics vary from sample to sample.

Solution: Quantifying how sample statistics vary provides a way to estimate the *margin of error* associated with our point estimate. A sampling distribution describes how these sample statistics vary.



Outline

1. Housekeeping

2. Main ideas

   Problem: We need a probability distribution of values that a sample statistic can take on to assess how "extreme" (or likely) a sample statistic is (given an assumption about the population).

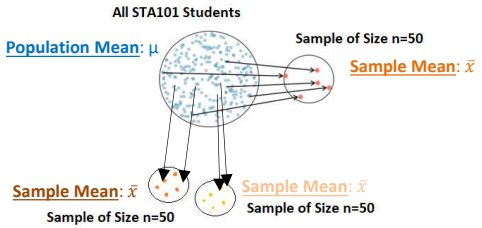
   One Solution: If you can, use a theoretical probability distribution of the sample statistic, using the Central Limit Theorem.

   1. Observation: ✋ 🔍 💻 Sample statistics vary from sample to sample.
   2. Important Theorem: CLT describes the shape, center, and spread of sampling distributions
   3. Necessary Conditions: CLT only applies when independence and sample size/skew conditions are met

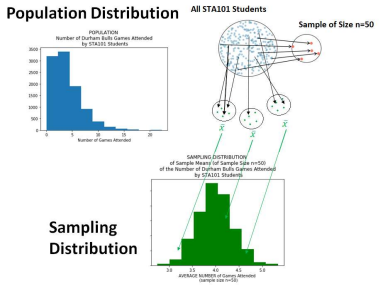
3. Summary

Outline

🔍Explore: How sample statistics *vary* from sample to sample.



How does the variability of observations from a population compare to the variability of sample statistics (generated from this population?



How does the variability of observations from a population compare to the variability of sample statistics (generated from this population?

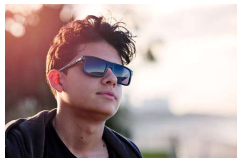
Population Dist. Standard Deviation: σ

Sampling Dist. Standard Deviation: $\sigma/\sqrt{n}$
(aka: Standard Error)

Variability of Observations in a Population

Would it be highly unlikely to randomly sample a US citizen of age **64 or older**?

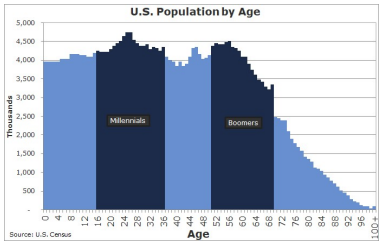
Would it be highly unlikely to randomly sample a US citizen of age **17 or younger**?



Variability of Observations in a Population

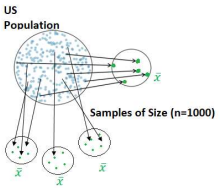
Would it be highly unlikely to randomly select a US citizen of age **64 or older**? – **No**!

Would it be highly unlikely to randomly select a US citizen of age **17 or younger**? – **No**!



Variability of Sample Statistics

Suppose we randomly sample 1,000 adults from each state in the US.



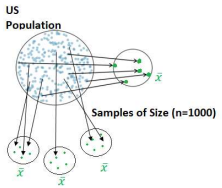
Variability of Sample Statistics

Suppose we randomly sample 1,000 adults from each state in the US.

Would it be highly unlikely to have a state with a sample mean that was **64 or higher**?

Would you expect the sample means of their ages from the different states to be:

- the same?
- just as variable as the observations from the population?
- LESS variable as the observations from the population?



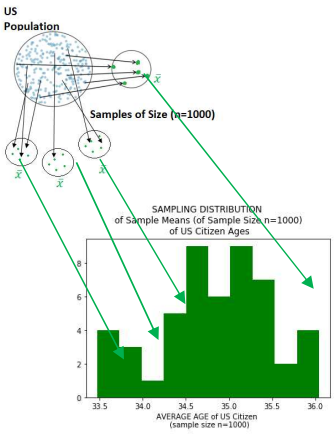
Variability of Sample Statistics

Suppose we randomly sample 1,000 adults from each state in the US.

Would it be highly unlikely to have a state with a sample mean that was **64 or higher**? **– YES!**

Would you expect the sample means of their ages from the different states to be:

- all the same?**–NO!**
- just as variable as the observations from the population?**–NO!**
- LESS variable as the observations from the population? **– YES!**

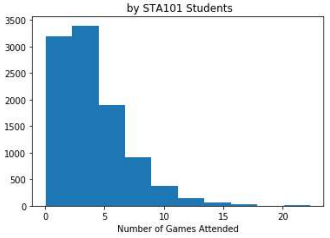


Variability of observations from a population

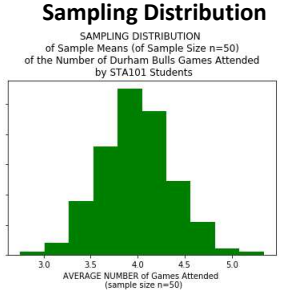
GREATER THAN

Variability of sample statistics.

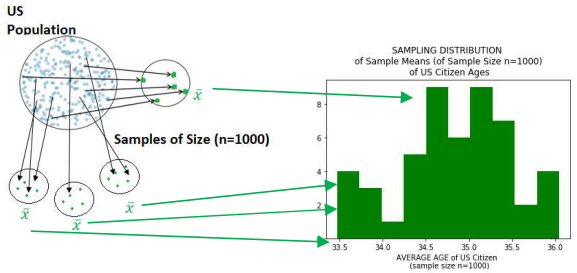
Population Distribution



Sampling Distribution



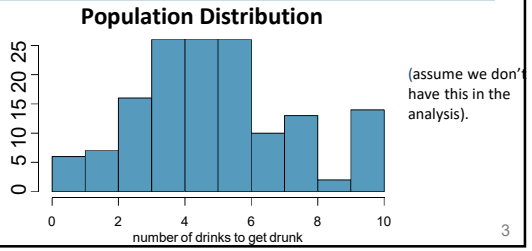
Let's make a SAMPLING DISTRIBUTION.



We would like to estimate the average number of drinks it takes students to get drunk.

- We will *assume* that our population is comprised of 146 students.
- *Assume* also that we don't have the resources to collect data from all 146, so we will take a sample of size n = 10.

If we randomly select observations from this data set, which values are most likely to be selected, which are least likely?



(assume we don't have this in the analysis).

RANDOMLY SAMPLE n=10 STUDENTS AND CALCULATE THE SAMPLE MEAN.

- Sample, with replacement, 10 student IDs:

```
> sample(1:146, size = 10, replace = TRUE)
[1]  59 121  88  46  58  72  82  81   5  10
```

▶ Sample, with replacement, 10 student IDs:

```
> sample(1:146, size = 10, replace = TRUE)
 [1]  59 121  88  46  58  72  82  81   5  10
```

A sampling distribution has random samples sampled **with** replacement.

Can still use it even if your random sample is sampled **without** replacement.

4

▶ Sample, with replacement, 10 student IDs:

```
> sample(1:146, size = 10, replace = TRUE)
 [1]  59 121  88  46  58  72  82  81   5  10
```

▶ Find the students with these IDs:

1	7	21	6	41	6	61	10	81	6	101	4	121	6	141	4
2	5	22	2	42	10	62	7	82	5	102	7	122	5	142	6
3	4	23	6	43	3	63	4	83	6	103	6	123	3	143	6
4	4	24	7	44	6	64	5	84	8	104	8	124	2	144	4
5	6	25	3	45	10	65	6	85	4	105	3	125	2	145	5
6	2	26	6	46	4	66	6	86	10	106	6	126	5	146	5
7	3	27	5	47	3	67	6	87	5	107	2	127	10		
8	5	28	8	48	3	68	7	88	10	108	5	128	4		
9	5	29	0	49	6	69	7	89	8	109	1	129	1		
10	6	30	8	50	8	70	5	90	5	110	5	130	4		
11	1	31	5	51	8	71	10	91	4	111	5	131	10		
12	10	32	9	52	8	72	3	92	0.5	112	4	132	8		
13	4	33	7	53	2	73	5.5	93	3	113	4	133	10		
14	4	34	5	54	4	74	7	94	3	114	9	134	6		
15	6	35	5	55	8	75	10	95	5	115	4	135	6		
16	3	36	7	56	3	76	6	96	6	116	3	136	6		
17	10	37	4	57	5	77	6	97	4	117	3	137	7		
18	8	38	0	58	5	78	5	98	4	118	4	138	3		
19	5	39	4	59	8	79	4	99	2	119	4	139	10		
20	10	40	3	60	4	80	5	100	5	120	8	140	4		

4

▶ Sample, with replacement, 10 student IDs:

```
> sample(1:146, size = 10, replace = TRUE)
 [1]  59 121  88  46  58  72  82  81   5  10
```

▶ Find the students with these IDs:

1	7	21	6	41	6	61	10	81	6	101	4	121	6	141	4
2	5	22	2	42	10	62	7	82	5	102	7	122	5	142	6
3	4	23	6	43	3	63	4	83	6	103	6	123	3	143	6
4	4	24	7	44	6	64	5	84	8	104	8	124	2	144	4
5	6	25	3	45	10	65	6	85	4	105	3	125	2	145	5
6	2	26	6	46	4	66	6	86	10	106	6	126	5	146	5
7	3	27	5	47	3	67	6	87	5	107	2	127	10		
8	5	28	8	48	3	68	7	88	10	108	5	128	4		
9	5	29	0	49	6	69	7	89	8	109	1	129	1		
10	6	30	8	50	8	70	5	90	5	110	5	130	4		
11	1	31	5	51	8	71	10	91	4	111	5	131	10		
12	10	32	9	52	8	72	3	92	0.5	112	4	132	8		
13	4	33	7	53	2	73	5.5	93	3	113	4	133	10		
14	4	34	5	54	4	74	7	94	3	114	9	134	6		
15	6	35	5	55	8	75	10	95	5	115	4	135	6		
16	3	36	7	56	3	76	6	96	6	116	3	136	6		
17	10	37	4	57	5	77	6	97	4	117	3	137	7		
18	8	38	0	58	5	78	5	98	4	118	4	138	3		
19	5	39	4	59	8	79	4	99	2	119	4	139	10		
20	10	40	3	60	4	80	5	100	5	120	8	140	4		

▶ Calculate the sample mean:

$(8 + 6 + 10 + 4 + 5 + 3 + 5 + 6 + 6 + 6)/10 = 5.9$

4

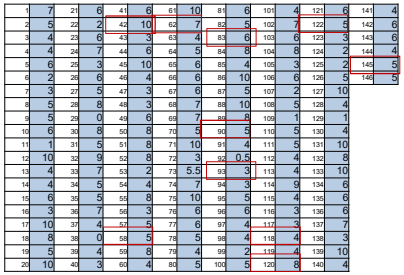
RANDOMLY SAMPLE n=10 STUDENTS AGAIN AND CALCULATE THE NEW SAMPLE MEAN.

4

- Sample, with replacement, 10 student IDs:

```
> sample(1:146, size = 10, replace = TRUE)
[1]  120 90  145 118 122 62  42  58  83 93
```

- Find the students with these IDs:

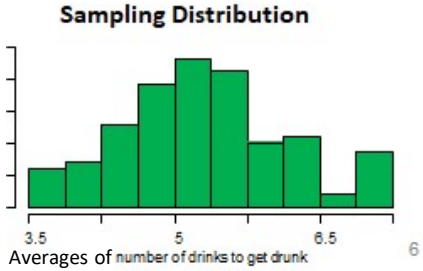


- Calculate the sample mean:
  (10+5+7+6+5+3+4+4+5+5)/10 = 5.4

4

Do this many many more times….

Plot all of your sample means in a histogram.



6

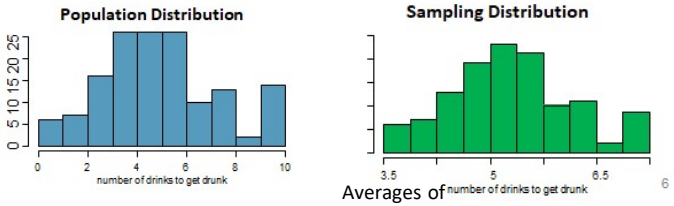
4

Sampling distribution

What you just constructed is called a *sampling distribution* (approximately)

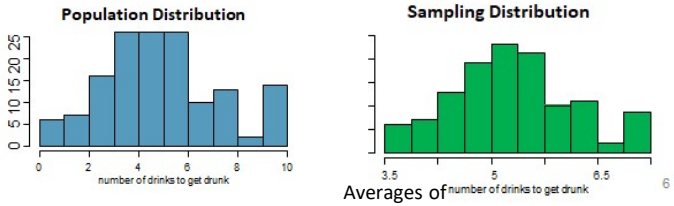
… not the sample distribution (this is the distribution of *observations* in a single *sample* from the population.)

What is the shape and center of this distribution. Based on this distribution what do you think is the true population average?



6

The sampling distribution mean is the population mean.



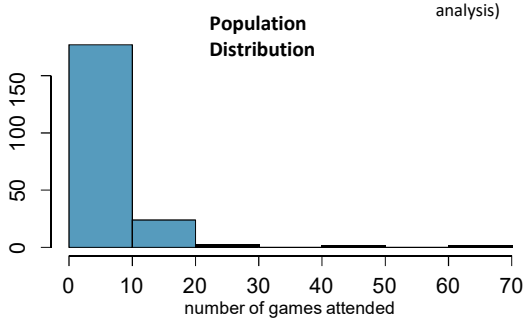
6

How does the variability (spread), shape, and mean of sampling distribution change as n (sample size) increases?

Average number of Duke games attended

Next let's look at the **population data** for the number of Duke basketball games attended:

(assume that we don't have this in a typical analysis)

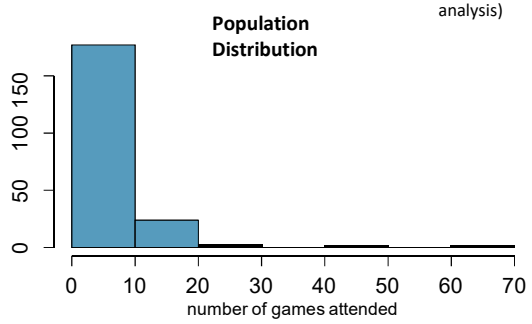


7

Average number of Duke games attended

Next let's look at the **population data** for the number of Duke basketball games attended:

(assume that we don't have this in a typical analysis)

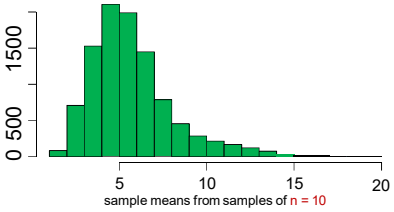


Center: mean between 0 and 10?
Shape: right skewed
Spread: range=70

7

Average number of Duke games attended (cont.)

Sampling distribution, of sample means from the population with samples of size **n = 10**.



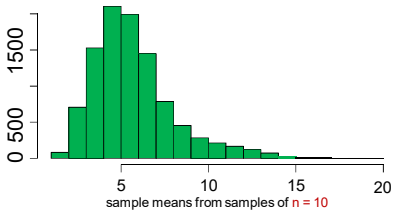
What does each observation in this distribution represent?

Is the variability of the sampling distribution smaller or larger than the variability of the population distribution?

8

Average number of Duke games attended (cont.)

Sampling distribution, of sample means from the population with samples of size **n = 10**.



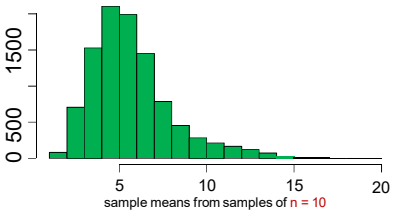
What does each observation in this distribution represent?

Sample mean, $\bar{x}$, of samples of size n = 10.

Is the variability of the sampling distribution smaller or larger than the variability of the population distribution?

Average number of Duke games attended (cont.)

Sampling distribution, of sample means from the population with samples of size **n = 10**.



What does each observation in this distribution represent?

Sample mean, $\bar{x}$, of samples of size n = 10.

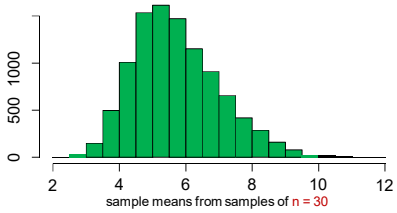
Is the variability of the sampling distribution smaller or larger than the variability of the population distribution?

Smaller, sample means will vary less than individual observations.

Center: mean between 6-7?
Shape: right skewed (less so)
Spread: range=20

Average number of Duke games attended (cont.)

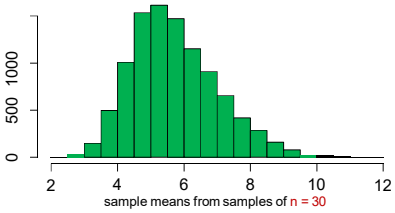
Sampling distribution, of sample means from the population with samples of size n = 30.



How did the shape, center, and spread of the sampling distribution change going from n = 10 to n = 30?

Average number of Duke games attended (cont.)

Sampling distribution, of sample means from the population with samples of size n = 30.



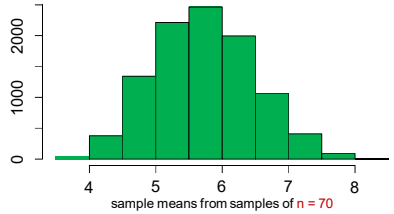
How did the shape, center, and spread of the sampling distribution change going from n = 10 to n = 30?

Shape is more symmetric, center is about the same, spread is smaller.

Center: mean around 6?
Shape: right skewed (even less so)
Spread: range=10

Average number of Duke games attended (cont.)

Sampling distribution, n = 70:



Center: mean around 6
Shape: mostly unimodal and symmetric now
Spread: range=5

10

- **Sampling distribution variability (spread) decreases as n (sample size) increases.**
- **Sampling distribution shape becomes more symmetric and unimodal as n (sample size) increases.**
- **Sampling distribution mean stays roughly the same n (sample size) increases.**

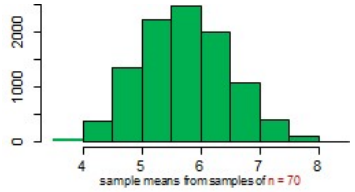
Outline

1. Housekeeping
2. Main ideas
   - Problem: We need a probability distribution of values that a sample statistic can take on to assess how "extreme" (or likely) a sample statistic is (given an assumption about the population).
   - One Solution: If you can, use a theoretical probability distribution of the sample statistic, using the Central Limit Theorem.
     1. Observation: Sample statistics vary from sample to sample.
     2. Important Theorem: CLT describes the shape, center, and spread of sampling distributions
     3. Necessary Conditions: CLT only applies when independence and sample size/skew conditions are met
3. Summary

- **Putting it all together:**

What happens to the center, spread, and shape of the sampling distribution when n (sample size) becomes "large"?

Can we prove this?



2. CLT describes the shape, center, and spread of sampling distributions

Central Limit Theorem

Under the right conditions, the distribution of the sample means (sampling dist) is well approximated by a normal distribution:

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$$

A cheat: If σ is unknown, use s.

2. CLT describes the shape, center, and spread of sampling distributions

Central Limit Theorem

Under the right conditions, the distribution of the sample means (sampling dist) is well approximated by a normal distribution:

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$$

A cheat: If σ is unknown, use s.

- So it wasn't a coincidence that the sampling distributions we saw earlier were symmetric.
- We won't go into the proving why $SE = \frac{\sigma}{\sqrt{n}}$, but note that as n increases SE _____.
- As the sample size increases we would expect samples to yield more consistent sample means, hence the variability among the sample means would be lower.

2. CLT describes the shape, center, and spread of sampling distributions

Central Limit Theorem

Under the right conditions, the distribution of the sample means (sampling dist) is well approximated by a normal distribution:

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$$

A cheat: If σ is unknown, use s.

- So it wasn't a coincidence that the sampling distributions we saw earlier were symmetric.
- We won't go into the proving why $SE = \frac{\sigma}{\sqrt{n}}$, but note that as n increases SE decreases.
- As the sample size increases we would expect samples to yield more consistent sample means, hence the variability among the sample means would be lower.

Outline

What are these conditions?

Central Limit Theorem

Under the right conditions, the distribution of the sample means (sampling dist) is well approximated by a normal distributio

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$$

3. CLT only applies when independence and sample size/skew conditions are met

1. *Independence:* Sampled observations must be independent.

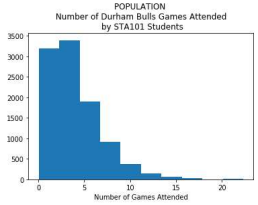
   This is difficult to verify, but is more likely if
   - random sampling/assignment is used, and,
   - if sampling without replacement, $n <$ 10% of the population.
2. *Sample size/skew:* Either
   - the population distribution is normal OR
   - *Sample size* $n >$ 30 and the population dist. is not extremely skewed, OR
   - *Sample size* $n >>$ 30 (approx. gets better as n increases).

13

Checking Skewness Conditions for Central Limit Theorem:

How can we guess what the shape of the population distribution is (without seeing it)?

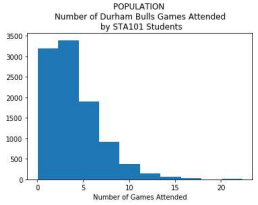
Population Distribution



Checking Skewness Conditions for Central Limit Theorem:

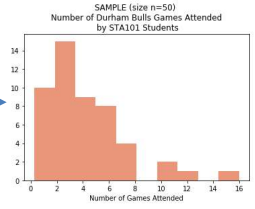
How can we guess what the shape of the population distribution is (without seeing it)?

Population Distribution



Shapes are roughly the same.

Sample Distribution



3. CLT only applies when independence and sample size/skew conditions are met

1. *Independence:* Sampled observations must be independent.

   This is difficult to verify, but is more likely if
   - random sampling/assignment is used, and,
   - if sampling without replacement, $n <$ 10% of the population.
2. *Sample size/skew:* Either
   - the population distribution is normal OR
   - $n > 30$ and the population dist. is not extremely skewed, OR
   - $n >> 30$ (approx. gets better as n increases).

   This is also difficult to verify for the population, but we can check it using the sample data, and assume that the sample distribution shape mirrors the population distribution shape.

13

3. CLT only applies when independence and sample size/skew conditions are met

Amongst other things, the central limit theorem is useful for

- constructing confidence intervals and
- conducting hypothesis tests.

14

Clicker question

Which of the below visualizations is not appropriate for checking the shape of the sample distribution of a numerical variable, and hence the population distribution?

(a) histogram
(b) boxplot
(c) normal probability plot
(d) mosaic plot

15

Clicker question

Which of the below visualizations is not appropriate for checking the shape of the sample distribution of a numerical variable, and hence the population distribution?

(a) histogram
(b) boxplot
(c) normal probability plot
(d) mosaic plot

15

Summary of main ideas

1. Sample statistics vary from sample to sample
2. CLT describes the shape, center, and spread of sampling distributions
3. CLT only applies when independence and sample size/skew conditions are met

16

Outline

1. Housekeeping

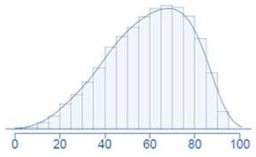
2. Main ideas
   1. Sample statistics vary from sample to sample
   2. CLT describes the shape, center, and spread of sampling distributions
   3. CLT only applies when independence and sample size/skew conditions are met

3. Summary

4. Exercises [time permitting]

Clicker question

Four plots: Determine which plot (A, B, or C) is which.
(1) At top: distribution for a population ($\mu = 60, \sigma = 18$),
(2) a single random sample of 500 observations from this population,
(3) a distribution of 500 sample means from random samples with size 18,
(4) a distribution of 500 sample means from random samples with size 81.

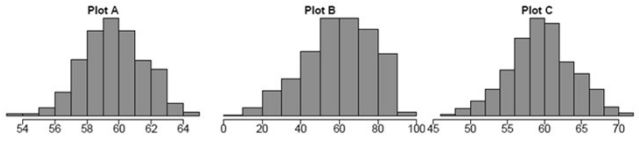


(a) (2) - B; (3) - A; (4) - C

(b) (2) - A; (3) - B; (4) - C

(c) (2) - C; (3) - A; (4) - D

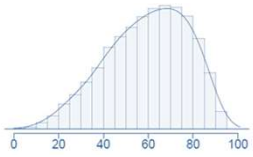
(d) (2) - B; (3) - C; (4) - A



17

Clicker question

Four plots: Determine which plot (A, B, or C) is which.
(1) At top: distribution for a population ($\mu = 60, \sigma = 18$),
(2) a single random sample of 500 observations from this population,
(3) a distribution of 500 sample means from random samples with size 18,
(4) a distribution of 500 sample means from random samples with size 81.

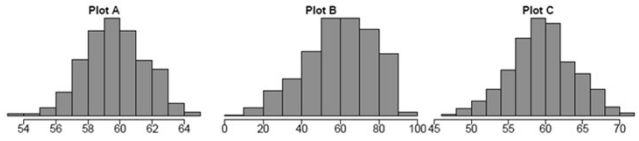


(a) (2) - B; (3) - A; (4) - C

(b) (2) - A; (3) - B; (4) - C

(c) (2) - C; (3) - A; (4) - D

(d) *(2) - B; (3) - C; (4) - A*



17

A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly $1.3 million with a standard deviation of $300,000. There were no houses listed below $600,000 but a few houses above $3 million.

Would you expect most houses in Topanga to cost more or less than $1.3 million? Hint: What is most likely the shape of this distribution?

A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly $1.3 million with a standard deviation of $300,000. There were no houses listed below $600,000 but a few houses above $3 million.

Would you expect most houses in Topanga to cost more or less than $1.3 million? Hint: What is most likely the shape of this distribution?

Since the distribution is probably right skewed, the median would be less than the mean, and a majority of observations would be lower than the mean.

A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly $1.3 million with a standard deviation of $300,000. There were no houses listed below $600,000 but a few houses above $3 million.

Clicker question

Can we estimate the probability that a randomly chosen house in Topanga costs more than $1.4 million using the normal distribution?

(a) yes
(b) no

A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly $1.3 million with a standard deviation of $300,000. There were no houses listed below $600,000 but a few houses above $3 million.

Clicker question

Can we estimate the probability that a randomly chosen house in Topanga costs more than $1.4 million using the normal distribution?

(a) yes
(b) no

The POPULATION (of houses) distribution is NOT normal.

$X \sim N(\mu = \$1.3m, \sigma = \$0.3m)$

A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly $1.3 million with a standard deviation of $300,000. There were no houses listed below $600,000 but a few houses above $3 million.

Clicker question

Can we estimate the probability that the mean of 60 randomly chosen houses in Topanga is more than $1.4 million?

(a) yes
(b) no

20

A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly $1.3 million with a standard deviation of $300,000. There were no houses listed below $600,000 but a few houses above $3 million.

Clicker question

Can we estimate the probability that the mean of 60 randomly chosen houses in Topanga is more than $1.4 million?

(a) yes
(b) no

The sampling distribution IS normal….

$\bar{x} \sim N(\mu = \$1.3m, SE = \frac{\$0.3m}{\sqrt{60}})$

…because:

- Random sampling is used.
- n=60<10% of Topanga, CA homes.
- n>30

20

A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly $1.3 million with a standard deviation of $300,000. There were no houses listed below $600,000 but a few houses above $3 million.

What is the probability that the mean of 60 randomly chosen houses in Topanga is more than $1.4 million?

In order to calculate $P(X > 1.4\ mil)$, we need to first determine the distribution of $\bar{X}$. According to the CLT,

21

A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly $1.3 million with a standard deviation of $300,000. There were no houses listed below $600,000 but a few houses above $3 million.

What is the probability that the mean of 60 randomly chosen houses in Topanga is more than $1.4 million?

In order to calculate $P(X > 1.4\ mil)$, we need to first determine the distribution of $\bar{X}$. According to the CLT,

$$\bar{X} \sim N\left(mean = \quad , SE = \quad\right)$$

21

A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly $1.3 million with a standard deviation of $300,000. There were no houses listed below $600,000 but a few houses above $3 million.

What is the probability that the mean of 60 randomly chosen houses in Topanga is more than $1.4 million?

In order to calculate $P(X > 1.4\ mil)$, we need to first determine the distribution of $\bar{X}$. According to the CLT,

$$\bar{X} \sim N\left(mean = 1.3, SE = \frac{0.3}{\sqrt{60}} = 0.0387\right)$$

Remember…
If a random variable □~N($mean_{□}$, $SD_{□}$), then
$\frac{□ - mean_{□}}{SD_{□}} \sim N(0,1)$

21

A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly $1.3 million with a standard deviation of $300,000. There were no houses listed below $600,000 but a few houses above $3 million.

What is the probability that the mean of 60 randomly chosen houses in Topanga is more than $1.4 million?

In order to calculate $P(X > 1.4\ mil)$, we need to first determine the distribution of $\bar{X}$. According to the CLT,

$\bar{X} \sim N\left(mean = 1.3, SE = \frac{0.3}{\sqrt{60}} = 0.0387\right)$

Remember…
If a random variable □~N($mean_{□}$, $SD_{□}$), then
$\frac{□ - mean_{□}}{SD_{□}} \sim N(0,1)$

$$P(\bar{X} > 1.4) = P\left(Z > \frac{1.4 - 1.3}{0.0387}\right)$$

Z-table

$$= P(Z > 2.58) \quad = 1 - P(Z \leq 2.58)$$

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

$$= 1 - 0.9951$$

21