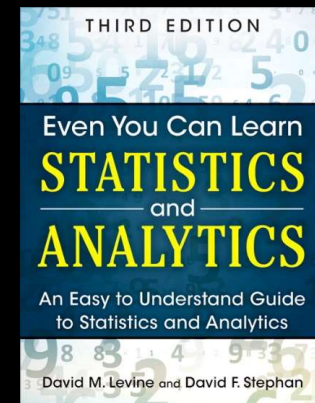


IMGD 2905

# Inferential Statistics

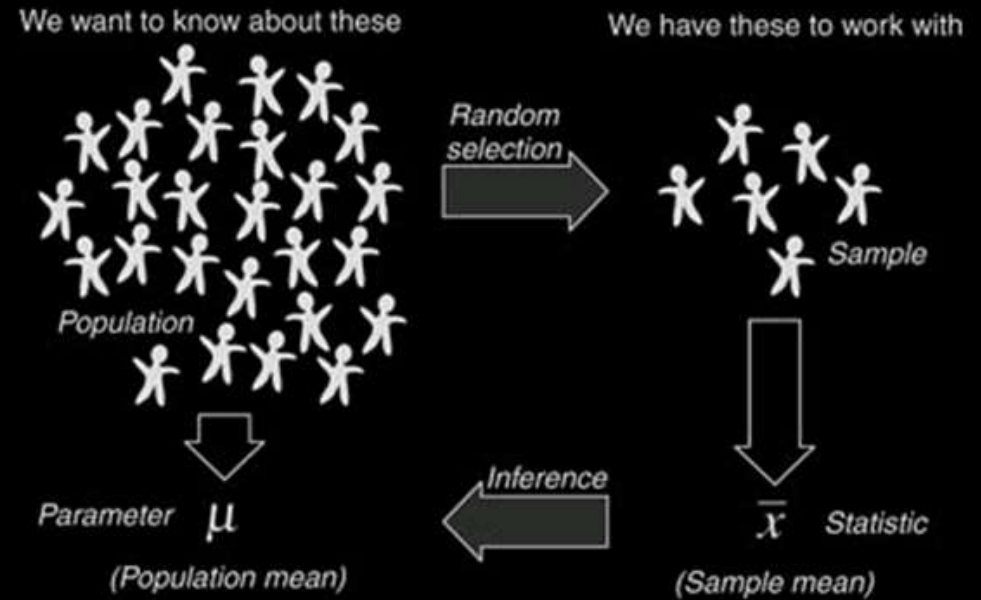
Chapter 6 & 7



# Overview



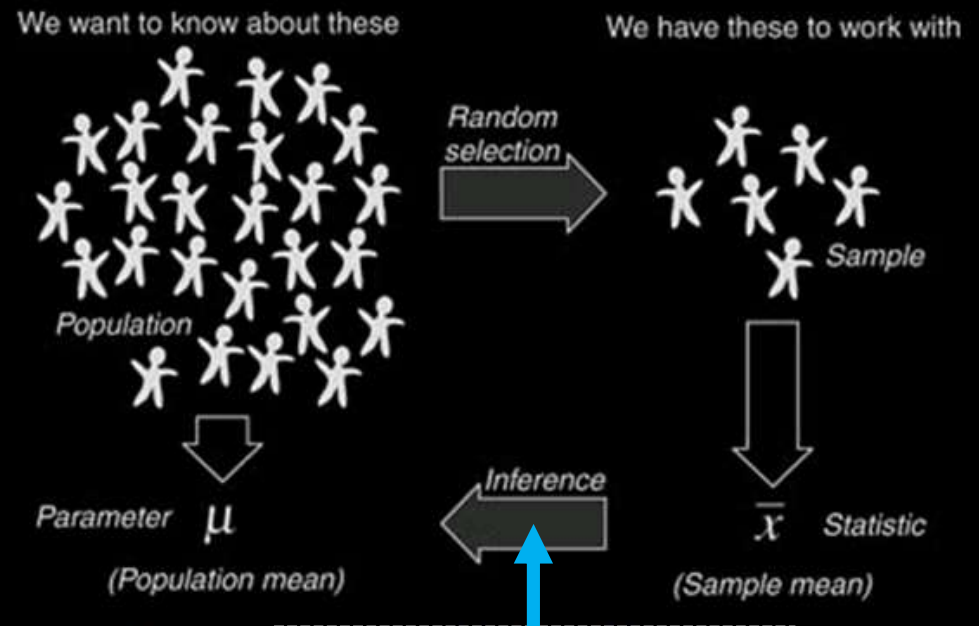
- Use statistics to infer population parameters



# Overview



- Use statistics to infer population parameters



<http://3.bp.blogspot.com/>

Inferential statistics

600/120909\_0159\_1.png

# Outline GA

- Overview (done)
- Foundation (next)
- Inferring Population Parameters
- Hypothesis Testing

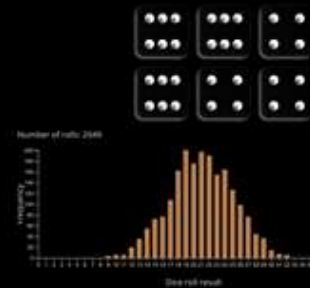


# Breakout 5



- Remember, *probability distribution* shows possible outcomes on x-axis and probability of each on y-axis.
  1. Describe probability distribution of 1 d6?
  2. Describe probability distribution of 2 d6?
  3. Describe probability distribution of 3 d6?
- Icebreaker, Groupwork, Questions

<https://web.cs.wpi.edu/~imgd2905/d21/breakout/breakout-5.html>



<https://academo.org/demos/dice-roll-statistics/>

## Dice Rolling (1 of 4)



- Have 1d6, sample (i.e., roll 1 die)
- What is probability distribution of values?

# Dice Rolling (1 of 4)



- Have 1d6, sample (i.e., roll 1 die)
- What is probability distribution of values?



Copyright © 2006 Investopedia.com

<http://www.investopedia.com/articles/06/probabilitydistribution.asp>

“Square”  
distribution

## Dice Rolling (2 of 4)

- Have 1d6, sample twice and sum (i.e., roll 2 dice)
- What is probability distribution of values?



# Dice Rolling (2 of 4) GA

- Have 1d6, sample twice and sum (i.e., roll 2 dice)
- What is probability distribution of values?



“Triangle”  
distribution

## Dice Rolling (3 of 5)

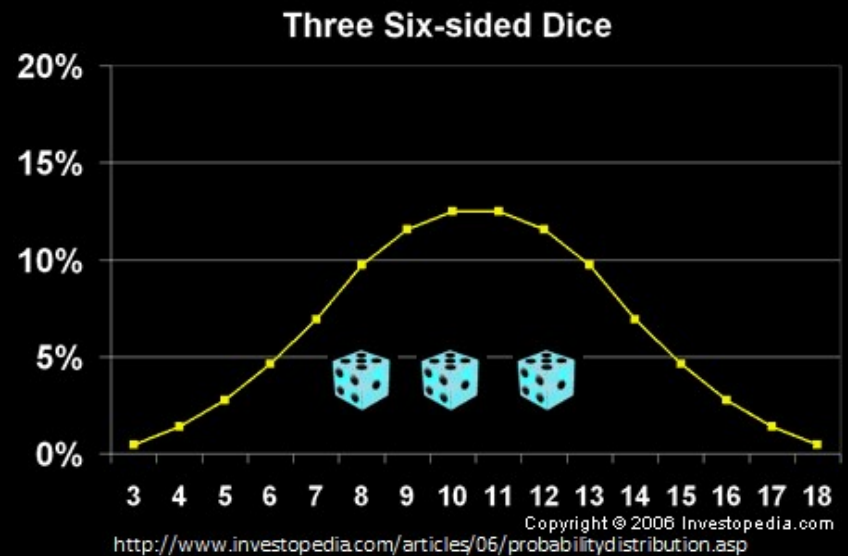
- Have 1d6, sample thrice and sum (i.e., roll 3 dice)
- What is probability distribution of values?

# Dice Rolling (3 of 5)



- Have 1d6, sample thrice and sum (i.e., roll 3 dice)
- What is probability distribution of values?

What's happening to the shape?



# Dice Rolling (4 of 5) GA

- Have d6, sample and sum, repeat
- What is probability distribution of values?

What's happening to the shape?



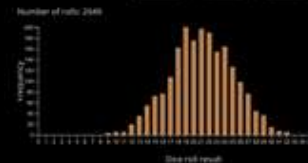
uniform distribution  
1 die



uniform sum distribution  
2 dice



uniform sum distribution  
3 dice

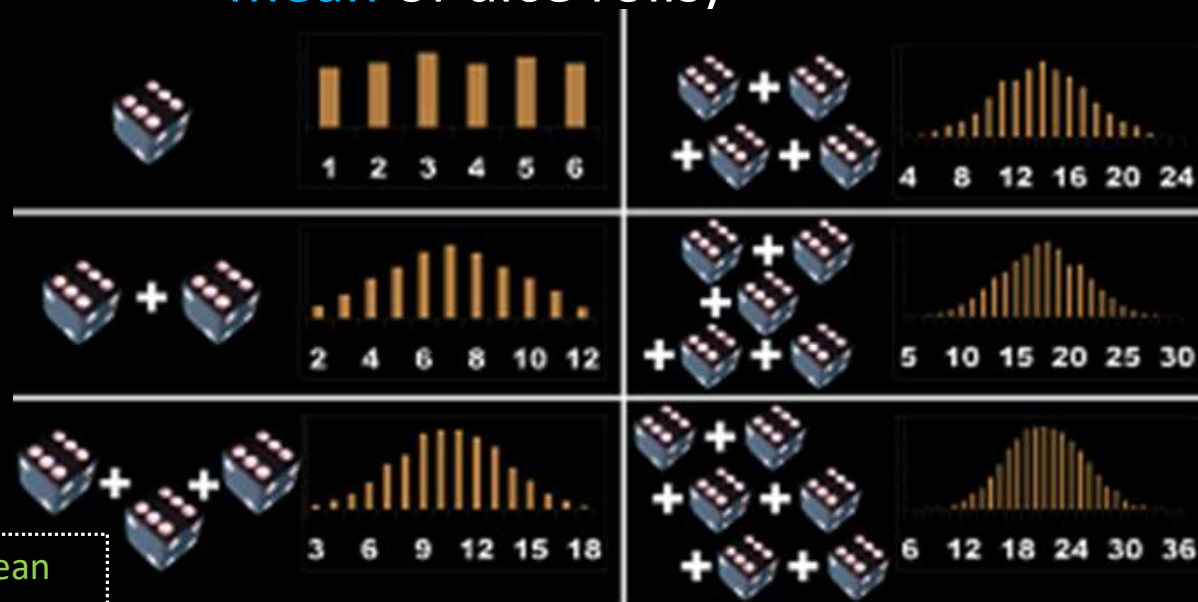


<https://academo.org/demos/dice-roll-statistics/>

Try rolling dice yourself!  
Observe shape over time

# Dice Rolling (5 of 5) GA

- Holds for general experiments with dice (i.e., observing **sample sum** and **mean** of dice rolls)



<http://www.muelaner.com/uncertainty-of-measurement/>

Resulting **sum/mean** follows a normal distribution  
→ Even though base distribution is uniform!

Ok, neat – for uniform (“square”) distributions (e.g., d6).  
But what about experiments with **other distributions**?

# Sampling Distributions

- Resulting **sum/mean** follows a normal distribution
- → Even though base distribution is uniform!

(b)  
Uniform



Parent Population



Sampling Distributions of  $x$  for  $N=2$



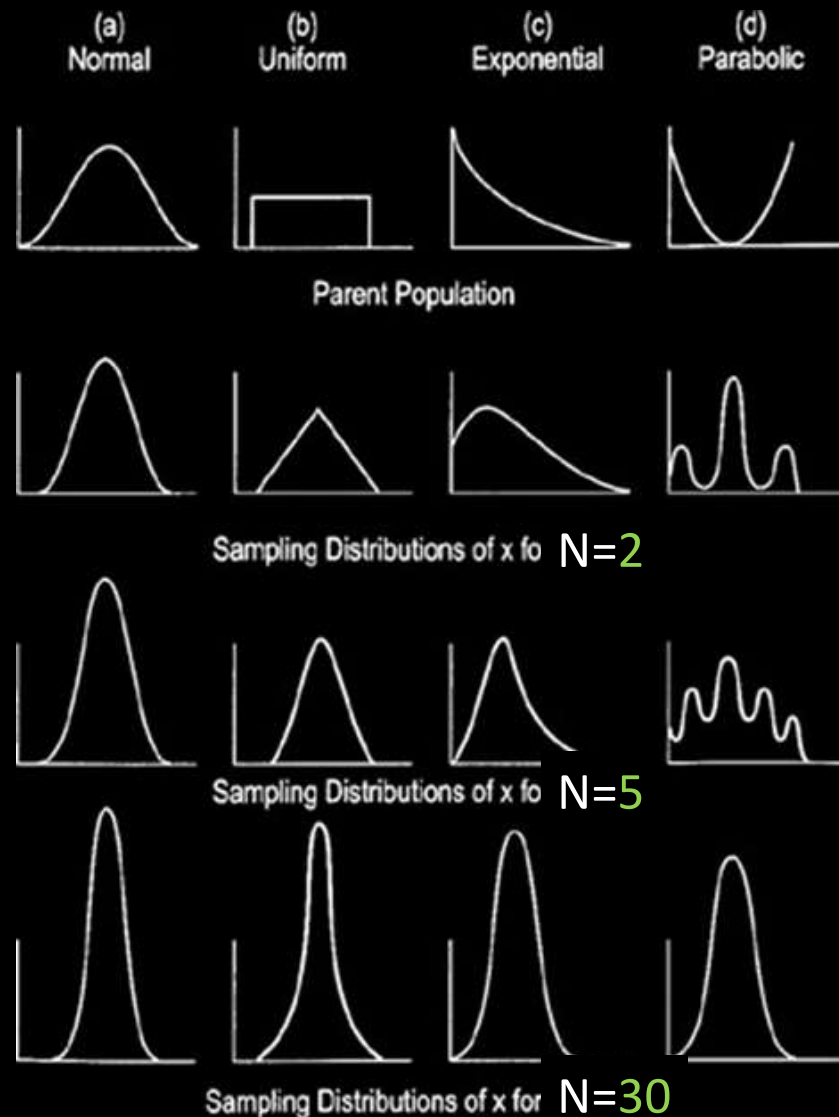
Sampling Distributions of  $x$  for  $N=5$



Sampling Distributions of  $x$  for  $N=30$

# Sampling Distributions

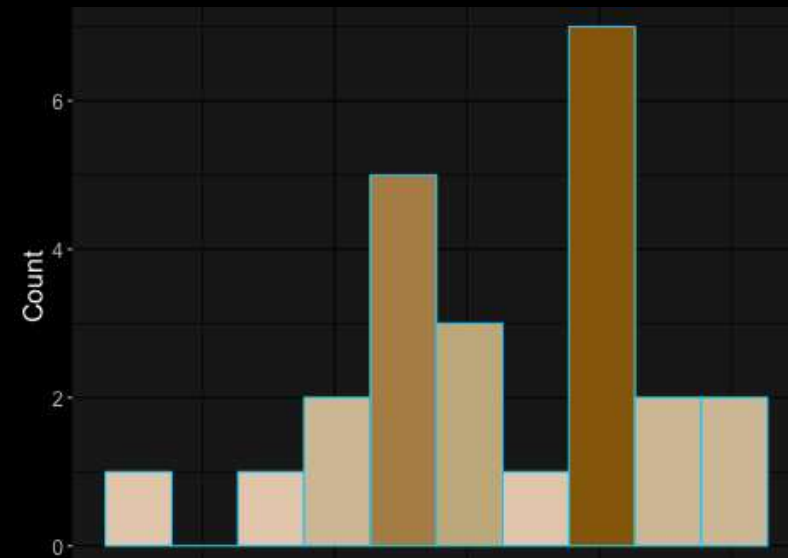
- With large “enough” sample size, looks “bell-shaped” → **Normal!**
- How many is large enough?
  - 30 (15 if symmetric distribution)
- **Central Limit Theorem**
  - Sum of independent variables tends towards **Normal distribution**



# Why do we care about sample means following Normal distribution?

GA

- What if we had only a **sample mean** and no measure of spread
  - e.g., mean rank for Overwatch is 50
- What can we say about **population mean**?



Sample mean

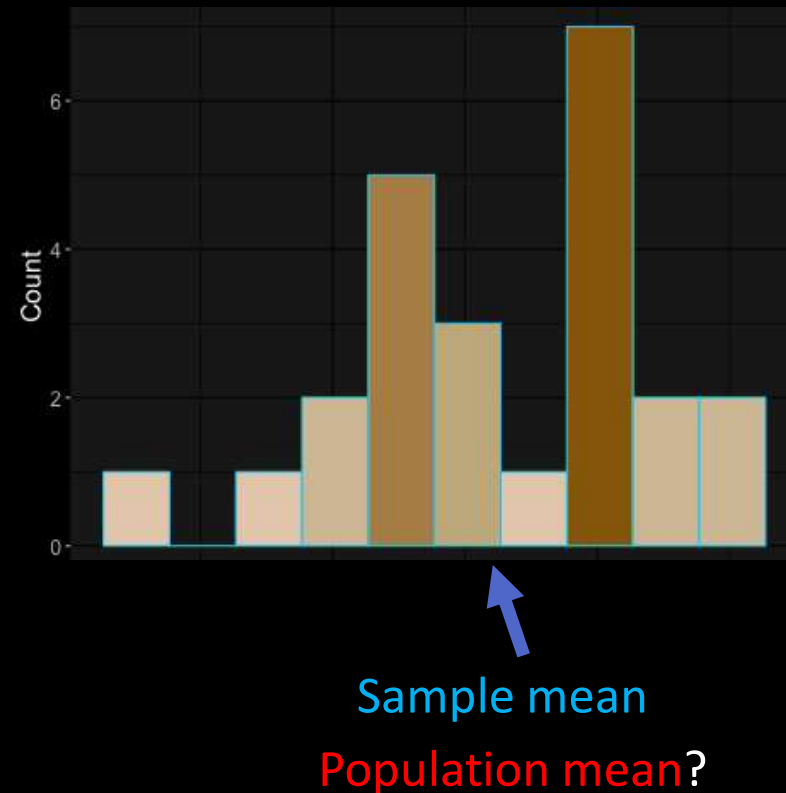
Population mean?



# Why do we care about sample means following Normal distribution?

GA

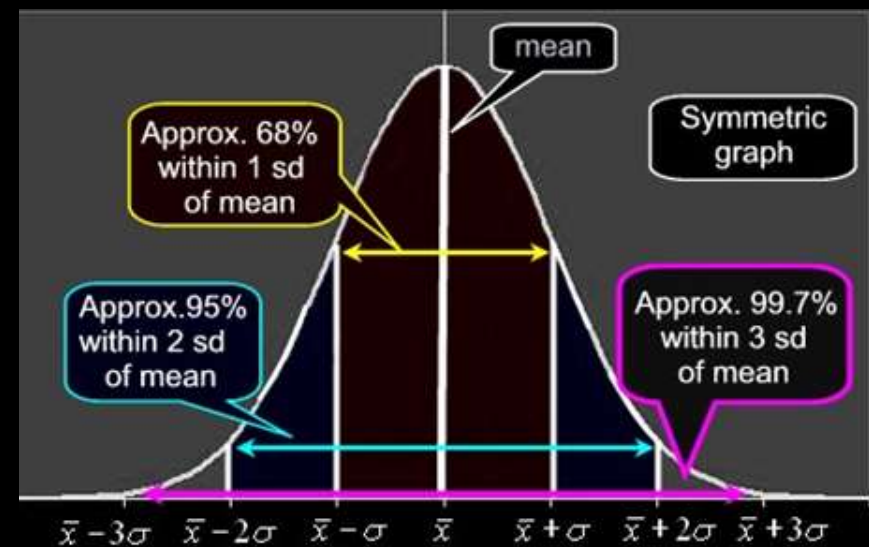
- What if we had only a **sample mean** and no measure of spread
    - e.g., mean rank for Overwatch is 50
  - What can we say about **population mean**?
    - Not a whole lot!
    - Yes, **population mean** could be 50. But could be 100. How likely are each?
- No idea!



# Why do we care about sample means following Normal distribution?

GA

Remember this?



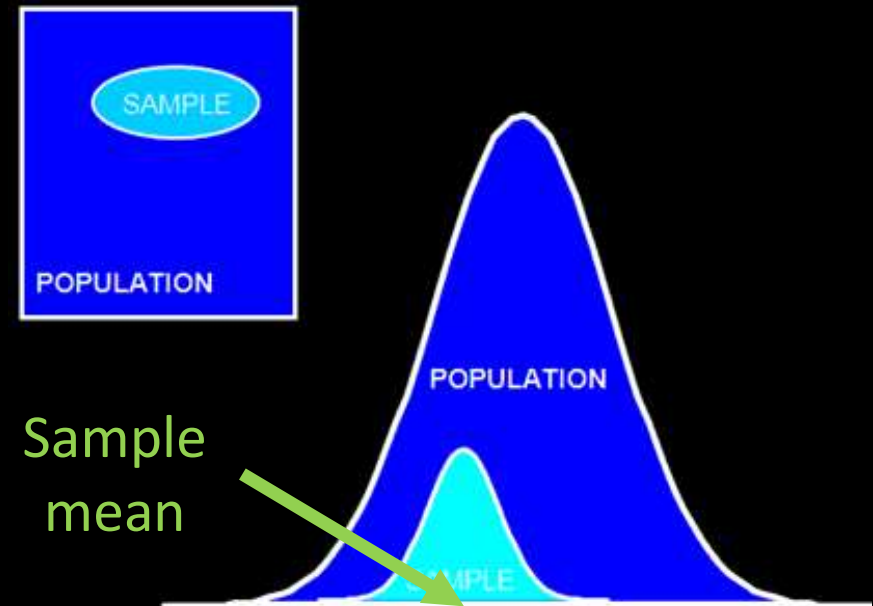
With mean and standard deviation



Allows us to predict range to bound population mean (see next slide)

# Why do we care about sample means following Normal distribution?

GA



<http://www.six-sigma-material.com/images/PopSamples.GIF>

Actual population mean (probably) in this range!

Probable range of population mean

# Outline

- Overview (done)
- Foundation (done)
- Inferring Population Parameters (next)
- Hypothesis Testing

# Estimating Population Mean



- Underlying data follows uniform probability distribution (d6)
  - But assume population mean unknown

Q: How do we estimate the population mean?

# Estimating Population Mean



- Underlying data follows uniform probability distribution (d6)
  - But assume population mean unknown

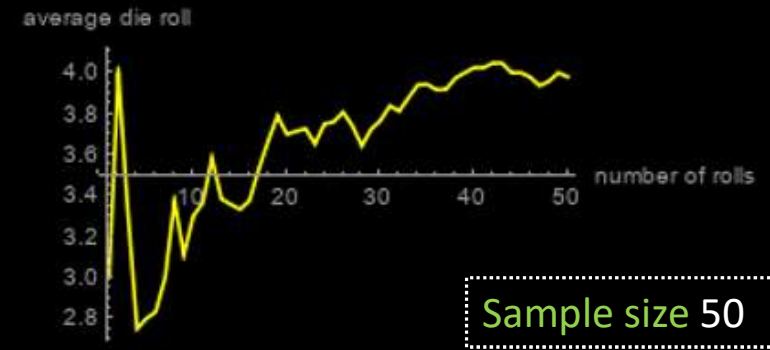
(Example)

Q: How do we estimate the population mean?

<u>Sample</u>	<u>Sample Mean</u>
1 d6	4.0
2 d6 (4 + 2) / 2 =	3.0
3 d6 (1 + 6 + 2) / 3 =	2.3
4 d6 (4 + 4 + 2 + 3) / 4 =	3.3

# Estimating Population Mean

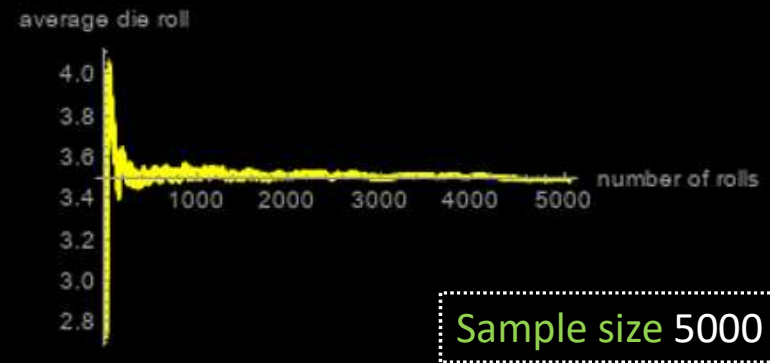
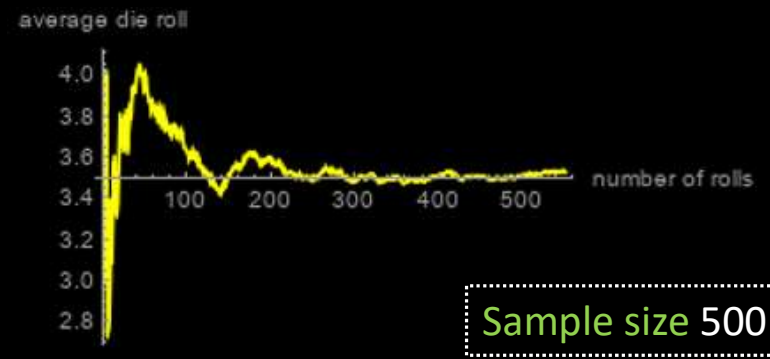
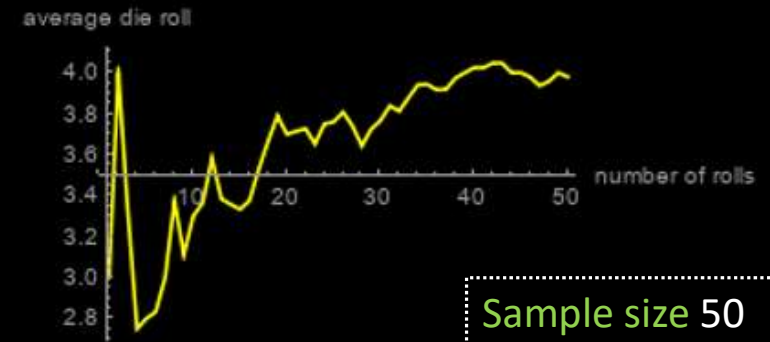
- Q: What happens as **sample size** increases?



# Estimating Population Mean

- Q: What happens as **sample size** increases?
- Q: How big a sample do we need?

POLL!

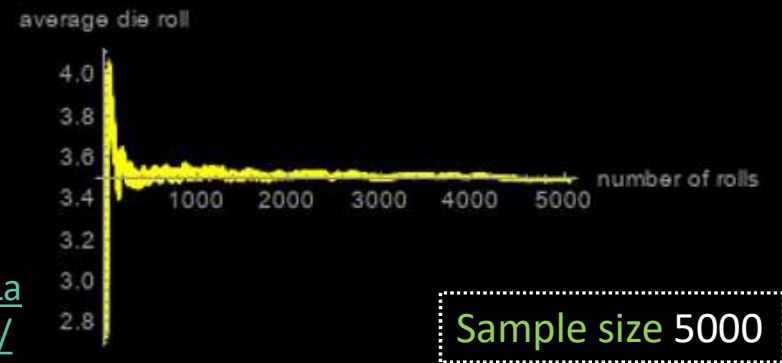
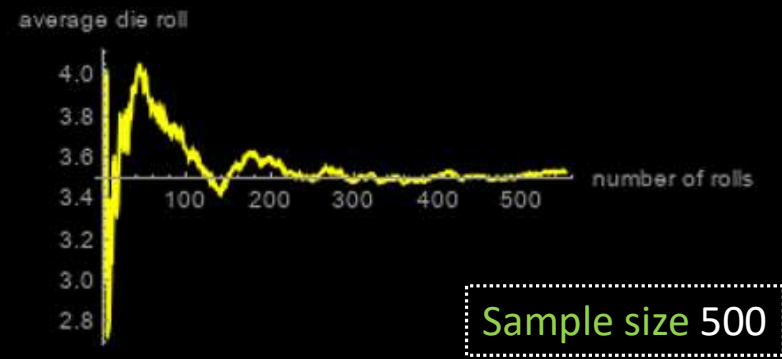
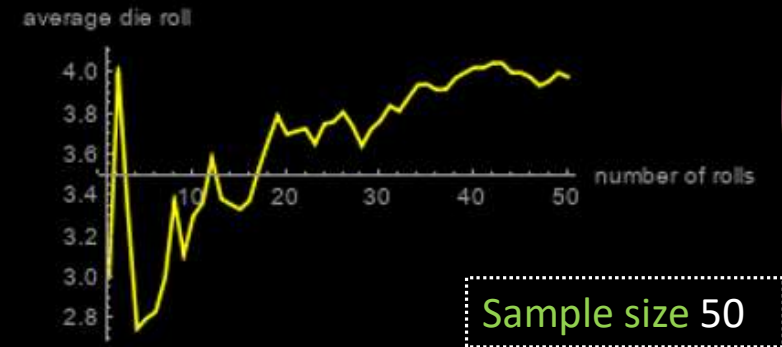




# Estimating Population Mean

- Q: What happens as **sample size** increases?
- Q: How big a sample do we need?
  - Depends upon how much varies
  - Depends upon how accuracy we want
- Values that are not the mean contribute to “error” → **sampling error**

<https://demonstrations.wolfram.com/LawOfLargeNumbersDiceRollingExample/>



# Sampling Error

- Error from estimating **population** parameters from **sample** statistics is **sampling error**
- Exact error often cannot be known (do not know population parameters)
- But *size* of error based on:
  - **Variation in population** ( $\sigma$ ) itself – more variation, more sample variation ( $s$ )
  - **Sample size** ( $N$ ) – larger sample, lower error

*Q: Can't we just make sample size super large?*
- How much does it vary?
  - **Standard error**

# Standard Error (1 of 2)

GA

- Amount **sample means** vary from experiment to experiment of same size
  - *Standard deviation of the sample means*
- Also, likelihood that sample statistic is near population parameter
  - Depends upon **sample size (N)**
  - Depends upon standard deviation (**s**)

(Example next)

standard error  $SE = \frac{s}{\sqrt{n}}$  sample size

Example:  
 $n = 5$   
 $s = 17$

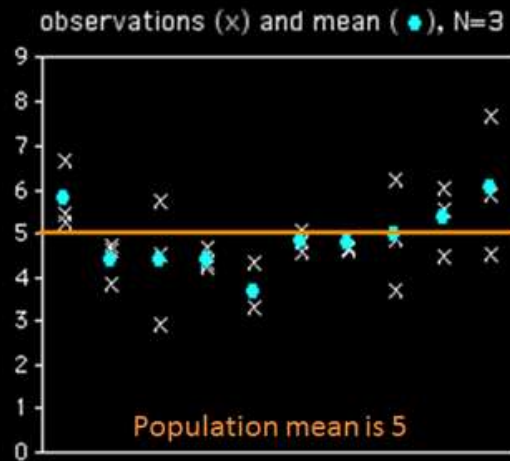
$$= \frac{17}{\sqrt{5}}$$

**SE = 7.6**

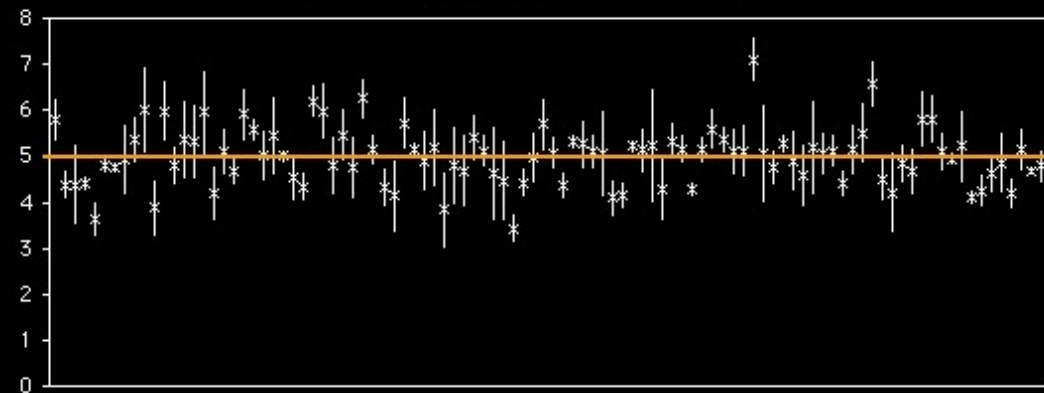
So what? Can reason about population mean  
e.g., **95% confident** that population mean is in  $\sim 2$  SE's  
(where does this come from?)

# Standard Error (2 of 2)

GA

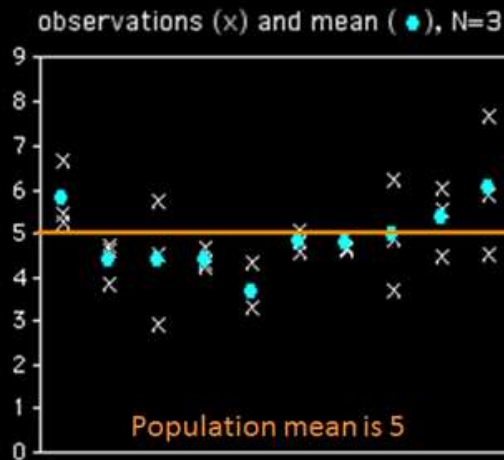


standard error, 100 experiments, N=3

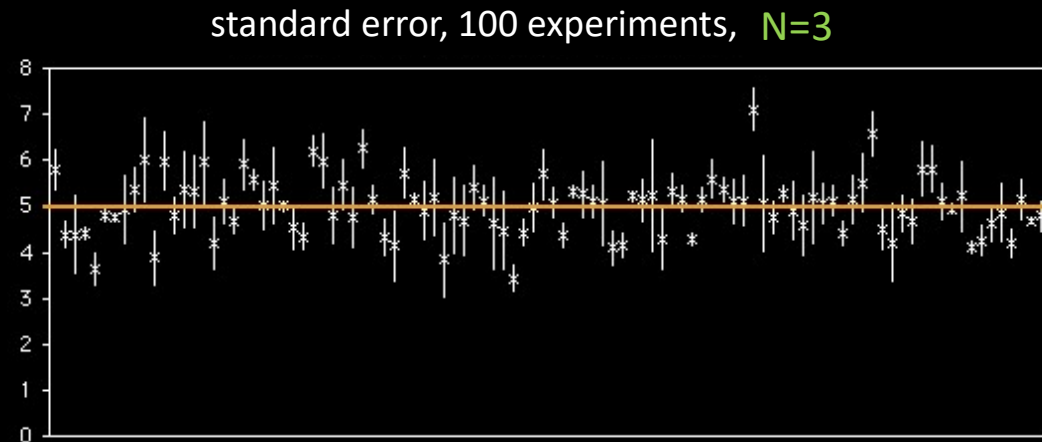


# Standard Error (2 of 2)

GA



If  $N = 20$ :  
What will happen to x's?  
What will happen to dots?



If  $N=20$ :  
What will happen to means?  
What will happen to bars?  
How many will cross the blue line?

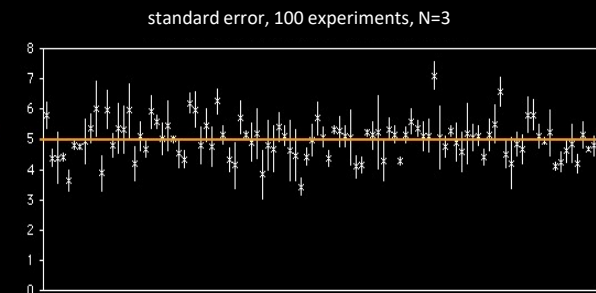
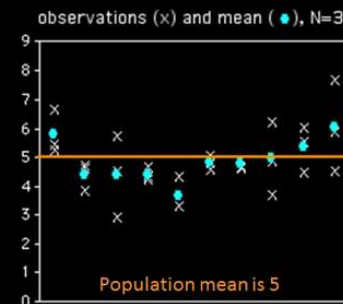
Breakout!

# Breakout 6



1. How many of the bars intersect?
  2. What do graphs look like  $N = 20$ ?
  3. Now, how many bars intersect?
- Icebreaker, Groupwork, Questions

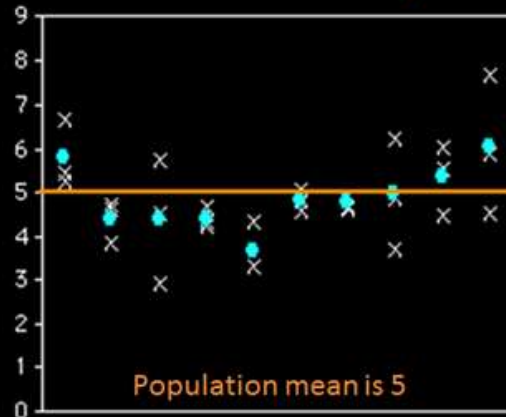
<https://web.cs.wpi.edu/~imgd2905/d21/breakout/breakout-6.html>



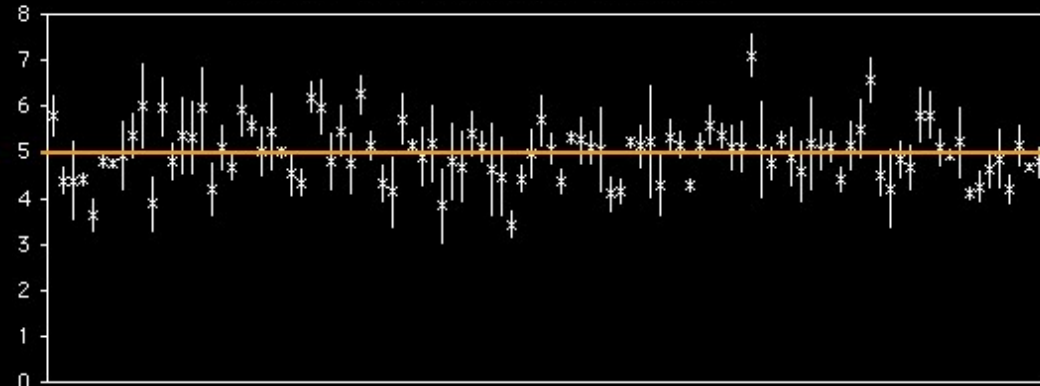


# Standard Error (2 of 2)

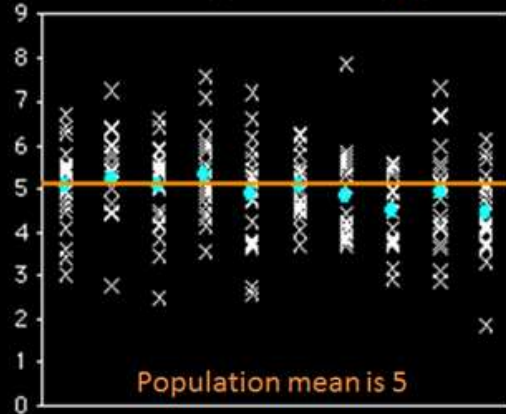
observations (x) and mean (●), N=3



standard error, 100 experiments, N=3



observations (x) and mean (●), N=20

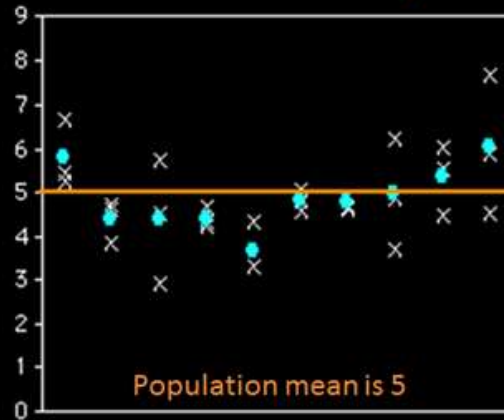


If N=20:  
What will happen to means?  
What will happen to bars?

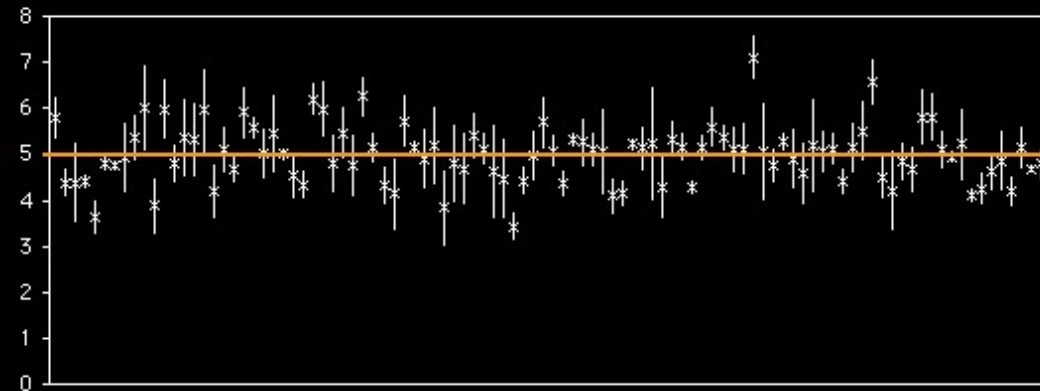


# Standard Error (2 of 2)

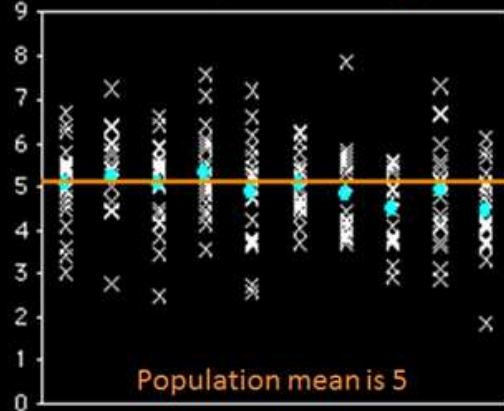
observations (x) and mean (●), N=3



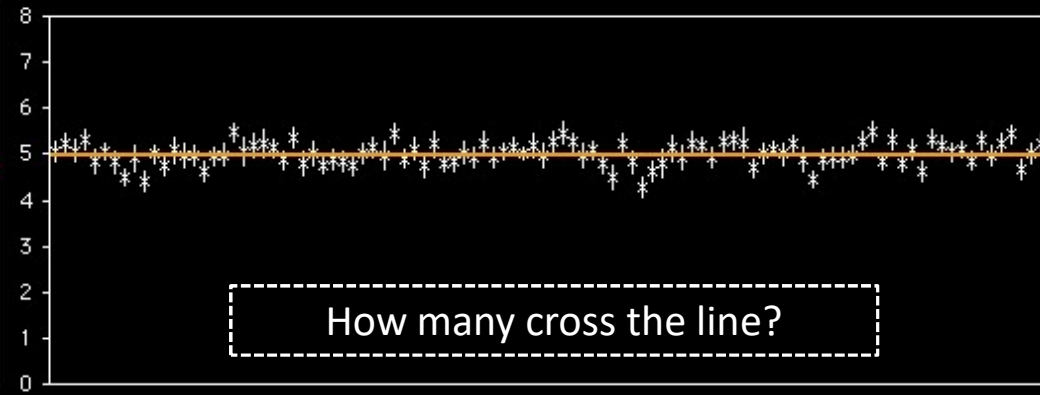
standard error, 100 experiments, N=3



observations (x) and mean (●), N=20



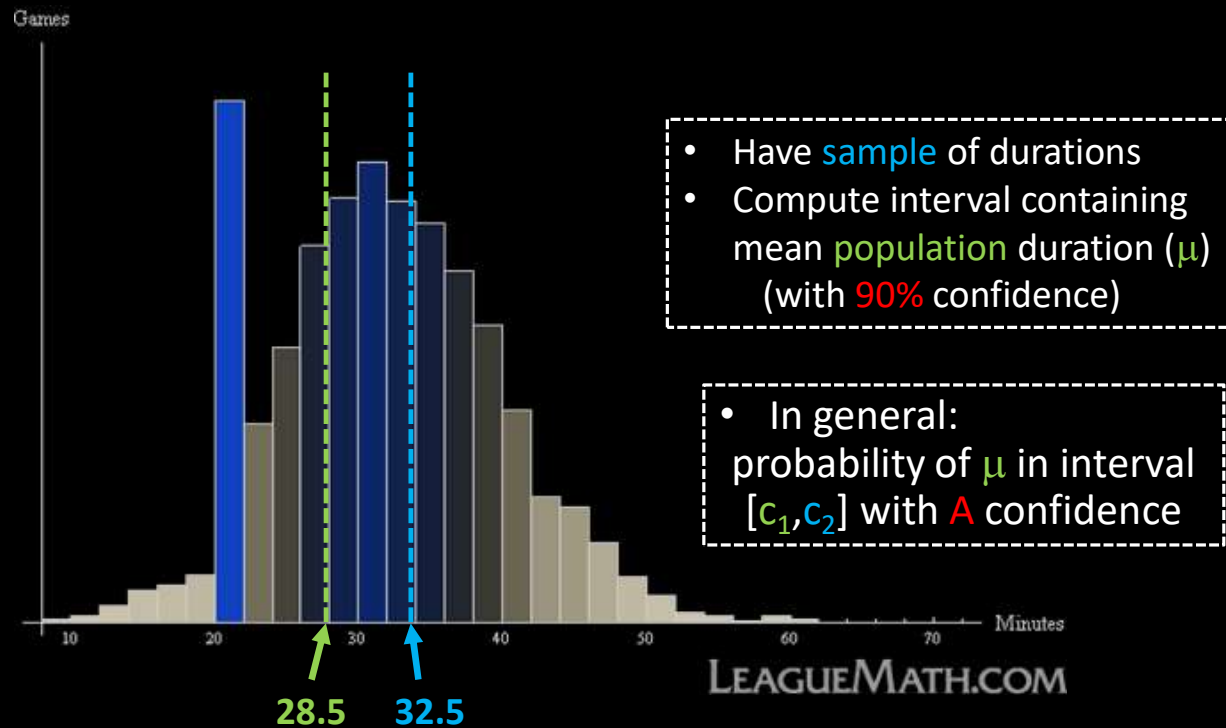
standard error, 100 experiments, N=20



Estimate population parameter → confidence interval

# Confidence Interval GA

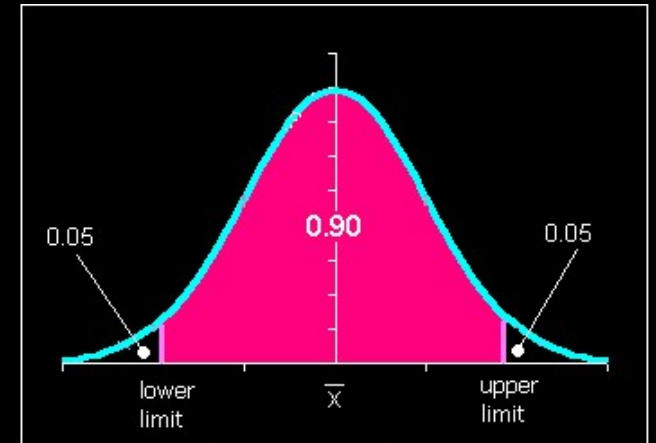
- Range of values with specific certainty that population parameter is within
  - e.g., 90% confidence interval for mean *League of Legends* match duration: [28.5 minutes, 32.5 minutes]



# Confidence Interval GA

- Probability of  $\mu$  in interval  $[c_1, c_2]$ 
  - $P(c_1 \leq \mu \leq c_2) = 1 - \alpha$   
 $[c_1, c_2]$  is *confidence interval*  
 $\alpha$  is *significance level*  
 $100(1 - \alpha)$  is *confidence level*
- Typically want  $\alpha$  small so confidence level **90%**, **95%** or **99%** (more on effect later)
- Say,  $\alpha = 0.1$ . Could do  $k$  experiments (size  $n$ ), find sample means, sort
  - Graph distribution
- Interval from distribution:
  - Lower bound: **5%**
  - Upper bound: **95%**
  - **90%** confidence interval

We have to do  $k$  experiments, each of size  $n$ ?



# Confidence Interval Estimate

GA

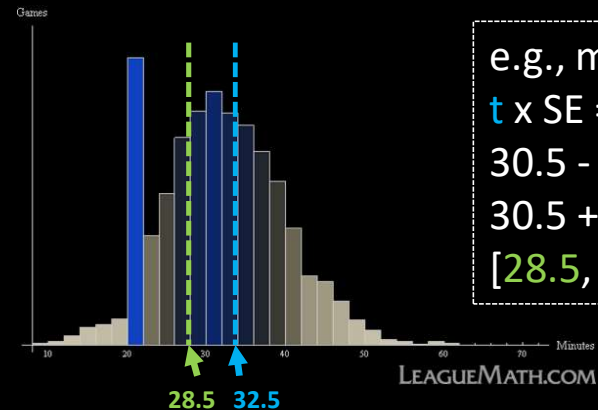
- Estimate interval from 1 experiment, size  $n$
  - Compute sample mean ( $\bar{x}$ ), sample standard error (SE)
  - Multiply SE by  $t$  distribution
  - Add/subtract from sample mean
- **Confidence interval**

$$\bar{X} \pm t \frac{s}{\sqrt{n}}$$



$$\left( \bar{x} - t \cdot \frac{s}{\sqrt{n}}, \bar{x} + t \cdot \frac{s}{\sqrt{n}} \right)$$

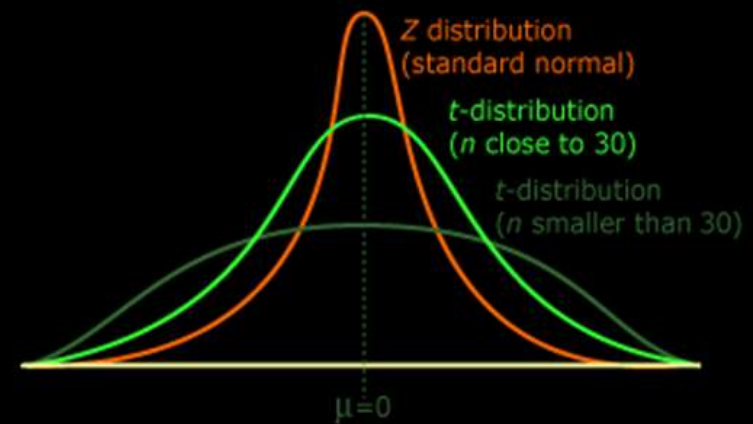
- Ok, what is  $t$  distribution?
  - Function, parameterized by  $\alpha$  and  $n$



e.g., mean 30.5  
 $t \times SE = 2$   
 $30.5 - 2 = 28.5$   
 $30.5 + 2 = 32.5$   
[28.5, 32.5]

# t distribution GA

- Looks like standard normal, but bit “squashed”
- Gets more squashed as  $n$  gets smaller
- Note, can use standard normal (z distribution) when large enough sample size ( $N = 30+$ )



[http://ci.columbia.edu/ci/premba\\_test/c0331/images/s7/6317178747.gif](http://ci.columbia.edu/ci/premba_test/c0331/images/s7/6317178747.gif)

aka **student's t distribution** (“student” was anonymous name used when published by William Gosset)

# Confidence Interval Example



(Unsorted)  
Game Time

4.4	3.9
3.8	3.2
2.8	4.1
4.2	3.3
2.8	2.8
2.9	4.2
1.9	3.1
5.9	4.5
3.9	4.5
3.2	4.8
4.1	4.9
5.3	5.1
3.6	3.7
5.1	3.4
2.7	5.6
3.9	3.1

- Suppose gathered game times in user study (e.g., for your MQP!)
  - Can compute *sample* mean, yes
  - But really want to know where *population* mean is
  - Can't know for sure, but can put within an interval
- Bound with **confidence interval**

# Confidence Interval Example



(Sorted)

Game Time	
1.9	3.9
2.7	3.9
2.8	4.1
2.8	4.1
2.8	4.2
2.9	4.2
3.1	4.4
3.1	4.5
3.2	4.5
3.2	4.8
3.3	4.9
3.4	5.1
3.6	5.1
3.7	5.3
3.8	5.6
3.9	5.9

- $\bar{x} = 3.90$ , stddev  $s=0.95$ ,  $n=32$
- A 90% confidence interval ( $\alpha$  is 0.1) for population mean ( $\mu$ ):

$$3.90 \pm \frac{1.696 \times 0.95}{\sqrt{32}} = [3.62, 4.19]$$

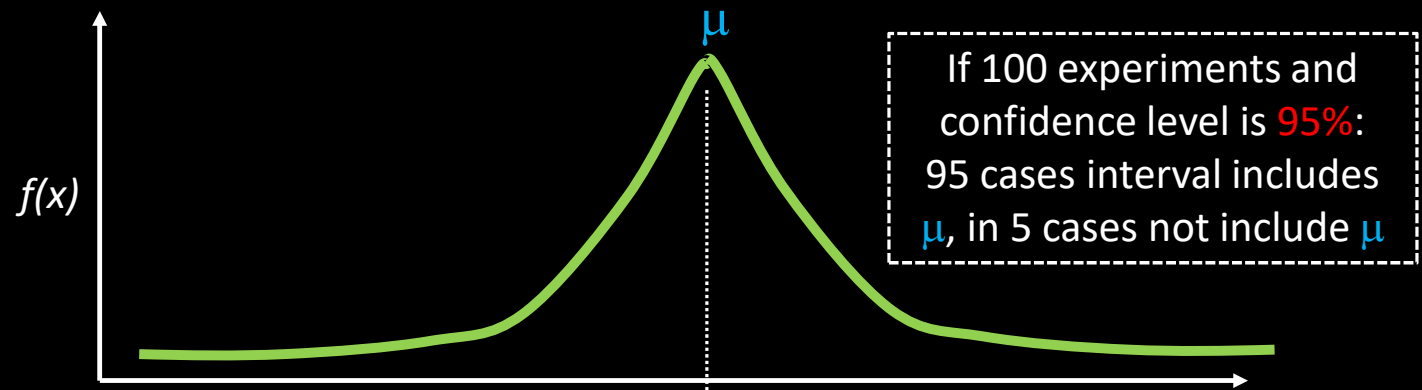
Lookup 1.645 in table, or  
=TINV(0.1, 31)



- With 90% confidence,  $\mu$  in that interval. Chance of error 10%.
- But ... what does that mean?

(See next slide for depiction of meaning)

# Meaning of Confidence Interval ( $\alpha$ )



Experiment/Sample

1



Includes  $\mu$ ?

yes

2



yes

3



no

...

100



yes

e.g.,

Total

yes  $\geq 100(1-\alpha)$

$\alpha = 0.05$

95

Total

no  $< 100\alpha$

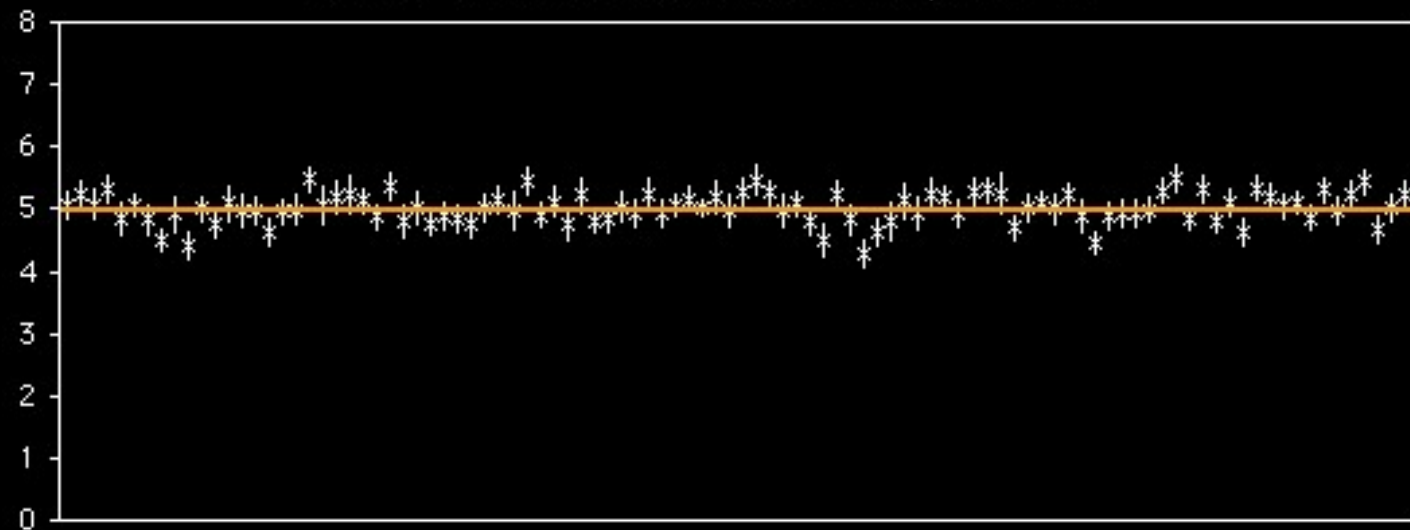
5



# Meaning of Confidence Interval ( $\alpha$ )



mean +/- 1 standard error of 100 samples, N=20



If 100 experiments and confidence level is 95%:  
95 cases interval includes  $\mu$ ,  
in 10 cases not include  $\mu$

# How does Confidence Interval Size Change?

GA

- With *sample size* ( $N$ )
- With *confidence level* ( $1-\alpha$ )

Look at each separately next

# How does Confidence Interval Change (1 of 2)?

GA

- What happens to confidence interval when *sample size* ( $N$ ) increases?
  - **Hint:** think about Standard Error

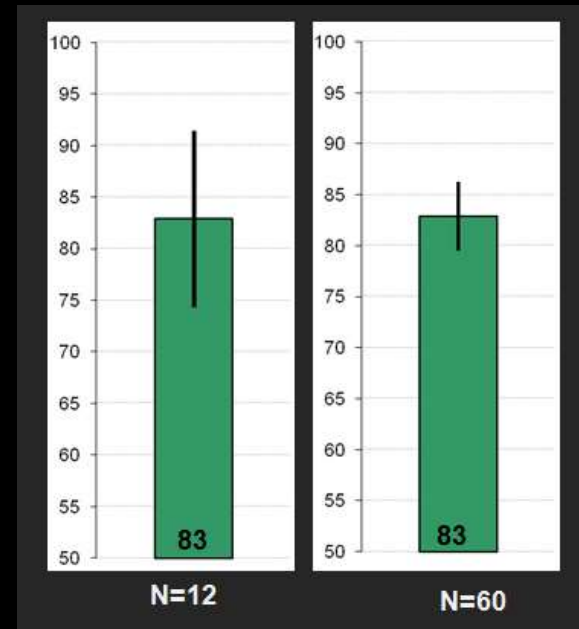
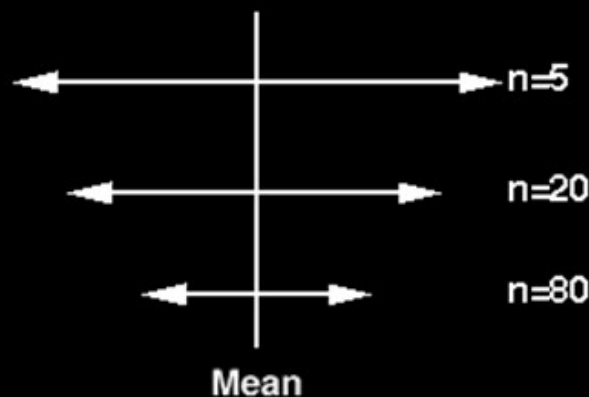
POLL!

# How does Confidence Interval Change (1 of 2)?

GA

- What happens to confidence interval when *sample size* ( $N$ ) increases?
  - **Hint:** think about Standard Error

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}} \quad \bar{X} \pm t \frac{s}{\sqrt{n}}$$



# How does Confidence Interval Change (2 of 2)?



- What happens to confidence interval when *confidence level* ( $1-\alpha$ ) increases?
- 90% CI = [6.5, 9.4]
  - 90% chance population value is between 6.5, 9.4
- 95% CI =

POLL!

# How does Confidence Interval Change (2 of 2)?

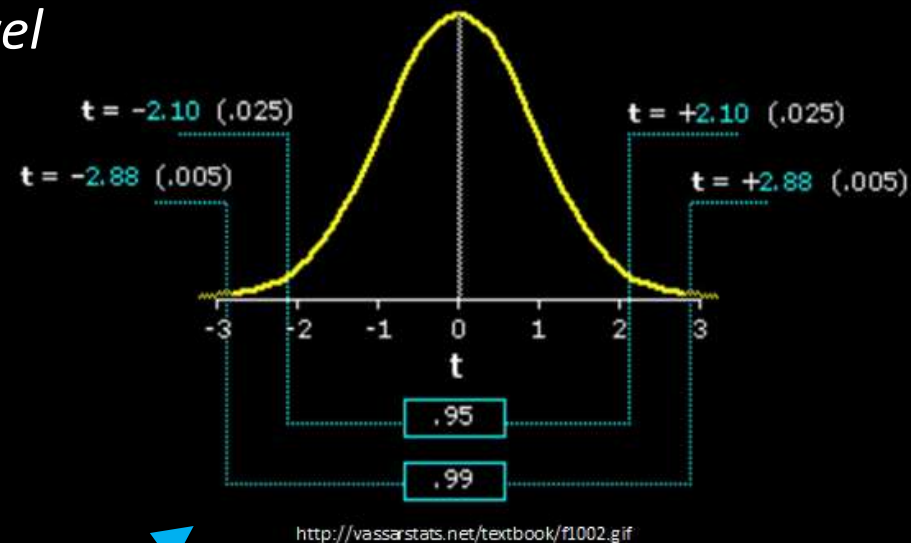


- What happens to confidence interval when *confidence level* ( $1-\alpha$ ) increases?
- 90% CI = [6.5, 9.4]
  - 90% chance population value is between 6.5, 9.4
- 95% CI = [6.1, 9.8]
  - 95% chance population value is between 6.1, 9.8
- Why is interval wider when we are “more” confident?

# How does Confidence Interval Change (2 of 2)?

GA

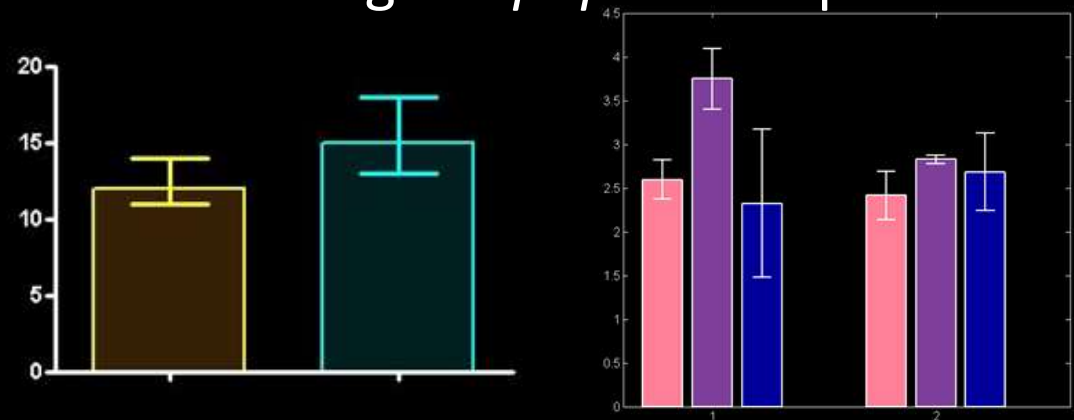
- What happens to confidence interval when *confidence level* ( $1-\alpha$ ) increases?
- **90% CI = [6.5, 9.4]**
  - 90% chance population value is between 6.5, 9.4
- **95% CI = [6.1, 9.8]**
  - 95% chance population value is between 6.1, 9.8
- Why is interval **wider** when we are “more” confident?



# Using Confidence Interval (1 of 3)

GA

- For charts, depict with **error bars**
  - CI more informative than standard deviation
    - Standard deviation doesn't change with **N**
- CI indicates range of *population* parameter

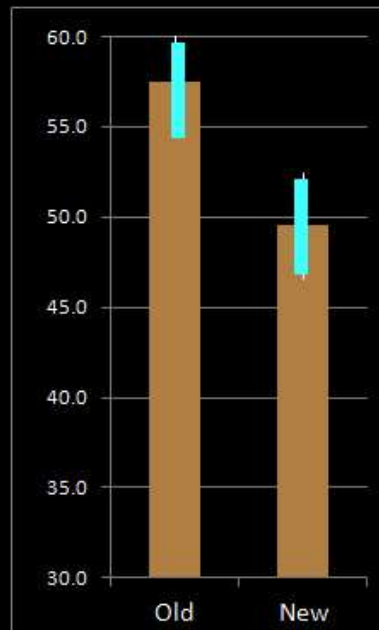


Make sure sample size **N=30+**  
(**N=15+** if somewhat normal.  
Any **N** if know distro is normal)

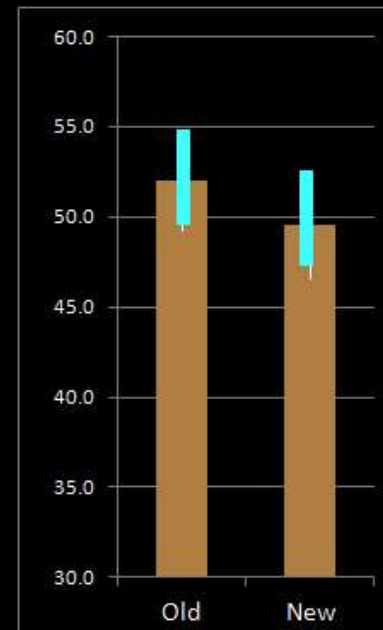


# Using Confidence Interval (2 of 3)

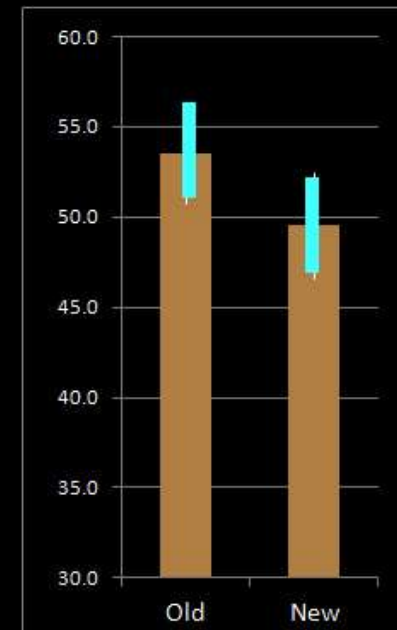
GA



No overlap



Large overlap



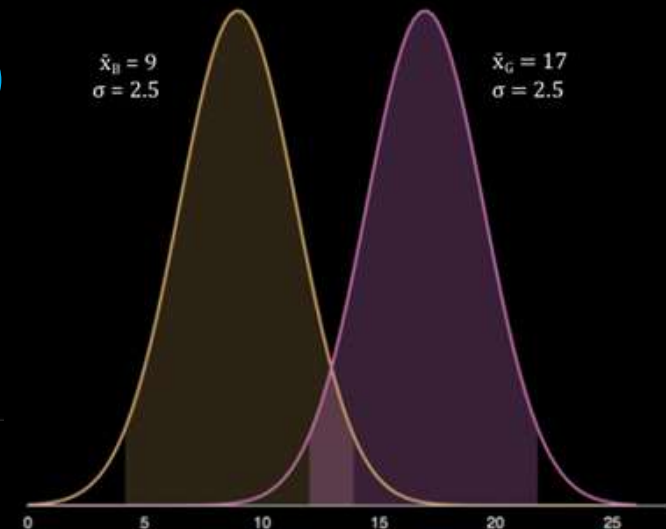
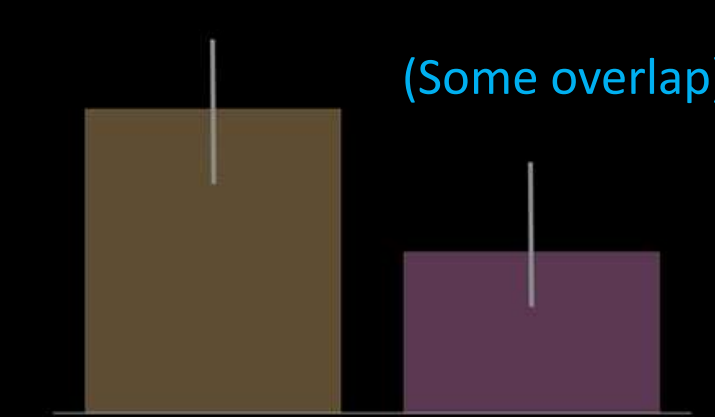
Some overlap

Compare two alternatives, quick check for statistical significance

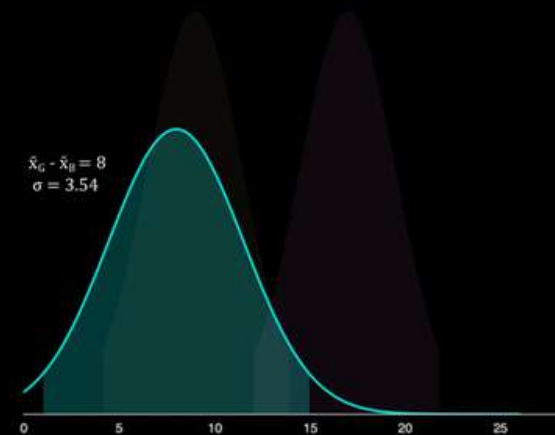
- No overlap? → 90% confident difference (at  $\alpha = 0.10$  level)
- Large overlap (50%+)? → No statistically significant diff (at  $\alpha = 0.10$  level)
- Some overlap? → more tests required

# Using Confidence Interval (3 of 3)

GA



(Here is the overlap)



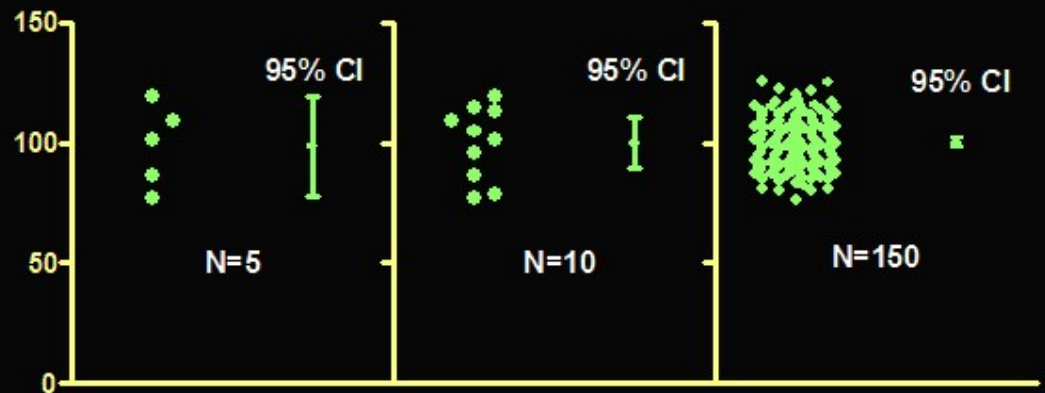
But if compute difference, then confidence interval may **not** cross 0!  
(Caused by error propagation)

# How *Not* to Use Confidence Intervals



*“The confidence intervals of the two groups overlap, hence the difference is not statistically significant” — Many people*

- Overlap – careful not to say no statistically significant difference (see previous slide)
- Do not quantify variability (e.g., 95% of values in interval)



# Statistical Significance versus Practical Significance (1 of 2)



**Warning:** may find statistically significant difference.  
That doesn't mean it is *important*.

**It's a Honey of an O**

**Latency can Kill?**

# Statistical Significance versus Practical Significance (1 of 2)



**Warning:** may find statistically significant difference.  
That doesn't mean it is *important*.

## It's a Honey of an O

- Boxes of Cheerios, Tostitos both target 12 oz.
- Measure weight of 18,000 boxes
- Using statistics:
  - Cheerio's heavier by 0.002 oz.
  - And statistically significant ( $\alpha=0.99$ )!
- But ... 0.0002 is only 2-3 O's. Customer doesn't care!

## Latency can Kill?

# Statistical Significance versus Practical Significance (2 of 2)



**Warning:** may find statistically significant difference.  
That doesn't mean it is *important*.

## It's a Honey of an O

- Boxes of Cheerios, Tostitos both target 12 oz.
- Measure weight of 18,000 boxes
- Using statistics:
  - Cheerio's heavier by 0.002 oz.
  - And statistically significant ( $\alpha=0.99$ )!
- But ... 0.0002 is only 2-3 O's. Customer doesn't care!

## Latency can Kill?

- Lag in League of Legends
- Pay \$\$ to upgrade Ethernet from 100 Mb/s to 1000 Mb/s
- Measure ping to LoL server for 20,000 samples
- Using statistics
  - Ping times improve 0.8 ms
  - And statistically significant ( $\alpha=0.99$ )!
- But ... below perception!

# Effect Size (1 of 2)



- Quantitative measure of strength of finding
  - Measures practical significance
- Emphasizes size of difference of relationship

$$\text{Effect size} = \frac{\text{Mean of experimental group} - \text{Mean of control group}}{\text{Standard deviation}}$$

(Cohen's d)

$$z = \frac{X - \bar{X}}{s}$$

Similar to Z-score

$$CV (\%) = \left( \frac{\text{Standard deviation}}{\text{Mean}} \right) \times 100$$

Similar to Coefficient of Variation

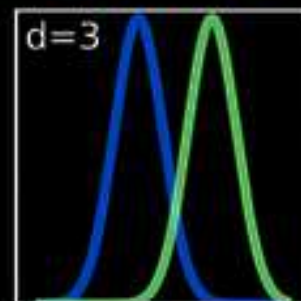
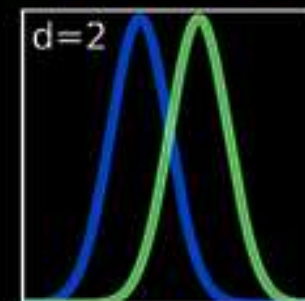
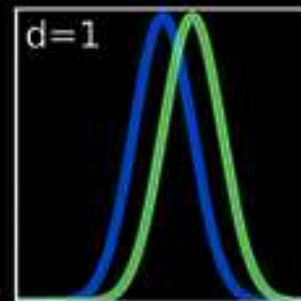
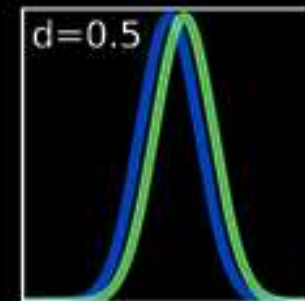
# Effect Size (2 of 2)

GA

- Quantitative measure of strength of finding
  - Measures practical significance
- Emphasizes size of difference of relationship

Relative size	Effect size	% of control group below the mean of experimental group
	0.0	50%
Small	0.2	58%
Medium	0.5	69%
Large	0.8	79%
	1.4	92%

<https://www.simplypsychology.org/cohen-d.jpg>





# What Confidence Level to Use (1 of 2)?

GA

- Often see 90% or 95% (or even 99%) used
- Choice based on **loss** if wrong (population parameter is outside), **gain** if right (parameter inside)
  - If **loss** is high compared to **gain**, use higher confidence
  - If **loss** is low compared to **gain**, use lower confidence
  - If **loss** is negligible, lower is fine
- Example (**loss** high compared to **gain**):
  - Hairspray, makes hair straight, but has chemicals
  - Want to be **99.9%** confident it doesn't cause cancer
- Example (**loss** low compared to **gain**):
  - Hairspray, makes hair straight, mainly water
  - Ok to be **75%** confident it straightens hair

# What Confidence Level to Use (2 of 2)?

GA

- Often see 90% or 95% (or even 99%) used
- Choice based on **loss** if wrong (population parameter is outside), **gain** if right (parameter inside)
  - If **loss** is high compared to **gain**, use higher confidence
  - If **loss** is low compared to **gain**, use lower confidence
  - If **loss** is negligible, lower is fine
- Example (**loss** negligible compared to **gain**):
  - Lottery ticket costs \$1, pays \$5 million
  - Chance of winning is  $10^{-7}$  (50% payout, so 1 in 10 million)
  - To win with **90%** confidence, need 9 million tickets
    - + No one would buy that many (\$9 mil to win \$5 mil)!
  - But many people happy with **0.0001%** confidence

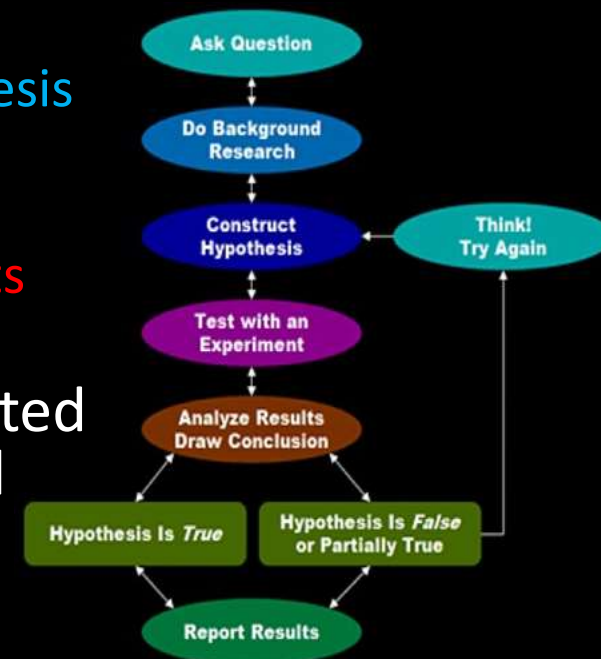
# Outline

- Overview (done)
- Foundation (done)
- Inferring Population Parameters (done)
- Hypothesis Testing (next)

# Hypothesis Testing

GA

- Term arises from science
    - State tentative explanation → hypothesis
    - Devise experiments to gather data
    - Data supports or rejects hypothesis
  - Statisticians have adopted to test using inferential statistics
- Hypothesis testing



[http://s1.hubimg.com/u/4205792\\_620.jpg](http://s1.hubimg.com/u/4205792_620.jpg)

Just brief overview here → *Conversant*  
Chapters 8 & 9 in book have more

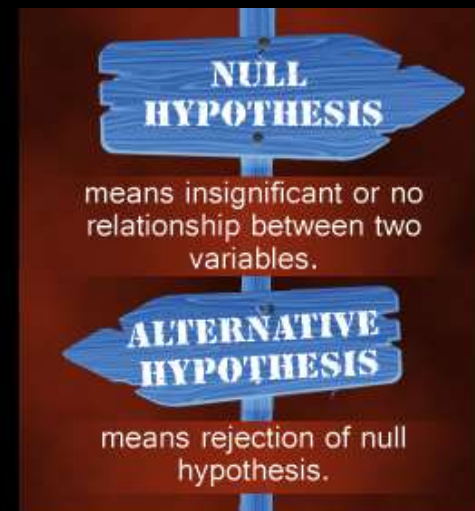
# Hypothesis Testing Terminology (1 of 2)



- **Null Hypothesis ( $H_0$ )** – no significance difference between measured and population (any observed difference due to error)
  - e.g., population mean for Riot to bring up NA servers 4 hours
- **Alternative Hypothesis** – hypothesis contrary to null hypothesis
  - e.g., population mean for Riot to bring up NA servers is *not* 4 hours
- Care about **Alternate**, but test **Null**
  - If data supports, **Alternate** likely not true
  - If data rejects, **Alternate** *may* be true
- Why **Null** and **Alternate**?
  - Remember, data doesn't “prove” hypothesis
  - Can only reject it (at certain significance)
  - So, reject **Null**

# Hypothesis Testing Terminology (2 of 2)

GA



<http://www.buzzle.com/img/articleImages/605910-49223-57.jpg>

- **P value** – smallest level that can reject  $H_0$ 
  - “If **p value** is low, then  $H_0$  must go”
- How “low” based on “risk” of being wrong (like conf. interval)

# Hypothesis Testing Steps

1. State hypothesis ( $H$ ) and null hypothesis ( $H_0$ )
2. Evaluate risks of being wrong (based on loss and gain), choosing significance ( $\alpha$ ) and sample size
3. Collect data ( $\text{sample}$ ), compute statistics
4. Calculate  $p$  value based on test statistic and compare to  $\alpha$
5. Make inference
  - Reject  $H_0$  if  $p$  value less than  $\alpha$ 
    - + So,  $H$  may be right
  - Do not reject  $H_0$  if  $p$  value greater than  $\alpha$ 
    - + So,  $H$  may not be right

# Hypothesis Testing Steps

## Example – 1 of 2

GA

- State hypothesis (**H**) and null hypothesis (**H<sub>0</sub>**)
  - **H**: Mario takes less than 5 minutes to complete
  - **H<sub>0</sub>**: Mario takes 5 minutes to complete (**H<sub>0</sub>** always has =)
- Evaluate risks of being wrong (based on loss and gain), choosing significance ( **$\alpha$** ) and sample size (**N**)
  - Player may get frustrated, quit game, so  **$\alpha$**  = 0.1
  - Not sure of normally distributed, so **30** (Central Limit Theorem)



# Hypothesis Testing Steps

## Example – 2 of 2

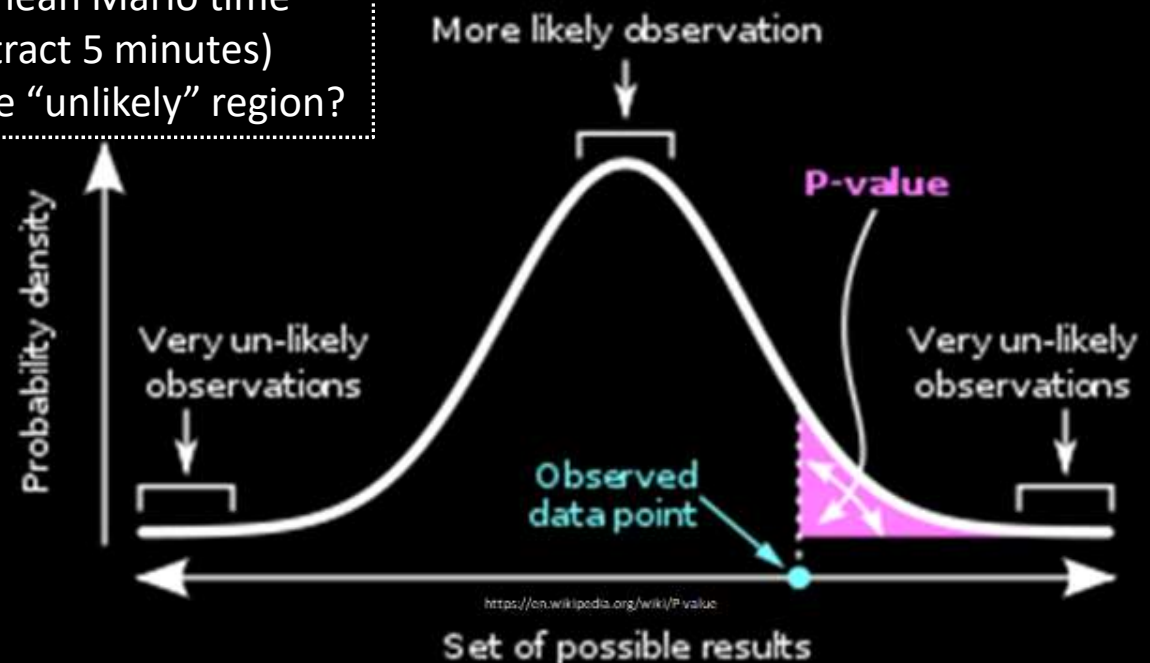
GA

- Collect data (**sample**), compute statistics
  - **30** people play level, compute average minutes (say, 7.2), compare to 5
- Calculate **p value** based on test statistic and compare to  $\alpha$ 
  - **p value** = 0.002,  $\alpha$  = 0.01
- Make inference
  - Here: **p value** less than  $\alpha$   $\rightarrow$  REJECT  $H_0$ , so **H** may be right
  - Note, would not have rejected  $H_0$  if **p value** greater than  $\alpha$  (**H** may have been wrong)

# Depiction of P Value

Probability density of each outcome, computed under **Null** hypothesis  
**p value** is area under curve past **observed data point** (e.g., sample mean)

E.g., mean Mario time  
(subtract 5 minutes)  
Is it in the “unlikely” region?



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.