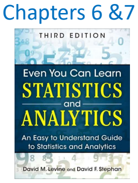
IMGD 2905

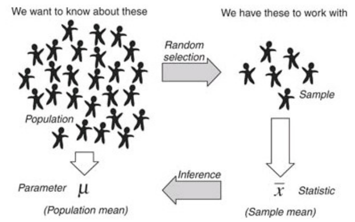
Inferential Statistics

Chapters 6 &7



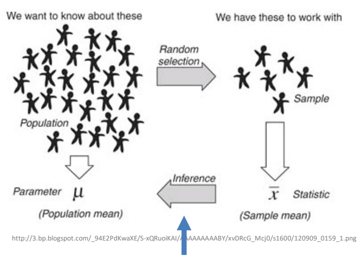
Overview

- Use statistics to infer population parameters



Overview

- Use statistics to infer population parameters



Inferential statistics

Outline

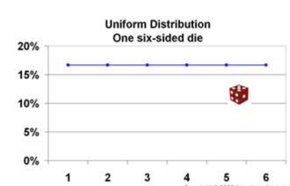
- Overview (done)
- Foundation (next)
- Confidence Intervals
- Hypothesis Testing

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?



"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)

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

"Triangle" distribution

Dice Rolling (3 of 4)

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

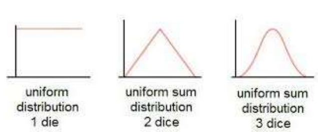
Dice Rolling (3 of 4)

- 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 (3 of 4)

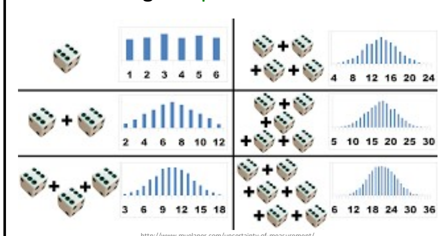
- Have 1d6, sample thrice and sum (i.e., roll 3 dice)
- What is probability distribution of values?



What's happening to the shape?

Dice Rolling (4 of 4)

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

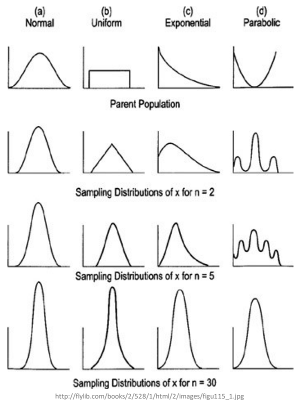


Resulting sum/mean follows a normal distribution
→ Even though base distribution is uniform!

Ok, neat – for "square" distributions.
But what about experiments with other distributions?

Sampling Distributions

- With "enough" samples, looks "bell-shaped" → Normal!
- How many is enough?
  - 30 (15 if symmetric distribution)
- Central Limit Theorem
  - Sum of independent variables tends towards Normal distribution



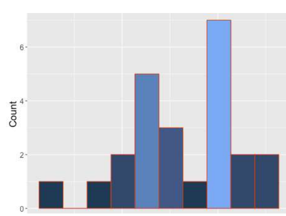
Why do we care about sample means following Normal distribution?

- 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?

Why do we care about sample means following Normal distribution?

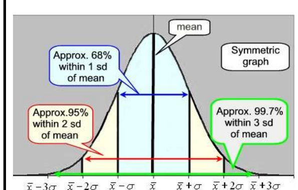
- What if we had only a sample mean and no measure of spread
  - e.g., mean rank for Overwatch is 50
- What can we say about population mean?
  - Not a whole lot!
  - Yes, population mean could be 50. But could be 100. How likely are each?

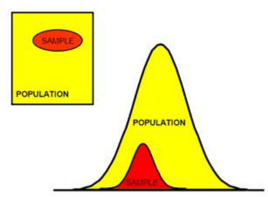
  → No idea!



Why do we care about sample means following Normal distribution?

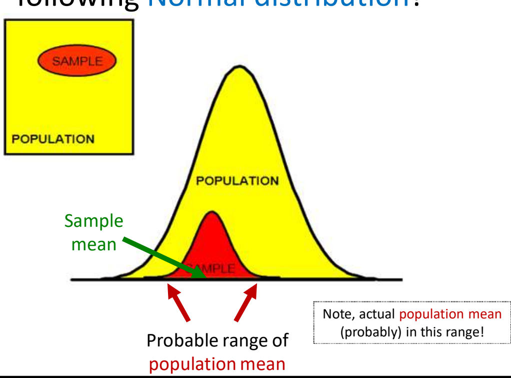
- Remember this?





With mean and standard deviation → Allows us to predict range to bound population mean

Why do we care about sample means following Normal distribution?



Outline

- Overview (done)
- Foundation (done)
- Confidence Intervals (next)
- Hypothesis Testing

Sampling Error (1 of 2)

- Population of 200 game durations
  - Mean μ = 69.637
  - Std Dev σ = 10.411
- Experiment w/20 samples
  - Each 15 game durations (with replacement)
  - Table on right has 20 experiments
- Observations?

Sample	Mean	Standard Deviation	Minimum	Median	Maximum	Range
1	66.12	9.21	47.20	65.00	87.00	39.80
2	73.30	12.48	52.40	71.10	101.10	48.70
3	68.67	10.78	54.00	69.10	85.40	31.40
4	69.95	10.57	54.50	68.00	87.80	33.30
5	73.27	13.56	54.40	71.80	101.10	46.70
6	69.27	10.04	50.10	70.30	85.70	35.60
7	66.75	9.38	52.40	67.30	82.60	30.20
8	68.72	7.62	54.50	68.80	81.50	27.00
9	72.42	9.97	50.10	71.90	88.90	38.80
10	69.25	10.68	51.10	66.50	85.40	34.30
11	72.56	10.60	60.20	69.10	101.10	40.90
12	69.48	11.67	49.10	69.40	97.70	48.60
13	64.65	9.71	47.10	64.10	78.50	31.40
14	68.85	14.42	46.80	69.40	88.10	41.30
15	67.91	8.34	52.40	69.40	79.60	27.20
16	66.22	10.18	51.00	66.40	85.40	34.40
17	68.17	8.18	54.20	66.50	86.10	31.90
18	68.73	8.50	57.70	66.10	84.40	26.70
19	68.57	11.08	47.10	70.40	82.60	35.50
20	75.80	12.49	56.70	77.10	101.10	44.40

Sampling Error (1 of 2)

- Population of 200 game durations
  - Mean μ = 69.637
  - Std Dev σ = 10.411
- Experiment w/20 samples
  - Each 15 game durations (with replacement)
  - Table on right has 20 experiments
- Observations?
  - Stats ($\bar{x}$, s) differ each time!
  - Sometimes higher, sometimes lower than population (μ , σ)
  - Sample range varies a lot more than sample standard deviation
  - Population mean (μ) always within sample range

Sample	Mean	Standard Deviation	Minimum	Median	Maximum	Range
1	66.12	9.21	47.20	65.00	87.00	39.80
2	73.30	12.48	52.40	71.10	101.10	48.70
3	68.67	10.78	54.00	69.10	85.40	31.40
4	69.95	10.57	54.50	68.00	87.80	33.30
5	73.27	13.56	54.40	71.80	101.10	46.70
6	69.27	10.04	50.10	70.30	85.70	35.60
7	66.75	9.38	52.40	67.30	82.60	30.20
8	68.72	7.62	54.50	68.80	81.50	27.00
9	72.42	9.97	50.10	71.90	88.90	38.80
10	69.25	10.68	51.10	66.50	85.40	34.30
11	72.56	10.60	60.20	69.10	101.10	40.90
12	69.48	11.67	49.10	69.40	97.70	48.60
13	64.65	9.71	47.10	64.10	78.50	31.40
14	68.85	14.42	46.80	69.40	88.10	41.30
15	67.91	8.34	52.40	69.40	79.60	27.20
16	66.22	10.18	51.00	66.40	85.40	34.40
17	68.17	8.18	54.20	66.50	86.10	31.90
18	68.73	8.50	57.70	66.10	84.40	26.70
19	68.57	11.08	47.10	70.40	82.60	35.50
20	75.80	12.49	56.70	77.10	101.10	44.40

This variation → Sampling error

Sampling Error (2 of 2)

- 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 (σ) itself – more variation, more sample statistic variation (s)
  - Sample size (N) – larger sample, lower error
    - *Q: Why can't we just make sample size super large?*
- How much does it vary? → Standard error

Standard Error (1 of 2)

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

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

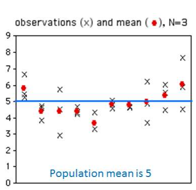
Example:
n = 5
σ = 17

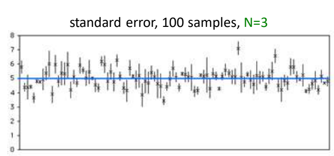
$= \frac{17}{\sqrt{5}}$

SE = 7.6

(Example next)

Standard Error (2 of 2)



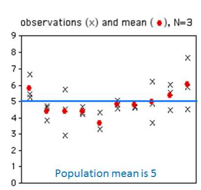


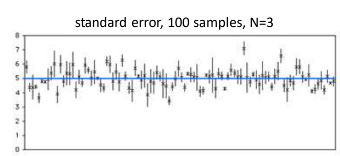
If N = 20:
What will happen to x's?
What will happen to dots?

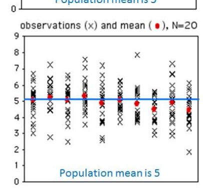
If N=20: what will happen to bars?

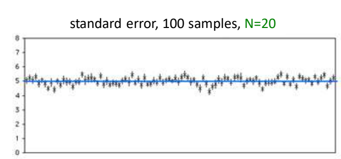
http://www.biostathandbook.com/standarderror.html

Standard Error (2 of 2)







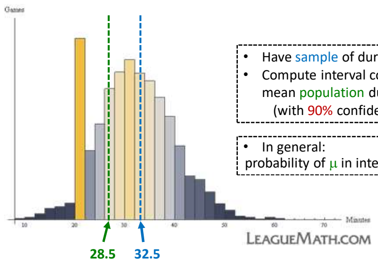


Estimate population parameter → confidence interval

http://www.biostathandbook.com/standarderror.html

Confidence Interval

- 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]



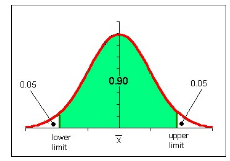
- Have sample of durations
- Compute interval containing mean population duration (μ) (with 90% confidence)

- In general: probability of μ in interval $[c_1,c_2]$

Confidence Interval for Mean

- Probability of μ in interval $[c_1,c_2]$
  - $P(c_1 \le \mu \le c_2) = 1-\alpha$
  - [c1, c2] is *confidence interval*
  - α is *significance level*
  - $100(1-\alpha)$ is *confidence level*
- Typically want α 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
  - Cumulative distribution
- Interval from distribution:
  - Lower bound: 5%
  - Upper bound: 95%
  - → 90% confidence interval

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



http://www.comfsm.fm/~dleeling/statistics/notes009_normalcurve90.png

Confidence Interval Estimate

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

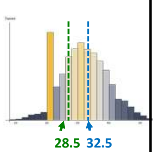
→ Confidence interval

- Ok, what is t distribution?
  - Function, parameterized by α and *n*

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

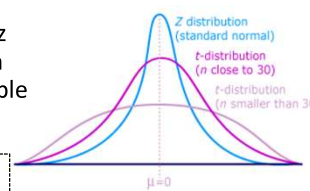
$$\left(\overline{x} - t\cdot\frac{s}{\sqrt{n}}, \overline{x} + t\cdot\frac{s}{\sqrt{n}}\right)$$

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



t distribution

- 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

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

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 (α is 0.1) for population mean (μ):

$$3.90 \pm \frac{1.696 \times 0.95}{\sqrt{32}}$$

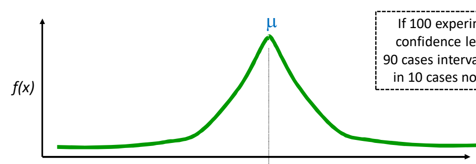
$$= [3.62, 4.19]$$

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

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

(See next slide for depiction of meaning)

Meaning of Confidence Interval (α)



If 100 experiments and confidence level is 90%: 90 cases interval includes μ, in 10 cases not include μ

Experiment/Sample	Includes μ?	
1	yes	
2	yes	
3	no	
…		e.g.,
100	yes	α =0.1
Total	yes ≥ 100 (1-α)	90
Total	no < 100 α	10

How does Confidence Interval Size Change?

- With *sample size* (N)
- With *confidence level* (α)

Look at each separately next

How does Confidence Interval Change (1 of 2)?

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

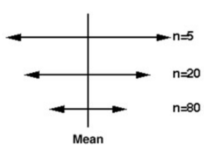
$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

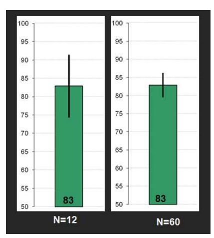
How does Confidence Interval Change (1 of 2)?

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

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

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



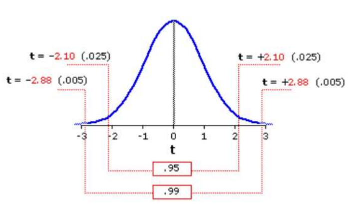


How does Confidence Interval Change (2 of 2)?

- 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)?

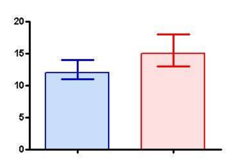
- 90% CI = [6.5, 9.4]
  - 90% chance population value is between 6.5, 9.4
- 95% CI = [6.1, 9.8]
  - 95% chance population value is between 6.1, 9.8
- Why is interval wider when we are "more" confident?

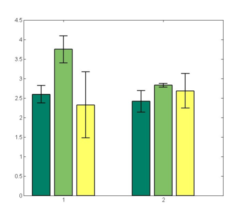


Using Confidence Interval (1 of 2)

- Indicator of spread → Error bars
- CI can be more informative than standard deviation

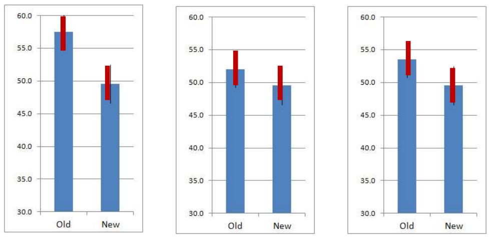
→ indicates range of *population* parameter (make sure sample size 30+!)





Using Confidence Interval (2 of 2)

https://measuringu.com/ci-10things/



No overlap | Large overlap | Some overlap

Compare two alternatives, quick check for statistical significance

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

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, Tastee-O's both target 12 oz.
- Measure weight of 18,000 boxes
- Using statistics:
  - Cheerio's heavier by 0.002 oz.
  - And statistically significant (α=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, Tastee-O's both target 12 oz.
- Measure weight of 18,000 boxes
- Using statistics:
  - Cheerio's heavier by 0.002 oz.
  - And statistically significant (α=0.95)!
- 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 (α=0.99)!
- But … humans cannot notice 1 ms difference!

What Confidence Level to Use (1 of 2)?

- 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 99.99% confident it doesn't cause cancer
- Example (loss low compared to gain):
  - Hairspray, makes hair straight, only uses water
  - Ok to be 75% confident it straightens hair

What Confidence Level to Use (2 of 2)?

- 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):
  - Lottery ticket $1, pays $5 million
  - Chance of winning is 10^{-7} (1 in 10 million)
  - To win with 90% confidence, need 9 million tickets
    - No one would buy that many tickets!
  - So, most people happy with 0.01% confidence

Outline

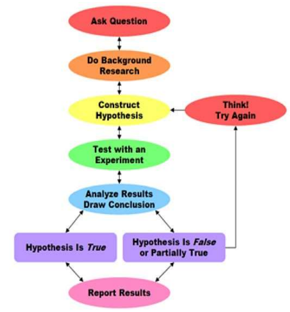
- Overview (done)
- Foundation (done)
- Confidence Intervals (done)
- Hypothesis Testing (next)

Hypothesis Testing

- 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

Just brief overview here. Next chapter in book has more.



http://s1.hubimg.com/u/4205792_f520.jpg

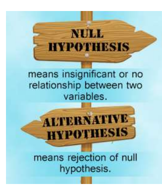
Hypothesis Testing Terminology

- Null Hypothesis (H_0) – hypothesis that no significance difference between measured value and population parameter (any observed difference due to error)
  - e.g., population mean time for Riot to bring up NA servers was 4 hours
- Alternative Hypothesis – hypothesis contrary to null hypothesis
  - e.g., population mean time for Riot to bring up NA servers was *not* 4 hours
- Care about alternate, but test null
  - If data supports, alternate 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

- P-value – smallest level that can reject H_0

"If p-value is low, then H_0 must go"

- How "low", consider s"risk" of being wrong



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

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 (α) and sample size
3. Collect data (sample), compute statistics
4. Calculate p-value based on test statistic and compare to α
5. Make inference
   - Reject H_0 if p-value less than α
   - Do not reject H_0 if p-value greater than α

Hypothesis Testing Steps (Example)

- State hypothesis (H) and null hypothesis (H_0)
  - H: Mario level takes less than 5 minutes to complete
  - H_0: Mario level takes 5 minutes to complete (H_0 always has =)
- Evaluate risks of being wrong (based on loss and gain), choosing significance (α) and sample size
  - Player may get frustrated, quit game, so $\alpha = 0.01$
  - Note sure of normally distributed, so 30 (Central Limit Theorem)
- Collect data (sample), compute statistics
  - 30 people play level, compute average time, compare to 5
- Calculate p-value based on test statistic and compare to α
  - p-value = 0.002, $\alpha = 0.01$
- Make inference
  - Reject H_0 if p-value less than α (REJECT H_0), so H may be right
  - Do not reject H_0 if p-value greater than α