Inferential Statistics

Chapter 6 & 7
Overview

• Use statistics to infer population parameters
Outline

• Overview (done)
• Foundation (next)
• Inferring Population Parameters
• Hypothesis Testing
Groupwork

Remember, *probability distribution* shows possible outcomes on x-axis and probability of each on y-axis.

1. Describe the probability distribution of 1 d6?
2. Describe the probability distribution of 2 d6?
3. Describe the probability distribution of 3 d6?

Icebreaker, Groupwork, Questions

[https://web.cs.wpi.edu/~imgd2905/d23/groupwork/6-prob-dist(handout.html)](https://web.cs.wpi.edu/~imgd2905/d23/groupwork/6-prob-dist(handout.html))

[https://academo.org/demos/dice-roll-statistics/](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?

http://www.investopedia.com/articles/06/probabilitydistribution.asp
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?

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

"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?

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

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)

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

- With “large enough” sample size, sum/mean looks “bell-shaped” → Normal!
Sampling Distributions

- With “large enough” sample size, \textit{sum/mean} looks “bell-shaped” \(\rightarrow\) \textbf{Normal}!
Sampling Distributions

- With “large enough” sample size, sum/mean looks “bell-shaped” → Normal!
Sampling Distributions

• With “large enough” sample size, sum/mean looks “bell-shaped” → Normal!

http://flylib.com/books/2/528/1/html/2/images/figu115_1.jpg
Sampling Distributions

- With “large enough” sample size, sum/mean looks “bell-shaped” → Normal!
Sampling Distributions

- With “large enough” sample size, sum/mean looks “bell-shaped” → Normal!
Sampling Distributions

- With “large enough” sample size, \textbf{sum/mean} looks “bell-shaped” \rightarrow \textbf{Normal}!
- How many is large enough?
  - 30 (15 if symmetric distribution)
Sampling Distributions

- With “large enough” sample size, sum/mean looks “bell-shaped” → Normal!
- How many is large enough?
  - 30 (15 if symmetric distribution)
- Central Limit Theorem
  - Sum/mean 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 score is 3
• What can we say about population mean?
  – Not a whole lot!
  – Yes, population mean could be 6. But could be 0. How likely are each?
  → No idea!

Sample mean Population mean?
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 (see next slide)
Why do we care about sample means following Normal distribution?

Sample mean

Probable range of population mean

Actual population mean (probably) in this range!
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?

<table>
<thead>
<tr>
<th>Sample</th>
<th>Sample Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 d6</td>
<td>4.0</td>
</tr>
<tr>
<td>2 d6 (4 + 2) / 2 =</td>
<td>3.0</td>
</tr>
<tr>
<td>3 d6 (1 + 6 + 2) / 3 =</td>
<td>2.3</td>
</tr>
<tr>
<td>4 d6 (4 + 4 + 2 + 3) / 4 =</td>
<td>3.3</td>
</tr>
</tbody>
</table>
Estimating Population Mean

Q: What happens as sample size increases?
Q: How big a sample do we need?
Estimating Population Mean

Q: What happens as sample size increases?

Q: How big a sample do we need?

- Depends upon how much varies
- Values that are not the mean contribute to “error” \(\rightarrow\) 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 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? $\rightarrow$ Standard error
Standard Error

- Amount **sample means** will vary from experiment to experiment of same size
  - *Standard deviation of the sample means*
- Also, likelihood that sample statistic is near population parameter

- What does the size of the standard error depend upon? (Hint: see formula above)

\[
\text{standard error} \quad SE = \frac{S}{\sqrt{n}} \quad \text{sample size}
\]

**Example:**
\[
\begin{align*}
  n &= 5 \\
  S &= 17 \\
  SE &= \frac{17}{\sqrt{5}} = 7.6
\end{align*}
\]

So what? Reason about population mean e.g., **95% confident** that sample mean is within ~ 2 SE’s
(where does this come from?)
Standard Error

- Amount **sample means** will 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 \((s)\)

So what? Reason about population mean e.g., 95% confident that sample mean is within \(\sim 2 \text{ SE’s}\)
(Example next)
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?

Groupwork!

http://www.biostathandbook.com/standarderror.html
1. How many of the bars intersect the blue?
2. What do graphs look like $N = 20$?
3. Now, how many bars intersect?

- Standard Error

[https://web.cs.wpi.edu/~imgd2905/d23/groupwork/7-std-error/handout.html](https://web.cs.wpi.edu/~imgd2905/d23/groupwork/7-std-error/handout.html)
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?

Population mean is 5
Standard Error (2 of 2)

Estimate population parameter → confidence interval

Population mean is 5

How many cross the blue line?
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 ($\mu$) (with 90% confidence)

- In general: probability of $\mu$ in interval $[c_1, c_2]$ with A confidence
Confidence Interval for Mean

- Probability of \( \mu \) in interval \([c_1, c_2]\)
  - \( P(c_1 \leq \mu \leq c_2) = 1-\alpha \)
  - \([c1, c2]\) 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%
  \( \rightarrow \) 90% confidence interval

So, do we have to do \( k \) experiments, each of size \( n \)?!
Confidence Interval Estimate

• 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

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

\[
\bar{X} \pm t \frac{S}{\sqrt{n}}
\]

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

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

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

aka student’s $t$ distribution (“student” was anonymous name used when published by William Gosset)
Computing a Confidence Interval – Example

• Suppose gathered game times in a user study (e.g., for your MQP)
• Can compute sample mean, yes
• But really want to know where population mean is

→ Bound with confidence interval

(Unsorted)

<table>
<thead>
<tr>
<th>Game Time</th>
<th>Game Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.4</td>
<td>3.9</td>
</tr>
<tr>
<td>3.8</td>
<td>3.2</td>
</tr>
<tr>
<td>2.8</td>
<td>4.1</td>
</tr>
<tr>
<td>4.2</td>
<td>3.3</td>
</tr>
<tr>
<td>2.8</td>
<td>2.8</td>
</tr>
<tr>
<td>2.9</td>
<td>4.2</td>
</tr>
<tr>
<td>1.9</td>
<td>3.1</td>
</tr>
<tr>
<td>5.9</td>
<td>4.5</td>
</tr>
<tr>
<td>3.9</td>
<td>4.5</td>
</tr>
<tr>
<td>3.2</td>
<td>4.8</td>
</tr>
<tr>
<td>4.1</td>
<td>4.9</td>
</tr>
<tr>
<td>5.3</td>
<td>5.1</td>
</tr>
<tr>
<td>3.6</td>
<td>3.7</td>
</tr>
<tr>
<td>5.1</td>
<td>3.4</td>
</tr>
<tr>
<td>2.7</td>
<td>5.6</td>
</tr>
<tr>
<td>3.9</td>
<td>3.1</td>
</tr>
</tbody>
</table>
Computing a Confidence Interval – Example

- $\bar{x} = 3.90$, stddev $s=0.95$, $n=32$
- A 90% confidence interval ($\alpha$ is 0.1) for population mean ($\mu$):
  \[
  3.90 \pm \frac{1.696 \times 0.95}{\sqrt{32}}
  \]
  \[
  = [3.62, 4.19]
  \]

- 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 ($\alpha$)

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

<table>
<thead>
<tr>
<th>Experiment/Sample</th>
<th>Includes $\mu$?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>2</td>
<td>yes</td>
</tr>
<tr>
<td>3</td>
<td>no</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>yes ($\alpha = 0.1$)</td>
</tr>
<tr>
<td>Total</td>
<td>yes $\geq 100$ (1-$\alpha$) 90</td>
</tr>
<tr>
<td>Total</td>
<td>no $&lt; 100\alpha$ 10</td>
</tr>
</tbody>
</table>
How does Confidence Interval Size Change?

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

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
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)?

• 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 = \(\)\)**
  – 95% chance population value is between
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? See distribution on the right

http://vassarstats.net/textbook/f1002.gif
Groupwork – Interpreting a Confidence Interval

Using Confidence Interval (1 of 3)

- For charts, depict with error bars.
- CI different than standard deviation.
  - Standard deviation show spread.
  - CI bounds population parameter (decreases 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)

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
But if compute difference, and then confidence interval does not cross 0! (Caused by error propagation)
How *Not* to Use Confidence Intervals (1 of 2)

“The confidence intervals of the two groups **overlap**, hence *the difference is not statistically significant*” — A lot of People

• Overlap – careful not to say no statistically significant difference (see previous slide)
How Not to Use Confidence Intervals (2 of 2)

“The 95% confidence interval goes from C1 to C2, so 95% of all observations are between C1 and C2. — A lot of People

• Do not quantify variability (e.g., 95% of values in interval)
Statistical Significance versus Practical Significance

It’s a Honey of an O

- Boxes of Cheerios, Tastee-O’s both target 12 oz.
- Measure weight of 18,000 boxes (large N!)
- 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 Internet from 100 Mb/s to 1000 Mb/s
- Measure ping to LoL server for 20,000 samples (large N!)
- Using statistics
  - Ping times improve 0.8 ms
  - And statistically significant ($\alpha=0.99$)!
- But ... below perception!

Warning: may find statistically significant difference. That doesn’t mean it is important.
Effect Size

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

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

<table>
<thead>
<tr>
<th>Relative size</th>
<th>Effect size</th>
<th>% of control group below the mean of experimental group</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>50%</td>
</tr>
<tr>
<td>Small</td>
<td>0.2</td>
<td>58%</td>
</tr>
<tr>
<td>Medium</td>
<td>0.5</td>
<td>69%</td>
</tr>
<tr>
<td>Large</td>
<td>0.8</td>
<td>79%</td>
</tr>
<tr>
<td></td>
<td>1.4</td>
<td>92%</td>
</tr>
</tbody>
</table>

Similar to Z-score

\[
z = \frac{X - \bar{X}}{s}
\]

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

(Cohen’s d)
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 99.9% 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)?

• 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 tickets ($9 mil to win $5 mil)!
  – So, most people happy with 0.0001% confidence
Outline

• Overview (done)
• Foundation (done)
• Inferring Population Parameters (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 → Conversant Chapters 8 & 9 in book have more
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 is 4 hours
- **Alternative Hypothesis** – hypothesis contrary to null hypothesis
  - e.g., population mean time for Riot to bring up NA servers is *not* 4 hours
- **Care about Alternative, but test Null**
  - If data supports, *Alternate* may not be true
  - If data rejects, *Alternate* *may* be true
- **Why Null and Alternative?**
  - Remember, data doesn’t “prove” hypothesis
  - Can only reject it at certain significance (e.g., there is probably a difference)
  - So, reject Null

- **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 confidence 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 ($N$)
3. Collect data (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)

• State hypothesis (H) and null hypothesis (H₀)
  – H: Mario level takes more than 5 minutes to complete
  – H₀: Mario level takes 5 minutes to complete (H₀ always has =)

• Evaluate risks of being wrong (based on loss and gain), choosing significance (α) and sample size (N)
  – Player may get frustrated, quit game, so α = 0.1
  – Without distribution analysis, 30 (Central Limit Theorem)

• Collect data (sample), compute statistics
  – 30 people play level, compute average minutes, compare to 5
  – E.g., mean of 6.1 minutes

• Calculate p value based on test statistic and compare to α
  – P value = 0.02, α = 0.1
  – “How likely is it that the true mean is 5 when measure 6.1?”

• Make inference
  – Here: p value less than α → REJECT H₀, so H may be right
  – Note, would not have rejected H₀ if p value greater than α
Depiction of P Value

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

E.g., Mario mean of 5, so is 6.1 in the “unlikely” region?
1. In Hypothesis testing, the Null Hypothesis
2. Game development team wants new model assessed. Steps?

https://web.cs.wpi.edu/~imgd2905/d23/groupwork/10-hypo-testing/handout.html
1. In Hypothesis testing, the Null Hypothesis (H0) is:
   a. sample mean is within a standard error of the population mean
   b. no significance difference between measured and population
   c. the sample mean equals the population mean
   d. all of the above
   e. none of the above
2. Your game development team wants to see if the new Hero model they created is played more often than the old Hero (10%). They task you with doing this assessment. What steps do you take?

a. Gather data
b. Compute sample mean
c. Set hypotheses
d. Test (compute $p$ value)
e. Analyze results to accept or reject