Inferential Statistics

Chapters 6 & 7

Overview

• Use statistics to infer population parameters

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 (2 of 4)

- Have 1d6, sample twice and sum (i.e., roll 2 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?

Dice Rolling (4 of 4)

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

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

What's happening to the shape?
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?
  - 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?
  - Allows us to predict range to bound population mean

Outline

- Overview (done)
- Foundation (done)
- Confidence Intervals (next)
- Hypothesis Testing
Sampling Error (1 of 2)

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

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 ($\sigma$) 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? → 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 (2 of 2)

- Standard error, 100 samples, $N=3$
- Standard error, 100 samples, $N=20$

Estimate population parameter → confidence interval
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, μ in that interval. Chance of error 10%.
  - But, what does that mean? (See next slide for depiction of meaning)

Confidence Interval Estimate

- Estimate interval from 1 experiment/sample, size n
- Compute sample mean (x̄), 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

Confidence Interval Example

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

- x̄ = 3.90, stddev = 0.95, n=32
- A 90% confidence interval (α is 0.1) for population mean (μ):
  - 3.90 ± \frac{1.645 \times 0.95}{\sqrt{32}}
  - \approx [3.62, 4.19]
- With 90% confidence, μ in that interval. Chance of error 10%.
- But, what does that mean? (See next slide for depiction of meaning)

Confidence Interval for Mean

- Probability of μ in interval [c₁, c₂]
  - \Pr(c₁ ≤ μ ≤ c₂) = 1-α
  - [c₁, c₂] is confidence interval
  - α is significance level
  - 100(1-α) is confidence level
  - Typically want α small so confidence level 90%, 95% or 99% (more on effect later)

- Say, α = 0.1. Could do k experiments (size n), find sample means, sort
- Interval from distribution:
  - Lower bound: 5%
  - Upper bound: 95% → 90% confidence interval

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

Meaning of Confidence Interval (α)

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

 Cumulative distribution

- Interval from distribution:
  - Lower bound: 5%
  - Upper bound: 95% → 90% confidence interval

- α was anonymous name used when published by William Gosset

- Note, can use standard normal (z distribution) when large enough sample size (N = 30+)

- Looks like standard normal, but bit “squashed”
- Gets more squashed as n gets smaller

- Function, parameterized by α and n

- t distribution
How does Confidence Interval Size Change?

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

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 (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?

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)**

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

[Graph showing no overlap, large overlap, some overlap]

**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 (\( \alpha = 0.99 \))
- But ... 0.0002 is only 2-3 O’s. Customer doesn’t care!

Latency can Kill?
- Lag in League of Legends
- Pay $5 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 ... humans cannot notice 1 ms difference!

**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 (\( \alpha = 0.95 \))
- But ... 0.0002 is only 2-3 O’s. Customer doesn’t care!

Latency can Kill?
- Lag in League of Legends
- Pay $5 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 ... 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 99.9% 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
  - So, most people happy with 0.01% confidence
Outlines

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

Hypothesis Testing Terminology

- **Null Hypothesis (H₀)** – 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 H₀
  - “If p-value is low, then H₀ must go”
  - “How low”, consider s”risk” of being wrong

Hypothesis Testing Steps

1. State hypothesis (H) and null hypothesis (H₀)
   - H: Mario level takes less than 5 minutes to complete
   - H₀: Mario level takes 5 minutes to complete (H₀ always has =)
2. Evaluate risks of being wrong (based on loss and gain), choosing significance (α) and sample size
   - Player may get frustrated, quit game, so α = 0.01
   - Note sure of normally distributed, so 30 (Central Limit Theorem)
3. Collect data (sample), compute statistics
   - 30 people play level, compute average time, compare to 5
4. Calculate p-value based on test statistic and compare to α
   - p-value = 0.002, α = 0.01
5. Make inference
   - Reject H₀ if p-value less than α
   - Do not reject H₀ if p-value greater than α

Hypothesis Testing Steps (Example)

- State hypothesis (H) and null hypothesis (H₀)
  - H: Mario level takes less 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
  - Player may get frustrated, quit game, so α = 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, α = 0.01
- Make inference
  - Reject H₀ if p-value less than α (REJECT H₀), so H may be right
  - Do not reject H₀ if p-value greater than α