IMGD 2905

Inferential Statistics

Chapters 6 & 7

Overview

• Use statistics to infer population parameters

We want to know about these

\[ \mu \] (Population mean)

We have these to work with

\[ \bar{x} \] (Sample mean)

Random selection

Inference

Population

Sample

Statistic

http://3.bp.blogspot.com/_94E2PdKwaXE/S-xQRuoiKAI/AAAAAAAAABY/xvDRcG_Mcj0/s1600/120909_0159_1.png
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?

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?

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

"Triangle" distribution

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

![Uniform Distribution](image1)
![Uniform Sum Distribution 2 Dice](image2)
![Uniform Sum Distribution 3 Dice](image3)

https://academo.org/demos/dice-roll-statistics/
Try rolling dice yourself!

11

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](image4)

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

Ok, neat – for “square” distributions.
But what about experiments with other distributions?
Sampling Distributions

• With “enough” samples, looks “bell-shaped” \( \rightarrow \) 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
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?

  ![Diagram](http://www.six-sigma-material.com/images/PopSamples.GIF)

  With mean and standard deviation → Allows us to predict range to bound population mean
Why do we care about sample means following Normal distribution?

Sample mean

Probable range of population mean

Note, actual population mean (probably) in this range!

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 $N=20$ 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 ($\mu$, $\sigma$)
  - Sample range varies a lot more than sample standard deviation
  - Population mean ($\mu$) always within sample range

This variation $\rightarrow$ 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 ($\sigma$) 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 (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 ($s$).

\[
\text{SE} = \frac{\sigma}{\sqrt{n}}
\]

Example:

\[
\begin{align*}
  n &= 5 \\
  \sigma &= 17
\end{align*}
\]

\[
\text{SE} = \frac{17}{\sqrt{5}} = 7.6
\]
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?

Estimate population parameter $\rightarrow$ 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]

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
  – Cumulative distribution

• Interval from distribution:
  – Lower bound: 5%
  – Upper bound: 95%
  \( \rightarrow \) 90% confidence interval

We have to do \( k \) experiments, each of size \( n \)?
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
  \( \bar{x} \pm t \cdot \frac{s}{\sqrt{n}} \)
  \( \left( \bar{x} - t \cdot \frac{s}{\sqrt{n}}, \bar{x} + t \cdot \frac{s}{\sqrt{n}} \right) \)
  
  e.g., mean 30.5
  \( t \times SE = 2 \)
  \( 30.5 - 2 = 28.5 \)
  \( 30.5 + 2 = 32.5 \)
  \( [28.5, 32.5] \)

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

\( t \) 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 (\( N = 30+ \))

aka student’s \( t \) distribution (“student” was anonymous name used when published by William Gosset)
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?

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 $\geq 100 (1-\alpha)$</td>
</tr>
<tr>
<td></td>
<td>no $&lt; 100 \alpha$</td>
</tr>
</tbody>
</table>
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

\[
SE_s = \frac{s}{\sqrt{n}}
\]
How does Confidence Interval Change (1 of 2)?

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

\[
SE = \frac{s}{\sqrt{n}} \\
\bar{X} \pm t \frac{s}{\sqrt{n}}
\]

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)

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

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

Latency can Kill?

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

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 $$ 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 99.99% 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.
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

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 (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 $\alpha$
   - Do not reject $H_0$ if p-value greater than $\alpha
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