Lecture 11:

A/B Testing
Today

- Updates
- General presentation advice
- A/B Testing
Friends Release Tuesday 11/3

- Fixes discussed during team meetings
- Fifteen levels
- Tutorial messages
- Music
- Sound effects
- Logging
Not needed for Friends Release

- A/B test
Friends Postmortems Tuesday 11/10
Engage the audience

• Use sufficient volume
Use sufficient volume
Engage the audience

- Use sufficient volume
- Organize and motivate
Motivate
Engage the audience

- Use sufficient volume
- Organize and motivate
- Minimize cognitive load
Avoid Death by Powerpoint

• People often put every word they are going to say on their slides
• It’s terrible
• Why?
  • No one can read this fast
  • It’s stressful to even have to look at this much text
  • No one can read and listen at the same time
    • If your audience is reading, *they aren’t listening to you*
Remove unnecessary information
Include necessary information
Use pictures
Fixing the Spikes

- Our data shows that spikes are bad.
- We will remove some of them.
Plan: Reduce Spikes
For bulleted lists

- Simplify each point
- Bring in one-by-one
- This minimizes cognitive load
Logging Goals & Summary

• We wanted to answer these questions:
  • How far did players get before giving up, and was this affected by the difficulty of specific levels?
  • In what ways did players circumvent the intended solution?

• Entries were logged for the following events:
  • Entering water
  • Gaining power from a gate
  • Getting the key
  • Finishing the level

• Data examined:
  • How often people played each level
  • How often people beat each level
  • Events on a per-level basis
Logging Goals & Summary

- We wanted to answer these questions:
  - How far did players get before giving up, and was this affected by the difficulty of specific levels?
  - In what ways did players circumvent the intended solution?

- Entries were logged for the following events:
  - Entering water
  - Gaining power from a gate
  - Getting the key
  - Finishing the level

- Data examined:
  - How often people played each level
  - How often people beat each level
  - Events on a per-level basis
Design Questions

- How far did players get?
- Did they circumvent the intended solution?
Logging Goals & Summary

• We wanted to answer these questions:
  • How far did players get before giving up, and was this affected by the difficulty of specific levels?
  • In what ways did players circumvent the intended solution?

• Entries were logged for the following events:
  • Entering water
  • Gaining power from a gate
  • Getting the key
  • Finishing the level

• Data examined:
  • How often people played each level
  • How often people beat each level
  • Events on a per-level basis
Events

1. Grab Key
2. Enter Gate
3. Enter Water
Don’t forget to summarize

- Use sufficient volume
- Organize and motivate
- Minimize cognitive load
Design decisions are a pain
A/B Testing

version A

version B

the world

data
GSN Games A/B Testing

Revenue +12.3%
GSN Games A/B Testing

Revenue +11\%
Battleship Numberline

Level 1 of 2

AVG. Accuracy 95.8%

Ship spotted at: 4/5

MENU
Optimal game type?
Optimal ship size?
Optimal time limit?
How much challenge?
Measures of engagement

• Engagement
  • Time played
  • Log(time * attempts)

• Challenge
  • Probability of success?
BrainPop
Experiment: 70,000 people

- Ship types: submarine and battleship
- Ship sizes: 4, 6, 8, 10, 16, 20, 24, 30, 40%
- Time limits: 2, 3, 4, 5, 8, 10, 15, 30 seconds
Results

- Clicking on target = more time played
- Bigger target = more time played
- Longer time limit = more time played
A/B testing logistics

- What conditions?
- How to track players?
- Players per condition?
A/B testing in this class

1. Wait for logging update from Kelvin
2. Call reportPageLoad first
3. Then call recordABTestingValue
4. **Use the return value** of recordABTestingValue to set the condition
recordABTestingValue

- **Goals:**
  - New condition for new players
  - Previous condition for returning players
- **Takes a single parameter:** proposed integer condition

```plaintext
var proposedCondition:int = Math.Floor(Math.Random() * 2) + 1;
var actualCondition:int = recordABTestingValue(proposedCondition);
if (actualCondition == 1) { ... }
else if (actualCondition == 2) { ... }
```
Assignment of players to condition

- 50% of players to each condition
Null hypothesis testing
Typical problem

- Version A
  - 100 players
  - Average time played: 120 seconds
  - Standard deviation: 5 points

- Version B
  - 100 players
  - Average time played: 105 seconds
  - Standard deviation: 5 points

- What is the probability of obtaining a result at least this extreme?
p-value

probability the null hypothesis is true
Common fallacies

- lower p value means stronger effect
- $p \geq 0.05$ means no effect
p-values (source: XKCD)
Fishing (source: XKCD)

| WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE ($P > 0.05$). |
| WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE ($P > 0.05$). |
| WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE ($P > 0.05$). |
| WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE ($P > 0.05$). |
| WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE ($P > 0.05$). |
| WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE ($P > 0.05$). |
| WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE ($P > 0.05$). |
| WE FOUND NO LINK BETWEEN GREEN JELLY BEANS AND ACNE ($P > 0.05$). |
| WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE ($P > 0.05$). |
| WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE ($P < 0.05$). |
| WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE ($P > 0.05$). |

---

News

GREY JELLY BEANS LINKED TO ACNE!

95% CONFIDENCE

95% CHANCE OF COINCIDENCE!
Student’s t-test

- Compares *means*
- Assumes normal distribution
Student’s t-test

![Bar chart showing comparison between Treatment and Control groups.](chart.png)
F-test

- Compares *variance*
- Assumes normal distribution
F-test

<table>
<thead>
<tr>
<th></th>
<th>Treatment</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
U-test

- Compares *ranks*
- No assumption of normality
Mann-Whitney U-test
(aka Wilcoxon-Kruskal-Wallis two-tailed test)
How to run U-test

- Download 30-day free trial of JMP
  - Can also pay $15 for one year Cornell student license
- Can also use Excel or SPSS or R or …
Reporting U-test results

- Medians
- Z-statistic
- p-value
Group activity

- Brainstorm A/B tests!
  - What is your hypothesis?
  - What do you need to measure?
  - What do you need to log?