A/B Testing & Bandit Based Solutions

Nick Sumner
wsumner@sfu.ca
How do you know that a change adds value?

● The scenario
  – You maintain a web site and are considering a change
  – You hypothesize that the change improves outcomes in some way
How do you know that a change adds value?

- The scenario
  - You maintain a web site and are considering a change
  - You hypothesize that the change improves outcomes in some way

- The problem
  - How can you find out whether one change (or many!) improves results?
How do you know that a change adds value?

- **The scenario**
  - You maintain a web site and are considering a change
  - You hypothesize that the change improves outcomes in some way

- **The problem**
  - How can you find out whether one change (or many!) improves results?
  - How can you do this without costing your company money?
How do you know that a change adds value?

- **The scenario**
  - You maintain a web site and are considering a change
  - You hypothesize that the change improves outcomes in some way

- **The problem**
  - How can you find out whether one change (or many!) improves results?
  - How can you do this without costing your company money?
How do you know that a change adds value?

● The scenario
   – You maintain a web site and are considering a change
   – You hypothesize that the change improves outcomes in some way

● The problem
   – How can you find out whether one change (or many!) improves results?
   – How can you do this without costing your company money?

You should already have an intuition for attacking this. What should you do?
How do you know that a change adds value?

• The scenario
  – You maintain a web site and are considering a change
  – You hypothesize that the change improves outcomes in some way

• The problem
  – How can you find out whether one change (or many!) improves results?
  – How can you do this without costing your company money?

• Solutions
  – *A/B Testing* uses different forms of hypothesis testing
  – Alternatively, you can use *multi-armed bandits* to attack the problem
How do you know that a change adds value?

● The scenario
  – You maintain a web site and are considering a change
  – You hypothesize that the change improves outcomes in some way

● The problem
  – How can you find out whether one change (or many!) improves results?
  – How can you do this without costing your company money?

● Solutions
  – \textit{A/B Testing} uses different forms of hypothesis testing
  – Alternatively, you can use \textit{multi-armed bandits} to attack the problem
  – Key idea: run controlled experiments live on the deployed software
How do you know that a change adds value?

● The scenario
  – You maintain a web site and are considering a change
  – You hypothesize that the change improves outcomes in some way

● The problem
  – How can you find out whether one change (or many!) improves results?
  – How can you do this without costing your company money?

● Solutions
  – A/B Testing uses different forms of hypothesis testing
  – Alternatively, you can use multi-armed bandits to attack the problem
  – Key idea: run controlled experiments live on the deployed software

● Caveat: We will not dive into a full stats background for these
  – We will discuss some common pitfalls that arise from misunderstandings
When might you want to know?

- Exploring ideas to improve usability
When might you want to know?

- Exploring ideas to improve usability
  - Or performance (throughput, latency, ...)

When might you want to know?

- Exploring ideas to improve usability
  - Or performance (throughput, latency, ...)
- Establishing the effectiveness of promotion before campaigns
When might you want to know?

- Exploring ideas to improve usability
  - Or performance (throughput, latency, ...)
- Establishing the effectiveness of promotion before campaigns
- Staged rollouts of major changes
When might you want to know?

- Exploring ideas to improve usability
  - Or performance (throughput, latency, ...)
- Establishing the effectiveness of promotion before campaigns
- Staged rollouts of major changes
  - Minimizing risk of: CD, fragmented configurations, ...
  - e.g. rolling out apps to the Android store
Simple A/B Testing

- You have:
  - two solutions, A and B (e.g., A is old, B is new)
  - A hypothesis (e.g. A will improve conversion over B by at least 5%)
Simple A/B Testing

- You have:
  - two solutions, A and B (e.g., A is old, B is new)
  - A hypothesis (e.g. A will improve conversion over B by at least 5%)

- Basic solution:
  - **Determine what data to collect** (choose population, metric, & size up front!!!)
Simple A/B Testing

- **You have:**
  - two solutions, A and B (e.g., A is old, B is new)
  - A hypothesis (e.g. A will improve conversion over B by at least 5%)

- **Basic solution:**
  - **Determine what data to collect** (choose population, metric, & size up front!!!)
  - Randomly provide(/serve) A to one population and B to another to collect predetermined stats
Simple A/B Testing

● You have:
  – two solutions, A and B (e.g., A is old, B is new)
  – A hypothesis (e.g. A will improve conversion over B by at least 5%)

● Basic solution:
  – **Determine what data to collect** (choose population, metric, & size up front!!!)
  – Randomly provide(/serve) A to one population and B to another to collect predetermined stats
  – Use a basic t-test to measure differences in the populations
Simple A/B Testing

• You have:
  – two solutions, A and B (e.g., A is old, B is new)
  – A hypothesis (e.g. A will improve conversion over B by at least 5%)

• Basic solution:
  – Determine what data to collect (choose population, metric, & size up front!!!)
  – Randomly provide(/serve) A to one population and B to another to collect predetermined stats
  – Use a basic t-test to measure differences in the populations

\[ \mu_1 < \mu_2 \]
Recalling T-tests

- Can be one-sided (tailed) or two sided (tailed)
  - distinguishing directed and undirected differences
Recalling T-tests

- Can be one-sided (tailed) or two sided (tailed)
  - distinguishing directed and undirected differences
- Assume (1) observation independence and (2) normal distribution
Recalling T-tests

- Can be one-sided (tailed) or two sided (tailed)
  - distinguishing directed and undirected differences
- Assume (1) observation independence and (2) normal distribution
- Distinguish 2 hypotheses (e.g.):
  - $H_0$: $\mu_1 - \mu_2 = 0$ (the null hypothesis – assumed true until disproven)
  - $H_1$: $\mu_1 < \mu_2$ (the alternative)
Recalling T-tests

- Can be one-sided (tailed) or two sided (tailed)
  - distinguishing directed and undirected differences
- Assume (1) observation independence and (2) normal distribution
- Distinguish 2 hypotheses (e.g.):
  - $H_0$: $\mu_1 - \mu_2 = 0$ (the null hypothesis – assumed true until disproven)
  - $H_1$: $\mu_1 < \mu_2$ (the alternative)
Recalling T-tests

- Can be one-sided (tailed) or two sided (tailed)
  - distinguishing directed and undirected differences
- Assume (1) observation independence and (2) normal distribution
- Distinguish 2 hypotheses (e.g.):
  - $H_0: \mu_1 - \mu_2 = 0$ (the null hypothesis – assumed true until disproven)
  - $H_1: \mu_1 < \mu_2$ (the alternative)
- **RECALL:**
  - We never prove a hypothesis!
  - We gather sufficient evidence to reject the null hypothesis and thus accept the alternative
Recalling T-tests

\[ t = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta}{\sqrt{\frac{S_1^2}{m}} + \sqrt{\frac{S_2^2}{n}}} \]

Where \( H_0: \mu_1 - \mu_2 = \Delta \)
Recalling T-tests

\[ t = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}} \quad \text{Where } H_0: \mu_1 - \mu_2 = \Delta \]

\[ H_a: \mu_1 - \mu_2 > \Delta \]

\[ H_a: \mu_1 - \mu_2 < \Delta \]

\[ H_a: \mu_1 - \mu_2 \neq \Delta \]
Recalling T-tests

\[ t = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}} \]

Where \( H_0: \mu_1 - \mu_2 = \Delta \)

- \( H_a: \mu_1 - \mu_2 > \Delta \) \( t > t_{\alpha, v} \)
- \( H_a: \mu_1 - \mu_2 < \Delta \) \( t < -t_{\alpha, v} \)
- \( H_a: \mu_1 - \mu_2 \neq \Delta \) \( |t| > t_{\alpha/2, v} \)
Recalling T-tests

\[ t = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}} \]

Where \( H_0: \mu_1 - \mu_2 = \Delta \)

\[ t > t_{\alpha, v} \quad p = P[T \geq t|H_0] \]

\[ t < -t_{\alpha, v} \quad p = P[T \leq t|H_0] \]

\[ |t| > t_{\alpha/2, v} \quad p = P[|T| \geq t|H_0] \]
Recalling T-tests

\[ t = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}} \]

Where \( H_0 : \mu_1 - \mu_2 = \Delta \)

\( H_a : \mu_1 - \mu_2 > \Delta \) \quad \Rightarrow \quad t > t_{\alpha, v} \quad p = P[T \geq t|H_0]

\( H_a : \mu_1 - \mu_2 < \Delta \) \quad \Rightarrow \quad t < -t_{\alpha, v} \quad p = P[T \leq t|H_0]

\( H_a : \mu_1 - \mu_2 \neq \Delta \) \quad \Rightarrow \quad |t| > t_{\alpha/2, v} \quad p = P[|T| \geq t|H_0]

\[ v = \frac{\left( \frac{S_1^2}{m} + \frac{S_2^2}{n} \right)}{\left( \frac{S_1^2}{m} \right)^2 \left( \frac{S_2^2}{n} \right)^2} \]

\[ v = \frac{m-1}{m-1} + \frac{n-1}{n-1} \]
Recalling T-tests

\[ t = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta}{\sqrt{\frac{S^2_1}{m} + \frac{S^2_2}{n}}} \]

Where \( H_0: \mu_1 - \mu_2 = \Delta \)

\[ H_a: \mu_1 - \mu_2 > \Delta \]
\[ t > t_{\alpha, v} \]
\[ p = P[T \geq t | H_0] \]

\[ H_a: \mu_1 - \mu_2 < \Delta \]
\[ t < -t_{\alpha, v} \]
\[ p = P[T \leq t | H_0] \]

\[ H_a: \mu_1 - \mu_2 \neq \Delta \]
\[ |t| > t_{\alpha/2, v} \]
\[ p = P[|T| \geq t | H_0] \]

\[ v = \left( \frac{S^2_1}{m} + \frac{S^2_2}{n} \right) \left( \frac{(S^2_1/m)^2}{m - 1} + \frac{(S^2_2/n)^2}{n - 1} \right)^{-1} \]

Where \( \alpha \) captures the level of confidence for a p-value
Recalling T-tests

\[ t = \frac{\left( \bar{x}_1 - \bar{x}_2 \right) - \Delta}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}} \]

Where \( H_0: \mu_1 - \mu_2 = \Delta \)

\[ v = \frac{\left( S_1^2 + S_2^2 \right)}{\left( \frac{S_1^2}{m} + \frac{S_2^2}{n} \right)} \]

\( m - 1 + \frac{\left( S_2^2 / n \right)^2}{n - 1} \)

But subtle challenges arise in practice!
Problem: Choosing and tagging populations

- The hypothesis in question may not apply to everyone
Problem: Choosing and tagging populations

- The hypothesis in question may not apply to everyone
  - Is there a specific user segment that it should apply to?
    (Users of features X,Y,Z? Users in a specific country? Early adopters?)
Problem: Choosing and tagging populations

- The hypothesis in question may not apply to everyone
  - Is there a specific user segment that it should apply to?
    (Users of features X,Y,Z? Users in a specific country? Early adopters?)

- The hypothesis might affect different subpopulations differently
Problem: Choosing and tagging populations

- The hypothesis in question may not apply to everyone
  - Is there a specific user segment that it should apply to? (Users of features X,Y,Z? Users in a specific country? Early adopters?)

- The hypothesis might affect different subpopulations differently
  - People familiar with workflow X
  - Different age groups
  - People speaking different languages
  - People using the software on different workdays
Problem: Choosing and tagging populations

- The hypothesis in question may not apply to everyone
  - Is there a specific user segment that it should apply to? (Users of features X,Y,Z? Users in a specific country? Early adopters?)

- The hypothesis might affect different subpopulations differently
  - People familiar with workflow X
  - Different age groups
  - People speaking different languages
  - People using the software on different workdays

- Possible factors in the results ought to be identified up front. Collecting them after the fact requires rerunning an experiment.
Problem: Choosing and tagging populations

- The hypothesis in question may not apply to everyone
  - Is there a specific user segment that it should apply to?
    (Users of features X,Y,Z? Users in a specific country? Early adopters?)

- The hypothesis might affect different subpopulations differently
  - People familiar with workflow X
  - Different age groups
  - People speaking different languages
  - People using the software on different workdays

- Possible factors in the results ought to be identified up front. Collecting them after the fact requires rerunning an experiment.

- Your sample ought to be representative.
Problem: False positives and negatives

There is always a risk of error

<table>
<thead>
<tr>
<th>Event</th>
<th>Formula</th>
<th>Type I error</th>
<th>Type II error</th>
</tr>
</thead>
<tbody>
<tr>
<td>P[fail to reject $H_0$</td>
<td>$H_0$]</td>
<td></td>
<td>$\beta$</td>
</tr>
<tr>
<td>P[fail to reject $H_0$</td>
<td>$\neg H_0$]</td>
<td>$\beta$</td>
<td></td>
</tr>
<tr>
<td>P[reject $H_0$</td>
<td>$H_0$]</td>
<td>$\alpha$</td>
<td></td>
</tr>
<tr>
<td>P[reject $H_0$</td>
<td>$\neg H_0$]</td>
<td></td>
<td>$\beta$</td>
</tr>
</tbody>
</table>
Problem: Choosing hypotheses

- Can you simply test any and all hypotheses?
  Can you run your tests and try many hypotheses later?
Problem: Choosing hypotheses

- Can you simply test any and all hypotheses? Can you run your tests and try many hypotheses later?
  - Define clear goals. Hypotheses not targeting goals are useless.
  - Testing many things increases the likelihood of *false positives* \( P(\text{reject } H_0 \mid H_0) \)
Problem: Choosing hypotheses

- Can you simply test any and all hypotheses? Can you run your tests and try many hypotheses later?
  - Define clear goals. Hypotheses not targeting goals are useless.
  - Testing many things increases the likelihood of false positives \( P[\text{reject } H_0 \mid H_0] \)

\[
p = P[\text{A sample is at least as extreme as observed } \mid H_0]
\]
Problem: Choosing hypotheses

- Can you simply test any and all hypotheses? Can you run your tests and try many hypotheses later?
  - Define clear goals. Hypotheses not targeting goals are useless.
  - Testing many things increases the likelihood of *false positive* $P(\text{reject } H_0 | H_0)$
  - The temptation (and management pressure) favors *p-hacking*
    \[ p = P(\text{A sample is at least as extreme as observed } | H_0) \]
Problem: Choosing hypotheses

- Can you simply test any and all hypotheses? Can you run your tests and try many hypotheses later?
  - Define clear goals. Hypotheses not targeting goals are useless.
  - Testing many things increases the likelihood of **false positives**
  - The temptation (and management pressure) favors **p-hacking**
    
    \[ p = P[\text{A sample is at least as extreme as observed} \mid H_0] \]

Suppose you run 5 tests with \( p = 0.1 \), What is the likelihood of a false positive?
Problem: Choosing hypotheses

- Can you simply test any and all hypotheses? Can you run your tests and try many hypotheses later?
  - Define clear goals. Hypotheses not targeting goals are useless.
  - Testing many things increases the likelihood of false positives
  - The temptation (and management pressure) favors p-hacking

- The more hypotheses you test, the greater your risk of false positives
  - This can be mitigated, but you should choose hypotheses well up front
Problem: Stopping criteria & confidence

- In order to test with a certain significance (e.g. \( \alpha = 0.05 \)), the size of a test campaign with T-tests must be set up front.
Problem: Stopping criteria & confidence

- In order to test with a certain significance (e.g. $\alpha=0.05$), the size of a test campaign with T-tests must be set up front.
  - Calculate the number of samples required first, then run the test.
In order to test with a certain significance (e.g. $\alpha=0.05$), the size of a test campaign with T-tests must be set up front.

- Calculate the number of samples required first, then run the test.
- Do not just observe the process and stop it “after significance reached”
Problem: Stopping criteria & confidence

• In order to test with a certain significance (e.g. \( \alpha=0.05 \)), the size of a test campaign with T-tests must be set up front.
  – Calculate the number of samples required first, then run the test.
  – **Do not** just observe the process and stop it “after significance reached”

• But then how many samples are required?
Problem: Stopping criteria & confidence

- In order to test with a certain significance (e.g. $\alpha=0.05$), the size of a test campaign with T-tests must be set up front.
  - Calculate the number of samples required first, then run the test.
  - **Do not** just observe the process and stop it “after significance reached”

- But then how many samples are required?
  - First determine the acceptable error probabilities, $\alpha$ and $\beta$ (often 5% & 20%)
Problem: Stopping criteria & confidence

- In order to test with a certain significance (e.g. $\alpha=0.05$), the size of a test campaign with T-tests must be set up front.
  - Calculate the number of samples required first, then run the test.
  - *Do not* just observe the process and stop it “after significance reached”

- But then how many samples are required?
  - First determine the acceptable error probabilities, $\alpha$ and $\beta$ (often 5% & 20%)
  - The *power* of a test is $(1-\beta)$. $P[\text{reject } H_0 \mid \neg H_0]$  
  - This can also be expressed as “minimum detectable effect size”
Problem: Stopping criteria & confidence

- In order to test with a certain significance (e.g. $\alpha = 0.05$), the size of a test campaign with T-tests must be set up front.
  - Calculate the number of samples required first, then run the test.
  - *Do not* just observe the process and stop it “after significance reached”

- But then how many samples are required?
  - First determine the acceptable error probabilities, $\alpha$ and $\beta$ (often 5% & 20%)
  - The *power* of a test is $(1-\beta)$. $P[\text{reject } H_0 \mid \neg H_0]$ 
  - This can also be expressed as “minimum detectable effect size”
  - If variance and sample sizes can differ, this is challenging, so most just use available sample size calculators based on $\alpha$ and $\beta$. 
Problem: Regression to the mean

- Following an extreme event, the next event is likely less extreme.
Problem: Regression to the mean

- Following an extreme event, the next event is likely less extreme.
- Suppose poorly performing students are put in a special program.
Problem: Regression to the mean

• Following an extreme event, the next event is likely less extreme.
• Suppose poorly performing students are put in a special program.
  – After completion of the program, they perform better.
  – *Is the program effective?*
Problem: Regression to the mean

- Following an extreme event, the next event is likely less extreme.
- Suppose poorly performing students are put in a special program.
  - After completion of the program, they perform better.
  - Is the program effective?
  - If they were already poor performers, improving was more likely anyway!
Problem: Regression to the mean

- Following an extreme event, the next event is likely less extreme.
- Suppose poorly performing students are put in a special program.
  - After completion of the program, they perform better.
  - Is the program effective?
  - If they were already poor performers, improving was more likely anyway!
  - This can be used to falsely justify punishment & rewards
**Problem: Regression to the mean**

- Following an extreme event, the next event is likely less extreme.
- Suppose poorly performing students are put in a special program.
  - After completion of the program, they perform better.
  - Is the program effective?
  - If they were already poor performers, improving was more likely anyway!
  - This can be used to falsely justify punishment & rewards
- **The illusion of significance**
Problem: Novelty effects

- Users are used to seeing a blue “buy” button and ignore it, so you change it to red.
Problem: Novelty effects

- Users are used to seeing a blue “buy” button and ignore it, so you change it to red.
  - Sales skyrocket. **Red is clearly better!**
  - Until a week later when sales return to normal...
Problem: Novelty effects

- Users are used to seeing a blue “buy” button and ignore it, so you change it to red.
  - Sales skyrocket. Red is clearly better!
  - Until a week later when sales return to normal...

- The novelty of the change for the sample may bias the underlying results of the study
Other forms of hypothesis testing

- T-tests are not the only approach and do not always apply
Other forms of hypothesis testing

- T-tests are not the only approach and do not always apply
  - Known variance?
  - Independence?
  - Normality?
  - Qualitative vs Quantitative measures? (does a relationship exist at all?)
  - Small sample sizes expected?
  - ...
Other forms of hypothesis testing

- T-tests are not the only approach and do not always apply
  - Known variance?
  - Independence?
  - Normality?
  - Qualitative vs Quantitative measures? (does a relationship exist at all?)
  - Small sample sizes expected?
  - ...

If the testing is important, you should be doing something obvious or consulting a statistician.
Other forms of hypothesis testing

- T-tests are not the only approach and do not always apply
  - Known variance?
  - Independence?
  - Normality?
  - Qualitative vs Quantitative measures? (does a relationship exist at all?)
  - Small sample sizes expected?
  - ...

- But what if even the notion of a predetermined campaign does not fit?
Other forms of hypothesis testing

- T-tests are not the only approach and do not always apply
  - Known variance?
  - Independence?
  - Normality?
  - Qualitative vs Quantitative measures? (does a relationship exist at all?)
  - Small sample sizes expected?
  - ... 

- But what if even the notion of a predetermined campaign does not fit?
  - Sequential hypothesis testing / Sequential analysis
  - Bandits