An introduction to A/B testing using a Google Optimize example

Juan M. Fonseca-Solís

https://juanfonsecasolis.github.com

October 14, 2019

Introduction

A/B testing is used for:

- Comparing *statistically* two or more variations and determine which one is better
- Measuring success in terms of key performance indicators (KPI)
- Offering periodical little increments to clients and obtain fast feedback

Anecdote: it is also a form of torture for developers by spending their time in functionalities that won't roll out.

Introduction

A/B testing is used for:

- Comparing *statistically* two or more variations and determine which one is better
- Measuring success in terms of key performance indicators (KPI)
- Offering periodical little increments to clients and obtain fast feedback

Anecdote: it is also a form of torture for developers by spending their time in functionalities that won't roll out.

Introduction

A/B testing is used for:

- Comparing *statistically* two or more variations and determine which one is better
- Measuring success in terms of key performance indicators (KPI)
- Offering periodical little increments to clients and obtain fast feedback

Anecdote: it is also a form of torture for developers by spending their time in functionalities that won't roll out.

Introduction

A/B testing is used for:

- Comparing *statistically* two or more variations and determine which one is better
- Measuring success in terms of key performance indicators (KPI)
- Offering periodical little increments to clients and obtain fast feedback

Anecdote: it is also a form of torture for developers by spending their time in functionalities that won't roll out.

Introduction (cont.)

- Some goals in A/B testing are [4]:
  - Increase the conversion rate
  - Increase the throughput
  - Increase the session time
  - Decrease the bounce rate
- In few slides we are going to present an example of an A/B test using Google Optimize

Introduction (cont.)

- Some goals in A/B testing are [4]:
  - Increase the conversion rate
  - Increase the throughput
  - Increase the session time
  - Decrease the bounce rate
- In few slides we are going to present an example of an A/B test using Google Optimize

Introduction (cont.)

- Some goals in A/B testing are [4]:
  - Increase the conversion rate
  - Increase the throughput
  - Increase the session time
  - Decrease the bounce rate
- In few slides we are going to present an example of an A/B test using Google Optimize

Introduction (cont.)

- Some goals in A/B testing are [4]:
  - Increase the conversion rate
  - Increase the throughput
  - Increase the session time
  - Decrease the bounce rate
- In few slides we are going to present an example of an A/B test using Google Optimize

Introduction (cont.)

- Some goals in A/B testing are [4]:
  - Increase the conversion rate
  - Increase the throughput
  - Increase the session time
  - Decrease the bounce rate
- In few slides we are going to present an example of an A/B test using Google Optimize

Introduction (cont.)

- Some goals in A/B testing are [4]:
  - Increase the conversion rate
  - Increase the throughput
  - Increase the session time
  - Decrease the bounce rate
- In few slides we are going to present an example of an A/B test using Google Optimize

Introduction (cont.)

- Some goals in A/B testing are [4]:
  - Increase the conversion rate
  - Increase the throughput
  - Increase the session time
  - Decrease the bounce rate
- In few slides we are going to present an example of an A/B test using Google Optimize

A word of caution!

A/B testing is useful only if you understand the objectives of your organization, so you must be able to answer things like [4]:

- Sales nature
- Target audience
- Revenue per customer

A word of caution!

A/B testing is useful only if you understand the objectives of your organization, so you must be able to answer things like [4]:

- Sales nature
- Target audience
- Revenue per customer

A word of caution!

A/B testing is useful only if you understand the objectives of your organization, so you must be able to answer things like [4]:

- Sales nature
- Target audience
- Revenue per customer

Background

Ok, let's talk about the example.

We want to increase the time that users spend reading an article called *Band limited interpolation for daily reference rates.*[1]

[1] Available at https://juanfonsecasolis.github.io/blog/JFonseca.interpolacion51.html

Background

Ok, let's talk about the example.

We want to increase the time that users spend reading an article called *Band limited interpolation for daily reference rates*.[1]

[1] Available at `https://juanfonsecasolis.github.io/blog/JFonseca.interpolacionBL.html`.

Background

Ok, let's talk about the example.

We want to increase the time that users spend reading an article called *Band limited interpolation for daily reference rates*.[1]

Band limited interpolation for daily reference rates

Juan M. Fonseca-Solís · Mar 2015 · 7 min read ★

Summary

In this *ipython notebook* we'll use data from daily reference rates, such as the London interbank interest rate (LIBOR) or the dollar exchange rate in Costa Rica offered monthly by the Central Bank of Costa Rica (BCCR), to explain linear and band-limited interpolation techniques.

History

In June 2012, when resolving a legal dispute, the Commodity Futures Negotiation Commission of the United States (CFTC) discovered a series of irregularities in the management of the LIBOR by the British multinational bank Barclays. The Financial Times newspaper later confirmed the manipulation of this rate since 1991, which caused an international scandal, since the LIBOR is used as a reference to determine the interest rate of the loans in foreign currency all over the world [4,5].

[1] Available at `https://juanfonsecasolis.github.io/blog/JFonseca.interpolacionBL.html`.

Background (cont.)

- The target audience is composed by data scientists, digital signal processing engineers, and machine learning engineers
- There is a section, approximately at 38%, were mathematical technical explanation makes the text harder to read
- We want to avoid people getting stuck in this section

Ok, being that said, let's begin with the experiment design...

Background (cont.)

- The target audience is composed by data scientists, digital signal processing engineers, and machine learning engineers
- There is a section, approximately at 38%, were mathematical technical explanation makes the text harder to read
- We want to avoid people getting stuck in this section

Ok, being that said, let's begin with the experiment design...

Background (cont.)

- The target audience is composed by data scientists, digital signal processing engineers, and machine learning engineers
- There is a section, approximately at 38%, were mathematical technical explanation makes the text harder to read
- We want to avoid people getting stuck in this section

Ok, being that said, let's begin with the experiment design...

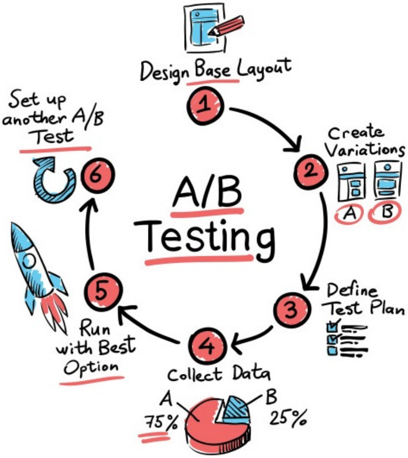
Background (cont.)

- The target audience is composed by data scientists, digital signal processing engineers, and machine learning engineers
- There is a section, approximately at 38%, were mathematical technical explanation makes the text harder to read
- We want to avoid people getting stuck in this section

Ok, being that said, let's begin with the experiment design...

Experiment design

Here are the steps:



https://venturebeat.com/wp-content/uploads/2016/02/ab-testing.jpg?w=930&strip=all.

Experiment design (cont.)

Opportunity: readers might get discouraged to continue at the 38% milestone, were the text becomes harder to digest

Hypothesis: if users had a progress bar, they would be encouraged to reach the 45% milestone —where the text becomes more understandable—

Experiment design (cont.)

Opportunity: readers might get discouraged to continue at the 38% milestone, were the text becomes harder to digest

Hypothesis: if users had a progress bar, they would be encouraged to reach the 45% milestone —where the text becomes more understandable—

Experiment design (cont.)

Before adding a progress bar other ideas were considered:

- Vary text content, size or font
- Change images
- Replace the one column layout by two columns
- Provide a lighter text
- etc...

Experiment design (cont.)

Before adding a progress bar other ideas were considered:

- Vary text content, size or font
- Change images
- Replace the one column layout by two columns
- Provide a lighter text
- etc...

Experiment design (cont.)

Before adding a progress bar other ideas were considered:

- Vary text content, size or font
- Change images
- Replace the one column layout by two columns
- Provide a lighter text
- etc...

Experiment design (cont.)

Before adding a progress bar other ideas were considered:

- Vary text content, size or font
- Change images
- Replace the one column layout by two columns
- Provide a lighter text
- etc...

Experiment design (cont.)

Before adding a progress bar other ideas were considered:

- Vary text content, size or font
- Change images
- Replace the one column layout by two columns
- Provide a lighter text
- etc...

Experiment design (cont.)

Goal: increase the session time to at least 5 min (less would mean that users are not reading)

Successful criteria: 5% conversion rate

Traffic allocation: 50% original and 50% variant

Duration: 2 weeks

Experiment design (cont.)

Goal: increase the session time to at least 5 min (less would mean that users are not reading)

Successful criteria: 5% conversion rate

Traffic allocation: 50% original and 50% variant

Duration: 2 weeks

Experiment design (cont.)

Goal: increase the session time to at least 5 min (less would mean that users are not reading)

Successful criteria: 5% conversion rate

Traffic allocation: 50% original and 50% variant

Duration: 2 weeks

Experiment design (cont.)

Goal: increase the session time to at least 5 min (less would mean that users are not reading)

Successful criteria: 5% conversion rate

Traffic allocation: 50% original and 50% variant

Duration: 2 weeks

Experiment design (cont.)

These are the target groups:

Platform	Name	Contacts
Facebook	ML group	1319
Linkedin	Personal contacts	120
Meetup	Machine Learning CR	1128
	Data Visualization & Analytics Costa Rica	956
	Data Latam	487
	Python CR	824
Skype	Internal company's chat	200
	Total	5034

Experiment implementation

And this is how we implemented the experiment:

- For the progress bar, we added a library called **VerLim.js**

```
<script src="dist/VerLim.min.js"></script>
<link rel="stylesheet"
    href="dist/themeNUIwithCounter.css">
```

Experiment implementation

- We created the **Google optimize** experiment and setup the page's header with the provided script:

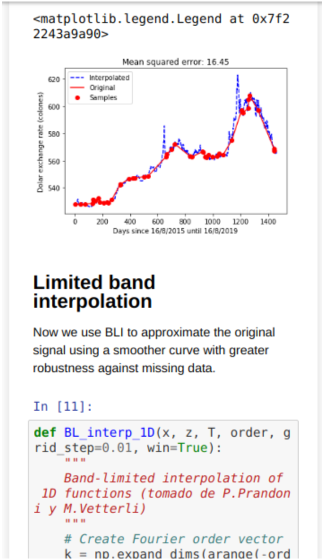
```
<script>
    (function(i,s,o,g,r,a,m)i['GoogleAnalyticsObject']
    =r;i[r]=i[r]||function()
    (i[r].q=i[r].q||[]).push(arguments),
    i[r].l=1*new Date();a=s.createElement(o),
    m=s.getElementsByTagName(o)[0];a.async=1;a.src=g;
    m.parentNode.insertBefore(a,m)
    )(window,document,'script',
    'https://www.google-analytics.com/analytics.js','ga');
    ga('create', '<UA-code-here>', 'auto');
    ga('require', '<GTM-code-here>');
    ga('send', 'pageview');
</script>
```

Experiment implementation (cont.)

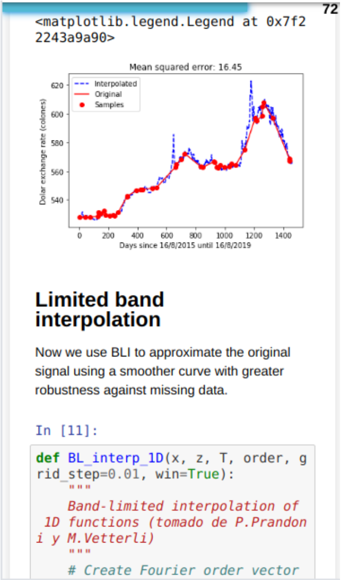
- We made Google Optimize inject the following JS code on variation to display the progress bar 50% of the times:

```
jQuery(document).ready(function ()
$(window).VerLim(
    autoHide: "on",
    autoHideTime: "2",
    theme: "off",
    position: "top",
    thickness: "10px",
    shadow: "on"
);)
```

The result in mobile view:



Original



Variant

We ran the experiment, and after two weeks we got this...

Results threw by Google Optimize

From August 25th - Sept. 7th of 2019:

Number of sessions: **48 (20 original, 28 variant)**

Improvement: **1.178%** on conversions with confidence of 87%

Median session time: **1:24** on original and **3:35** on variant ($\Delta t = 2:11$, max. 9 min)

Results threw by Google Optimize

From August 25th - Sept. 7th of 2019:

Number of sessions: **48 (20 original, 28 variant)**

Improvement: **1.178%** on conversions with confidence of 87%

Median session time: **1:24** on original and **3:35** on variant ($\Delta t = 2:11$, max. 9 min)

Results threw by Google Optimize

From August 25th - Sept. 7th of 2019:

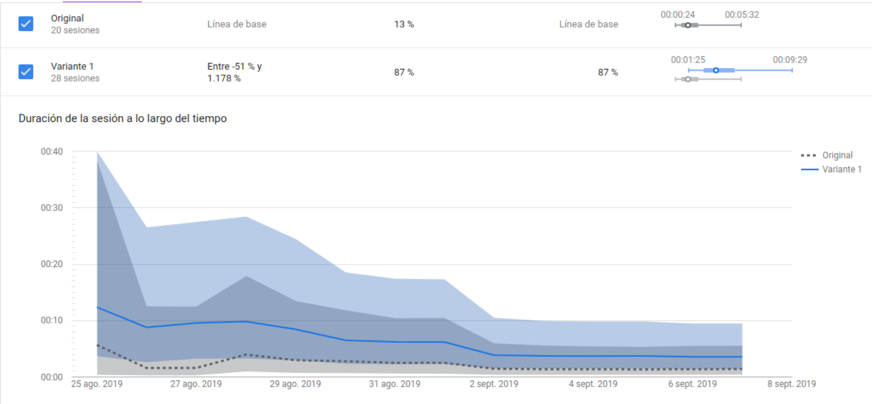
Number of sessions: **48 (20 original, 28 variant)**

Improvement: **1.178%** on conversions with confidence of 87%

Median session time: **1:24** on original and **3:35** on variant ($\Delta t = 2:11$, max. 9 min)

Results threw by Google Optimize (cont.)

This is how it looked in Google Analytics:



Results threw by Google Optimize (cont.)

- So, adding a progress bar didn't make a big different
- But, can we trust in the results by having only 48 sessions?

  Google: Unlike frequentist approaches, Bayesian inference doesn't need a minimum sample. If your conversion rates are really consistent (and consistently different) with low traffic, you can still find actionable results.

  `https://support.google.com/optimize/answer/7404625?hl=en`

- **R/yes**, but... what is Bayes inference and why it doesn't need a minimum size?

Results threw by Google Optimize (cont.)

- So, adding a progress bar didn't make a big different
- But, can we trust in the results by having only 48 sessions?

> *Google: Unlike frequentist approaches, Bayesian inference doesn't need a minimum sample. If your conversion rates are really consistent (and consistently different) with low traffic, you can still find actionable results.*

`https://support.google.com/optimize/answer/7404625?hl=en`

- **R/yes**, but... what is Bayes inference and why it doesn't need a minimum size?

Results threw by Google Optimize (cont.)

- So, adding a progress bar didn't make a big different
- But, can we trust in the results by having only 48 sessions?

Google: Unlike frequentist approaches, Bayesian inference doesn't need a minimum sample. If your conversion rates are really consistent (and consistently different) with low traffic, you can still find actionable results.

https://support.google.com/optimize/answer/7404625?hl=en

- **R/yes**, but... what is Bayes inference and why it doesn't need a minimum size?

Results threw by Google Optimize (cont.)

- So, adding a progress bar didn't make a big different
- But, can we trust in the results by having only 48 sessions?

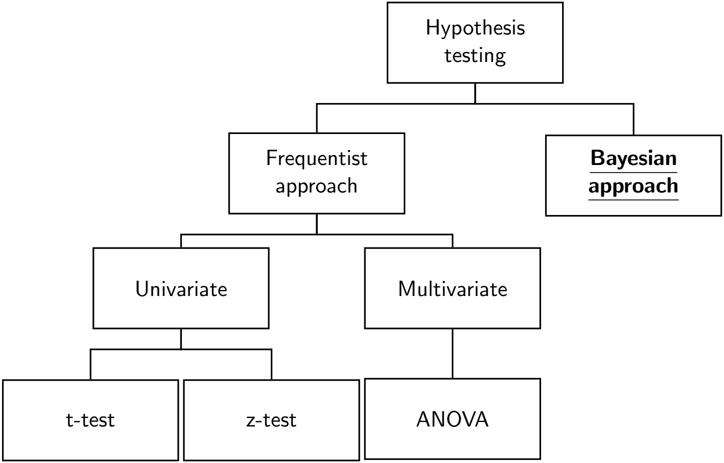
Google: Unlike frequentist approaches, Bayesian inference doesn't need a minimum sample. If your conversion rates are really consistent (and consistently different) with low traffic, you can still find actionable results.

`https://support.google.com/optimize/answer/7404625?hl=en`

- **R/yes**, but... what is Bayes inference and why it doesn't need a minimum size?

Bayesian testing

Let's have a parenthesis...



Bayesian testing (cont.)

Just in case you haven't heard about Thomas Bayes:

United Kingdom 1702 D.C., mathematician, *"An Essay towards solving a Problem in the Doctrine of Chances"*

https://en.wikipedia.org/wiki/Thomas_Bayes.

Bayesian testing (cont.)

Bayes formula:

$$P(\text{fact}|\text{evidence}) = \frac{P(\text{evidence}|\text{fact})P(\text{fact})}{P(\text{evidence})}$$

Where:

- $P(\text{fact})$ is the probability "a priori"
- $P(\text{evidence}|\text{fact})$ is the conditional probability
- $P(\text{evidence})$ is the total probability
- $P(\text{fact}|\text{evidence})$ is the probability "a posteriori" (**our target**)

Bayesian testing (cont.)

Bayes formula:

$$P(\text{fact}|\text{evidence}) = \frac{P(\text{evidence}|\text{fact})P(\text{fact})}{P(\text{evidence})}$$

Where:

- $P(\text{fact})$ is the probability "a priori"
- $P(\text{evidence}|\text{fact})$ is the conditional probability
- $P(\text{evidence})$ is the total probability
- $P(\text{fact}|\text{evidence})$ is the probability "a posteriori" (**our target**)

Bayesian testing (cont.)

Bayes formula:

$$P(\text{fact}|\text{evidence}) = \frac{P(\text{evidence}|\text{fact})P(\text{fact})}{P(\text{evidence})}$$

Where:

- $P(\text{fact})$ is the probability "a priori"
- $P(\text{evidence}|\text{fact})$ is the conditional probability
- $P(\text{evidence})$ is the total probability
- $P(\text{fact}|\text{evidence})$ is the probability "a posteriori" (**our target**)

Bayesian testing (cont.)

Bayes formula:

$$P(\text{fact}|\text{evidence}) = \frac{P(\text{evidence}|\text{fact})P(\text{fact})}{P(\text{evidence})}$$

Where:

- $P(\text{fact})$ is the probability "a priori"
- $P(\text{evidence}|\text{fact})$ is the conditional probability
- $P(\text{evidence})$ is the total probability
- $P(\text{fact}|\text{evidence})$ is the probability "a posteriori" (**our target**)

Bayesian testing (cont.)

Bayes formula:

$$P(\text{fact}|\text{evidence}) = \frac{P(\text{evidence}|\text{fact})P(\text{fact})}{P(\text{evidence})}$$

Where:

- $P(\text{fact})$ is the probability "a priori"
- $P(\text{evidence}|\text{fact})$ is the conditional probability
- $P(\text{evidence})$ is the total probability
- $P(\text{fact}|\text{evidence})$ is the probability "a posteriori" (**our target**)

Bayesian testing (cont.)

Bayes formula:

$$P(\text{fact}|\text{evidence}) = \frac{P(\text{evidence}|\text{fact})P(\text{fact})}{P(\text{evidence})}$$

Where:

- $P(\text{fact})$ is the probability "a priori"
- $P(\text{evidence}|\text{fact})$ is the conditional probability
- $P(\text{evidence})$ is the total probability
- $P(\text{fact}|\text{evidence})$ is the probability "a posteriori" (**our target**)

Bayesian testing (cont.)

So in our example it means [2]:

$$P(\theta|48\,\text{visitors}\,,\Delta t = 2:11) = \frac{P(48\,\text{visitors}\,,\Delta t = 2:11|\theta)P(\theta)}{P(48\,\text{visitors}\,,\Delta t = 2:11)}$$

The *"a priori"* probability $P(\text{fact})$ can either be known using an uniform or gamma distribution:

Bayesian testing (cont.)

So in our example it means [2]:

$$P(\theta|48\,\text{visitors}\,,\Delta t = 2:11) = \frac{P(48\,\text{visitors}\,,\Delta t = 2:11|\theta)P(\theta)}{P(48\,\text{visitors}\,,\Delta t = 2:11)}$$

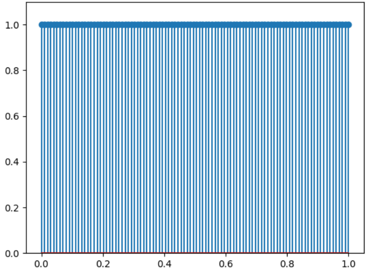
The *"a priori"* probability $P(\text{fact})$ can either be known using an uniform or gamma distribution:

Bayesian testing (cont.)

So in our example it means [2]:

$$P(\theta|48\,\text{visitors}\,,\Delta t = 2:11) = \frac{P(48\,\text{visitors}\,,\Delta t = 2:11|\theta)P(\theta)}{P(48\,\text{visitors}\,,\Delta t = 2:11)}$$

The *"a priori"* probability $P(\text{fact})$ can either be known using an uniform or gamma distribution:

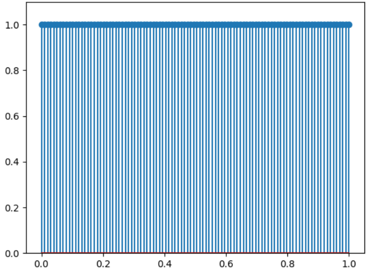


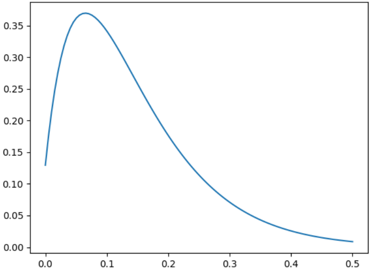
Bayesian testing (cont.)

So in our example it means [2]:

$$P(\theta | 48\,\text{visitors}\,, \Delta t = 2:11) = \frac{P(48\,\text{visitors}\,, \Delta t = 2:11 | \theta) P(\theta)}{P(48\,\text{visitors}\,, \Delta t = 2:11)}$$

The *"a priori"* probability $P(\text{fact})$ can either be known using an uniform or gamma distribution:





Bayesian testing (cont.)

- The rest of probabilities can be known also using the gamma distribution like explained by [2]

Ok, nice... but... what is the frequentist approach?

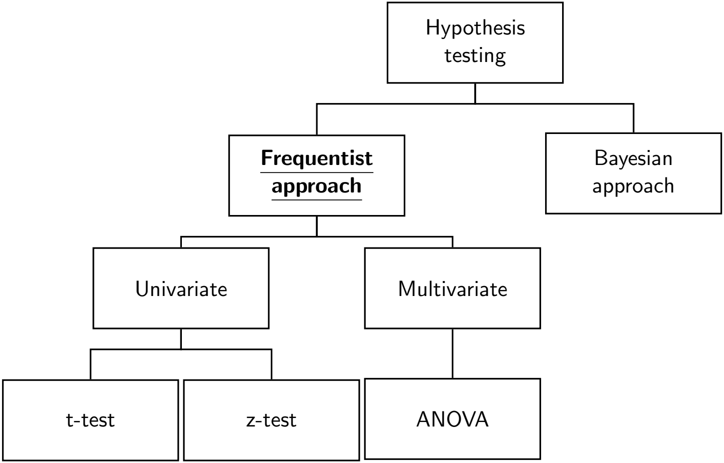
Bayesian testing (cont.)

- The rest of probabilities can be known also using the gamma distribution like explained by [2]

Ok, nice... but... what is the frequentist approach?

Frequentist testing

Let's have another parenthesis...



Frequentist testing (cont.)

- It's the term used to group all the tests that depend on n, the sample size
- It allows to find the probability of getting a certain sample mean $\bar{x}$, for instance, 3:35 (var)
- Some types of frequentist tests are:
  - Z-test
  - T-test (when the sample is not normally distributed)
  - ANalysis Of VAriance (ANOVA)

Frequentist testing (cont.)

- It's the term used to group all the tests that depend on n, the sample size
- It allows to find the probability of getting a certain sample mean $\bar{x}$, for instance, 3:35 (var)
- Some types of frequentist tests are:
  - Z-test
  - T-test (when the sample is not normally distributed)
  - ANalysis Of VAriance (ANOVA)

Frequentist testing (cont.)

- It's the term used to group all the tests that depend on n, the sample size
- It allows to find the probability of getting a certain sample mean $\bar{x}$, for instance, 3:35 (var)
- Some types of frequentist tests are:
    - Z-test
    - T-test (when the sample is not normally distributed)
    - ANalysis Of VAriance (ANOVA)

Frequentist testing (cont.)

- It's the term used to group all the tests that depend on n, the sample size
- It allows to find the probability of getting a certain sample mean $\bar{x}$, for instance, 3:35 (var)
- Some types of frequentist tests are:
  - Z-test
  - T-test (when the sample is not normally distributed)
  - ANalysis Of VAriance (ANOVA)

Frequentist testing (cont.)

- It's the term used to group all the tests that depend on n, the sample size
- It allows to find the probability of getting a certain sample mean $\bar{x}$, for instance, 3:35 (var)
- Some types of frequentist tests are:
  - Z-test
  - T-test (when the sample is not normally distributed)
  - ANalysis Of VAriance (ANOVA)

Frequentist testing (cont.)

- It's the term used to group all the tests that depend on n, the sample size
- It allows to find the probability of getting a certain sample mean $\bar{x}$, for instance, 3:35 (var)
- Some types of frequentist tests are:
  - Z-test
  - T-test (when the sample is not normally distributed)
  - ANalysis Of VAriance (ANOVA)

Frequentist testing (cont.)

For instance, a z-test would look like this [3]:[2]

- Define null and alternative sample hypothesis H_0 and H_a
- Choose a significance level for evaluating the null-hypothesis (e.g. $\alpha = 0.05$)
- Compute the z value:

$$z = \frac{\bar{x} - \mu_a}{\frac{\sigma}{\sqrt{n}}} = \frac{2:11}{\frac{\sigma}{\sqrt{48}}}$$

- Where:
  - $\bar{x}$: mean of the sample
  - σ : standard deviation of the population
  - n : sample size

[2] In other words, how many standard deviations is μ_a from $\bar{x}$.

Frequentist testing (cont.)

For instance, a z-test would look like this [3]:[2]

- Define null and alternative sample hypothesis H_0 and H_a
- Choose a significance level for evaluating the null-hypothesis (e.g. $\alpha = 0.05$)
- Compute the z value:

$$z = \frac{\bar{x} - \mu_a}{\frac{\sigma}{\sqrt{n}}} = \frac{2:11}{\frac{\sigma}{\sqrt{48}}}$$

- Where:
  - $\bar{x}$: mean of the sample
  - σ : standard deviation of the population
  - n : sample size

[2] In other words, how many standard deviations is μ_a from $\bar{x}$.

Frequentist testing (cont.)

For instance, a z-test would look like this [3]:[2]

- Define null and alternative sample hypothesis H_0 and H_a
- Choose a significance level for evaluating the null-hypothesis (e.g. $\alpha = 0.05$)
- Compute the z value:

$$z = \frac{\bar{x} - \mu_a}{\frac{\sigma}{\sqrt{n}}} = \frac{2:11}{\frac{\sigma}{\sqrt{48}}}$$

- Where:
  - $\bar{x}$: mean of the sample
  - σ : standard deviation of the population
  - n : sample size

[2] In other words, how many standard deviations is μ_a from $\bar{x}$.

Frequentist testing (cont.)

For instance, a z-test would look like this [3]:[2]

- Define null and alternative sample hypothesis H_0 and H_a
- Choose a significance level for evaluating the null-hypothesis (e.g. $\alpha = 0.05$)
- Compute the z value:

$$z = \frac{\bar{x} - \mu_a}{\frac{\sigma}{\sqrt{n}}} = \frac{2:11}{\frac{\sigma}{\sqrt{48}}}$$

- Where:
  - $\bar{x}$: mean of the sample
  - σ : standard deviation of the population
  - n : sample size

[2] In other words, how many standard deviations is μ_a from $\bar{x}$.

Frequentist testing (cont.)

For instance, a z-test would look like this [3]:[2]

- Define null and alternative sample hypothesis H_0 and H_a
- Choose a significance level for evaluating the null-hypothesis (e.g. $\alpha = 0.05$)
- Compute the z value:

$$z = \frac{\bar{x} - \mu_a}{\frac{\sigma}{\sqrt{n}}} = \frac{2:11}{\frac{\sigma}{\sqrt{48}}}$$

- Where:
  - $\bar{x}$: mean of the sample
  - σ : standard deviation of the population
  - n : sample size

[2] In other words, how many standard deviations is μ_a from $\bar{x}$.

Frequentist testing (cont.)

For instance, a z-test would look like this [3]:[2]

- Define null and alternative sample hypothesis H_0 and H_a
- Choose a significance level for evaluating the null-hypothesis (e.g. $\alpha = 0.05$)
- Compute the z value:

$$z = \frac{\bar{x} - \mu_a}{\frac{\sigma}{\sqrt{n}}} = \frac{2:11}{\frac{\sigma}{\sqrt{48}}}$$

- Where:
  - $\bar{x}$: mean of the sample
  - σ : standard deviation of the population
  - n : sample size

[2] In other words, how many standard deviations is μ_a from $\bar{x}$.

Frequentist testing (cont.)

For instance, a z-test would look like this [3]:[2]

- Define null and alternative sample hypothesis H_0 and H_a
- Choose a significance level for evaluating the null-hypothesis (e.g. $\alpha = 0.05$)
- Compute the z value:

$$z = \frac{\bar{x} - \mu_a}{\frac{\sigma}{\sqrt{n}}} = \frac{2 : 11}{\frac{\sigma}{\sqrt{48}}}$$

- Where:
  - $\bar{x}$: mean of the sample
  - σ : standard deviation of the population
  - n : sample size

[2] In other words, how many standard deviations is μ_a from $\bar{x}$.

Frequentist testing (cont.)

- Set the **state decision rule:** one or two tails test
- Then find the p-value that matches $1 - \alpha$ (area under the curve) using the z-table:[3]

https://www.dummies.com/wp-content/uploads/451654.image0.jpg

[3] X-axis on the table is the second decimal place of p.

Frequentist testing (cont.)

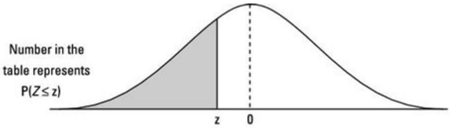
- Set the **state decision rule:** one or two tails test
- Then find the p-value that matches $1 - \alpha$ (area under the curve) using the z-table:[3]

https://www.dummies.com/wp-content/uploads/451654.image0.jpg

[3] X-axis on the table is the second decimal place of p.

Frequentist testing (cont.)

- Set the **state decision rule:** one or two tails test
- Then find the *p*-value that matches $1 - \alpha$ (area under the curve) using the z-table:[3]



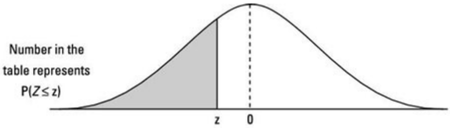
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
–3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
–3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
–3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
–3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
–3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
–3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
–3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
–2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
–2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019

https://www.dummies.com/wp-content/uploads/451654.image0.jpg

[3] X-axis on the table is the second decimal place of *p*.

Frequentist testing (cont.)

- Set the **state decision rule:** one or two tails test
- Then find the *p*-value that matches $1 - \alpha$ (area under the curve) using the z-table:[3]



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
−3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
−3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019

`https://www.dummies.com/wp-content/uploads/451654.image0.jpg`

[3] X-axis on the table is the second decimal place of p.

Frequentist testing (cont.)

- So if $z < |p|$ (for a two tails test), reject the null-hypothesis, if not discard the alternative one
- $1 - \alpha$ is the upper bound for the cumulative probability distribution

Frequentist testing (cont.)

- So if $z < |p|$ (for a two tails test), reject the null-hypothesis, if not discard the alternative one
- $1 - \alpha$ is the upper bound for the cumulative probability distribution

Discussion

As we saw, the Bayesian approach doesn't use the sample size n, whereas the frequentist approach does:

- The z-value depends of n
- The *"a posteriori"* probability $P(\text{fact}|\text{evidence})$, does not
- That's why Google says that we can still have significant results with low traffic

We can breath in peace! finally.

Discussion

As we saw, the Bayesian approach doesn't use the sample size n, whereas the frequentist approach does:

- The z-value depends of n
- The *"a posteriori"* probability $P(\text{fact}|\text{evidence})$, does not
- That's why Google says that we can still have significant results with low traffic

We can breath in peace! finally.

Discussion

As we saw, the Bayesian approach doesn't use the sample size n, whereas the frequentist approach does:

- The z-value depends of n
- The *"a posteriori"* probability $P(\text{fact}|\text{evidence})$, does not
- That's why Google says that we can still have significant results with low traffic

We can breath in peace! finally.

Discussion

As we saw, the Bayesian approach doesn't use the sample size n, whereas the frequentist approach does:

- The z-value depends of n
- The *"a posteriori"* probability $P(\text{fact}|\text{evidence})$, does not
- That's why Google says that we can still have significant results with low traffic

We can breath in peace! finally.

Discussion

Oh, by the way, Google Optimize is not the only tool in the market:

Optimizely

mixpanel

Conclusions

What have we learned?

- A/B testing requires knowledge about the business domain
- You can't have good results if you don't design good experiments with a reasonable hypothesis (it's an art)
- Google Optimizely allows you to implement experiments easily
- The Bayesian approach does not depend on the sample size

Conclusions

What have we learned?

- A/B testing requires knowledge about the business domain
- You can't have good results if you don't design good experiments with a reasonable hypothesis (it's an art)
- Google Optimizely allows you to implement experiments easily
- The Bayesian approach does not depend on the sample size

Conclusions

What have we learned?

- A/B testing requires knowledge about the business domain
- You can't have good results if you don't design good experiments with a reasonable hypothesis (it's an art)
- Google Optimizely allows you to implement experiments easily
- The Bayesian approach does not depend on the sample size

Conclusions

What have we learned?

- A/B testing requires knowledge about the business domain
- You can't have good results if you don't design good experiments with a reasonable hypothesis (it's an art)
- Google Optimizely allows you to implement experiments easily
- The Bayesian approach does not depend on the sample size

References I

Alex Birkett
Bayesian vs Frequentist A/B Testing – What's the Difference?.
CXL, 2015. https://conversionxl.com/blog/bayesian-frequentist-ab-testing

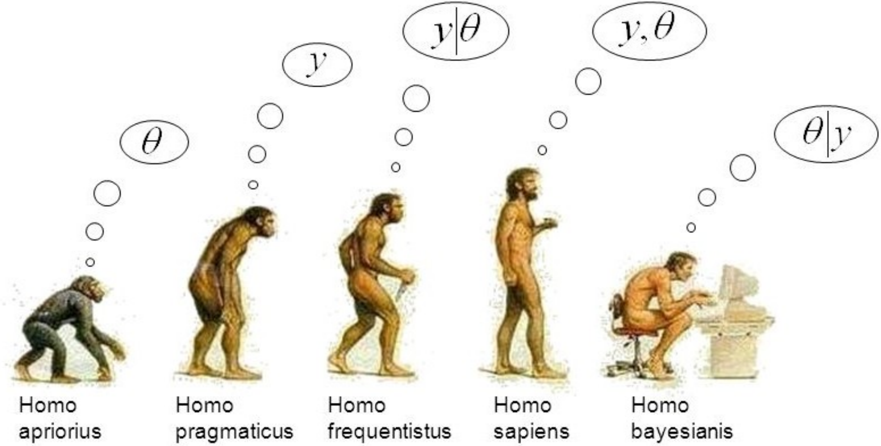
Chris Stucchio
Analyzing conversion rates with Bayes Rule.
Personal webpage, 2013. https://www.chrisstucchio.com/blog/2013/bayesian_analysis_conversion_rates.html

Muhammad Anas
Z-test with examples.
Linkedin Slideshare, 2017. https://es.slideshare.net/MuhammadAnas96/ztest-with-examples

References II

Anil Batra
A/B Testing and Experimentation for Beginners.
Udemy, 2019. https://www.udemy.com/course/ab-testing-and-experimentation-for-websites-and-marketing/

Questions?



https://images.slideplayer.com/3/780091/slides/slide_29.jpg