L3 Metrics. Notes

Jakub Kuźnicki 2022 · 79 Lub

A/B Testing

Michael Sliney

- Introduction
  - 1.1 When did people first start to use A/B testing?
  - 1.2 Why is it important that users are A/B testing?
  - 1.3 Note on metrics
  - 1.4 Binomial Distribution
  - 1.5 Confidence Interval
  - 1.6 Hypothesis Testing
  - 1.7 Comparing two samples
  - 1.8 Practical significance
  - 1.9 Size vs Power trade-off
  - 1.10 Conclusion
- Metrics
  - 2.1 Characteristics of a metric
  - 2.2 Categories of metrics
  - 2.3 Absolute vs Relative
  - 2.4 Calculating variability
  - 2.5 Non-parametric methods
- 3 Designing an Experiment
  - 3.1 Hypothesis
  - 3.2 Population
  - 3.3 Randomization
  - 3.4 Sample Size
  - 3.5 Duration
  - 3.6 Sanity Checks
  - 3.7 Analysis
  - 3.8 Decision
  - 3.9 Follow-Up
  - 3.10 Pitfalls
- Conclusion
- References

Lab Testing, Helping Change Minds. Trial Results.

Start Experiment | Edit | Stop Experiment

Generated with Jupyter Book

1 Introduction

A/B testing is a statistical approach to validating the value of a product or design by comparing two versions against each other. In the context of A/B testing, it is a critical tool to ensure that a change to a product is actually improving the user experience and not just a random fluctuation or a bug.

1.1 When did people first start to use A/B testing?

A/B testing is not quite the same as the scientific method, but it shares many of the same principles. It has different between critical tools and control of testing. It is a critical tool to ensure that a change to a product is actually improving the user experience and not just a random fluctuation or a bug.

1.2 Why is it important that users are A/B testing?

A/B testing is important because it is testing the value of a product or design by comparing two versions against each other.

1.3 Note on metrics

The metrics that are used in A/B testing are critical to the success of the test, and potentially what you need to know about.

1.4 Binomial Distribution

The binomial distribution with parameters n and p is the discrete probability distribution of the number of successes in a sequence of n independent experiments, each asking a yes–no question, and each with its own Boolean-valued outcome: success (with probability p) or failure (with probability $q = 1 - p$). A single success/failure experiment is also called a Bernoulli trial or Bernoulli experiment, and a sequence of outcomes is called a Bernoulli process; for a single trial, i.e. $n = 1$, the binomial distribution is a Bernoulli distribution. The binomial distribution is the basis for the popular binomial test of statistical significance.

The binomial distribution is frequently used to model the number of successes in a sample of size n drawn with replacement from a population of size N. If the sampling is carried out without replacement, the draws are not independent and so the resulting distribution is a hypergeometric distribution, not a binomial one. However for N much larger than n, the binomial distribution is a good approximation, and widely used.

$$\Pr(k \text{ successes in } n \text{ trials}) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

$$\Pr(k \text{ successes in } n \text{ trials}) = \int_0^1 p^k (1-p)^{n-k} \, dp$$

1.5 Confidence Interval

A confidence interval indicates the range within which the mean is expected to fall a multiple times of standard deviation.

For a sample of n observations $\bar{x}$ and the sample mean, $\bar{x}$ is the standard of value x and an arbitrary $\bar{x}$ binomial distribution (the capacity N is $\bar{x}$ and $\bar{x}$ $\bar{x}$). The margin of error is given by

$$\bar{x} \pm \frac{\sigma}{\sqrt{n}}$$

For a 95% confidence interval, $\sigma = 1.96$.

1.6 Hypothesis Testing

Hypothesis testing is a statistical method that is used in making statistical decisions using experimental data. In statistics, a well-established method for making decisions is to confirm the null hypothesis H_0 by using the null hypothesis:

$$H_0: \mu_1 = \mu_2 = 0$$

The alternative hypothesis is that

$$H_1: \mu_1 \neq 0$$

1.7 Comparing two samples

For a statistical test, assuming unbiased data, we can use the two-sample t-test to determine if two samples x_1 and x_2 are the current estimate of some true mean, and we can construct a value for the current value μ_1 and μ_2, μ_1 and μ_2 are the estimated mean of the two samples x_1 and x_2:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}}$$

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

1.8 Practical significance

$$\Pr(|T| \geq |t| \mid H_0)$$

$$\Pr(|T| \geq |t| \mid \mu_1 - \mu_2 = 0)$$

The two-sample t-test demonstrates the analysis of a A/B test.

```
# Import the library
import numpy as np, scipy.stats as st
np.random.seed(12345678)

# Fix random seed for reproducibility
rvs = st.norm(loc=5, scale=10).rvs(size=500)

# Perform the one-sample t-test
st.ttest_1samp(rvs, 5.0)
```

```
Ttest_1samp(statistic=0.5489963725317436, pvalue=0.5833766058314146)
```

```
# Print the sample mean
rvs.mean()
```

```
5.08615393959798
```

```
# Perform the one-sample t-test
st.ttest_1samp(rvs, 5.0)[0]
```

```
0.5489963725317436
```

```
# Print the p-value
st.ttest_1samp(rvs, 5.0)[1]
```

```
0.5833766058314146
```

The two-sample t-test demonstrates the analysis of a A/B test.

- $\mu_1 = \mu_2 = 0$ (the null hypothesis)
- $\mu_1 \neq \mu_2$ (the alternative hypothesis)
- $\mu_1 > \mu_2$ (the alternative hypothesis)
- $\mu_1 < \mu_2$ (the alternative hypothesis)

The two-sample t-test demonstrates the analysis of a A/B test.

```
# Import the library
import numpy as np, scipy.stats as st
np.random.seed(12345678)

# Fix random seed for reproducibility
rvs1 = st.norm(loc=5, scale=10).rvs(size=500)
rvs2 = st.norm(loc=5.0, scale=10).rvs(size=500)

# Perform the two-sample t-test
st.ttest_ind(rvs1, rvs2)
```

```
Ttest_ind(statistic=0.5489963725317436, pvalue=0.5833766058314146)
```

```
# Print the sample mean
rvs1.mean()
```

```
5.08615393959798
```

```
# Print the sample mean
rvs2.mean()
```

```
5.0
```

```
# Perform the two-sample t-test
st.ttest_ind(rvs1, rvs2)[0]
```

```
0.5489963725317436
```

```
# Print the p-value
st.ttest_ind(rvs1, rvs2)[1]
```

```
0.5833766058314146
```

1.9 Size vs Power trade-off

Size and power are two critical metrics that need to be considered when designing a test. The size is the probability of rejecting the null hypothesis when it is actually true. The power is the probability of rejecting the null hypothesis when it is false. The power is the probability of rejecting the null hypothesis when it is false.

$$\Pr(\text{reject } H_0 \mid H_0 \text{ is true})$$

$$\Pr(\text{reject } H_0 \mid H_0 \text{ is false})$$

$$\Pr(\text{reject } H_0 \mid H_0 \text{ is true}) = \