Contents

List of Figures

List of Tables

Hypothesis & AB Testing

0.1 Basics - Hypothesis Testing

Suppose we have a coin that is designed so that head is shown 50% of the time when flipped. Since the coin lands on either heads or tails, we call this a Bernoulli trial (exactly two outcomes):

$$Pr(X = \mathrm{H}) \;=\; 1 \;-\; Pr(X = \mathrm{T}) \tag{1}$$

If the probability of heads being shown is 50%, then the probability of tails is also 50%.

Assume you have a set of observations O from flipping the coin. We would like to know the probability of those observations given some hypothesis H_0:

$$P(O|H_0) \tag{2}$$

The hypothesis we want to consider is whether the observations in the set O are consistent with a coin whose probability of showing heads (or tails) is 50% when flipped. We call this the *null* hypothesis and denote it as H_0. The alternate hypothesis (H_A) is that the probability is not 50%.

$$\begin{aligned} H_0 &: \; p_0 = 0.50 \\ H_A &: \; p_0 \neq 0.50 \end{aligned} \tag{3}$$

We usually reject the null hypothesis if the probability of these observations are less than 0.05 — this is equivalent to saying that we are 95% confident in the alternate hypothesis.

$$P(O|H_0) \leq 1 - 0.95 \Rightarrow \text{reject } H_0 \tag{4}$$

So, how do we figure out $P(O|H_0)$? Let's begin with some observations:

Coin	Num Flips	% Heads
Coin 1	100	51%
Coin 2	120	42%
Coin 3	105	20%

Table 0.1: Results of flipping 3 coins

We estimate the standard deviation of p:

$$\sigma = \sqrt{\frac{p(1-p)}{N}} \tag{5}$$

We use this equation to determine the standard deviation for the probability of "heads" for each coin: This gives us:

Coin	% Heads	σ
Coin 1	51%	0.050
Coin 2	42%	0.045
Coin 3	20%	0.038

Table 0.2: Results of flipping 3 coins

Once we know the standard deviation, we can compute the Z-score; the Z-Score tells how much the sample mean differs from the expected mean given in the null hypothesis. We compute the Z-Score:

$$z = \frac{p - 0.50}{\sigma} \tag{6}$$

Using this formula, we determine the Z-Score for each of our three coins:

Coin	% Heads	σ	Z-Score
Coin 1	51%	0.050	0.20
Coin 2	42%	0.045	-1.78
Coin 3	20%	0.038	-7.89

Table 0.3: Results of flipping 3 coins

The following figure shows we expect 95% of observations to be within 1.96 standard deviations.

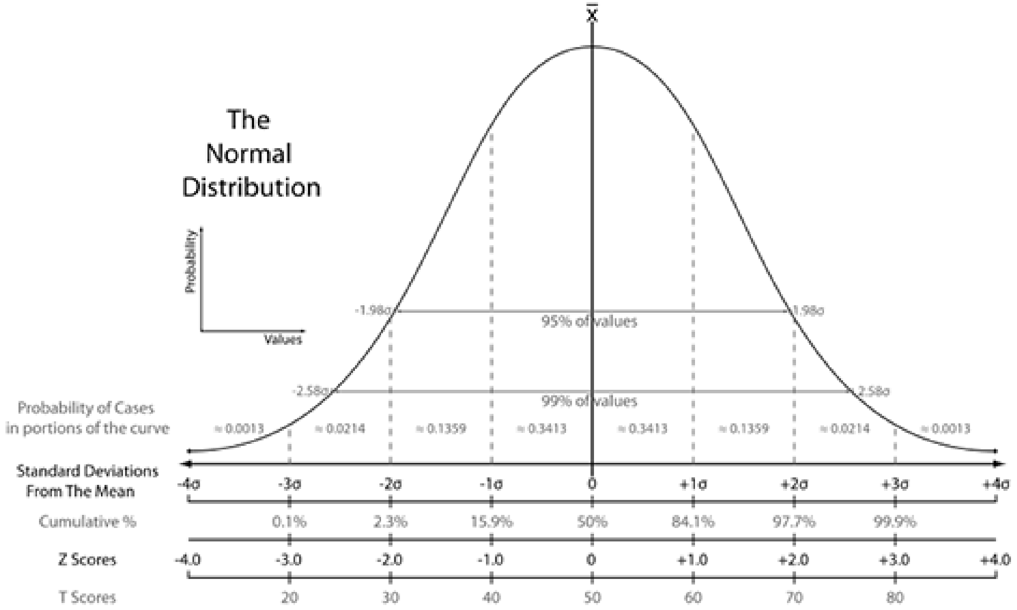


Figure 0.1: Bell Curve

In the case of our coins, the null hypothesis is that each coin is fair — that the observed probability of showing heads is not different from 0.50, the expected probability for a fair coin.

A Z-Score between +1.96 and −1.96, inclusively, identifies that the observed mean is "expected." More exactly: A Z-Score larger than +1.96 or less than −1.96 is expected in less than 5% of observed cases; thus, in the case of Coin 3, we conclude that the coin is **not** fair.

We conclude, also, that Coin 1 and 2 are fair. Coin 3's Z-Score tell us that is not fair because the observed mean probability of showing heads is unlikely to be observed in a fair coin given the number trials (flips). Coin 1 and 2's Z-Score tell us that they are fair because the observed mean probability of showing is likely to be observed in a fair coin given the number of trials.

0.2 A/B Testing

A/B testing is a way to determine whether a change to a system had the effect you intended. For example, suppose you have an announcement that you'd like people to read (perhaps the announcement is an advertisement). Since announcements are rather long, you send a "header" message that shows up as the subject and/or preview of the message. What you'd like to determine is which of two headers (A/B) is likely to entice your audience to read the announcement.

When people talk about A/B tests, they mean an experiment that runs concurrently (to eliminate any time related sensitivities) and that entities are randomly assigned a treatment group.

A/B testing is like a clinical trial where the effects of one thing are measured by comparing the behavior of a control group with a treatment group. A/B testing differs from a clinical trial in at least two ways:

1. Clinical trials involve at most hundreds to thousands of people; A/B testing involves millions of users.
2. We generally know quite a bit about each person in a clinical trial whereas in A/B testing we know very little.

Suppose that we have an existing header message (header A) for our announcement. We'll call the new header message header B. Our goal is that header B increases announcement conversion (% announcements read) by 10%.

Let's pause for a moment and discuss the previous paragraph. It's not really enough to "know" that header B announcement conversion differs from header A's. And it's not really enough to know that header B's announcement conversion is greater than header A's. We want to say that based on the observed announcement conversion rates for header A and header B groups, we are 95% confident that the header B group has an announcement conversion rate at least 10% greater than the header A group.

To measure conversion rates, we use click-through-probability (CTP):

$$\frac{\text{Number of unique visitors who read}}{\text{Number of unique visitors sent}} \tag{7}$$

If we sent our announcement to 2 people and one of them read the announcement 4 times, the CTP is 50%.

0.3 Confidence Intervals

Suppose we run a test on March 15$^{\text{th}}$ and measure the CTP as 10% (100 out of 1000). The next day we measure the CTP as 10.1% (202 out of 2000). And we continue measuring for several days:

Date	CTP	# Read	# Sent
March 15	10.0%	100	1000
March 16	10.1%	202	2000
March 17	7.0%	098	1400
March 18	11.0%	242	2200
March 19	10.5%	184	1749

Table 0.4: CTP Daily Scores

The CTP for March 17$^{\text{th}}$ seems like an aberration; perhaps something happened that day (servers down, bad weather, etc.)? Before investigating what happened, we need to know whether 7% is indeed "surprising" — is 7% surprising when every other similar observation is relatively near 10%?

As a general rule, we expect values within the 95% confidence interval and are "surprised" when we see a value outside of that interval. To determine whether the observed values are within the 95% confidence interval we use the binomial distribution (Figure 0.1) and its standard deviation (Equation 5). For March 17$^{\text{th}}$, we have:

$$\begin{aligned} p &= 0.07 \\ n &= 1400 \\ \sigma &= 0.0068 = \sqrt{\frac{p(1-p)}{n}} \end{aligned}$$

A quick glance at the binomial distribution tells us that 95% of values are within ±1.96 σ of the mean:

$$\begin{aligned}\text{LowerBound} &= 0.0566 = (0.07 - 1.96 \times 0.0068)\\ \text{UpperBound} &= 0.0834 = (0.07 + 1.96 \times 0.0068)\end{aligned}$$

These results tell that the March 17th 7% CTP is surprising given we are expecting a CTP of roughly 10%. Using our March 15th results, we have:

$$\begin{aligned}p &= 0.10\\ n &= 1000\\ \sigma &= 0.0095\end{aligned}$$

LCB and UCB:

$$\begin{aligned}\text{LowerBound} &= 0.0814 = (0.10 - 1.96 \times 0.0095)\\ \text{UpperBound} &= 0.1186 = (0.10 + 1.96 \times 0.0095)\end{aligned}$$

Again, the March 17th 7% CTP is surprising.

section summary. Confidence intervals formalize our notion of surprising. we use the binomial distribution because it is central limit theorem. and we can use it when there are two outcomes, previous trials don't affect future ones and when the probabilities don't change .

0.4 Methodology & Interpretation

When we run an A/B test, we want to determine whether A is better than B or vice versa. To make this determination we need to compare the CTP of A with that of B. However, just because A's CTP is larger than B's doesn't mean that it is statistically better. Recall that because of confidence intervals, it is possible that A's true value is smaller than our observation and that B's is bigger than what we observe.

Let us say that B is better than A by at least 3%. We call this the *statistical significance*. But suppose that, from a business perspective, we cannot justify B unless the difference is 5% or more. This is called the *practical significance*. So, if the practical significance is 3% and the statistical significance is 2%, the business decision won't support making the improvement. On the other hand, if the practical significance is 2% and the statistical significance is 3%, then the business ought to make the improvement.

Suppose that we ran an A/B test with the following observations:

Test	CTP	# Sent	# Read
A	9.69%	10,099	979
B	12.64%	9,891	1,250

Table 0.5: A/B Test Results

The A/B test results look encouraging. A quick check tells us that observed CTP's for each test group don't overlap (i.e., the upper bound of one doesn't overlap the lower bound of the other).

$$\begin{aligned}\text{Upper Bound A} &= 0.0969 + (1.96 * \sigma_A)\\ \text{Lower Bound B} &= 0.1264 - (1.96 * \sigma_B)\end{aligned}$$

But that they don't overlap, doesn't tell us how different they are from each other. What we want to know is that difference between CTP_A and CTP_B is at least d with 95% confidence; a 95% confidence means that we will will say the difference is at least d, 5% of the time, when in fact the difference is less than d. Another way to say all this is that we are using wrongly reject the null hypothesis at most 5% of the time.

Let's work through this example. The overall CTP of A and B is called the *pooled* CTP:

$$p_{\text{pool}} \quad = \quad \frac{A_{\text{Read}} + B_{\text{Read}}}{A_{\text{Sent}} + B_{\text{Sent}}} \quad = \quad \frac{979 + 1250}{10099 + 9891} \quad = \quad 0.1115 \tag{8}$$

Next, we determine the *pooled* standard error:

$$SE_{\text{pool}} \quad = \quad \sqrt{p_{\text{pool}}(1 - p_{\text{pool}})(\frac{1}{A_{\text{Sent}}} + \frac{1}{B_{\text{Sent}}})} \quad = \quad 0.0045 \tag{9}$$

Now, d is the difference in observed CTP's; so,

$$d = 0.1264 - 0.0969 = 0.0295$$

Given a pooled standard error of 0.0045 and an observed difference in CTP of 2.95%, with 95% confidence we can say that the true difference lies between:

$$\begin{aligned} d_{\text{lower}} \quad &= \quad 0.0207 \quad = \quad 0.0295 - (1.96 * 0.0045) \\ d_{\text{upper}} \quad &= \quad 0.0383 \quad = \quad 0.0295 + (1.96 * 0.0045) \end{aligned}$$

Our conclusion is that B is better than A with a statistical significance of approximately 2.1%. The business decision to use B depends on the practical significance.

0.5 Launch Decision

Once an A/B test has been completed, we need to decide whether to "launch" the change. Let us say that the practical significance to launch is d (e.g., d = 3%). Suppose further that the result of our A/B test is that the proposed change has an observed difference of 3.5%. Whether we go with the proposed change or keep the existing system and features in place depends on the confidence interval!

Figure 0.2 details some common situations.

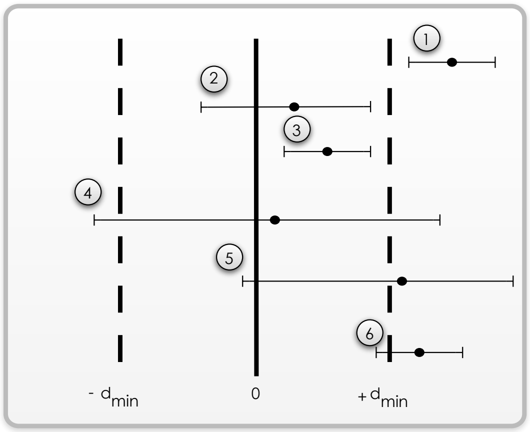


Figure 0.2: Can We Launch?

Case 1:

> In this case, the observed difference as well as the lower confidence limit is greater than the practical significant. We should launch this change!

Case 2:

In this case, the observed difference as well as both the upper and lower confidence intervals are less than the practical significance. We should **NOT** launch this change!

Case 3:

In this case, the observed difference as well as both the upper and lower confidence intervals are less than the practical significance. We should **NOT** launch this change!

Case 4:

In this case, the observed difference as well as both the upper and lower confidence intervals are beyond the *edges* of the practical significance interval. Our recommendation is that further testing is needed; we don't have enough information to make a recommendation.

Case 5:

In this case, the observed difference as well as both the upper and lower confidence intervals are beyond the *edges* of the practical significance interval. Our recommendation is that further testing is needed; we don't have enough information to make a recommendation.

Case 6:

In this case, part of the lower portion of the confidence interval lies below the practical significance. Our recommendation is that further testing is needed.

0.6 Why Sample Size Is Important

As sample size increases, we gain confidence that our observed CTP is the real CTP. To be very precise, as sample size increases the difference between the upper and lower confidence bounds becomes less. But there is another side to sample size.

Our approach to deciding whether to accept or reject the null hypothesis is based on minimizing the probability that we reject null hypothesis when it is actually true. Rejecting the null hypothesis when it is true is called a **Type I** error. It is common practice to denote the probability of making a Type I error by the Greek letter α. It is also common practice to set $\alpha = 0.05$ so that the likelihood of concluding there is an effect when there is none (Type I error) does not exceed 5%.

When we get a result and the probability of that result given that the null hypothesis is true is less than α, then the probability of rejecting the null hypothesis when it is in true is less than α. For this reason, α is referred to as the *statistical significance.*

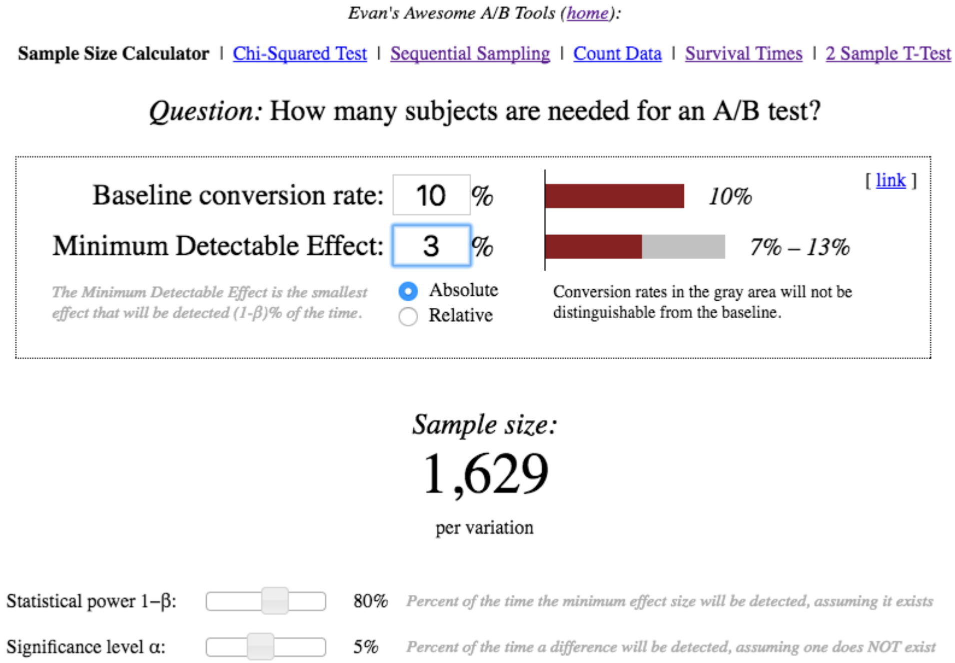
When we conclude that an effect exists even when there is no such effect, we make a Type I error. When we conclude that no effect exists when there is an effect, we make a Type II error. The probability of making a Type II error is denoted by the Greek letter β. The probability of not making a Type II error is called *statistical power*; statistical power is the probability of detecting an effect when effect exists. If the power is high, the probability of wrongly concluding that no effect occurred when, in fact, one occurred is low.

It is common practice to set power to 80%. Here's why. In 1988, Jacob Cohen, suggested that most researchers considered Type I errors as being 4 more serious than Type II errors. Since α is almost always set to 0.05, then β ought to be set to 0.20; thus, power $(1 - \beta)$ at 0.80.

Both significance and power increase as the sample size is increased. For a given sample size, power increases as the effect is increased. The relationship between significance, power, effect and sample size means that knowing any three lets us determine the other. Most people use a tool to compute the sample size needed to measure a given effect at a particular significance and power. We suggest using *"Evan's Awesome A/B Tools"* which can be found on the web:

`http://www.evanmiller.org/ab-testing/`

For example, suppose you want to measure a 3% improvement (this is called the Minimum Detectable Effect) to an existing (conversion rate) CTP of 10%. As shown in Figure 0.3, setting these parameters along with $\alpha = 0.05$, $1 - \beta = 0.80$ and "Absolute" in the sample size calculator found on *"Evan's Awesome A/B Tools"* website will identify that you need a sample size of $1,629$.



Evan's Awesome A/B Tools (home):

Sample Size Calculator | Chi-Squared Test | Sequential Sampling | Count Data | Survival Times | 2 Sample T-Test

Question: How many subjects are needed for an A/B test?

[link]

Baseline conversion rate: 10 % — *10%*

Minimum Detectable Effect: 3 % — *7% – 13%*

The Minimum Detectable Effect is the smallest effect that will be detected (1-β)% of the time.

● Absolute
○ Relative

Conversion rates in the gray area will not be distinguishable from the baseline.

Sample size:

1,629

per variation

Statistical power 1−β: 80% *Percent of the time the minimum effect size will be detected, assuming it exists*

Significance level α: 5% *Percent of the time a difference will be detected, assuming one does NOT exist*

Figure 0.3: Using Evan's Awesome A/B Tools to determine sample size