Conversion Rate Quality

Jonathan Bown

Developed in association with Perfect Storm Media

Abstract

Finding the right way to measure conversion rate quality can be a burdensome task for decision makers in search engine marketing. There are many different methods of measuring how good one conversion rate is relative to another or several others with the most popular being the A/B test. Too often methods of testing are unable to arrive at a conclusion because either there isn't enough data or assumptions aren't properly met. This can be decieving to business owners, marketers, and decision makers. In this document we present several simple methods of measuring conversion rate quailty to assist in the decision making process that are required to meet minimal statistical assumptions. These methods are meant to tell a story of how a conversion rate is doing at the moment it is analyzed to help users better understand their conversion rates.

Contents

1 Issues with A/B testing

A/B tests are prone to many statistical errors due to the nature of the test. A/B testing relies on the statistical framework of hypothesis testing in order to find 'statistical significance'. Hypothesis testing is not the problem but rather the application and scope can often miss the mark relative to what the assumptions of the test are and where it can fail. Hypothesis tests are prone to two specific types of errors namely 'false positives' and 'false negatives'. Meaning that if you find a statistically significant result, there is a non-zero probability that it is actually not significant. Conversely if you conclude a lack of statistical significance there is still a non-zero probability that the result is actually statistically significant. The occurence of these errors rely heavily on underlying assumptions about the distribution of variables being tested such as parameters chosen or the type of test statistic.

Once a test is chosen and hypotheses determined we often see that these tests are run until a significant result is chosen. This leads to an error of application called *repeated significance testing errors* which means that as we run multiple tests the significance level actually changes without realizing it. Repeated significance testing will always increase the rate of false positives. A false positive would indicate that insignificant results are significant. More details on this error can be found at [4]. It is worth pointing out that you are limited on how many times you can run this test in order to meet assumption and actually find a valid result. Because of the volatile nature of conversion rates a more ideal approach should allow the user to run a test whenever information is required.

A/B testing often doesn't give rates with no conversions any value. In our view this is a huge misstep because more often than not conversion rates for different search engine campaigns will have a zero conversion rate. These rates still contain information that is valuable for decision making. If you have two zero conversion rates with one that has received 1000 clicks versus one that has 10 clicks you should be able to tell that adjustments need to be made with the first campaign before the second.

The methods we have developed resolve these issues and present the user with useful ways of detecting these differences without violating any test assumptions.

2 Objectives and Constraints

The design of conversion rate quality scores sought to meet specific criteria, namely:

- Simple - easy to use, run, and replicate.
- Interpretable - it shouldn't take a data scientist or statistician to explain the score to the client.
- Scaleable - the method shouldn't depend on how many conversion rates are being compared at a given time and should work regardless of what values are measured.
- Theoretically sound - mathematical properties and assumptions should be reasonably met.
- Reliable - should handle special conversion rate cases well, i.e. (0% and 100%)
- Perceptible - should be able to detect small differences in conversion rate parameters to produce different scores.
- Time independence - While controlling for time is important when comparing these rates, it shouldn't be a factor in the measures of quality.

Overall the methods we reference and introduce should be parts of a whole picture of conversion rate quality. We want to inform the business or client with these scores in an objective manner rather than letting the scores make all the decisions. In the following sections we will show how we have met each of these objectives.

3 Statistical Framework

3.1 Random Variables

A random variable is a variable that takes on values determined by a random process. Random variables are denoted by a capital letter, say X, to indicate that the value of X is unknown but can take on a range of values that follow some probability distribution. Once this random process has taken place and a value has been determined we denote this realized value by x. Consider flipping a coin, define the random variable X to be 0 if the coin lands tails and 1 if heads. If the coin is flipped and lands on heads then we would write $x = 1$. If the coin lands on tails we would write $x = 0$. This notation is used throughout this document to distinguish random variables from realized values. The joining of a random variable's possible outcomes with their respective likelihoods is constructing the probability distribution of the random variable.

3.2 Discrete Random Variables

The first step in identifying a proper statistical method to apply is to consider what type of random variable we are trying to measure. Ideal methods for conversion rate quality wouldn't just score based on the conversion rate but rather the number of clicks and conversions relative to the average conversion rate. Throughout this document we will denote conversions by X, clicks by Y, and the conversion rate by $R = X/Y$. We also denote the average conversion rate by p The constraints on these variables are as follows,

$$\begin{aligned} &X \in \{0, 1, 2, ...\}, \;\; 0 \leq X \leq Y \\ &Y \in \{0, 1, 2, ...\}, \;\; X \leq Y \\ &R \in [0, 1]. \end{aligned}$$

Here we are showing that X, Y are discrete integer valued random variables. R is a continuous random variable that can only take on values between 0 and 1 inclusive. This is important to identify because it limits what types of distributions can be used to measure possible outcomes.

3.3 Conditional Probability Distributions

Conditional probability distributions are important in the study of conversion rates. There is some dependence on the number of clicks with the number of conversions that would indicate that information about one or the other would inform the distribution of the other. For example, if the number of clicks of a conversion rate are known but the conversions are unknown one could reason out what they might be based on other conversion rates with a similar amount of clicks. The scores we have developed rely on conditional distributions because we assume one of the variables X or Y is known and the other is a random quantity. If we want the distribution of conversions given clicks we write the probability that X takes on a certain value x as

$$P(X = x|y),$$

or similarly if we want clicks given conversions

$$P(Y = y|x).$$

The vertical line together with the lower case letter indicates a conditional probability distribution with the known quantity to the right and the random quantity to the left.

3.4 Bernoulli Trials

A Bernoulli trial is a random experiment with exactly two possible outcomes. Each trial has the same probability of success. When we take a collection of these trials for an experiment the distribution of possible values will change overall but each trial individually will still be one Bernoulli trial. Whether someone clicks on a search engine ad or not can be thought of as a simple Bernoulli trial with a relatively consistent probability of success. Whether someone converts after that click or not can also be thought of as a Bernoulli trial with a consistent probability of success. We assume that conversion rates contain many Bernoulli trials which allows us to use the Binomial and Negative Binomial distribution.

3.5 Binomial Distribution

The following notation is used to make formulas more concise:

- $X, x =$ conversions
- $p =$ average conversion rate
- $Y, y =$ clicks
- $y - x =$ number of clicks that failed to convert
- $q = 1 - p =$ probability of a click failing to convert
- $i =$ index used to denote a row number in excel

The binomial distribution is another discrete probability distribution of the number of successes in a sequence of y independent experiments each resulting in success or failure.

Suppose there is a sequence of independent coin tosses (Bernoulli trials). Each trial or coin toss has two potential outcomes called "success" (1) and "failure" (0). In each trial the probability of success is p and the probability of failure is $(1 - p)$. The number of successes X we see in these independent trials will have a binomial distributions.

The probability mass function f of the binomial distribution in the context of conversion rates is written as

$$f(x; y, p) = P(X = x) = \binom{y}{x} p^x (1 - p)^{y-x}.$$

The mean or expected value of this distribution is

$$\mathrm{E}[X] = yp,$$

which says the number of expected conversions is found by taking the average conversion rate and multiplying it by the number of clicks received.

The variance of this distribution is

$$\mathrm{Var}(X) = yp(1 - p),$$

which gives a standard or expected deviation from the mean

$$\mathrm{SD} = \sqrt{\mathrm{Var}(X)} = \sqrt{yp(1 - p)}.$$

3.6 Negative Binomial Distribution

One of the quality scores we introduce relies on the negative binomial distribution. The negative binomial distribution is a discrete probability distribution that measures the number of trials in a sequence of independent Bernoulli trials before a certain number of successes occur.

Suppose there is a sequence of independent coin tosses (Bernoulli trials). Each trial or coin toss has two potential outcomes called "success" (1) and "failure" (0). In each trial the probability of success is p and the probability of failure is $(1-p)$. The random number of trials we have seen, Y, will have a negative binomial distribution.

The probability mass function f of the negative binomial distribution is written as

$$f(y;x,p) = P(Y = y|x) = \binom{y-1}{x-1} p^x (1-p)^{y-x}.$$

The mean or expected value or expected total number of trials until x conversions of this distribution is

$$\mathrm{E}[X] = x/p$$

which is the average number of clicks it takes to get x conversions.

The variance of this distribution is

$$\mathrm{Var}(X) = x(1-p)/p^2,$$

which gives a standard or expected deviation from the mean

$$\mathrm{SD} = \sqrt{\mathrm{Var}(X)} = \sqrt{x(1-p)/p^2}.$$

3.7 Assumptions

The main assumption we are making is that each Bernoulli trial of a conversion or no conversion is independent of the previous. We also assume that the average conversion rate is robust enough to be used as the probability of one click resulting in a conversion. These two assumptions allow us to construct the quality scores in a theoretically sound manner and maintain simplicity. Each score also assumes that one of the conversion rate variables X or Y is known and the other is random. The non-random variable is assigned as a parameter of either the negative binomial or binomial distributions.

4 Next Conversion Spread (NC Spread) Score

NC Spread is a way of checking how far (or close) a conversion rate would fall from the expectation if another conversion were to happen on the next click. We would like conversions to happen sooner as the number of clicks increase. Consider a conversion rate with one conversion and fifty clicks when the known average conversion rate is 0.71%. Our current conversion rate is $1/50 = 0.02$ or 2%. Immediately we would consider this a better than average conversion rate. But the Negative Binomial Distribution gives an expectation of $1/0.0071 = 140.845$ meaning we would expect on average to wait 140 clicks before one conversion. Thus the current conversion rate is giving much better results than average so naturally we should somehow score this higher. However, one standard deviation of this distribution is also 140 ($\sqrt{(1-0.0071)/0.0071^2}$) clicks. This tells us that the current rate is actually not much better than average given the typical volatility of these variables. This should also be taken into account in a scoring method.

One issue we run into with the negative binomial distribution is that the parameter x that represents conversions can't be zero. To solve this, we always increment the number of conversions by one to score it

based on a hypothetical next conversion. This gives a forward looking behavior to the metric rather than having it depend only on past performance and allows us to use the proper distribution for these variables. We also assume that the number of conversions x is a given parameter of the negative binomial rather than a random variable.

Definition 1 *The NC Spread of a conversion rate with the number of conversions denoted by X and number of clicks denoted by Y with probability of conversion (average conversion rate) p is the ratio of deviation from the mean to standard deviation of the next conversion, more formally*

$$NC_{spread} = (-1)\frac{(Y + 1 - (X + 1)/p)}{\sqrt{(X + 1)(q/p^2)}}$$

Note that the (-1) is included to flip the sign of the output score to make the positive scores indicate better performance than negative scores.

To simplify NC_{spread}, recall the definitions of the negative binomial distribution. The quantity

$$(X + 1)/p = \mu$$

is the mean of the negative binomial or mean number of clicks required to obtain $X + 1$ conversions assuming a conversion on the next click. The quantity

$$\sqrt{(X + 1)(q/p^2)} = \sigma$$

is the standard deviation from the mean of the negative binomial assuming a conversion on the next click.

4.1 Interpretation

The NC Spread Score is meant to tell you if you have a more favorable conversion rate when compared to the average. The NC Spread will typically range between 3 and -3. The interpretation will fall into four categories:

- $1 < NC_{\text{spread}} < \infty$: conversions are happening much faster than expected
- $0 \leq NC_{\text{spread}} \leq 1$: conversions are happening about as fast as expected but slightly faster
- $-1 \leq NC_{\text{spread}} < 0$: conversions are occurring about as fast as expected but slightly slower
- $-\infty < NC_{\text{spread}} < -1$: conversions are happening much slower than expected

This information is useful but it doesn't show of how good a conversion rate is for bidding adjustment decisions. Consider two conversion rates, $R_1 = 1/134$ and $R_2 = 0/29$. R_1 should show to be more robust than the second rate because it has more clicks but the NC Spread is 0.74 and 0.796 respectively, assuming an average conversion rate of 0.0071. All the NC Spread is telling you is that if each of these were to convert on the next click both rates would be converting slightly faster than expected but more so with the second rate. This is NOT telling you that you are 5% more confident in the second rate. In the coming sections we introduce another score that shows degree of confidence to use in conjunction with NC Spread for an overall perspective on how to judge rates in this manner.

4.2 Expected Click Confidence Interval

To construct a simple interval for the expected number of clicks required for $x + 1$ conversions we can take the mean and add/subtract one standard deviation.

$$\left(\text{MAX}\left(0,\ (x+1)/p - \sqrt{(x+1)(1-p)/p^2}\right),\ \ (x+1)/p + \sqrt{(x+1)(1-p)/p^2}\right)$$

There are other more complicated methods for calculating a more precise confidence interval for this distribution but given the context this will perform well enough. This is a one standard deviation confidence interval and can be interpreted as there is better than a 50% chance that the true number of clicks required for $x+1$ conversions is within this interval. If the current number of clicks lies far below this interval then it is indicating as with NC Spread that your conversion rate is performing very well compared to average.

4.3 Click Stopping Time (CST)

The number of clicks required to wait before seeing a certain change in conversion rate is a simple calculation that only involves the mean of the negative binomial distribution. Recall that to find the expected number of clicks for one conversion is found by $1/p$. We can abstract this to a situation when we want to calculate the amount of clicks to wait for a certain number of conversions based on how much we want to see the conversion rate change for a given keyword.

First let p_i denote the conversion rate for the i^{th} keyword, $0 \leq i \leq$ Keyword Count. Also let Δ represent the change in conversion rate we want to see.

The number of clicks to wait for the given change is calculated as

$$\text{CST}_i = \left\lceil \frac{(p_i + \Delta)Y}{p_i} \right\rceil.$$

Note here that $\lceil c \rceil$ denotes the ceiling function which takes a number c and rounds up to the nearest integer. One key property here is that $\text{CST} > Y$ because $(p_i + \Delta)/p_i > 1$.

Consider a conversion rate $R = 13/730 = 0.0178$ with which we would like to see a 1% lift. Perhaps there are changes being considered to this current ad campaign or keyword and we want to see if those changes cause this 1% change. We set $p_i = 0.0178$, $\Delta = 0.01$, and $Y = 730$. Then we calculate the number of clicks to stop at

$$\text{CST} = \left\lceil \frac{(0.027) * 730}{0.017} \right\rceil = 1141$$

When we take the conversions we had (13) and the conversions we need ($\lceil (0.027) * 730 \rceil = 20$) and calculate the new conversion rate $R = 33/1160 \approx 0.0289$. Thus we should wait until the number of clicks has reached 1141 to see a 1% lift in the conversion rate and if this desired change is not met decisions can be made with more confidence.

The only edge case is when $x_i = 0$ (no conversion yet) which implies that $p_i = 0$. This can be easily remedied by setting $p_i = p$, or the average conversion rate. This is included as a case statement in the implementation.

4.3.1 CST Interval

A simple interval around the number of clicks to wait can be easily calculated using the new mean and standard deviation.

$$\left(\text{MAX}\left(0,\ \text{CST}_i - \sqrt{(1-p_i)\text{CST}_i/p_i}\right),\ \ \text{CST}_i + \sqrt{(1-p_i)\text{CST}_i/p_i}\right)$$

This one-standard-deviation confidence interval would give an upper and lower bound for the typical number of clicks that would be expected for a certain change to occur.

5 Conversion Rate Confidence Score (CRC)

Conversion Rate Confidence Score (CRC) looks at the probability of obtaining the conversion rate you arrived at given your clicks, conversions, and average conversion rate. This score makes no assumptions on future conversions and only looks at the historical quality of a given conversion rate. The value produced by CRC can be thought of similar to a p-value. A p-value in a typical hypothesis or A/B test is the probability of obtaining the sample statistic you obtained or something more extreme. Given your data, if your outcome is unlikely based on your null hypothesis or initial guess then this is evidence in favor of the alternative hypothesis or contradiction to your initial guess. We base this score off of the binomial distribution because the random number of successes X given the number of trials y fits the definition of a binomial random variable better than a negative binomial.

Definition 2 *The Conversion Rate Confidence Score of a conversion rate with conversions X, clicks y, and probability of conversion (average conversion rate) p is the compliment of the probability of obtaining this value along the binomial distribution. More formally*

$$CRC = P(X \neq x|y) = 1 - P(X = x|y) = 1 - \binom{y}{x} p^x (1-p)^{y-x}.$$

Note that we use y instead of Y becasue we are treating the number of clicks as a parameter of the binomial distribution rather than a random variable.

5.1 Interpretation

The CRC Score will always give a score between 0 and 1. The basic interpretation is what is the probability of this conversion rate not occurring given the information we have. If this probability is high we can say that we are confident we have a more robust conversion rate where adjustments can be made. When the probability is low we can interpret that this particular conversion rate is more likely and is perhaps not robust enough to make adjustments to. CRC works in conjunction with NC Spread to give two different perspectives. NC Spread could be above 1 but CRC could give a low number indicating that although a particular conversion rate is doing well above average there is not enough information in the conversion rate to make adjustments.

Consider the same conversion rates used previously, $R_1 = 1/134$ and $R_2 = 0/29$. Recall that the NC Spread showed that R_1 is closer to the expectation than R_2. CRC for these rates is 0.63 and 0.18 respectively. Now we have a better idea of what is happening with both of these rates relative to the average. CRC says that although the NC Spread for R_2 was higher we are 45% less confident in this conversion rate than R_1.

6 Excel Implementation

The excel formulas for each metric presented are given below. Note that each variable is a specific cell reference in excel.

- X: Conversions
- Y: Clicks
- p: Average Conversion Rate
- p_i: Conversion rate for keyword i
- Δ: Desired change in conversion rate

6.1 NC Spread

$$= -(Y + 1 - (X + 1)/p)/\text{SQRT}((X + 1) * (1 - p)/(p^2))$$

6.2 Expected Click Interval

Upper:

$$= (X + 1)/p + \text{SQRT}((X + 1) * (1 - p)/(p^2))$$

Lower:

$$= \text{MAX}\left((X + 1)/p - \text{SQRT}((X + 1) * (1 - p)/(p^2)), 0\right)$$

6.3 CRC

$$= 1 - \text{BINOMDIST}(X, Y, p, \text{FALSE})$$

6.4 CST

$$= \text{IF}(X_i = 0, \text{CEILING}\left((p + \Delta) * Y_i/p, 1\right),\ \text{CEILING}\left((p_i + \Delta) * Y_i/p_i, 1\right))$$

6.5 CST Interval

Upper:

$$= \text{IF}(X_i = 0, \text{MAXCST}_i + \text{SQRT}\left((1 - p) * \text{CST}_i/p\right),\ \text{CST}_i + \text{SQRT}\left((1 - p_i) * \text{CST}_i/p_i\right))$$

Lower:

$$= \text{IF}\left(X_i = 0, \text{MAX}\left(\text{CST}_i - \text{SQRT}\left((1 - p)\text{CST}_i/p\right), 0\right),\ \text{MAX}\left(\text{CST}_i - \text{SQRT}\left((1 - p_i)\text{CST}_i/p_i\right), 0\right)\right)$$

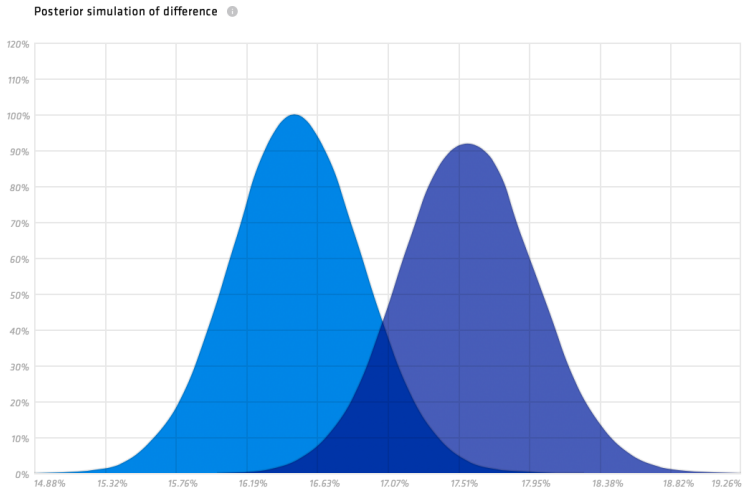
7 Bayesian A/B Testing

An area of growing popularity in A/B testing revolves around Bayesian statistics. Bayesian statistics assumes some prior knowledge about a variable and then accumulates more knowledge about the true distribution of a variable as more information comes in. A Bayesian framework tends to have a more intuitive and interpretable approach that generates more reliable information.

A really great online Bayesian conversion rate test tool by Dynamic Yield is available at the link in source [2]. There are also great articles linked on their page that go into more detail than we want to here about how these tests and metrics work. The differences between the two methodologies of A/B testing are summarized well by the table below.

	Hypothesis Testing	Bayesian A/B Testing
Knowledge of Baseline Performance	Required	Not Required
Intuitiveness	Less, as p-value is a convoluted term	More, as we directly calculate the probability of A being better than B
Sample size	Pre-defined	No need to pre-define
Peeking at the data while the test runs	Not allowed	Allowed (with caution)
Quick to make decisions	Less, as it has more restrictive assumptions on distributions	More, as it has less restrictive assumptions
Representing uncertainty	Confidence Interval (again, a convoluted interpretation which is often misunderstood)	Highest Posterior Density Region – highly intuitive interpretation
Declaring a winner	When sample size is reached and p-value is below a certain threshold	When either "probability to be best" goes above a threshold or the expected loss is below a threshold (in which case a "tie" can be declared between multiple variations)

This specific method of A/B testing provides a "Probability to be Best" which determines a conversion rate as a winner after the probability passes say 95%. They also provide an "Expected Loss" metric which basically says how much of this conversion rate will be lost in the long run if it is declared as a winner. So if the loser is chosen instead then the conversion rate is expected to decrease by "Expected Loss" amount. The Posterior simulation of difference graphic shows the distribution for each conversion rate that is calculated using prior knowledge about conversion rates together with what has been collected so far. A taller more narrow peak indicates that the true distribution of conversion rate is being achieved. The further apart these two distributions are the more of a clear difference is found between them. The two example conversion rates on the site give the following posterior simulations based on 8,500 samples.



These two distrubutions are both fairly symmetric which indicates that each conversion rate found so far is converging to their true average conversion rate. The purple distribution centers around an average that is higher than the blue distribution which is more evidence to suggest that the campaign in purple is truly better at converting.

8 Conclusion

Conversion rates are used to make many meaningful business decisions and should be analyzed in ways that are not detrimental to this process. These methods introduced and cited should be used to gain a better perspective of which conversion rates are in need of adjustment around strategy. They are simple ways of arriving at conclusions that are based in solid statistical theory, reliable for use at any time, scaleable to be used across all conversion rates or a few, and they detect small differences in conversion rates better than any A/B testing tool could. The metrics are not meant to be used to say one conversion rate is the "Best" or a "Winner" like the online Bayesian A/B testing tool.

9 Sources

[1] Key Properties of a Negative Binomial Random Variable. (n.d.). Retrieved from https://newonlinecourses.science.psu.edu/stat414/node/79/

[2] Bayesian A/B Testing Calculator. (n.d.). Retrieved from https://marketing.dynamicyield.com/bayesian-calculator/?_ga=2.101583233.1428414581.1564956547-1932974290.1564956547

[3] Michaeli, I. (2019, June 28). Going Bayesian: A Fresh Approach to A/B Testing. Retrieved from https://www.dynamicyield.com/blog/bayesian-testing/?_ga=2.134040718.1428414581.1564956547-1932974290.1564956547

[4] E. M. (n.d.). How Not To Run an A/B Test. Retrieved from http://www.evanmiller.org/how-not-to-run-an-ab-test.html