17

The Statistics behind Online Controlled Experiments

Smoking is one of the leading causes of statistics
– *Fletcher Knebel*

***Why you care**: Statistics are fundamental to designing and analyzing experiments.*

We introduced several statistical concepts. This chapter goes deeper on the Statistics critical to experimentation, including hypothesis testing and statistical power (Lehmann and Romano 2005, Casella and Berger 2001, Kohavi, Longbotham et al. 2009).

Two-Sample t-Test

Two-sample t-tests are the most common statistical significance tests for determining whether the difference we see between Treatment and Control is real or just noise (Student 1908; Wasserman 2004). Two-sample t-tests look at the size of the difference between the two means relative to the variance. The significance of the difference is represented by the p-value. The lower the p-value, the stronger the evidence that the Treatment is different from the Control.

To apply the two-sample t-test to a metric of interest Y (e.g., queries-per-user), assume that the observed values of the metric for users in the Treatment and Control are independent realizations of random variables, Y^t and Y^c. The Null hypothesis is that Y^t and Y^c have the same mean; the alternative hypothesis is that they do not (see Equation 17.1):

$$H_0 : mean(Y^t) = mean(Y^c)$$
$$H_A : mean(Y^t) \neq mean(Y^c) \tag{17.1}$$

The two-sample t-test is based on the t-statistic, T:

$$T = \frac{\Delta}{\sqrt{var(\Delta)}} \tag{17.2}$$

where, $\Delta = \overline{Y^t} - \overline{Y^c}$ is the difference between the Treatment average and the Control average, an unbiased estimator for the shift of the mean. Because the samples are independent:

$$var(\Delta) = var\left(\overline{Y^t} - \overline{Y^c}\right) = var\left(\overline{Y^t}\right) + var\left(\overline{Y^c}\right) \tag{17.3}$$

The t-statistic T is just a normalized version of Δ.

Intuitively, the larger the T, the less likely it is that the means are the same. In other words, you are more likely to reject the Null hypothesis. How do we quantify this?

p-Value and Confidence Interval

Now that you have the t-statistic T, you can compute the p-value, which is the probability that T would be at least this extreme if there really is no difference between Treatment and Control. By convention, any difference with a p-value smaller than 0.05 is considered "statistically significant," though there are ongoing debates calling for lower p-values by default (Benjamin et al. 2017). A p-value less than 0.01 is considered very significant.

Even though p-value is one of the most well-known statistical terms, it is often misinterpreted. One common misinterpretation is that the p-value captures the probability that the Null hypothesis is true given the data observed. This is a reasonable interpretation on the surface as most experimenters would expect to get a probability on whether their Treatment has impact. However, the correct interpretation is almost the opposite, which is the probability of observing the delta, or a more extreme delta, if the Null hypothesis is true. To see how these two interpretations are different yet related, you can break it down using Bayes rule:

$$\begin{aligned} P(H_0 \text{ is true } | \Delta \text{ observed}) &= \frac{P(\Delta \text{ observed } | H_0 \text{ is true}) P(H_0 \text{ is true})}{P(\Delta \text{ observed})} \\ &= \frac{P(H_0 \text{ is true})}{P(\Delta \text{ observed})} * P(\Delta \text{ observed } | H_0 \text{ is true}) \\ &= \frac{P(H_0 \text{ is true})}{P(\Delta \text{ observed})} * pvalue \end{aligned} \tag{17.4}$$

As indicated in the equation, to know whether the Null hypothesis is true based on data collected (posterior probability), you not only need a p-value but also the likelihood that the Null hypothesis is true.

Another way to examine whether the delta is statistically significant is to check whether the confidence interval overlaps with zero. Some people find confidence intervals a more intuitive way to interpret the noise and uncertainty around the observed delta than the p-value. A 95% confidence interval is the range that covers the true difference 95% of the time and has an equivalence to a p-value of 0.05; the delta is statistically significant at 0.05 significance level if the 95% confidence interval does not contain zero or if the p-value is less than 0.05. In most cases, the confidence interval for the delta centers around the observed delta with an extension of about two standard deviations on each side. This is true for any statistics that (approximately) follow the normal distribution, including the percent delta.

Normality Assumption

In most cases we compute p-values with the assumption that the t-statistic T follows a normal distribution, and under the Null hypothesis the distribution has a mean 0 and variance 1. The p-value is just the area under the normal curve, as highlighted in Figure 2.1 in Chapter 2. Many people misinterpret the normality assumption to be an assumption on the sample distribution of the metric Y, and consider it a poor assumption because almost none of the metrics used in practice follow a normal distribution. However, in most online experiments the sample sizes for both Control and Treatment are at least in the thousands. While the sample distribution of Y does not follow normal distribution, the average $\bar{Y}$ usually does because of the *Central Limit Theorem* (Billingsly 1995). Figure 17.1 illustrates the convergence with samples Y drawn from a beta distribution. As the sample size increases, the distribution of the mean $\bar{Y}$ becomes more normally distributed.

One rule-of-thumb for the minimum number of samples needed for the average $\bar{Y}$ to have normal distribution is $355s^2$ for each variant (Kohavi, Deng et al 2014), where s is the skewness coefficient of the sample distribution of the metric Y defined as in Equation 17.5:

$$s = \frac{E[Y - E(Y)]^3}{[Var(Y)]^{3/2}}. \tag{17.5}$$

Some metrics, especially revenue metrics, tend to have a high skewness coefficient. One effective way to reduce skewness is to transform the metric or

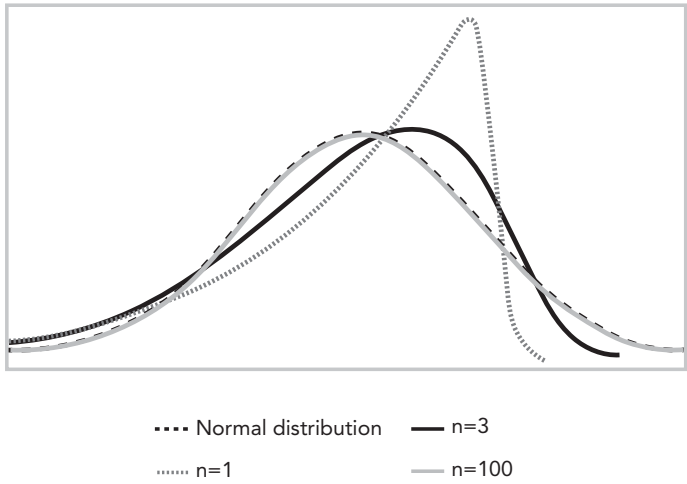

Figure 17.1 Distribution of the mean becomes increasingly normal as the sample size n increases

cap the values. For example, after Bing capped Revenue/User to $10 per user per week, they saw skewness drop from 18 to 5, and the minimum sample needed drop tenfold from 114k to 10k. This rule-of-thumb provides good guidance for when $|s| > 1$ but does not offer a useful lower bound when the distribution is symmetric or has small skewness. On the other hand, it is generally true that fewer samples are needed when skewness is smaller (Tyurin 2009).

For two-sample t-tests, because you are looking at the *difference* of the two variables with similar distributions, the number of samples needed for the normality assumption to be plausible tends to be fewer. This is especially the case if Treatment and Control have equal traffic allocations (Kohavi, Deng et al 2014), as the distribution of the difference is approximately symmetric (it is perfectly symmetric with zero skewness under the Null hypothesis).

If you ever wonder whether your sample size is large enough to assume normality, test it at least once with offline simulation. You can randomly shuffle samples across Treatment and Control to generate the null distribution and compare that distribution with the normal curve using statistical tests such as Kolmogorov–Smirnov and Anderson-Darling (Razali and Wah 2011). As the tail distribution is of interest in hypothesis testing, you can also increase test sensitivity by only focusing on whether the Type I error rate is bounded by the preset threshold, for example, 0.05.

When the normality assumption fails, you can then do a permutation test (Efron and Tibshriani 1994) and see where your observation stands relative to

the simulated null distribution. Note that even though a permutation test is very expensive to run at scale, occasions when it is needed are often with small sample sizes, so it works out nicely in practice.

Type I/II Errors and Power

With any test there are errors. In hypothesis testing, we care about Type I and Type II errors. A Type I error is concluding that there is a significant difference between Treatment and Control when there is no real difference. A Type II error is when we conclude that there is no significant difference when there really is one. You control Type I error rates at 0.05 by concluding statistical significance only if the p-value < 0.05. Clearly, there is a tradeoff between these two errors. Using a higher p-value threshold means a higher Type I error rate but a smaller chance of missing a real difference, therefore a lower Type II error rate.

The concept of Type II errors is better known as *power*. Power is the probability of detecting a difference between the variants , that is, rejecting the null, when there really is a difference(see Equation 17.6):

$$\text{Power} = 1 - \text{Type II error} \tag{17.6}$$

Power is typically parameterized by delta, δ, the minimum delta of practical interest. Mathematically, assuming the desired confidence level is 95%, the equation is as in Equation 17.7:

$$\text{Power}_{\delta} = P(|T| \geq 1.96 \,|\text{true diff is } \delta). \tag{17.7}$$

The industry standard is to achieve at least 80% power in our tests. Therefore, it is common to conduct power analysis before starting the experiment to decide how many samples are needed to achieve sufficient power. Assuming Treatment and Control are of equal size, the total number of samples you need to achieve 80% power can be derived from the power formula above, and is approximately as shown in Equation 17.8 (van Belle 2008):

$$n \approx \frac{16\sigma^2}{\delta^2} \tag{17.8}$$

where, σ^2 is the sample variance, and δ is the difference between Treatment and Control. A common question people ask is that how would they know δ before they run the experiment? It is true that we do not know the true δ and that is the reason to run the experiment to begin with.

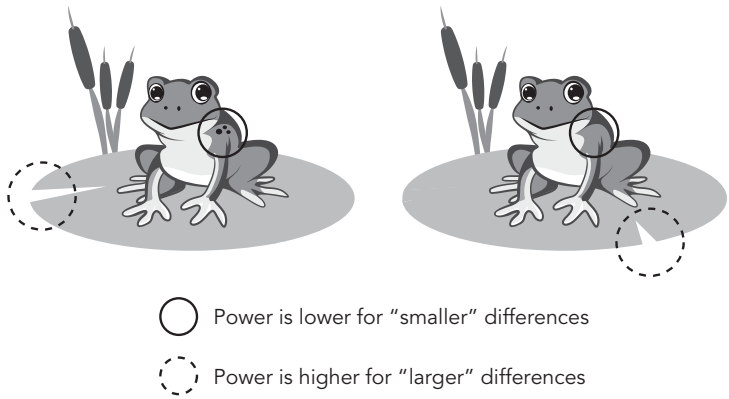

Figure 17.2 Analogy of statistical power with the game "Spot the difference." Power is higher for detecting a larger difference

However, we know the size of δ that would matter in practice, in other words, that of *practical* significance. For example, you could miss detecting a difference of 0.1% in revenue and that's fine, but a drop of 1% revenue is not fine. In this case, 0.1% is not practically significant while 1% is. To estimate the required minimum sample size, use the smallest δ that is practically significant (also called the *minimum detectable effect*).

For online experiments, sample size estimation is more complex because online users visit over time, so the duration of the experiment also plays a role in the actual sample size of an experiment. Depending on the randomization unit, the sample variance σ^2 can also change over time. Another challenge is that with triggered analysis (see Chapter 20), the values σ^2 and δ change as the trigger conditions change across experiments. For these reasons, we present a more practical approach in Chapter 15 for deciding traffic allocation and the duration for most online experiments.

We want to highlight a common misinterpretation of the concept of statistical power. Many people consider power an absolute property of a test and forget that it is relative to the size of the effect you want to detect. An experiment that has enough power to detect a 10% difference does not necessarily have enough power to detect a 1% difference. A good analogy is the game "spot the difference." Figure 17.2 demonstrates that relative to difference in the spots (solid circle), it is easier to detect the difference on the lily pads (dashed circle) as it is a larger difference.

As you can tell, power analysis is deeply coupled with Type I and II errors. Gelman and Carlin (2014) argue that for small sample size settings, it is also

important to calculate a) the probability of an estimate being in the wrong direction (Type S [sign] error), and b) the factor by which the magnitude of an effect might be overestimated (Type M [magnitude] error or exaggeration ratio).

Bias

In experiment results, bias arises when the estimate and the true value of the mean are systematically different. It can be caused by a platform bug, a flawed experiment design, or an unrepresentative sample such as company employee or test accounts. We discuss several examples and recommendations for prevention and detection in Chapter 3.

Multiple Testing

With hundreds of metrics computed for each experiment, we commonly hear from experimenters "Why is this irrelevant metric significant?" Here is a simplified way to look at it. If you compute 100 metrics for your experiment, how many metrics would you see as statistically significant even if your feature does nothing? With the significance level at 5%, the answer is around five (assuming that the metrics are independent). The problem worsens when examining hundreds of experiments and multiple iterations per experiment. When testing multiple things in parallel, the number of false discoveries increases. This is called the "multiple testing" problem.

How can we ensure that Type I and Type II errors are still reasonably controlled under multiple testing? There are many well studied approaches; however, most approaches are either simple but too conservative, or complex and hence less accessible. For example, the popular Bonferroni correction, which uses a consistent but much smaller p-value threshold (0.05 divided by the number of tests), falls into the former category. The Benjamini-Hochberg procedure (Hochberg and Benjamini 1995) uses varying p-value thresholds for different tests and it falls into the latter category.

So, what should you do when a metric is unexpectedly significant? Here's a simple two-step rule-of-thumb:

1. Separate all metrics into three groups:
   - First-order metrics: those you expect to be impacted by the experiment
   - Second-order metrics: those potentially to be impacted (e.g., through cannibalization)

- Third-order metrics: those unlikely to be impacted.

2. Apply tiered significance levels to each group (e.g., 0.05, 0.01 and 0.001 respectively).

These rules-of-thumb are based on an interesting Bayesian interpretation: How much do you believe the Null hypothesis (H_0) is true before you even run the experiment? The stronger the belief, the lower the significance level you should use.

Fisher's Meta-analysis

We discuss how to identify patterns, create and utilize institutional memories based on meta-analysis on historical experiments in Chapter 8. In this section, we are particularly interested in combining results from multiple experiments that test on the *same* hypothesis. For example, it is a common technique to replicate an experiment that had surprising results. Replication is done using either orthogonal randomization or users who were not allocated to the original round of the experiment. These two experiments, the original and the replication, both produce p-values independent of each other. Intuitively, if both p-values are less than 0.05, that's stronger evidence that the Treatment has an impact than if only one p-value is less than 0.05. Fisher formalizes this intuition in his meta-analysis method (Fisher 1925), saying that we can combine p-values from multiple independent statistical tests into one test statistic as shown in Equation 17.9:

$$X_{2k}^2 = -2\sum_{i=1}^{k} ln\,(p_i) \tag{17.9}$$

where p_i is the p-value for the ith hypothesis test. If all k Null hypothesis are true, this test statistic follows a chi-squared distribution with $2k$ degrees of freedom. Brown (1975) extends Fisher's method to cases when the p-values are not independent. There are other p-value combination methods, such as Edgington (1972), Volumne 80 (2) and Mudholkar and George (1979). See Hedges and Olkin (2014) for more discussions.

In general, Fisher's method (or any other meta-analysis technique) is great for increasing power and reducing false-positives. You may have an experiment that is underpowered even after applying all power-increasing techniques, such as maximum power traffic allocation (see Chapter 15) and variance reduction (see Chapter 22). In this case, you can consider two or more (orthogonal) replications of the same experiment (one after another) and achieve higher power by combining the results using Fisher's method.