A/B Testing Using the Negative Binomial Distribution in an Internet Search Application

Saharon Rosset and Slava Borodovsky

Tel Aviv University

Saharon@post.tau.ac.il

A/B testing plays an important role in modern applications, particularly on the internet, as it helps businesses to optimize their user experience to maximize usage and profits. In this paper we discuss A/B testing on counts (such as number of searches by a user), and point out the importance of using appropriate distributions for statistical analysis. We discuss the use of the negative binomial (NB) distribution to evaluate performance instead of the commonly used Poisson distribution. Our motivating application is in A/B testing on number of searches of users of an internet search engine made by the company SweetIM. We demonstrate the inappropriateness of the standard Poisson assumption for these data, and show that the conclusions from analyses of specific A/A and A/B tests run in this application with NB differ from those with an incorrect Poisson assumption. Using a normal approximation, we describe a general property of NB tests – the existence of a bound on testing power, which is independent of mean expected usage (or length of time running the test) under typical assumptions. This leads to a disconcerting conclusion that such tests cannot guarantee high statistical power for identifying a small difference between the A and B groups, no matter how long they are run. This is in sharp contrast to "standard" tests with binomial, normal or Poisson assumptions, where any desired power can be attained by running the test long enough, as long as the A and B groups differ. We also describe and apply a permutation test as a non-parametric approach for testing. In our view, the non-parametric approach is an important complement to parametric tests like NB inference, because it is valid for testing the most general null hypotheses of equality between A and B distributions, without assuming anything about the form of these distributions.

Keywords: A/B testing, negative binomial, internet, hypothesis testing.

Introduction

Controlled experiments, also known as A/B testing, are an important general framework for assessing the impact of changes on performance. Controlled experiments have a long history in medicine, but in recent years A/B testing is increasingly employed in e-commerce, marketing and other consumer-centric domains (Ash 2008; Keppel et al. 1997).

In particular, in applications where a decision maker can control the experience each user or customer encounters, it often makes sense to try out changes in user

experience by presenting some (randomly chosen) subset of users with the new "B" experience, while others still encounter the old "A" experience. The vendor can then objectively examine the effects of the change in user experience on user behavior, in buying or usage patterns. This mode of experimentation is especially common on the internet, and is widely used by online vendors and service providers to examine effectiveness of their platforms (Ash 2008; Kohavi et al. 2007, Tang et al. 2010).

A key component of A/B testing planning, execution and usage is obviously statistical analysis and inference (Box et al. 2005; Myers and Well 2003). In planning, this can be used to determine how many users should be taken, how long the experiment should run, etc. Once data are gathered, statistical analysis is used to evaluate the results and examine whether the effects of the "A" and "B" experiences differ. Proper statistical inference is widely acknowledged as a critical and often non-trivial aspect of A/B testing (Crook et al. 2009).

Many A/B tests examine binary outcomes (like buy/don't buy), and typically use binomial distributions for statistical inference. Others examine continuous traits (like time spent on a site) and most often use normal distributions, though sometimes other distributions are more appropriate (Andersen 1997). In this paper we are concerned with A/B tests on counts (like number of searches by a user) and the appropriate statistical approaches for analyzing them.

The most commonly used distribution in analyzing counts – in A/B tests as well as other applications – is the Poisson distribution (Andersen, 1997; Johnson et al. 2004; McCullagh and Nelder 1989). Assuming the number of actions by a random user in the A or B groups is Poisson distributed is equivalent to assuming that actions arrive from all users at a constant rate in a "memoryless"' manner, i.e., that each user may act in the next instant with equal probability regardless of who acted when in the past. This is often a reasonable assumption, especially if all users are of the same "type". However if the users differ significantly (for example, if some are internet worms or bots that constantly act, while others are occasional users that rarely do), the Poisson assumption may well be inappropriate and lead to incorrect inference and conclusions. In this paper we demonstrate that this is indeed the case in our domain of interest. Some authors have advocated identifying and removing robots in A/B tests, but have also acknowledged that this can be difficult or impossible (Crook et al. 2009). Even if this is successfully done, users

may still differ significantly, for example heavy internet users compared to occasional users.

An alternative in the case that users differ significantly is to use the Negative Binomial distribution (henceforth NB, also known as over-dispersed Poisson) (Thall and Vail 1990). This is equivalent to assuming that each user still acts as a "memoryless" Poisson process, but their rates are now allowed to differ (see mathematical derivation in the next section). The NB distribution can be fitted to data, and used for inference, and this has been done extensively in many scientific and engineering domains, like genetics (Efron and Thisted 1976; Kimura 1994; Tamura and Nei 1993; Thall and Vail 1990), traffic modeling (Gerlough et al. 1971; Greenshields et al. 1978; McGee et al. 2003), marketing (Schmittlein et al., 1985), criminology (Berk and MacDonald, 2008) and others. Thus, this basic idea is well known and widely used. However, we are not aware of A/B testing methodology using the NB distribution, and to our knowledge some of the calculations we present here on power limitations against specific alternatives, have not been adequately addressed by previous work in other domains.

An interesting alternative that is commonly used in practice is to use a “t-test” for testing difference between populations, based on a normal approximation for mean usage in the population, with estimated variance from the data. As Aban et al. (2008) show, this approach is inferior to NB-based analysis (overly conservative, low power), when the NB distribution is appropriate. We discuss this issue further below.

Another approach that is very similar to NB is the quasi-Poisson model (McCulagh and Nelder, 1989), which takes slightly different assumptions, but typically leads to very similar conclusions in simple situations like those we face (Ver Hoef and Boveng, 2007).

Our motivating application in this paper comes from the actual A/B experimentation experience of the company SweetIM, which provides content and search services for instant messaging applications (IMs), Social Networks and Webmail. An important goal of these experiments is to identify changes that increase the number of searches of SweetIM users. Our analysis below shows that use of the Poisson distribution for planning and analyzing A/B tests on number of searches results in incorrect conclusions. For example, A/A tests, where the exact same experience is provided to both groups, often result in decisive rejection of the

"null hypothesis" that the groups do not differ, which should not happen by definition if correct inference is applied.

Our main contributions are:

- Demonstration that the NB distribution is indeed more appropriate for analysis of A/B tests on SweetIM data, with estimation of the appropriate NB parameters and comparison of actual test results with Poisson and NB assumptions. We show that NB inference leads to correct non-rejection of the null in A/A tests, and to some rejections in A/B tests examined. We also verify validity of our results by comparing them to assumption-free permutation tests.
- Statistical analysis of the implications of using NB in A/B testing. In particular, we discuss the statistical power of A/B tests (i.e., their ability to identify "real" differences) and its dependence on the duration of the tests, or equivalently the expected number of searches per user. Based on a normal approximation, we show that NB tests cannot guarantee arbitrary power by running the test longer, in contrast to other commonly used distributions like binomial, normal or Poisson. Instead, power in NB tests is bounded from above by a quantity that does not depend on the duration. The only way to increase the power further is to add users to the A and B groups. This has significant implications on the planning and execution of A/B tests in the setting where NB inference is appropriate. This problem is not purely academic, as expressed in SweetIM's own experience, where large usage differences of several percents are hard to assert with very large samples because of this power limitation. This problem is shared by giants running huge A/B tests like Google and Microsoft (Tang et al. 2010, Kohavi et al. 2007).

Statistical Inference for AB Tests

Poisson and Negative Binomial Distributions in AB Tests

A random variables X has a Poisson distribution, denoted $X \sim Pois(\mu)$, if $P(X=k) = e^{-\lambda} \mu^k / k!$, $k=0,1,2,\ldots$ X has mean and variance both equal to the Poisson parameter μ (Johnson et al. 2004).

When events are collected over time, it is common to assume that the Poisson rate is μ for a time unit (typically a day in A/B testing) and therefore if the experiment is run for t days, we get X~Pois(t·μ).

In an A/B testing situation, assume we have n users in the A group and m users in the B group, and denote by $x_1,x_2,\ldots,x_n \sim Pois(t\cdot\mu_A)$ i.i.d the number of searches in the A group and by $y_1,y_2,\ldots,y_m \sim Pois(t\cdot\mu_B)$ i.i.d the corresponding numbers in the B group. Then the A/B test for difference between the groups can be formally stated as testing a null of no difference in rates against the proper alternative (we'll assume a one sided alternative since we are only interested to learn whether B is better than A, i.e., leads to more searches):

$H_0: \mu_A=\mu_B$ vs $H_A: \mu_A<\mu_B$.

Our proposed approach for hypothesis testing on this hypothesis uses the difference in mean activity between the A and B groups as a test statistics, and calculates a p value using a normal approximation:

$$\bar{X}-\bar{Y} \sim N(t\mu_A - t\mu_B, t\mu_A/n + t\mu_B/m)$$

Where under H_0 the mean is 0 and the variance $t\mu_A\cdot (m+n)/(mn)$.

So the p-value for testing H_0 against H_A (one sided) is:

$$p_{Pois} = \Phi\left(\frac{\bar{X}-\bar{Y}}{(\mu t(n+m)/(nm))^{1/2}}\right);$$

where Φ is the standard normal cumulative distribution and $\mu=\mu_A=\mu_B$. In the common case that μ is not known it is replaced by $\hat{\mu}=(\sum X_i+\sum Y_i)/[t(m+n)]$, its pooled estimator.

An important aspect of planning an experiment is power calculations (Kendall and Stuart 1976): how do the number of users and the duration of the test affect our chances of detecting "real" differences between the A and B groups and rejecting the null hypothesis?

Given a known, hypothesized, or learned from past data value for μ_A (actions per day), the power to detect an increase of δ (i.e., $H_A: \mu_B = (1+\delta)\mu_A$) as a function of the time *t* and number of users in each group *n* (assume *n=m* for simplicity) is:

$$\pi(\mu_A, \delta, t, n) = P_{H_A}(p \le 0.05) = P_{H_A}\left[\left(\frac{\overline{X} - \overline{Y}}{(2t\mu_A / n)^{1/2}}\right) \le Z_{0.05}\right] =$$

$$= \Phi\left(\frac{\sqrt{2/n} \cdot Z_{0.05} + \delta\sqrt{t\lambda_A}}{\sqrt{(2+\delta)/n}}\right) \approx \Phi\left(Z_{0.05} + \delta\sqrt{\mu_A/2}\sqrt{tn}\right),$$

where the last approximation assumes that $\delta << 2$.

This standard formula is used to determine what combinations of *(t,n)* will give the desired power. It is a critically important feature of this calculation that the power can be driven to 1 either by increasing *t* or *n*. In mathematical notation we have that, as long as $\mu_A > 0$, $\delta > 0$: $\pi(\mu_A, \delta, n, t) \to 1$ as either $n \to \infty$ or $t \to \infty$.

This important property is shared by other "standard" distributions used in A/B testing, including the normal and binomial distributions.

Another way to think about the Poisson model is as a generalized linear model (GLM, McCullagh and Nelder 1989), and use likelihood-based inference like Wald tests or generalized likelihood ratio tests for inference. In the case of Poisson, because the maximum likelihood estimate of the Poisson rate is the observed mean, it is easy to show that the Wald approach and the heuristic normal approximation we described above are very closely related (basically, amounting to normal approximations on different scales). Other likelihood based approaches are likely to yield very similar results (Aban et al., 2008). Our choice of presenting our inference using a normal approximation is based on the desire to present simple power calculations, which will easily generalize to the NB case we discuss next. We discuss alternative testing approaches in more detail in the context of NB inference below.

The NB distribution is defined by two parameters, $X \sim NB(\alpha, p)$ if (Johnson et al. 2004):

$$P(X = k) = \binom{\alpha + k - 1}{k}(1-p)^{\alpha} p^{k} .$$

If α is natural, a common interpretation of $NB(\alpha,p)$ is the number of failures until α successes occur in a Bernoulli sequence, when the probability of success is *p*. However, the NB distribution is defined for any non-negative α and $p \in (0,1)$.

The NB distribution can also be thought of as a generalization of Poisson, where now instead of assuming a fixed rate λ, we assume λ is random and drawn from a Gamma distribution (Johnson et al. 2004):

$\lambda \sim \Gamma(\alpha,\beta)$, $X|\lambda \sim Pois(\lambda)$.

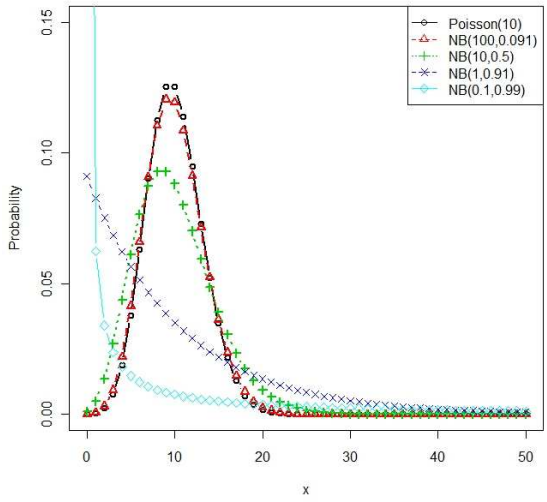

Figure 1. Comparison of Poisson distribution with mean 10 (black line) and NB distributions with the same mean for various values of the shape parameter α. We can see the long tail of distributions with small α.

In this situation, it can be shown that the unconditional (marginal) distribution of X is $X \sim NB(\alpha, \beta/(1+\beta))$. The Poisson distribution is the limit of NB as $\alpha \to \infty$ (and the Gamma distribution becomes concentrated at a point). The mean of X is $\mu = \alpha p/(1-p) = \alpha/\beta$, and the variance of X is $\sigma^2 = \alpha p/(1-p)^2 = \mu(\mu+\alpha)/\alpha$ (decreasing with α for fixed mean).

Figure 1 demonstrates a Poisson distribution with mean $\mu=10$ and several NB distributions with the same mean. We can see that as α gets smaller the right tail of the NB distribution becomes more extreme, and hence we would expect to observe big outliers (such as bots or crawlers in an internet search setting, which perform a continuous stream of searches). On the other hand, the vast majority of users concentrate on the smallest end of the scale, corresponding to the bulk of "casual" users.

In the A/B testing context, it makes sense to assume that α is common to both the A and B groups and known (i.e., the skew of the distribution is similar in both groups) and they differ in μ only. Like in the Poisson setting, we can now take μ as the expected number of searches per day, and so the means of the two groups will be $t\mu_A$ and $t\mu_B$, and the A/B hypothesis testing problem can be written as: $X \sim NB(\alpha, t\mu_A/\alpha)$, $Y \sim NB(\alpha, t\mu_B/\alpha)$ and $H_0: \mu_A=\mu_B$ vs $H_A: \mu_A<\mu_B$.

Using a normal approximation again (whose validity has to be carefully monitored when α is small because of the long tail), we get a p value calculation:

$$p_{NB,\alpha} = \Phi\left(\frac{\overline{X} - \overline{Y}}{\left(\mu t(\mu t + \alpha)/\alpha \cdot (n+m)/(nm)\right)^{1/2}} \right),$$

where again $\mu=\mu_A=\mu_B$ can be replaced by the pooled estimator.

It is easy to see that for a fixed set of data, $p_{NB,\alpha}$ is monotone decreasing in α for fixed μ, in particular it is always larger than p_{Pois} (corresponding to $\alpha=\infty$). Thus, if we get a p value that is small enough and reject H_0 for NB, we would reject for Poisson as well, but not vice versa. As we will see later, this is the main problem with erroneously using Poisson to analyze data better fitted with NB – over-rejection of H_0, including in A/A tests, where H_0 is known to be true.

As in the Poisson case, for power calculation we assume μ_A and $\mu_B = \mu_A(1+\delta)$ are both known and calculate:

$$\pi(\mu_A, \delta, t, n) = P_{H_A}\left[\left(\frac{\overline{X} - \overline{Y}}{\left(2\mu_A t(\mu_A t + \alpha)/(\alpha n)\right)^{1/2}} \right) \le Z_{0.05} \right] =$$

$$= \Phi\left(\frac{\sqrt{2(1+\alpha/(\mu_A t))} \cdot Z_{0.05} + \delta\sqrt{\alpha n}}{\sqrt{2(1+\alpha/(\mu_A t)) + 2\delta + \delta^2 + \delta\alpha/(\mu_A t)}} \right), \qquad (1)$$

using this expression we demonstrate next that $\pi(\mu_A, \delta, t, n) < \Phi(Z_{0.05} + \delta\sqrt{\alpha n}) < 1$, when n is fixed, for any value of t. For small enough δ and n, this bound may be quite low, as we demonstrate empirically. Thus, for NB A/B tests, we cannot guarantee arbitrarily high power by running the experiments longer, and have to accept that some differences δ are "undiscoverable" unless we are able to increase the number of test participants n.

Power Limitation of NB Inference

Our main result is the following:

Proposition 1

For a NB hypothesis test, where we observe n users in each group and μ_A is given, the power of the test based on the normal approximation for:

H_0: $\mu_A=\mu_B$ vs H_A: $(1+\delta)\mu_A=\mu_B$

is bounded as follows, independently of the time t for which the experiment is run:

$$\forall t, \pi(\mu_A, \delta, t, n) < \max\left(0.5, \Phi(Z_{0.05} + \delta\sqrt{\alpha n/2})\right) < 1.$$

Proof: Consider the expression inside the parentheses of Eq. (1). $Z_{0.05}<0$, so for small t this expression will be negative and $\pi<0.5$, but for large enough values of t it will be positive. Once the expression is positive, making it larger in absolute value increases the power because Φ is a monotone increasing function. Thus assuming t is large enough that the numerator is positive we have:

$$\frac{\sqrt{2(1+\alpha/(\mu_A t))}\cdot Z_{0.05} + \delta\sqrt{\alpha n}}{\sqrt{2(1+\alpha/(\mu_A t)) + 2\delta + \delta^2 + \delta\alpha/(\mu_A t)}} \leq \frac{\sqrt{2(1+\alpha/(\mu_A t))}\cdot Z_{0.05} + \delta\sqrt{\alpha n}}{\sqrt{2(1+\alpha/(\mu_A t))}} =$$

$$= Z_{0.05} + \delta\sqrt{\frac{1}{(1+\alpha/(\mu_A t))}}\sqrt{\alpha n/2} < Z_{0.05} + \delta\sqrt{\alpha n/2}$$

$$\Rightarrow \pi(\mu_A, \delta, t, n) < \Phi(Z_{0.05} + \delta\sqrt{\alpha n/2})$$

■

Looking carefully at the expressions, it is possible to refine this bound slightly further, and make it tight, i.e., we can in fact prove that:

$$.\pi(\mu_A, \delta, t, n) \leq \Phi\left(\sqrt{\frac{1}{1+\delta+\delta^2/2}}\,[Z_{0.05} + \delta\sqrt{\alpha n/2}]\right),$$

$$\pi(\mu_A, \delta, t, n) \xrightarrow{t\to\infty} \Phi\left(\sqrt{\frac{1}{1+\delta+\delta^2/2}}\,[Z_{0.05} + \delta\sqrt{\alpha n/2}]\right),$$

but this slight improvement in bounds has little bearing on our main message, which is that the power is bounded from above for any t, with the largest possible power decreasing with α (as we move further away from Poisson).

Figure 2 illustrates this result by explicitly calculating (1) and showing the resulting power curves as a function of t for the same five distributions as in Figure 1, with *δ=0.02* (2% increase), *n*=1000 users in each group (left) and n=100,000 users in each group (right). As we can see, after 10 days the power of the Poisson test is close to 1. For the NB with n=1000,*α=100*, the power is also good, but for the smaller values of α the power is low, with the graph for *α=0.1* not going above power 0.07 for a test at level 0.05, so there is almost no difference between the distributions of the A and B group. Indeed, the bound for power at *α=0.1* gives *π<0.069*. With 100,000 users the power converges to 1 for *α≥1*, but at α=0.1 the power is again bounded at 0.4, no matter how long we run the experiment. Thus, even with 100,000 users in each group, we cannot guarantee high power if the tail is long enough! The flip side of this result is that

if one mistakenly uses Poisson when NB is appropriate, the relation $p_{Pois} < p_{NB,\alpha}$ implies probability of a false rejection of H_0 becomes large, as our experiments below demonstrate in A/A tests.

Alternative Inference Approaches for the NB Model

As we mentioned above, the normal approximation we employed for NB inference can be replaced with several different and widely used approaches. The most statistically sound methods use formal likelihood-based inference using the theory of generalized linear models (McCullagh and Nelder, 1989). The two commonly used approaches are Wald's test (as implemented in the R package glm, for example) and the Chi-square generalized likelihood ratio test (as implemented in the R function anova.glm, for example). (Aban et al., 2008). Note that in our normal approximation we utilized the maximum likelihood estimates of NB parameters in our mean and variance estimates. Because the normal approximation is very reasonable for large data like we have, we expect the results of our inference to be very similar to likelihood-based inference, and we confirm that in the experimental section below. Our choice of the normal approximation was motivated by the need for simple power and sample size calculations, which are facilitated by the normality assumption.

Another alternative approach is to use the "moment" estimator of variance in a normal approximation, by utilizing the data's empirical variance as the estimate of variance. This is an approach commonly referred to as "t-test" (Aban et al., 2008). Aban et al. (2008) have performed extensive experimentation and simulation and have shown that this approach is inferior to likelihood-based inference for NB, in terms of power. Although their simulations are relatively small, this result is consistent and appears to be independent of sample size. Together with our observation that our normal approximation gives results that are very similar to likelihood-based inference, we believe this justifies our concentration on the normal approximation as our primary parametric inference tool.

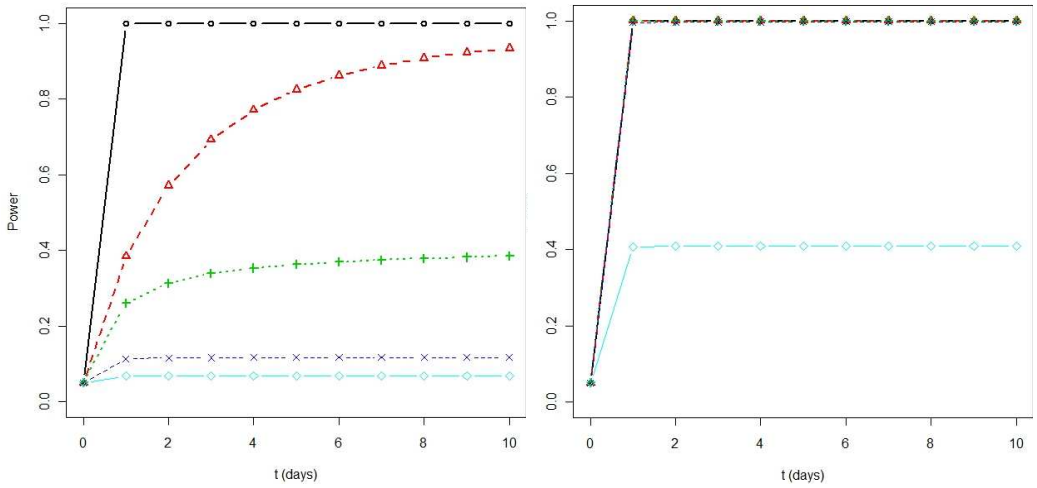

Figure 2. Power curves for δ=0.02 (2% increase in B compared to A), μ=10 per day, n=1000 (left), n=100,000 (right). As in Fig. 1, the black curve is for Poisson, reaching power of almost 1 after 10 days at n=1000, while the others are in decreasing order for α={100,10,1,0.1}. We can see that for small α there is practically no power to detect such a difference at n=1000, and even at n=100,000 the power is bounded when α=0.1, though it converges quickly to 1 for bigger α .

Testing for Appropriateness of Poisson vs Negative Binomial Distribution

Our results so far have demonstrated that it is critical to choose correctly whether the statistical analysis should be done with the Poisson distribution or the NB distribution. Choosing Poisson when NB is appropriate will lead to incorrect rejection of the null (type-I error too big), while choosing NB when Poisson is appropriate will result in loss of power (increased type-II error). In the next section we will also demonstrate the first effect clearly on the real A/A and A/B testing data of SweetIM.

An important question is thus, how can we determine empirically whether Poisson or NB is appropriate for our data? Because the Poisson(μ) distribution is nested within the family of NB(α,μ/(μ+α)) models, being the limit as $\alpha\to\infty$, we can use the standard generalized likelihood ratio (GLR) test for nested hypothesis (Casella and Berger 2001). This test relies on the fact that under H_0: $\alpha=\infty$, twice the difference of log-likelihoods of the maximum likelihood fits of the NB and Poisson model has approximately a χ^2_1 distribution (chi square with one degree of freedom).

This result can be used to test the appropriateness of assuming Poisson vs the alternative that NB is needed. Intuitively, if NB fits much better it should be chosen and GLR tests give us an approach to identify whether this is the case.

Beyond Parametric Modeling: Permutation Tests

While the methodology described so far allows us to test whether NB is more appropriate than Poisson (which it is, in all our examples below), it is clear that parametric distributions like Poisson or NB will never exactly describe our data. For example, even if the rate varies between individuals, there is no reason to assume the rates are Gamma distributed (and hence overall usage has an NB distribution).

An alternative which is almost assumption free is to use non-parametric permutation tests for inference (Efron and Tibshirani 1994, Chapter 15). Briefly, if we have n individuals in the A group and m individuals in the B group, permutation testing involves randomly dividing the union of the two groups into subsets of n and m individuals, and documenting the difference in average usage between the two resulting groups. This is done repeatedly, giving a "null" distribution of differences in averages when the groups do not differ. If the actual difference in the true A and B groups is extreme given this null distribution, then we can reject the null hypothesis that the groups do not differ.

In all analyses below, we give the permutation test p value in addition to the Poisson and NB p values.

Of course, the non-parametric approach cannot be used for planning A/B tests, calculating needed sample sizes, etc., because no analysis can be performed before data are observed. Thus, it does not answer some of the major statistical needs of A/B testing. Even if permutation tests are used for inference, it is critical to also have a "reasonable" parametric approach like NB to address the planning needs.

Case Study – A/B Tests on SweetIM Data

SweetIM is a company offering various services to enhance online communications, including add-ons to instant messaging (IM) services, a web browser toolbar, and an internet search tool. Naturally, these are all designed with an aim of increasing exposure to new users and usage by existing users. To that extent, it is a critical aspect of the company's research and development effort to optimize user experience and maximize usage. Two of the most important measures of usage employed by the company are searches per user and number of

SweetIM generated content items sent per user. Both of these measures are counts, and we will concentrate on them for the rest of this article.

SweetIM has developed its own internal A/B testing environment, and it implements within this environment numerous tests to examine the effect of various features and aspects of the user experience on usage measures. In addition, it also runs occasional A/A tests, where both groups get the exact same experience, as a “sanity check” that all aspects of the test generation system are working properly, in which case these tests should “typically” (in 95% of cases, when testing at level $\alpha=0.05$) result in a non-rejection of H_0, i.e. a conclusion of "no detected difference".

Environment

SweetIM's A/B testing environment consists of a distribution mechanism, data collection and an analysis / reporting system. Each new test starts from the definition of sample size and time the test will be active. This first step is critical because a wrong design can easily destroy the whole test no matter how accurate the data and analysis are. In this phase we define the sample size for the test groups based on the nature of performance counters that we want to compare. In this paper we are focusing on count data like number of searches for user, so the experimental design is based on the negative binomial distribution and sample size calculation is derived from equations like (1). The distribution period (in days) is defined to be long enough to reach the sample size, and in most cases it will be rounded to a multiple of seven for neutralizing the day-of-week effect. The observation period is less critical in terms of power limitation of negative binomial inference, but still is important to detect long-term changes in users’ behavior.

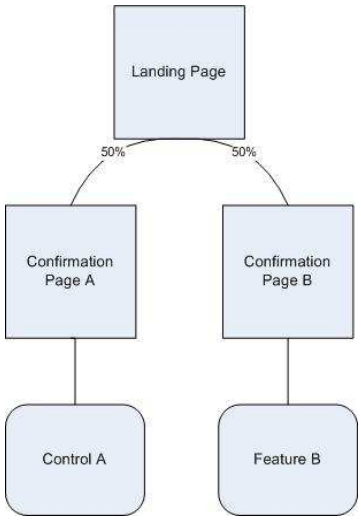

Figure 3. A sample flow of traffic diversion for new SweetIM users.

We separate all tests on 2 groups based on the test population – tests on new users that just installed SweetIM product and tests on all audience of active users. All potential new users that reached our Landing Page are divided randomly to two (or more) groups and receive test and group cookie according to their group. Based on this cookie the user will receive the regular application (Control group) or the application with a new feature (group B) as described in Figure 3.

For active users we are able to take unique ID modulo number of groups we want to check in the test, for example all active users with ID whose mod will be equal to 3 will receive group 3 solution and will be grouped in the analysis under the corresponding test group. It is important to note that in our case the unique ID of the user is an absolutely random number and not an increment - this allows us to use this type of distribution approach.

Once the user starts her/his participation in the test, we track related activity automatically from our web logs, until the test ends or the user removes cookies (in case of new users test). Automatic business intelligence (BI) processes are responsible for delivery and aggregation of the test data from web logs to the internal reporting system and OLAP database.

Web activity of performance counters like number of searches per test participant can be affected by Internet robots and worms that have nothing in common with real user behavior. These robots can introduce significant skew into estimates, enough to render assumptions invalid. We have seen cases where robots caused the performance counter to be significantly different from control group, when, in fact there was no difference between the groups. While the NB assumption takes

care of some forms of non-uniformity of users, these extreme deviations can still be devastating. For the purpose of experimentation, it is especially important to remove some types of robots, those that interact with the user ID and simulate proper User-Agent. Since many robots have the same characteristics as human users, it is difficult to clearly delineate between the two. Benign or simple robots can often be filtered by basic characteristics but many modern robots use sophisticated techniques to escape detections and filtering.

We didn't develop any automatic filtering solution in our A/B test processes; however, the first step in each A/B test analysis is to check the groups for exceptional user behavior and try to filter non-human test participants.

Once the test reached sufficient sample size it can be analyzed through the internal reporting system. In addition to standard frequency tables and daily trends, results of significance tests are also available.

Tests on Search

We present here two A/A tests and two A/B tests on search activity (tests 1-5 in Tables 1, 2). Tests 2, 4 and 5 are actual A/B tests run at SweetIM to test new features. Test 1 is an actual A/A test run by randomly assigning users to groups A1 and A2 and following their search activity. Test 3 is an A/A test derived by us from A/B test 2 by randomly dividing the users in the A group into two groups. The different search tests 1,2,4,5 differ slightly in the time they were run, their duration and the target audience, and so they are not necessarily expected to be similar between them in terms of number of searches or the NB shape parameter α.

Number	Type	Group A			Group B (or other A)		
		# Users	*#Activity*	*Average (μ)*	*# Users*	*#Activity*	*Average (μ)*
1	Search A/A	61,719	1,274,279	20.65	61,608	1,288,333	20.91
2	Search A/B	130,062	2,420,670	18.61	129,286	2,432,957	18.82
3	Search A/A[a]	65,031	1,201,612	18.48	65,031	1,219,058	18.75
4	Search A/B	80,690	1,100,560	13.64	80,174	1,135,716	14.17
5	Search A/B	17,287	297,485	17.21	17,295	300,843	17.39
6	Content A/A	61,719	3,142,766	50.92	61,608	3,165,208	51.38
7	Content A/B	17,568	770,436	43.85	17,843	930,066	52.12

[a]Random splitting of group A of experiment 2.

TABLE I. EXPERIMENT DETAILS

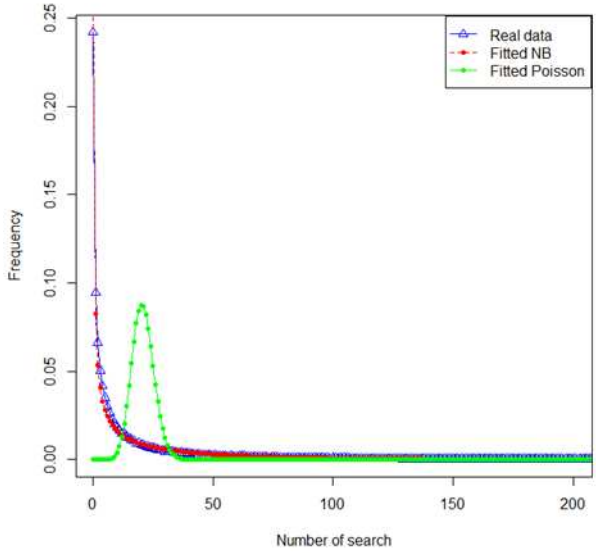

Figure 4. Graphical demonstration of the much better fit of the NB distribution to the search data of test 1, compared to the Poisson distribution.

In the A/A tests, the NB parameter α was estimated from the entire dataset, while in the A/B tests it was estimated from the A group only. As we can see from Table 2, the α estimates from the different datasets ranged between 0.19 and 0.53. However all of these estimates imply a very long tailed, non-Poisson distribution. This is illustrated in the p-values of the GLR tests for Poisson vs NB, which are rejected with p-value of 0 (i.e., smaller than expressible by the software) for all five tests. In Figure 4 we also show the empirical distribution of search number for test 1, compared to the fitted Poisson and NB distributions. As can be seen, the NB estimated distribution fits the empirical distribution well, while the Poisson is completely off. Similar plots for tests 2-5 yield similar conclusions. Table 2 reports the difference in mean searches between the two groups, and the p-value of the tests we described above for testing H_0: $\mu_A=\mu_B$ vs HA: $\mu_A<\mu_B$, under the Poisson and NB assumptions.

As can be seen, Poisson inference (which is incorrect, as the GLR p values indicate) strongly rejects H_0 for all tests, including the two A/A tests. One way to understand this result is to consider the property of Poisson distribution, that the variance is equal to the mean. For overdispersed distributions like NB, on the other hand, the variance is much bigger than the mean. Thus, incorrectly assuming a Poisson distribution amounts to underestimating the variance, and hence it leads to misjudging the magnitude of "expected" differences under the null. NB based

inference rejects H_0 for one of the A/B tests only, where the difference in indeed big (almost 4%) – test 4. In this test we checked different sizes of the search box on SweetIM search page and its influence on users' activity in search. We can see

Number	Type	NB fit		Difference between groups %	Difference tests, p-value				
		Estimated α	*GLR p-value*		*Poisson*	*Permutation Test*	*Our NB Approx.*	*Wald Test*	*T-test*
1	Search A/A	0.31	0.00E+00	1.27%	**8.06E-25**	3.67E-01	1.10E-01	1.07E-01	1.82E-01
2	Search A/B	0.24	0.00E+00	1.10%	**2.27E-34**	3.49E-01	9.00E-02	8.66E-02	1.76E-01
3	Search A/A[a]	0.24	0.00E+00	1.40%	**4.61E-04**	1.40E-01	3.55E-01	3.55E-01	6.00E-01
4	Search A/B	0.19	0.00E+00	3.90%	**1.50E-176**	**3.24E-02**	**5.82E-04**	**5.81E-04**	**1.39E-02**
5	Search A/B	0.53	0.00E+00	1.08%	**2.16E-03**	7.90E-01	3.11E-01	3.11E-01	6.04E-01
6	Content A/A	0.33	0.00E+00	0.98%	**2.13E-29**	4.20E-01	1.85E-01	1.85E-01	2.05E-01
7	Content A/B	0.45	0.00E+00	8.68%	**0.00E+00**	**0.00E+00**	**5.59E-27**	**3.33E-27**	**2.20E-16**

[a]Random splitting of group A of experiment 2.
* significant p-values (< 0.05) were bolded

TABLE II. EXPERIMENT INFERENCE

that group B that received bigger search window has significant advantage over the A (Control) group. Based` on this test results the search page was changed to version B, leading to an overall increase of search of about 5% – this increased the company's revenue directly by 5%. Note that the p value of the NB test is not overwhelming in spite of the big difference – this is a consequence of the power issue we discussed above. In addition to NB p values based on our normal approximation we also report p-values of the Wald test obtained using the function glm.nb in R, which are almost identical as we discussed above; and those of a "t-test" based on moment estimators, which are less extreme (reflecting their reduced power as demonstrated by Aban et al., 2008). We also report in Table 2 p values obtained from permutation tests, and we see that their conclusions closely match those of the NB inference. In general, all non-Poisson approaches seem to lead to reasonable and consistent inference.

Tests on Content

The last two tests in Tables 1 and 2 are tests run on counts of content items sent by users rather than searches. Test 6 is an actual A/A test, while test 7 is an actual A/B test. The conclusions are all similar to the search tests. Note the very significant rejection of the A/B test even under the NB assumption, with a p value of practically zero. This is due to the combination of the higher α compared to the search test (recall that the power deteriorates with α) and the very large effect size. In this test different types of content delivery were checked. In general, only

SweetIM users can communicate with each other using fun content, i.e. if SweetIM user will send content to non-SweetIM user, the other side will receive only an invitation to install the product and not the content itself. In this test we checked the option of two-sided content; content that will be visible for both sides. Again, based on the significant results this feature was implemented for all users. Group B results increase content usage by ~9%, this feature improves user experience and increases user lifetime value.

Conclusions

In this paper we have demonstrated both theoretically and empirically the importance of avoiding incorrect use of the Poisson distribution assumption in A/B testing. We have pointed out the limitation of NB inference in attaining power in testing. Although we have rigorously shown this for our normal approximation only, this conclusion clearly applies to all non-Poisson tests in Table II. It should be reiterated, however, that power considerations must not lead us to forsake correct inference in favor of more powerful approaches (like Poisson or others), since if NB or a similar long tailed distribution is the appropriate distribution for modeling, these more powerful approaches are inevitably going to lead to incorrect conclusions, in particular rejection of the null hypothesis of no difference when it is true, as was demonstrated in the A/A tests we discussed.

Ultimately, one has to acknowledge the limitations of correct inference and adapt the testing procedures to those. In the case of power limitations of NB, the obvious solution, as we showed in our statistical analysis, is to increase the number of users in the A and B groups, thus increasing power to identify smaller effects.

A key question relates to the range of applicability of NB assumptions in A/B tests on counts. Is our example of SweetIM data, where Poisson inference is clearly incorrect and one should resort to NB-like approaches, anecdotal or is it representative of a large class of A/B testing scenarios? We believe the latter alternative is true and non-Poisson testing should prove the rule rather than the exception in A/B testing, as the user population is almost always non-homogeneous, and the long-tailed nature of user behavior is a typical profile in most interesting cases. Perhaps a more important practical question is, to what extent can we assume that NB distribution is a proper approximation of the true

distribution and use NB-based inference, or should the rule be to use assumption-free approaches like the permutation test we describe. We believe the latter to be true, and thus in our view a proper approach for testing combines parametric approaches like NB before the data are obtained for calculating power limitations and required sample sizes, and non-parametric, assumption-free approaches like permutation test as the ultimate testing procedure – at least as a complement to the parametric methods.

References

Aban, I. B., Cutter, G. R., Mavinga, N. 2008. “Inferences and Power Analysis Concerning Two Negative Binomial Distributions with An Application to MRI Lesion Counts Data”. *Comput Stat Data Anal.* 53(3): 820–833.

Andersen, E. B. 1997. “Introduction to the statistical analysis of categorical data”. Springer-Verlag Berlin.

Berk, R., MacDonald, J. M. 2008. “Overdispersion and Poisson Regression”. *Journal of Quantitative Criminology*, Volume 24, Number 3, 269-284,

Ash, T. 2008. “Landing Page Optimization. The Definitive Guide to Testing and Tuning for Conversions”. John Wiley & Sons, Inc.

Box, G., Hunter, J. S. and Hunter, W. G. 2005. “Statistics for Experimenters: Design, Innovation, and Discovery”. 2nd. s.l. John Wiley & Sons, Inc.

Casella, G., Berger, R. L. 2001. “Statistical Inference”. Duxbury Press, second edition.

Crook, T., Kohavi, R., Longbotham, R., Frasca, B. "Seven Pitfalls to Avoid when Running Controlled Experiments on the Web." at KDD 2009

Efron, B., Thisted, R. 1976. “Estimating the number of unseen species: How many words did Shakespeare know?”. *Biometrika*, 63, 435-447.

Efron, B., Tibshirani, R. 1994. "An Introduction to the bootstrap". Chapman and Hall / CRC.

Gerlough, D. L., Barnes, F. C., Schuhl A. 1971. “Poisson and other distributions in traffic: the Poisson and other probability distributions in highway traffic”. *Eno Foundation for Transportation*.

Greenshields, B. D., Weida, F. M., Gerlough, D. L., Huber M., J. 1978. “Statistics, with applications to highway traffic analyses”. *Eno Foundation for Transportation*.

Johnson, N. L., Kemp, A. W., Kotz, S. 2004. “Univariate discrete distributions”. John Wiley & Sons, Inc.

Kendall, M. G., Stuart, A. 1976. “The Advanced Theory of Statistics: Design and analysis, and time-series”. Hafner Pub. Co.

Keppel, G, Saufley, W. H and Tokunaga, H. 1992. “Introduction to Design and Analysis. 2nd". s.l. : W.H. Freeman and Company.

Kimura, M. 1994. “Population genetics, molecular evolution, and the neutral theory”. The University of Chicago Press.

Kohavi, R., Henne, R. and Sommerfield, D. 2007. “Practical Guide to Controlled Experiments on the Web: Listen to Your Customers not to the HiPPO”. http://exp-platform.com/hippo.aspx

McCullagh, P., Nelder, J. A. 1989. “Generalized Linear Models, Second Edition”. Taylor and Francis.

McGee, H. W., Taori, S., Persaud, B. N. 2003. “Crash experience warrant for traffic signals”.pp19-20. *National Cooperative Highway Research Program*, Report 491.

Myers, J. L., Well, A. 2003. “Research design and statistical analysis”. V 1. Lawrence Erlbaum Associates.

Schmittlein, D. C., Bemmaor, A. C., Morrison, D. G. 1985. “Why does the nbd model work? robustness in representing product purchases, brand purchases and imperfectly recorded purchases.” *Marketing Science*, Vol. 4, No. 3, 255-266.

Tamura, K., Nei, M. 1993. “Estimation of Nucleotide Substitutions in the Control Region of Mitochondrial DNA in Humans and Chimpanzees”. *Mol.Biol.Evol*, 10 (3): 512-526.

Tang, D., Agarwal, A., O'Brien, D., Meyer, M. 2010. “Overlapping Experiment Infrastructure: More, Better, Faster Experimentation.” *Proceedings 16th Conference on Knowledge Discovery and Data Mining*, 2010, pp. 17-26.

Thall, P. F., Vail, S. C. 1990. “Some covariance models for longitudinal count data with overdispersion”. *Biometrics, Sep*; 46(3):657-71

Ver Hoef , J. M., Boveng, P. l. 2007. “Quasi-poisson vs. Negative binomial regression: How should we model overdispersed count data?” *Ecology*, 88(11), pp. 2766–2772