Zagreb International Review of Economics & Business, Vol. 21, No. 1, pp. 95-104, 2018

ISSN 1331-5609; UDC: 33+65
DOI: 10.2478/zireb-2018-0004
CONFERENCE PAPER

Are Multi-Armed Bandits Susceptible to Peeking?

Markus Loecher *

Abstract: *A standard method to evaluate new features and changes to e.g. Web sites is A/B testing. A common pitfall in performing A/B testing is the habit of looking at a test while it's running, then stopping early. Due to the implicit multiple testing, the p-value is no longer trustworthy and usually too small. We investigate the claim that Bayesian methods, unlike frequentist tests, are immune to this "peeking" problem. We demonstrate that two regularly used measures, namely posterior probability and value remaining are severely affected by repeated testing. We further show a strong dependence on the prior probability of the parameters of interest.*

Keywords: multiple comparisons; A/B testing; Bayesian decision theory

JEL Classification: C4, C6

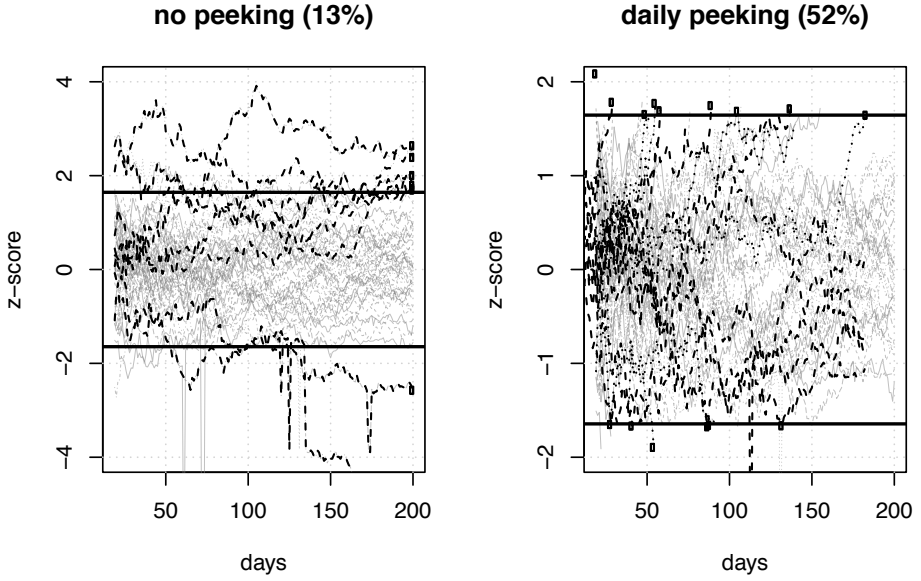
Introduction

The sequential testing procedure of repeatedly evaluating the significance of an observed difference between two (or more) treatments is often referred to as "optional stopping", or more colloquially as "peeking". One consequence is that the type-I error rate is no longer controlled at significance level α.

As an example, Fig. 1 shows the temporal evolution of p-values (translated into a z-score for better visual perception) over the course of a 200-"day" experiment. Patiently waiting to the end of the experiment results in a number of false positives that is consistent with the set significance level . "Daily peeking", i.e. stopping the experiment as soon as the p-value dips below a threshold, on the other hand leads to a severely inflated type-I error rate.

* Markus Loecher is at Berlin School of Economics and Law, Berlin, Germany.

Figure 1: left: waiting until day 200 to make a decision leads to a nominal type-I error rate indicated by the thick dashed lines; right: peeking daily yields many false positives and an overall type-I error rate of 50%. Only the thick dotted lines correspond to paths that would have led to a rejection of the NULL after 200 days.



An often-proposed alternative to traditional A/B testing is to rely on a Bayesian procedure rather than frequentist hypothesis testing. There are many claims in the literature that those approaches allow you to peek at your test while it's running and stop once you've collected enough evidence (Rouder 2014, Sanborn and Hills (2014), Erica et al. (2014)). In this paper we use simulations to demonstrate that the Bayesian procedure outlined in (Scott 2012) is not robust to the implications of multiple comparison. While the conclusions w.r.t. type-I errors are similar to (Robinson 2015), the following important differences are relevant: (i) the Bayesian procedure adapted in this paper is (also known as "Thompson Sampling") in combination with a beta-binomial model. (ii) the effect of peeking on type-II errors is also investigated.

Multi-Armed Bandits

The name "multi-armed bandit" describes a hypothetical experiment where you face several slot machines ("one-armed bandits") with potentially different expected payouts The goal of this sequential experiment is to produce the largest reward. In the

typical setup there are K actions or "arms". Arm i is associated with an unknown quantity υ_i giving the "value" of that arm. The goal is to choose the arm providing the greatest value, and to accumulate the greatest total reward in doing so. The name "multi-armed bandit" is an allusion to a row of slot machines (colloquially known as "one armed bandits") with different reward probabilities. For a thorough overview of the vast literature of the dynamic research on multi-armed bandits we refer the reader to (Scott 2010) and references therein.

A trademark of sequential testing is the balancing of the so called: while one wants to find the slot machine with the best payout rate, the cost of the experiments need to be "minimized" at the same time. The fundamental tension is between "exploiting" arms that have performed well in the past and "exploring" new or seemingly inferior arms in case they might perform even better.

Google Analytics Content Experiments

The following examples and specific Bayesian procedures describe the multi-armed bandit approach to managing online experiments taken at Google Analytics as described in (Scott 2012). This section summarizes the claims made in and heavily borrows from (Scott 2012).

> *Suppose you've got a conversion rate of 4% on your site. You experiment with a new version of the site that actually generates conversions 5% of the time. You don't know the true conversion rates of course, which is why you're experimenting, but let's suppose you'd like your experiment to be able to detect a 5% conversion rate as statistically significant with 95% probability. A standard power calculation1 tells you that you need 22,330 observations (11,165 in each arm) to have a 95% chance of detecting a .04 to .05 shift in conversion rates. Suppose you get 100 visits per day to the experiment, so the experiment will take 223 days to complete. In a standard experiment you wait 223 days, run the hypothesis test, and get your answer.*
>
> *Now let's manage the 100 visits each day through the multi-armed bandit. On the first day about 50 visits are assigned to each arm, and we look at the results. We use Bayes' theorem to compute the probability that the variation is better than the original2. One minus this number is the probability that the original is better. Let's suppose the original got really lucky on the first day, and it appears to have a 70% chance of being superior. Then we assign it 70% of the traffic on the second day, and the variation gets 30%. At the end of the second day we accumulate all the traffic we've seen so far (over both days), and recompute the probability that each arm is best. That gives us the serving weights for day 3. We repeat this process until a set of stopping rules has been satisfied.*
>
> *Figure 2 shows a simulation of what can happen with this setup. In it, you can see the serving weights for the original (the solid line) and the variation (the*

Figure 2: Two simulations of the optimal arm probabilities for a simple two-armed experiment. These posterior probabilities are also the fractions of the traffic allocated to each arm on each day. The true success rates are 0.05 (dotted) and 0.04 (solid) respectively.

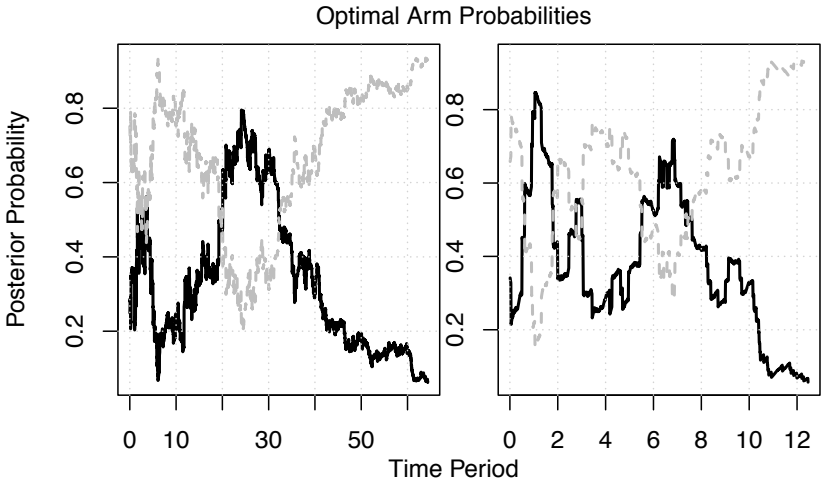
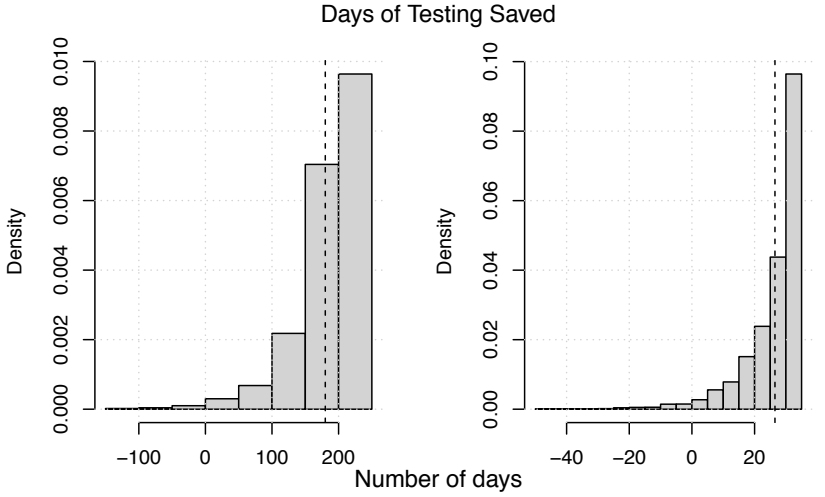


Figure 3: Number of days saved for conversion probabilities $p_1 = 0.05$, $p_2 = 0.04$ (left panel) and $p_1 = 0.05$, $p_2 = 0.08$ (right panel), respectively. ''Days saved'' is measured w.r.t. a classical experiment planned by a power calculation. Note the negative values that correspond to runs where the binomial bandit took longer. The average savings are 180 from a max of 223 (left panel) and 26.5 out of 35 days (right panel), respectively.

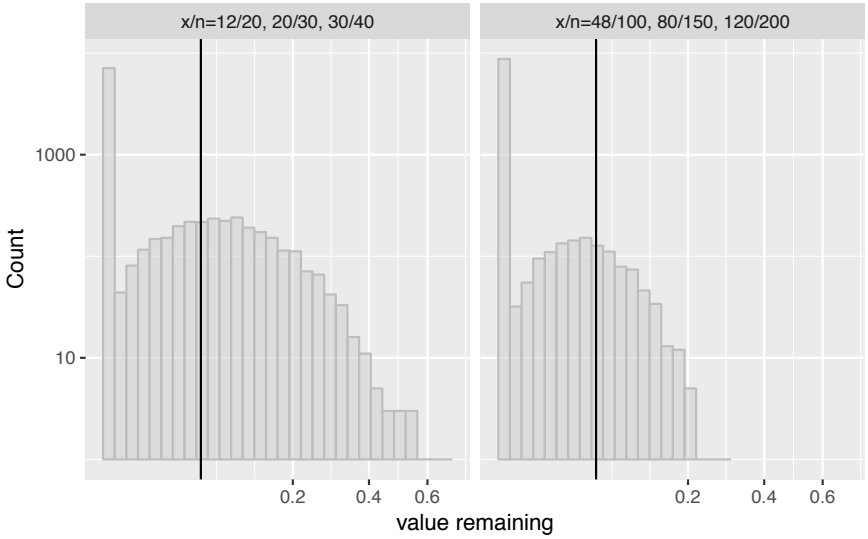


dotted line), essentially alternating back and forth until the variation eventually crosses the line of 95% confidence. (The two percentages must add to 100%, so when one goes up the other goes down). The experiment finished in 66 days, so it saved you 157 days of testing.

The distributions of saved days for two choices of parameters are shown in Figure 3. On average the tests in the left panel ended 175 days sooner than the classical test based on the power calculation. The gains are much less pronounced when the difference in proportion is larger (right panel).

Stopping Criteria

Figure 4: Distribution of VR with its th percentile as a vertical line (left panel). We increased the sample size by a factor of 4 (right panel). Note the nonlinear scaling of the axis (log10 for y and sqrt for the x-axis).



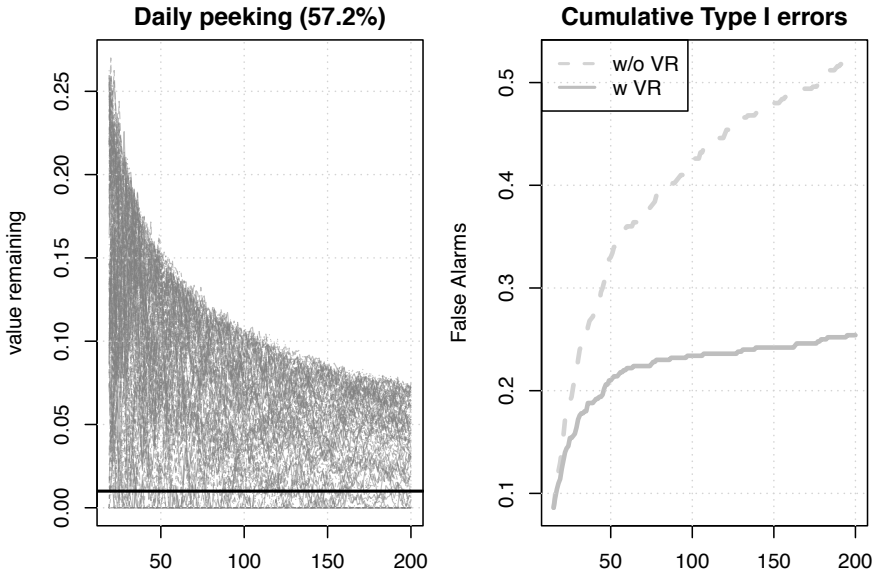
The second metric being monitored is the *potential value remaining in the experiment*, which is particularly useful when there are multiple arms. At any point in the experiment there is a "champion" arm believed to be the best. If the experiment ended "now", the champion is the arm you would choose. The **value remaining** in an experiment is the amount of increased conversion rate you could get by switching away from the champion. Google Analytics ends the experiment when there is at least a 95% probability that the value remaining in the experiment is less than 1% of the ''champions'' conversion rate.

Figure 4 illustrates the distribution of the value remaining (VR) metric for the particular observed outcomes of three arms with 20, 30, and 40 sessions that have generated 12, 20, and 30 conversions. The right panel in Figure shows what happens to the value-remaining distribution as the experiment progresses (three arms with 100, 150, and 200 sessions that have generated 48, 80, and 120 conversions).

Type I-Error and Multiple Testing

We now address the question whether Bayesian Methods are immune to multiple comparisons w.r.t to false positive rates. Our simulations assume just two arms, each having a 5% conversion rate, $p_1 = p_2 = 0.05$. For sake of clarity we use instead of Thomson sampling for all experiments in this paper, i.e. each arm is presented with a constant number ($N = 100$) of impressions each day. The motivation is to decouple the Bayesian early stopping method from the additional regret minimization due to randomized probabilistic matching (RPM). This paper isolates the effects of the former method. All simulations are run in R (R Core Team 2016) using the bandit package (Lotze and Loecher 2014) with an initial "burn-in" period of 14 days during which no early stopping decisions are attempted.

Figure 5: The left panel shows the cumulative effects of declaring ties based on the value remaining. Right panel: The upper curve shows the cumulative effect of declaring a winner based on posterior probability. The lower curve combines both stopping Criteria



The left panel of Figure 5 shows the effect of multiple testing solely by declaring ties based on the value remaining metric. In this case, we in fact observe an increase of correct decisions since the remaining value indeed is zero!

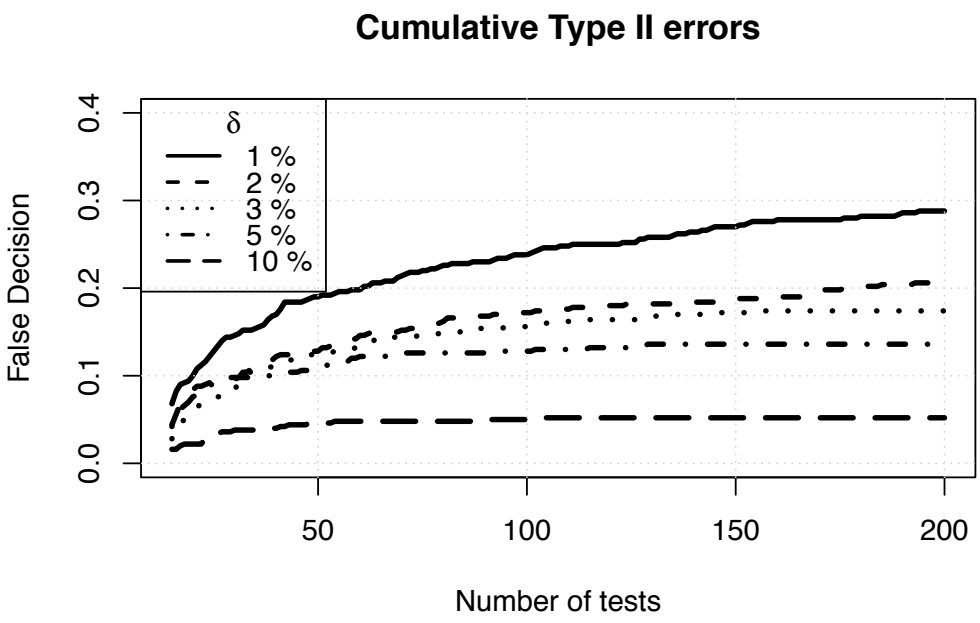
The message sent by the right panel of Figure 5, however, is alarming. When applying both stopping criteria (posterior probability AND value remaining) at the same time, the lower curve shows that the type-I error rate is increasing from its nominal level to about 25%. For sake of completeness, we overlaid the upper line which applies only the former stopping criteria, i.e. no tie declarations.

Type II-Errors and Multiple Testing

It is often said that Bayesian methods do not claim to control the type I error rate. They instead set a goal about the expected loss. For the example of ad campaign selection, one could argue that a type-I error hardly matters. In that light, the results above could be dismissed as irrelevant and missing the point.

We also investigate the probability of declaring the wrong arm when there are actual differences, i.e. the type-II error rate for the scenario $p_1 = 0.05$, $p_2 > 0.05$. A false decision in this case clearly can have significant consequences for businesses.

Figure 6: As the percentage difference between p_1 and p_2 decreases, the cumulative effects on the type-II error rate by repeatedly applying the optional stopping rules, increases.

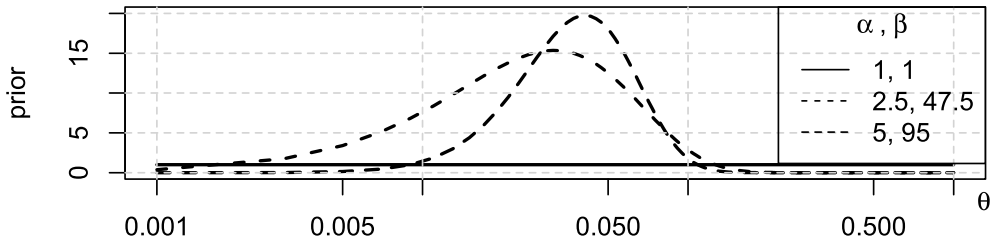


In particular, we vary the percentage difference δ between 1% (the threshold the value remaininng metric claims to be able to detect.) and 2, 3, 5, 10%: $p_2 = (1 + \delta) \cdot p_1$. Figure 6 shows increasingly severe effects of applying the optional stopping rules as δ decreases. The type-II error rate is controlled at 95% as claimed. For δ= 1%, the cumulative error rate after 200 "days" is nearly 30%!

Free of Parameters?

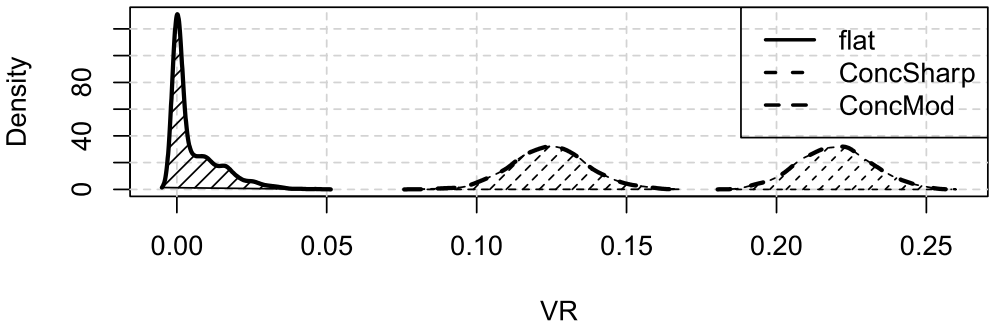
A particular appealing aspects of Bayesian methods is that apparently no tuning parameters have to be chosen at the onset of an experiment -unlike in classical testing where one has to fix the smallest difference in proportion to be detected at a certain level.

Figure 7: beta distribution for various parameter combination as prior for the binomial distribution. The probability density goes from a flat, uniform to a sharply concentrated probability mass around 0.05.



While this is not strictly true, since the prior distribution for the binomial parameter has to be chosen, a common default assumption seems to be to assume the uniform distribution as a prior. Sometimes this choice is incorrectly described as "making no assumptions", hence suggesting a vritually parameter free method. Figure 7 illustrates the uniform density along with two alternative beta distributions, which we refer to as "moderately" and "sharply" concentrated around 0.05 from here on.

Figure 8: Fluctuations of the 95% quantile value remaining distributions and their dependence on the prior distribution. We assume two arms arms with 100 sessions each that have generated 7 and 2 conversions, respectively. Each quantile was generated from 10,000 simulations of the binomial posterior distribution.



Our point is that the choice of prior distribution is extremely important for both stopping metrics. In particular, Figure 8 shows the distribution of the 95% quantile of the value remaining for the three prior distributions from above. As the prior concentrates more sharply around 0.05%, the VR density moves to the right; the value remaining is 10 times as high for the more concentrated prior!

Another insight is gained from inspecting the widths of the distributions in Figure 8. As there exists no closed analytic form for the value remaining metric, the 95% quantile is typically obtained from simulations of the posterior distributions of sample size e.g. $n = 10^4$ or 10^5. Accurate estimation of higher quantiles requires extremely high sample sizes though, which is reflected in the additional uncertainty of the $Q_{0.95} - VR$ metric. That in turn results in a higher rate of false decisions based upon an incorrectly estimate of $Q_{0.95} - VR$.

Conclusion

We conclude that the claim of "Bayesian testing is unaffected by early stopping" is simply too strong. Type-I errors may or may not incur additional cost: it seems more reasonable that switching campaigns for no good reason should be avoided. The effect of priors is strong and not easily understood by A/B testing. We also show that the fluctuations of the value remaining metric add uncertainty that needs to be accounted for in statements on significance.

At the same time, it is true that the cost of an experiment can be substantially reduced by deploying randomized probabilistic matching bandits. Instead of pre-com-

mitting to a fixed sample size, sequential testing allows to rapidly detect large differences and not waste resources sticking to an unnecessarily rigid protocol. The optimization attempted by Bayesian bandits is much less concerned with the null hypothesis. Instead, its focus is on minimizing the posterior expected loss: the average amount we would lose by switching from A to B. The magnitude of a type-II error plays no role in frequentist hypothesis testing whereas for campaign optimization, it is clearly highly relevant. However, gains in the exploratory phase could be wiped out by slow losses during long running campaigns.

REFERENCES

Yu, E.C., Sprenger, A.M., Thomas, R.P., & Dougherty, M.R. (2014). When Decision Heuristics and Science Collide. *Psychonomic Bulletin & Review* 21 (2) (pp. 268–82). Maryland, USA: Springer.

Lotze, T., & Loecher, M. (2014). Bandit: Functions for Simple a/B Split Test and Multi-Armed Bandit Analysis. https://CRAN.R-project.org/package=bandit.

R Core Team. (2016). *R*: *A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved August 2017 from https://www.R-project.org/.

Robinson, D. (2015). *Is Bayesian a/B Testing Immune to Peeking? Not Exactly*. Retrieved August 2017 from http://varianceexplained.org/r/bayesian-ab-testing/.

Rouder, J. N. (2014). Optional Stopping: No Problem for Bayesians. *Psychonomic Bulletin & Review* 21 (2) (pp. 301–8). Missouri, USA: Springer

Sanborn, A. N, & Hills, T. T. (2014). The Frequentist Implications of Optional Stopping on Bayesian Hypothesis Tests. *Psychonomic Bulletin & Review* 21 (2) (pp. 283-300). Coventry, UK: Springer.

Scott, S. L. (2010). A Modern Bayesian Look at the Multi-Armed Bandit. *Applied Stochastic Models in Business and Industry* 26 (6) (pp. 639–58). Wiley Online Library

Google Analytics (2012). *Google Analytics Help Page*. Retrieved August 2017 from https://support.google.com/analytics/answer/2844870?hl=en.