# Practical Aspects of Sensitivity in Online Experimentation with User Engagement Metrics

**Conference Paper** · October 2015

DOI: 10.1145/2806416.2806496

CITATIONS
14

READS
498

**3 authors:**



Alexey V. Drutsa
Lomonosov Moscow State University
**29** PUBLICATIONS **147** CITATIONS
SEE PROFILE

Anna Ufland
University of Tartu
**2** PUBLICATIONS **14** CITATIONS
SEE PROFILE

Gleb Gusev
Yandex
**44** PUBLICATIONS **289** CITATIONS
SEE PROFILE



Short Remarks:

* We use different statistics in order to find treatment effect if it exists.
* To do so we apply stat-tests to each statistics separatly
* If choise of stat-test / p-value is inappropriate we, in some cases, will get higher FPR than desired.
  Thus given statistics is unreliable to detect treatment effect.
* Need to make some adjustments.

# Practical Aspects of Sensitivity in Online Experimentation with User Engagement Metrics

Alexey Drutsa
Yandex
Moscow, Russia
adrutsa@yandex.ru

Anna Ufland
Yandex
St.Petersburg, Russia
anna.uflyand@gmail.com

Gleb Gusev
Yandex
Moscow, Russia
gleb57@yandex-team.ru

## ABSTRACT

Online controlled experiments, e.g., A/B testing, is the state-of-the-art approach used by modern Internet companies to improve their services based on data-driven decisions. The most challenging problem is to define an appropriate online metric of user behavior, so-called Overall Evaluation Criterion (OEC), which is both interpretable and sensitive. A typical OEC consists of a key metric and an evaluation statistic. Sensitivity of an OEC to the treatment effect of an A/B test is measured by a statistical significance test. We introduce the notion of Overall Acceptance Criterion (OAC) that includes both the components of an OEC and a statistical significance test. While existing studies on A/B tests are mostly concentrated on the first component of an OAC, its key metric, we widely study the two latter ones by comparison of several statistics and several statistical tests with respect to user engagement metrics on hundreds of A/B experiments run on real users of Yandex. We discovered that the application of the state-of-the-art Student's t-tests to several main user engagement metrics may lead to an underestimation of the false-positive rate by an order of magnitude. We investigate both well-known and novel techniques to overcome this issue in practical settings. At last, we propose the entropy and the quantiles as novel OECs that reflect the diversity and extreme cases of user engagement.

**Categories and Subject Descriptions:** H.1.2 [User/Machine Systems]: Human information processing; H.5.2 [User interface]: Evaluation/methodology

**General Terms:** Measurement, Experimentation

**Keywords:** User engagement; online controlled experiment; Overall Acceptance Criterion; A/B test; sensitivity; quality metrics; evaluation statistic; significance level; p-value

## 1. INTRODUCTION

Nowadays, online controlled experiments, e.g., A/B testing, is the state-of-the-art approach utilized by modern Internet companies (such as web search engines [35, 6, 5], social networks [1], etc.) to improve their services based on data-driven decisions [21]. An A/B test compares two variants of a service at a time, usually its current version (control) and a new one (treatment), by exposing them to two groups of users. The aim of controlled experiments is to detect the causal effect of service updates on its performance relying on an *Overall Evaluation Criterion (OEC)* [22], a user behavior metric that is assumed to correlate with the quality of the service, i.e., the value of the OEC must have a clear *interpretation*. When the treatment effect exists, the OEC has to detect the difference of the two versions of the service at a high level of statistical significance in order to distinguish the treatment effect from the noise observed when the effect does not exist. This property is referred to as the *sensitivity* of the OEC [22].

A common OEC consists of two main components [22]: (a) a *key metric* (e.g., the number of sessions, the number of queries, etc.), calculated for each event (entity, *experimental unit*) of a certain type observed in the behavioral data (e.g., a query, a user, a session, etc); and (b) the *evaluation statistic* of the key metric over the experimental units (e.g., the average value, the median, etc.). A third (c) component of any A/B test is the *statistical significance test* (e.g., t-test, bootstrap, etc.), which examines weather the evaluation statistic of the key metric over the two groups of users coincide. We refer to these three components (a)–(c) as an *Overall Acceptance Criterion* (*OAC*).

The first two components (i.e., the OEC) are responsible for the interpretation of the OAC, and all of them directly affect its sensitivity. On the one hand, one important part of existing studies on interpretation of an OEC concerns only development or improvement of key metrics that correlate with different aspects of the system quality [12, 33, 10]. On the other hand, the large set of studies devoted to the sensitivity of an OEC relates either to experimental units (e.g., increasing of the experiment duration or of the user population participated in the experiment [22, 35, 5]) or to the key metric as well (e.g., variation reduction techniques [22, 7, 6], metric transformation [21, 11]). At the same time, to the best of our knowledge, the statistical significance test and the statistic used in an OAC are understudied areas in online evaluation of web services.

In the current study, we focus on the impact of the statistics and statistical tests chosen for an OAC on the sensitivity of different key metrics. First, we show that a key metric is most effective in combination with an appropriate statistical test, which is individual for each key metric. Utilization of the usual t-test with a default p-value threshold may lead to


*CIKM'15,* October 19–23, 2015, Melbourne, Australia.

DOI: http://dx.doi.org/10.1145/2806416.2806496.

wrong conclusions on the performance of some state-of-the-art key metrics. Second, we show that each combination of a key metric and a statistical test requires its individual p-value threshold to control the false-positive rate at a predefined level.

We conduct our study for the case of user engagement metrics, since they are often considered to be most appropriate for online evaluation. User engagement reflects how often the user solves her needs (e.g., to search something) by means of the considered service (e.g., a search engine). On the one hand, these metrics are measurable in the short-term experiment period, and, on the other hand, they are predictive of the long-term success of the company [19, 20, 21, 27]. In this study, we pay special attention to the metrics that reflect the loyalty aspect of engagement: the state-of-the-art *number of sessions per user* metric [19, 33] (which is accepted as the "North-star" for A/B testing evaluation in major search engine companies like Bing [20, 21]) and the recently proposed *absence time* metric [12].

The state-of-the-art evaluation statistic of the key metric over experimental units is the average value, which is used in vast majority of studies on A/B testing [4, 18, 35, 19, 33, 20, 21, 5, 10, 11]. Nonetheless, there are other statistics that have intelligible interpretation for user engagement metrics. The *median* is a popular statistic within statistical tests [34, 31] and it measures engagement of a typical user of the web service. Other *quantiles* quantify the extreme cases [14]: users whose behaviors are far from the mean (e.g., they are able to describe users that are engaged with the service less or more than an average user) [24, 16]. *Entropy* [30] measures the *diversity* of users against their engagement with a web service. Hence, these statistics could be utilized to detect the treatment effect in an evaluated update for such user behavior aspects as well[1]. We consider all these statistics in our OACs and compare them in terms of sensitivity. We found that each of the described above evaluation statistics noticeably outperforms the state-of-the-art mean statistic for some metrics in terms of the treatment detection rate with a fixed false-positive rate.

The widely applicable statistical significance test is the unpaired Student's t-test [7, 33, 6, 10, 11]. However, a key metric may not follow assumptions underlying this test (such as the normality of the metric's distribution or the independence of experimental units). This makes the statistical test inappropriate for such key metrics, and its utilization may lead to an underestimate of the false-positive rate. In our study, we show that application of the t-test to several different user engagement key metrics leads to the underestimation of this rate by an order of magnitude. Moreover, we demonstrate practical efficiency of both well-known and novel techniques (Bootstrapping, p-value adjustment, and others) to overcome this issue.

Different statistical significance tests were widely studied and compared in offline evaluation of information retrieval (IR) systems [28, 31, 38, 32, 2, 36, 29]. Unfortunately, statistical tests in offline evaluation are almost always paired. Hence, the results presented in these studies could not be straightforwardly applied to the case of A/B testing. In our study, we compare 5 statistical significance tests (Student's t-test, Bootstrapping, Mann Whitney U test, Tarone-Ware test, and Logrank test) and show how their application to different user engagement metrics results in different sensitivity of the OECs.

Despite the largest web services have designed special experimental platforms that allow them to run A/B tests at large scale (e.g., hundreds of concurrent experiments per day) [35, 20], to the best of our knowledge, the existing studies on A/B testing evaluate their approaches on dozens real online experiment runs (e.g., the largest numbers are 21 A/B tests in [3], 32 A/B tests and 18 A/A tests in [11]). In our study, we conduct our experimental analysis on 169 A/B experiments and 98 A/A ones run on real users of Yandex (www.yandex.com) with duration from one to several weeks. This should make the results of our study more valuable for practical use in modern web companies.

To sum up, our study considers the problem, which coincides with the *emerging Internet companies' needs*: to use more sensitive and interpretable OECs in online controlled experiments and to get their results on a proper false-positive level. Specifically, the major contributions of our work include:

- The practical comparison of 5 statistical significance tests (Student's t-test, Bootstrapping, Mann Whitney U test, Tarone-Ware test, and Logrank test) w.r.t. 16 commonly used engagement metrics and finding the most effective combinations of them.
- Investigation of the entropy and the quantiles as novel OECs that quantify the diversity and extreme cases of user engagement.
- Practical refinement and validation of several existing observations and pitfalls on the basis of hundreds of large-scale A/B tests run on real users of Yandex.

The rest of the paper is organized as follows. In Section 2, the related work on A/B testing and user engagement is discussed. In Section 3, A/B testing background is described and our key metrics, evaluation statistics and statistical significance tests are presented. In Section 4, we show and discuss the results of applying OACs to the set of A/B experiments. In Section 5, the study's conclusions and our plans for the future work are presented.

## 2. RELATED WORK

We compare our research with other studies in three aspects. The first one relates to the common methodology of conducting online controlled experiments. The second one concerns user engagement metrics used in web services and, particularly, search engines. The third one relates to statistical tests used to measure significance level in information retrieval evaluation.

**Online controlled experiment studies.** The theoretical details of the online controlled experiment methodology were widely described in the existing works [26, 22], and we conduct our experiments in accordance with them. A number of practical lessons learned from the applications of this methodology to different evaluation cases in many web companies was recently described in [18, 35, 20]. These studies concern, inter alia, experiments with different components of a web service (e.g., the user interface [18, 10, 25] and ranking algorithms [33, 10, 25]), large-scale experimental infrastructure [35, 20], different aspects of user behavior and interaction with a web service (clicks [19, 21], speed [21], abandonment [21], periodicity [10, 9]), and so on. These experiments show that system updates of different types may

[1]Some similar statistics were utilized in risk-sensitive optimization and offline evaluation of retrieval systems [37, 8].

affect various key metrics differently. Some of the existing works focused on the study of the trustworthiness of the results of an A/B test. Various pitfalls and puzzling outcomes of online controlled experimentation were shared in [4, 19] and several "rules of thumb" were discussed in [21].

Some other studies focused on sensitivity improvement techniques for an online controlled experiment. The simplest ones [22] include increasing of the experimental time period or expanding the user population participating in the experiment; investigation of evaluation metric with lower variance [10, 9]; and elimination of users who were not affected by the service change in the treatment group [33]. More sophisticated sensitivity improvement techniques include the stratification and the usage of control variates that are defined on the basis of pre-experiment data [7]; the reduction of skewness of the evaluation metric [21] (e.g., transformation of the metric or capping its values); the sensitivity improvement for two-stage online controlled experiments [6]; the future user behavior prediction approach [11]; and diluted treatment effect estimation for trigger analysis [5]. To the best of our knowledge, the existing studies on A/B testing evaluated their approaches on dozens real online experiment runs (e.g., maximal are 21 A/B tests in [3], 32 A/B tests and 18 A/A tests in [11]). In our study, we experiment with a diverse and huge set of system changes: 169 large-scale A/B tests and 98 A/A experiments based on actual interactions of hundreds of thousands of real users (we use no artificially simulated data).

This allowed our study to produce more essential results on the relative sensitivity of different metrics. We believe, these results are more valuable in practice than the previous studies did, since the largest modern web companies run up to hundreds of concurrent A/B tests per day [35, 20].

**Evaluation metrics.** *User engagement metrics* are popular in A/B testing practice of many companies, because user engagement reflects how often the user solves her needs (e.g., to search something) by means of the considered service (e.g., a search engine) [19, 20, 21]. Hence, on the one hand, these metrics are measurable in the short-term experiment period, and, on the other hand, they are predictive of long-term goals of the company [19, 20, 21]. The most well-known metrics of loyalty aspect of user engagement are the state-of-the-art *number of sessions per user* [33] and the *absence time* [12] metrics. There are also several widely used metrics of activity aspect of user engagement like the number of clicks per user, shows per user, queries per user [22, 7], etc. All these metrics were utilized to evaluate different changes in search engines [3, 10, 11, 9]. However, to the best of our knowledge, these metrics were not thoroughly compared on the same data: e.g., the number of sessions and the absence time were studied together in [11], but their absence time is a measure over users as experimental units, while its original definition [12] considers it as a measure over periods of absence between consecutive sessions of a user (as in [3]). In our work, we compare sensitivity of these and other main user engagement metrics (as in [33, 12, 3, 11]) in Overall Acceptance Criteria (OACs) with different statistics (the mean, the median, etc.) and statistical tests (T-test, bootstrap, etc.).

**Statistical tests in information retrieval.** Different statistical significance tests were widely studied and compared in *offline evaluation* of information retrieval (IR) systems [28, 31, 38, 32, 2, 36, 29]. Unfortunately, statistical tests in offline evaluation are almost always paired, since the systems are compared using the same test data set. Hence, the results presented in these studies could not be straightforwardly adopted to the case of A/B testing, where two systems are compared using *different* sets of users, and thus unpaired tests are needed. Our study addresses this understudied area of *online evaluation* of web services w.r.t. user engagement metrics. We compare the Student's t-test, which is the state-of-the-art in A/B testing [7, 33, 6, 10, 11]; Mann Whitney U, Tarone-Ware, Logrank tests, which are widely used in survival analysis [34]; and Bootstrapping whose different variants (including unpaired ones) were shown to be very useful for both evaluation of IR systems and for comparing the sensitivity of IR metrics (like nDCG or Q-measure) [28, 31, 32, 36]. Bootstrapping was applied to A/B tests in [1] and was also mentioned in [4, 21] as a method to estimate the variance, when the key metric in A/B testing is very skewed or its experimental unit does not coincide with the randomization one.

**Table 1: User engagement metrics and corresponding experimental units used in our study.**

| UE metric | Exp. unit | Description |
|---|---|---|
| **Sessions** | | |
| S | user | the number of sessions of a user |
| sS | "user-2" | the number of sessions of a user |
| **Absence time** | | |
| AT | absence | the duration of an absence in seconds |
| logAT | absence | the logarithm of the duration of an absence |
| ATpA | "user-2" | the average duration of a user's absences |
| log(ATpA) | "user-2" | the log. of the aver. duration of a user's absences |
| log(AT)pA | "user-2" | the average log. of duration of a user's absences |
| **Queries** | | |
| Q | user | the number of queries of a user |
| logQ | user | the logarithm of the number of queries of a user |
| **Clicks** | | |
| C | user | the number of clicks of a user |
| logC | user | the logarithm of the clicks of a user |
| **Clicks per query** | | |
| CpQ | user | the average number of clicks of a user's queries |
| **Presence time** | | |
| PT | session | the duration of a session in seconds |
| logPT | session | the logarithm of the duration of a session |
| sumPT | user | the sum of durations of a user's sessions |
| logsumPT | user | the logarithm of the presence time of a user |

# 3. FRAMEWORK

## 3.1 A/B testing background

In A/B testing, users participated in an experiment, are randomly exposed to one of the two variants of a service (the control (A) and the treatment (B), e.g., the current production version of the service and its update) in order to compare their performance [22, 19, 21]. The difference between the variants is quantitatively measured by an *Overall Evaluation Criterion* $\xi$ (*OEC*, also known as the evaluation metric, the online service quality metric, etc. [22]). In the classical methodology of A/B testing, this OEC is usually an *evaluation statistic* $\xi = \xi_\Omega(\mathtt{X})$ (e.g., *the average value*) of a *key metric* $\mathtt{X}(\omega)$ over the *events* (*entities*) $\omega \in \Omega$. More precisely, for each user group $v \in \{A, B\}$, we have the observations of the metric $\mathtt{X}$ over the experimental units $\Omega_v$. Then, the evaluation statistics $\xi_v = \xi_{\Omega_v}(\mathtt{X})$, $v \in \{A, B\}$, are

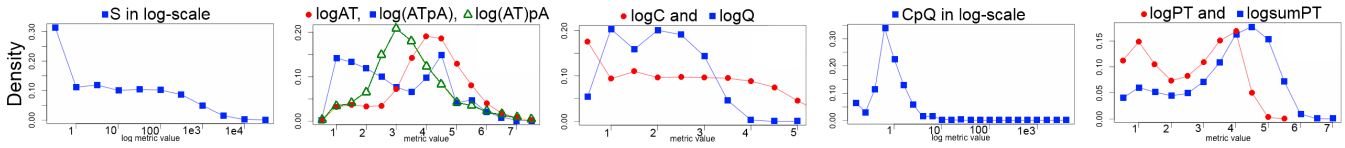


**Figure 1: The distributions of 9 representative key metrics over their experimental units.**

used as OEC, and their difference $\Delta = \xi_B - \xi_A$ is calculated to quantify the sign and the magnitude of the change in the OEC.

Nonetheless, the quantity $\Delta$ could not serve itself as an indicator of positive or negative consequences of the evaluated changes of the service. The difference between the evaluation metrics over groups should be controlled by a statistical significance test (e.g., Student's t-test) that calculates the probability (also known as *p-value*) to observe the difference under *the null hypothesis*, which assumes that the observed difference is caused by random fluctuations, and the variants of the system are not actually different in terms of user experience. If the p-value is lower than the threshold $p_{\text{val}} \leq \alpha$ ($\alpha = 0.05$ is commonly used [22, 19, 7, 33, 21, 10, 11]), then the test rejects the null hypothesis, and the difference $\Delta$ is accepted as statistically significant. We refer to the triplet (a key metric, an evaluation statistic, a statistical test) as an *Overall Acceptance Criterion* (*OAC*). The additional details of the A/B testing framework could be found in the survey and practical guide [22]. All key metrics, evaluation statistics, and statistical significance tests studied in this paper are specified further in the next subsections.

## 3.2 User engagement metrics

In this work, we study several popular engagement metrics and their modifications, which are calculated based on actions of users[2] of Yandex, one of the most popular global search engines. Following common practice [15, 19, 12, 33, 3, 10, 11], we define a *session* as a sequence of user actions whose *dwell times*[3] are less than 30 minutes. The *number of sessions* `S` (as in [33, 10, 11]) for each user in an A/B experiment is our first metric. In our study, we also consider this metric for the reduced set of users that have at least 2 sessions during the experiment (such experimental unit is referred to as "user-2"). The time between two consecutive user sessions (i.e., the duration of an absence) is the *absence time* (as in [12, 3]), which is referred to as `AT` and is calculated for each pair of consecutive sessions of each user. The experimental unit for `AT` is the *period of absence* between a pair of sessions, and, in order to match absence time to a user (as an experimental unit), we also consider the average duration of user's absences `ATpA` for each "user-2"[4]. We logarithmically transform[5] these metrics and obtain `logAT`, `log(ATpA)`, and `log(AT)pA`.

All above described metrics represent the *loyalty aspect* of user engagement, whereas, further, we present metrics that are associated with the *user activity* [27]. They are: the number of queries `Q` [10, 11], the number of clicks `C` [10, 11], and the number of clicks per query `CpQ` (i.e., the CTR of the search engine result pages [11]) for each user. The presence time is considered both for each user's session (i.e., its duration `PT` in seconds) and for each user (i.e., the sum of the durations of her sessions `sumPT` [10, 11]). We logarithmically transform these metrics and obtain `logQ`, `logC`, `logPT`, and `logsumPT`. Thereby, we study 16 key user engagement metrics in total, whose descriptions and corresponding experimentation units are summarized in Table 1. The shape of distributions of 9 representative key metrics over their experimental units are shown in Fig. 1 (their values are given in arbitrary units for confidentiality reasons). Short analysis of their persistence across time and relationships between almost all of them are studied in [10, 11].

## 3.3 Evaluation statistics

**Mean and standard deviation.** The average value[6] (AVG) is the state-of-the-art evaluation statistic of the key metric over experimental units. The vast majority of studies on A/B testing [4, 18, 35, 19, 33, 20, 21, 5, 10, 11, 9] utilize it, including the fact that the mean describes the typical behavior of a web service user (an "average" user). The second commonly used statistic in probability theory is the standard deviation (SD). It measures how the key metric `X` vary across the experimental units $\Omega$. In many studies on A/B testing [4, 7, 5], the standard deviation is assumed to be equal for the control variant and the treatment one, when a statistical test is applied. We consider this statistic in our analysis and show that this statement is not valid for some metrics (see Section 4).

**Median and quantiles.** The $\gamma$-*quantile* (also known as the $\gamma$-th population quantile [14]) is defined as the value $q_\gamma$ such that $\mathbb{P}(\mathtt{X} \leq q_\gamma) \geq \gamma$ and $\mathbb{P}(\mathtt{X} > q_\gamma) \leq 1 - \gamma$, $\gamma \in [0, 1]$, where $\mathbb{P}(\mathtt{X})$ is the distribution of the key metric `X` over the experimental units $\Omega$. For a finite sample of observations of the key metric, we use the traditional estimator of the $\gamma$-quantile, presented in the equation (1) in [14][7]. In our work, we have considered 19 different $\gamma$-quantiles with $\gamma = 0.05i$, $i = 1, \ldots, 19$, but, due to the space constraints, we present the results only for the 5 most representative ones. Their $\gamma$ are 0.5 (*the median*), 0.25 (the first quantile), 0.75 (the third quantile), 0.05, and 0.95.

The median is one of the popular statistics: its meaning is similar to the one of the mean value AVG, but the median is better suited for skewed distributions to quantify behavior of a typical user since it is much more robust and sensible (the mean value is highly influenced by outliers). There are a lot of paired [31] and unpaired [34] statistical tests specialized on the median as well. Contrariwise, a quantile with $|\gamma - 0.5| \gg 0$ could be useful in measuring the extreme cases [24, 16]: it quantifies user behaviors that are far from the mean.

[2]Users are identified by means of Cookie IDs [22] as done in other studies on user engagement [12, 23, 33, 10, 11, 9].
[3]i.e., times between two consecutive actions of a user [10]
[4]A user with only 1 session has no absence times `AT` by the definition above.
[5]It should increase the power of OACs if the key metric has a skewed distribution [17].
[6]A "rate" OEC, which is a percent of experimental units with some properties over all experimental units (e.g., the percent of users who click on a link) could be formalized via averaging as well [4].
[7]Subtler estimators (like in [14]) could be also considered, we left it for the future work.

**Table 2: Comparison of OACs w.r.t. the number of A/B and A/A experiments with detected treatment effect with the constant threshold $\alpha = 0.05$.**

| $\alpha =$ **0.05** | Success rate \| False-positive rate [# of the 169 A/B tests] \| [# of the 98 A/A tests] | | | | |
|---|---|---|---|---|---|
| Stat.test: | T-test | BS-AVG | MW test | TW test | LR test |
| S | 18 \| 2 | 23 \| 2 | 15 \| 5 | 21 \| 3 | 24 \| 2 |
| sS | 20 \| 2 | 20 \| 2 | 17 \| 2 | 19 \| 1 | 22 \| 2 |
| AT | 63 \| 20 | 20 \| 8 | 74 \| 27 | 77 \| 27 | 76 \| 24 |
| logAT | 72 \| 26 | 19 \| 9 | 74 \| 27 | 77 \| 27 | 76 \| 24 |
| ATpA | 11 \| 4 | 16 \| 5 | 24 \| 3 | 24 \| 3 | 20 \| 4 |
| log(ATpA) | 20 \| 3 | **28** \| 3 | 24 \| 3 | 24 \| 3 | 20 \| 4 |
| log(AT)pA | 22 \| 8 | 29 \| 7 | 29 \| 6 | 26 \| 9 | 26 \| 9 |
| Q | 34 \| 3 | 41 \| 4 | 35 \| 4 | 39 \| 5 | 38 \| 2 |
| logQ | 34 \| 5 | 39 \| 4 | 35 \| 4 | 39 \| 5 | 38 \| 2 |
| C | 52 \| 3 | **58** \| 3 | 46 \| 3 | 54 \| 3 | 57 \| 2 |
| logC | 49 \| 3 | 53 \| 3 | 46 \| 3 | 54 \| 3 | 57 \| 2 |
| CpQ | 80 \| 9 | 81 \| 9 | 96 \| 10 | 94 \| 8 | 90 \| 8 |
| PT | 70 \| 22 | 27 \| 3 | 82 \| 26 | 80 \| 23 | 79 \| 24 |
| logPT | 80 \| 24 | 50 \| 5 | 82 \| 26 | 80 \| 23 | 79 \| 24 |
| sumPT | 26 \| 3 | 34 \| 3 | 30 \| 3 | 34 \| 1 | 36 \| 1 |
| logsumPT | 27 \| 2 | 27 \| 3 | 30 \| 3 | 34 \| 1 | 36 \| 1 |

**Table 3: Comparison of OACs w.r.t. the number of A/B and A/A experiments with detected treatment effect with the constant threshold $\alpha = 0.01$.**

| $\alpha =$ **0.01** | Success rate \| False-positive rate [# of the 169 A/B tests] \| [# of the 98 A/A tests] | | | | |
|---|---|---|---|---|---|
| Stat.test: | T-test | BS-AVG | MW test | TW test | LR test |
| S | 11 \| 0 | **18** \| 0 | 8 \| 0 | 14 \| 0 | 15 \| 0 |
| sS | **18** \| 0 | 16 \| 0 | 12 \| 0 | 13 \| 0 | 17 \| 0 |
| AT | 43 \| 12 | 9 \| 0 | 49 \| 14 | 55 \| 15 | 56 \| 15 |
| logAT | 51 \| 14 | 10 \| 2 | 49 \| 14 | 55 \| 15 | 56 \| 15 |
| ATpA | 6 \| 0 | 11 \| 1 | 13 \| 0 | 13 \| 0 | 13 \| 0 |
| log(ATpA) | 8 \| 0 | 15 \| 0 | 13 \| 0 | 13 \| 0 | 13 \| 0 |
| log(AT)pA | 8 \| 0 | 14 \| 2 | 13 \| 0 | 14 \| 0 | 13 \| 0 |
| Q | 19 \| 0 | 27 \| 1 | 20 \| 1 | 24 \| 0 | 28 \| 0 |
| logQ | 19 \| 0 | 20 \| 0 | 20 \| 1 | 24 \| 0 | 28 \| 0 |
| C | 35 \| 0 | 38 \| 0 | 38 \| 0 | 44 \| 0 | **49** \| 0 |
| logC | 37 \| 0 | 41 \| 0 | 38 \| 0 | 44 \| 0 | **49** \| 0 |
| CpQ | 73 \| 3 | 73 \| 2 | 78 \| 2 | 77 \| 4 | 80 \| 4 |
| PT | 53 \| 14 | 13 \| 1 | 70 \| 13 | 63 \| 11 | 61 \| 15 |
| logPT | 69 \| 13 | 32 \| 2 | 70 \| 13 | 63 \| 11 | 61 \| 15 |
| sumPT | 12 \| 1 | 20 \| 1 | 16 \| 0 | 18 \| 0 | 19 \| 0 |
| logsumPT | 21 \| 1 | 21 \| 0 | 16 \| 0 | 18 \| 0 | 19 \| 0 |

Hence, such quantile may help a web service to detect the treatment effect of an evaluated update on users either with lower, or higher engagement than an "average" user (e.g., the 0.25-quantile of the number of sessions S describes users that use the service rarely, while the 0.75-quantile of it quantifies frequent users). This may be definitely important for web companies that fight for new, rare users in a competitive environment or choose to preserve and increase engagement of loyal, regular users. The $\gamma$-quantile is referred to as Q-$\gamma$.

**Entropy.** The *entropy* statistic (also known as the Shannon entropy) is defined as the expectation $\mathrm{E}(\mathtt{X}) = \mathbb{E}(\mathrm{I}(\mathtt{X}))$ of the information $\mathrm{I}(\mathtt{X}) = -\ln \mathbb{P}(\mathtt{X})$ [30]. Entropy is a measure of "chaos" or unpredictability of information content. In our case, entropy is a measure of *diversity* of user engagement with a web service w.r.t. the metric X over $\Omega$ (e.g., it quantifies the diversity of users w.r.t. their number of sessions for $\mathtt{X} = \mathtt{S}$). Hence, the entropy statistic may help the web service to detect the treatment effect of an evaluated update in terms of such diversity: whether the update increases or decreases the variety of different types of user behavior. *We propose to use the entropy as a novel OEC in A/B testing.*

## 3.4 Statistical significance tests

**Unpaired two-sample Student's t-test.** The *Student's t-test* (T-test) is the most popular statistical test for comparing the mean values of unpaired data samples, and, hence, it is the widely used one in A/B testing [7, 33, 6, 10, 11]. Its popularity could be explained, first, by its computational costs. Indeed, this test is based on *t-statistic*:

$$\frac{\mathrm{AVG}_B(\mathtt{X}) - \mathrm{AVG}_A(\mathtt{X})}{\sqrt{\mathrm{SD}^2_A(\mathtt{X}) \cdot |\Omega_A|^{-1} + \mathrm{SD}^2_B(\mathtt{X}) \cdot |\Omega_B|^{-1}}}$$

whose calculation takes $3n + o(n)$ arithmetic operations, where $n = \max\{|\Omega_A|, |\Omega_B|\}$. Second, despite the normality of the key metric's distribution is one of the assumptions underlying the test, it may be applicable in practice to some key metrics that do not follow a normal distribution even approximately. Further in our study, our experimental results confirm applicability of the statistical test in the case of such metrics. On the other hand, we show that violation of independence of experimental units (which is another assumption underlying the test) results in inapplicability of the T-test due to underestimation of its false-positive rate.

**Mann–Whitney, Tarone–Ware, and Logrank tests.** This family of rank-based statistical tests is widely used in survival analysis [34] and is based on the $\chi^2_{FH}$ statistic:

$$\chi^2_{FH} = \frac{\left(\sum_{j=1}^{K} \omega_j (d_{j,B} r_j - r_{j,B} d_j) r_j^{-1}\right)^2}{\sum_{j=1}^{K} \omega_j^2 r_{j,B} r_{j,A} d_j (r_j - d_j) r_j^{-2} (r_j - 1)^{-1}},$$

where the values of $r_{j,A}, r_{j,B}, r_j, d_{j,A}, d_{j,B}, d_j$, and $K$ are calculated in the following way. Given two samples of the key metrics $\{X_\omega\}_{\omega \in \Omega_v}$, $v = A, B$, we merge them into one increasing sequence of values. Among these values, there are $K$ unique: $y_1 < \ldots < y_K$, $K \le |\Omega_A| + |\Omega_B|$. Then, for each value $j = 1, \ldots, K$, the following values are computed: $d_{j,v}$ is the number of observations equal to $y_j$, and $r_{j,v}$ is the number of observations not less than $y_j$ in the sample $v = A, B$. The values $d_j$ and $r_j$ are defined analogously using the union of the two samples.

For $\omega_j = r_j$, we get the Gehan tests, which is equal to the *Mann–Whitney U test* (MW test), also known as the *Wilcoxon rank-sum test*; for $\omega_j = \sqrt{r_j}$, we get the *Tarone–Ware test* (TW test); and, for $\omega_j = 1$, we get the *Logrank test*[8] (LR test). Calculation of $\chi^2_{FH}$ usually takes $O(n \log n)$ operations, since one needs to sort the samples $\Omega_A$ and $\Omega_B$ by the values of the key metric. These tests could be considered as tests of medians with the additional assumption that the distributions for the samples have the same shape.

**Bootstrapping.** As opposed to the previous ones, this test can be applied to any evaluation statistic $\xi$. Therefore, the *Bootstrap test* is referred to as BS-$\xi$ with the statistic $\xi$ whose null hypothesis is tested (e.g., BS-AVG means that the Bootstrap test is applied to the mean values of the samples). We use the Bootstrap test as described in [13,

[8]The Logrank test is one of the main statistical tests used [34] in the Cox Model, which was applied to the absence time metric in [12].

The only requirement:
analysis unit = randomization unit.

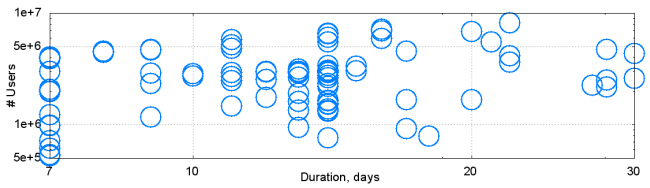


**Figure 2: The distribution of 169 A/B tests w.r.t. their duration and the size of their sample of users.**

Alg. 16.1] with *resampling of users*[9]. This statistical test is the most computational expensive among test we consider, since it takes $O(NB)$ operations, where $N$ is the number of operations needed to calculate the statistic $\xi$ (e.g., $N = O(n)$ for the mean) and $B$ is the number of bootstrap iterations. In our study, $B = 1000$ as in [28, 25].

Summarizing, *we study 16 key user engagement metrics and apply* 26 *evaluation-statistic–statistical-test pairs: 2 statistical tests (T-test and BS) for AVG, 4 tests (MW test, TW test, LR test, and BS) for Q-0.5 (the median), 1 test (BS) for all other considered statistics.*

# 4. EXPERIMENTATION

## 4.1 Set of our A/B tests

In our paper, we consider 169 large-scale A/B experiments conducted on *real users* of Yandex with duration from 7 to 30 days. The user samples used in our A/B tests are all uniformly randomly selected, and the control and the treatment groups are of approximately the same size (at least, hundreds of thousands of users). The distribution of these 169 A/B tests w.r.t. their duration and the size of their sample of users is presented in Figure 2. Each experiment evaluates a change in one of the main components of the search engine (including the ranking algorithm, the user interface, the server efficiency, etc). Each of these changes is either an artificial deterioration of a component[10] [20], or its update, which is evaluated before being shipped. In our experimentation, we also consider 98 control experiments: 70 1-week and 28 2-week so-called A/A tests that compare the same versions of the service [22, 4]. Both A/B and A/A tests were conducted during the period from January, 2013 to November, 2014.

Additionally, we generate several thousands of synthetic A/A experiments (like in [1, 7]) by randomly splitting users from one of our control experiments. We find that the results for these synthetic A/A experiments (p-value distributions, false-positive rares, etc.) are similar to the ones for real A/A experiments. Therefore, due to the space constraints, we present the results only for our 98 real control experiments.

## 4.2 Success and false-positive rates

The straightforward way to compare sensitivity of OACs is to apply them to a series of A/B experiments and compare the number of A/B tests whose treatment effect is detected by each OAC for a selected p-value threshold $p_{\text{val}} \leq \alpha$ (as in [10, 11, 9, 25]). This number is also referred to as the *success*

[9]Thus, we correctly bootstrap dependent data when the experimental unit of a considered key metric is not a user [1], e.g., in the case of PT, two sessions in a sample can be of the same user, and, thus, their durations are dependent.

[10]like the swap of the second and the fourth results in the ranked list formed by the current ranking [10, 25].

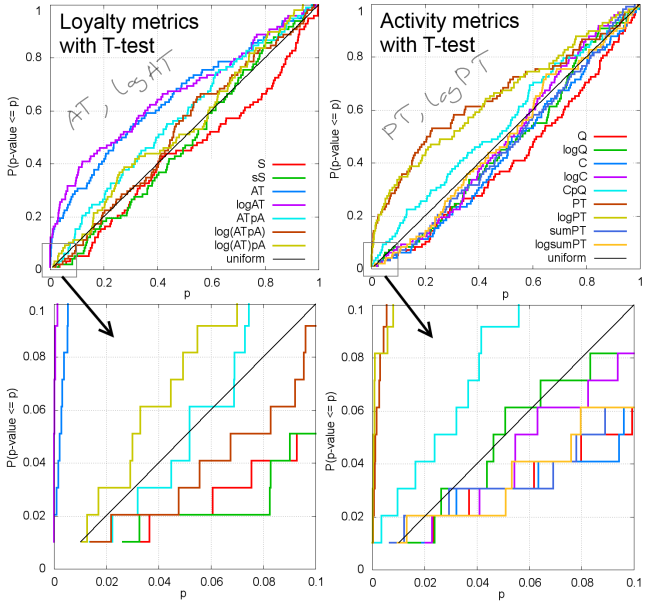


**Figure 3: The CDFs of the p-value of the T-test for all loyalty and activity metrics over 98 A/A tests.**

*rate* in IR offline evaluation [36]. The threshold $\alpha = 0.05$ is commonly used in existing studies on A/B testing [22, 19, 7, 33, 21, 10, 11]. However, smaller thresholds are also considered (e.g., $\alpha = 0.01, 0.001, 0.0001$) in order to focus on strong signals [20], since this is aligned with emerging needs of modern web companies to be more confident in the effect of evaluated updates on the quality of their services. But this is a double-edged sword for OAC comparison as well, since the lower the threshold $\alpha$ (i.e., the lower the false-positive rate), the lower the success rate. We expect that comparison of OACs at different p-value thresholds may lead to different results. Hence, in order to reveal it, we present our results both for the state-of-the-art threshold $\alpha = 0.05$, and for the stronger one $\alpha = 0.01$.

We compare all studied key metrics (7 loyalty and 9 activity ones) applied with 5 main statistical tests (T-test, Bootstrap test for AVG, MW test, TW test, and LR test) with respect to the success rate of their OACs in Table 2 and Table 3 (the first number in a cell) for the *constant threshold* $\alpha = 0.05$ and 0.01 respectively. We see that the success rates of the key metrics AT and logAT[11] with the Tarone–Ware test for $\alpha = 0.05$ (77 detected treatments in 169 A/B tests, i.e., 46%) and with the Logrank test (33%) for $\alpha = 0.01$ are the highest ones among all loyalty OACs, moreover, the success rates of the OACs of these metrics with T-test, MW test, TW test, and LR test dominate among other loyalty OACs by a large margin. The activity metrics CpQ, PT, and logPT (applied with almost all presented statistical tests) dominate other OACs w.r.t. the success rate both for $\alpha = 0.05$ and 0.01 (with the success rates up to 57% and 47% respectively).

In A/B testing, correctness of an experimentation is verified by control experiments (i.e., A/A tests) [22, 4]. Each of them compares two identical versions of the service and

[11]Note that logarithmic transformation is monotone, and, hence, the p-value results for the non-transformed and log-transformed variants of the same key metric for one of MW, TW, and LR tests are the same.

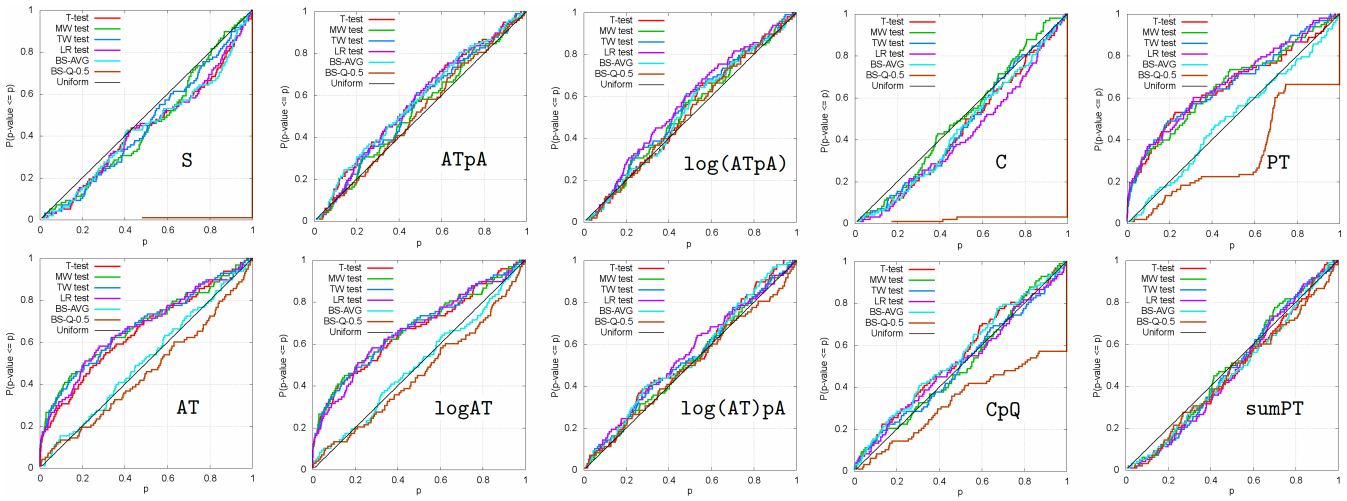


**Figure 4: Comparison of the CDFs of the p-value for different statistical test of the mean and the median values over 98 A/A tests.**

should be failed (i.e, the treatment effect is wrongly detected) in not more than $\alpha \cdot 100\%$ of cases (e.g., 5% for $\alpha = 0.05$), if the experimentation platform is correct and the OAC is valid. The number of failed A/A tests is referred to as the *false-positive rate* (also known as the type I error) and is reported as the second number in the cell for each OACs in Tables 2, 3. The cells of OACs that have inappropriate false-positive rates ($> 4$ and $> 0$ for $\alpha = 0.05$ and 0.01, respectively) are highlighted in red color. Such OACs should be rejected in the classical approach, since their statistical tests underestimate the standard deviation of their OEC [4]. According to this, all the previously top-rated key metrics `AT`, `logAT`, `CpQ`, `PT`, and `logPT` are rejected. Hence, *we conclude that the metrics* `log(ATpA)` *and* `C` *for* $\alpha = 0.05$ *(*`S`*,* `sS`*,* `C` *and* `logC` *for* $\alpha = 0.01$*) with suitable statistical tests (their cells are highlighted in* **boldface***) are the best ones w.r.t. the success rate among OACs with a valid false-positive rate.*

## 4.3 P-value distributions

In order to understand the dramatic underestimation of the false-positive rate of several OACs by their p-values (see the previous subsection), we consider the empirical Cumulative Distribution Function (CDF) of the p-value of an OAC over the 98 A/A experiments. If a key metric's distribution satisfies the assumptions that underlay a statistical test, then the p-value of this OAC should be uniformly distributed over $[0, 1]$ [22, 4].

We present the empirical CDFs of the p-value of the T-test for all loyalty and activity metrics in Fig. 3. We see that only the CDFs of the key metrics `AT`, `logAT`, `PT`, and `logPT` are noticeably higher than the uniform CDF, and this becomes even more apparent near 0, where commonly used p-value thresholds ($\alpha = 0.01$ and 0.05) are situated (see also Tables 2 and 3). These metrics are united by the fact that their experimental units do not coincide with the randomization one, i.e., a user (see Table 1). This finding is in line with the observations made in [4, 35], where Boostrapping and Delta methods are suggested to estimate the variance of non-per-user metrics.

We compare the CDFs of the p-value of the 6 statistical tests (T-test, MW test, TW test, LR test, BS-AVG, and BS-Q-0.5) that are utilized for the mean and the median of some representative engagement metrics[12] over 98 A/A experiments (see Fig. 4). We see that the observation made above for the metrics with non-user experimental units holds for all considered statistical tests except for the Bootstrapping for the mean value (BS-AVG), whose CDF is approximately uniform for all these metrics. Note that their analogs that are mapped to a user (i.e., the metrics `ATpA`, `log(ATpA)`, `log(AT)pA`, `sumPT`, and `logsumPT`) do not have such convexity in their CDFs, and this results in noticeably lower false-positive rates of their OACs (see Tables 2 and 3). Therefore, we conclude that, *for a non-per-user metric, its transformation to a per-user one or utilization of the Bootstrap test noticeably reduce the dramatic underestimation of the false-positive rate (up to the correct estimation).*

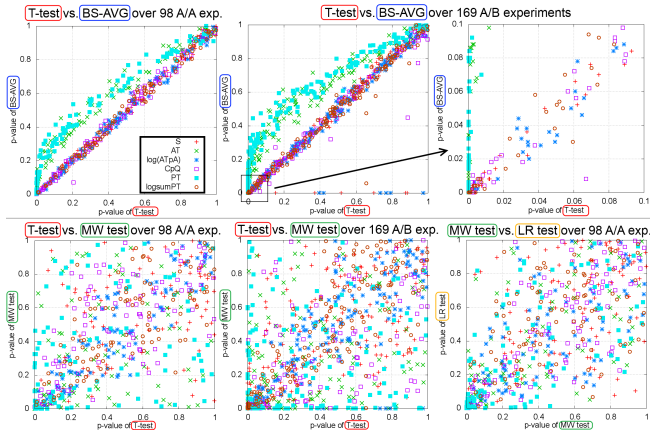


**Figure 5: The joint distributions of our A/A and A/B experiments w.r.t. the pairs of the p-values of two statistical tests for some representative engagement metrics.**

From Fig.3 and 4, we also make an interesting observation: the metrics `S`, `sS`, `Q`, `C`, `sumPT`, and `logsumPT` have a noticeable overestimation of false-positive rate. This observation prompts that their real success rates for the real

[12]From here on in this paper, due to the space constraints, we present results only for the best or representative metrics, statistics, and statistical tests, and we pay more attention to the loyalty aspect of user engagement.

One more proof that analysis unit = session gives smaller Pvals than it should

PT, AT - T-test in such case

Because of violated assumptions of T-test.

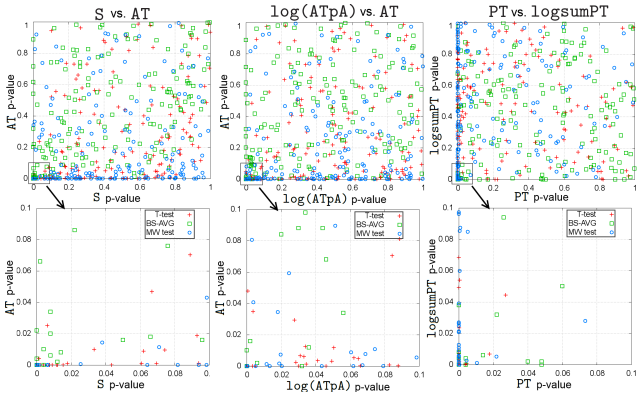


**Figure 6: The joint distributions of our A/B experiments w.r.t. the pairs of the p-values of statistical tests for two engagement metrics.**

false-positive rates of 5% or 1% are greater than the ones presented in Tables 2 and 3 (see the next subsection).

We also find that there is a strong relationship between the p-values of the Student's t-test and the Bootstrap test for the mean (BS-AVG) for all metrics, including the metrics with significantly skewed CDFs like `AT` or `PT`. The relationship of the tests is not linear for such metrics and is linear for the other metrics (see the top of Fig. 5). The similarity between the Student's t-test and bootstrapping is also observed for IR metrics in [32, 36]. The relation for each of other pairs of statistical tests (see examples of joint distributions in the bottom of Fig. 5) is noticeably more chaotic than the ones described above. There is no relation between the p-value of the same test applied to different metrics as well (e.g., see Fig. 6). The relation of T-test and BS-AVG infers that the false-positive level of an OAC could be improved by utilization of the latter one instead of the former one if the metric is skewed, since the Bootstrapping has approximately uniform distribution of its p-value for such metrics.

## 4.4 P-value adjustment

Motivated by the observations made in the previous subsection, we estimate the p-value threshold $\alpha$ for each OAC individually, such that its real false-positive rate is 5 failed A/A tests of the 98 ones from our set of experiments. This estimator equals to the 5-th smaller p-value from the observations. It is the traditional estimator of the 5/98-quantile, presented in the equation (1) in [14]. The subtler estimators from this study could also be considered, but we left it for the future work. The obtained estimations of p-value thresholds are referred to as $\alpha_{5:98} = \alpha_{5:98}(\text{OAC})$ and are presented in Table 4. The highest values for each row are highlighted in **boldface**, the lowest ones are underlined, the highest and the lowest values for each column both for loyalty and activity metrics are highlighted in green color and blue color, respectively. For instance, we see that the Mann–Whitney U test is the frequent one, whose false-positive rate has the worst underestimation, while the Logrank test (from the same family as MW test), contrariwise, frequently shows the highest overestimation for the activity metrics.

We utilize the individually adjusted thresholds $\alpha = \alpha_{5:98}$ to get the success rates of our OACs, that are presented in Table 5. Comparing these results with the ones from Table 2, we see that, on the one hand, the previously top-rated absence-time metrics (e.g., `AT` and `ATpA`) have now moderate success rates (similar to or lower than the ones of other loyalty metrics). On the other hand, the metrics `S` and `sS`, whose false-positive rates have been overestimated, demonstrate now higher success rates and the best ones among the loyalty metrics (for all statistical tests except the MW one). For activity metrics, we observe decay in the success rates of `PT` and `logPT`, while `CpQ` remains the leadership. Hence, we conclude that *the adjustment of the p-value threshold to a desired false-positive rate could significantly change the state of affairs in evaluation of Overall Acceptance Criteria: the success rates of ones can noticeably increase, other OACs can become usable in evaluation of a web service, since their false-positive rates reduce to an acceptable level.*

**Table 4: The thresholds $\alpha = \alpha_{5:98}$ for each OAC.**

| Stat.test: | T-test | BS-AVG | MW test | TW test | LR test |
|---|---|---|---|---|---|
| `S` | **0.1439** | 0.126 | 0.0726 | 0.1007 | 0.1144 |
| `sS` | 0.1004 | 0.098 | 0.0747 | 0.1025 | **0.1105** |
| `AT` | 0.0029 | **0.04** | 0.0002 | 0.0003 | 0.0005 |
| `logAT` | 0.0002 | **0.022** | 0.0002 | 0.0003 | 0.0005 |
| `ATpA` | 0.0519 | 0.05 | **0.0791** | 0.0729 | 0.0702 |
| `log(ATpA)` | **0.0828** | 0.082 | 0.0791 | 0.0729 | 0.0702 |
| `log(AT)pA` | 0.033 | 0.03 | **0.0498** | 0.0362 | 0.0333 |
| `Q` | 0.0993 | 0.06 | 0.0537 | 0.0503 | **0.138** |
| `logQ` | 0.0508 | 0.062 | 0.0537 | 0.0503 | **0.138** |
| `C` | 0.0961 | 0.108 | 0.0724 | 0.0839 | **0.127** |
| `logC` | 0.0632 | 0.058 | 0.0724 | 0.0839 | **0.127** |
| `CpQ` | **0.0322** | 0.026 | 0.0223 | 0.0304 | 0.0193 |
| `PT` | 0.0017 | **0.066** | 0.0009 | 0.0017 | 0.0028 |
| `logPT` | 0.0008 | **0.058** | 0.0009 | 0.0017 | 0.0028 |
| `sumPT` | 0.089 | 0.102 | 0.0784 | 0.1004 | **0.1234** |
| `logsumPT` | 0.0796 | 0.09 | 0.0784 | 0.1004 | **0.1234** |

**Table 5: Comparison of OACs w.r.t. the number of A/B and A/A experiments with detected treatment effect with the adjusted threshold $\alpha = \alpha_{5:98}$.**

| $\alpha =$ $\alpha_{5:98}$ | Success rate \| False-positive rate [# of the 169 A/B tests] \| [# of the 98 A/A tests] | | | | |
|---|---|---|---|---|---|
| Stat.test: | T-test | BS-AVG | MW test | TW test | LR test |
| `S` | 37 \| 5 | **41** \| 5 | 19 \| 5 | 35 \| 5 | 35 \| 5 |
| `sS` | 33 \| 5 | 34 \| 5 | 20 \| 5 | 26 \| 5 | 33 \| 5 |
| `AT` | 29 \| 5 | 20 \| 5 | 27 \| 5 | 31 \| 5 | 33 \| 5 |
| `logAT` | 23 \| 5 | 10 \| 5 | 27 \| 5 | 31 \| 5 | 33 \| 5 |
| `ATpA` | 11 \| 5 | 16 \| 5 | 33 \| 5 | 29 \| 5 | 24 \| 5 |
| `log(ATpA)` | 27 \| 5 | 32 \| 5 | 33 \| 5 | 29 \| 5 | 24 \| 5 |
| `log(AT)pA` | 14 \| 5 | 20 \| 5 | 29 \| 5 | 24 \| 5 | 20 \| 5 |
| `Q` | 39 \| 5 | 43 \| 5 | 35 \| 5 | 39 \| 5 | 53 \| 5 |
| `logQ` | 35 \| 5 | 41 \| 5 | 35 \| 5 | 39 \| 5 | 53 \| 5 |
| `C` | 60 \| 5 | 65 \| 5 | 53 \| 5 | 61 \| 5 | 67 \| 5 |
| `logC` | 51 \| 5 | 54 \| 5 | 53 \| 5 | 61 \| 5 | 67 \| 5 |
| `CpQ` | 79 \| 5 | 80 \| 5 | 85 \| 5 | **88** \| 5 | 83 \| 5 |
| `PT` | 39 \| 5 | 29 \| 5 | 54 \| 5 | 55 \| 5 | 48 \| 5 |
| `logPT` | 52 \| 5 | 51 \| 5 | 54 \| 5 | 55 \| 5 | 48 \| 5 |
| `sumPT` | 31 \| 5 | 39 \| 5 | 34 \| 5 | 43 \| 5 | 47 \| 5 |
| `logsumPT` | 36 \| 5 | 37 \| 5 | 34 \| 5 | 43 \| 5 | 47 \| 5 |

## 4.5 Other evaluation statistics

In order to compare different evaluation statistics (considered in Sec. 3.3), we select 6 most representative key metrics and present the success and false-positive rates in Table 6 for their OACs. These rates are calculated both for the *constant* threshold $\alpha = 0.05$ and for the *adjusted* one $\alpha = \alpha_{5:98}$ (the rates are highlighted as in the previous subsections).

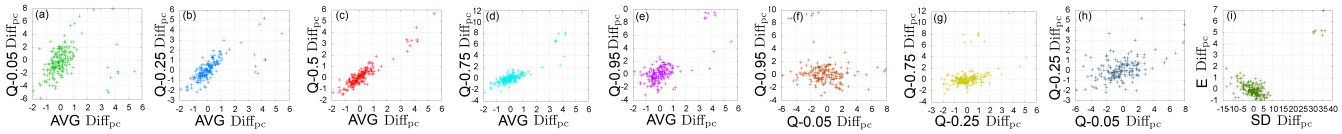


**Figure 7: The distributions of A/B tests w.r.t. $\text{Diff}_{\text{pc}}$ of some pairs of evaluation statistics for `log(ATpA)`.**

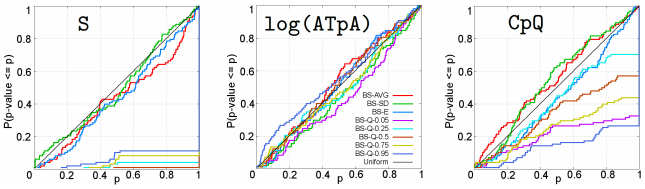


**Figure 8: The CDFs of the p-values for all statistics for the metrics `S`, `log(ATpA)`, and `CpQ` obtained from the Bootstrap test over 98 A/A experiments.**

Since the considered statistics have different meaning w.r.t. user behavior (see Sec. 3.3), we group OACs in such a way that their statistics have similar meaning and their key metrics have the same type of engagement (activity or loyalty). The highest success rate in each group is highlighted in blue color. In Fig. 8, we compare different statistics w.r.t. the CDFs of the p-values for several metrics.

In order to compare the meaning of our statistics, we calculate the scaled relative difference $\text{Diff}_{\text{pc}} = \varkappa\Delta/\xi_A$[13] of them for each of 169 A/B experiments. In Fig. 7, we plot joint distributions of these experiments w.r.t. $\text{Diff}_{\text{pc}}$ of some pairs of OECs for the metric `log(ATpA)`, whose OACs demonstrate the most consistent CDFs for different evaluation statistics (see Fig. 8). First, the joint distribution for the mean and the median (the plot (c) in Fig. 7) reveals that they have similar meaning. Therefore, we conclude that *both the mean-aware and the median-aware statistical tests could be rivals in the area of typical user behavior evaluation.*

Second, the mean value also coincides with the $\gamma$-quantile (in terms of $\text{Diff}_{\text{pc}}$) for $|\gamma - 0.5| \approx 0$, and this correlation reduces while the quantile becomes far from the median (i.e., $|\gamma - 0.5| \gg 0$). We demonstrate it by the joint distributions for the mean and the $\gamma$-quantile, $\gamma = 0.05$, 0.25, 0.75, and 0.95, in Fig. 7 (the plots (a), (b), (d), and (e) respectively). The correlation between two quantiles (in terms of $\text{Diff}_{\text{pc}}$) also depends on how far the values of their $\gamma$ are (see in Fig. 7). Therefore, we conclude that *quantiles correspond to really independent components of user engagement that represent different extreme cases of user behavior.* Note that the distributions of the metrics `S`, `Q`, and `C` are discrete, and a significant fraction of users has the same metric value (e.g., several percents of users have 1 session during an experiment). Hence, quantiles of these metrics do not change during most A/B experiments, and the p-values for the corresponding OACs have very skewed distributions (see Fig. 8 and Table 6).

Third, we see that the entropy statistic (E) has a weak correlation with the standard deviation (SD), whose meaning seems similar to diversity. Hence, we conclude that *entropy could be considered as a very promising statistic since it encodes a novel independent feature of engagement (diversity) in case of A/B testing, and OACs with entropy have very high sensitivity level (see Table 6).* Finally, note that the standard deviations (SD) for the control and the treatment variants of the service are significantly different for a noticeable number of A/B experiments (see the success rates for SD in Table 6). Therefore, we conclude that *the assumption on equality of SD for the variants A and B (as in [4, 7, 5]) should be used carefully and should be regularly validated for each OEC.*

[13]The factor $\varkappa$ is randomly chosen once in our study in order to hide real values for confidentiality reasons.

## 5. CONCLUSIONS AND FUTURE WORK

In our work, we considered a huge set of hundreds of large-scale A/B tests. We focused on the impact of different statistics and statistical tests of an OAC on the sensitivity of different key metrics. We utilized our set of experiments to evaluate dozens of key metrics of user engagement with dozens of statistical tests. First, we demonstrated that a key metric is most effective in combination with an appropriate statistical test, which is individual for each key metric. We found that utilization of the common Student's t-test with a default p-value may lead to wrong conclusions on the performance of some state-of-the-art key metrics. Second, we shown that each combination of a key metric and a statistical test requires its individual p-value threshold to control the false-positive rate at a predefined level. Third, we also proposed the entropy and the quantiles as novel OECs that quantify the diversity and extreme cases of user engagement.

**Future work.** First, we can extend the set of evaluation statistics or statistical tests by investigating more sophisticated ones. Second, we can improve the technique of adjusting the p-value threshold by utilizing more precise statistics. Third, we can study the sign of the treatment effect and the confidence interval for each of our OACs.

## 6. ACKNOWLEDGMENTS

The authors would like to thank Nikita Povarov and Pavel Serdyukov for useful discussions.

## 7. REFERENCES

[1] E. Bakshy and D. Eckles. Uncertainty in online experiments with dependent data: An evaluation of bootstrap methods. In *KDD'2013*, pages 1303–1311, 2013.

[2] B. A. Carterette. Multiple testing in statistical analysis of systems-based information retrieval experiments. *ACM Transactions on Information Systems (TOIS)*, 30(1):4, 2012.

[3] S. Chakraborty, F. Radlinski, M. Shokouhi, and P. Baecke. On correlation of absence time and search effectiveness. In *SIGIR'2014*, pages 1163–1166, 2014.

[4] T. Crook, B. Frasca, R. Kohavi, and R. Longbotham. Seven pitfalls to avoid when running controlled experiments on the web. In *KDD'2009*, pages 1105–1114, 2009.

[5] A. Deng and V. Hu. Diluted treatment effect estimation for trigger analysis in online controlled experiments. In *WSDM'2015*, pages 349–358, 2015.

[6] A. Deng, T. Li, and Y. Guo. Statistical inference in two-stage online controlled experiments with treatment selection and validation. In *WWW'2014*, pages 609–618, 2014.

**Table 6: Comparison of statistics w.r.t. the number of A/B and A/A experiments with detected treatment effect with the constant threshold $\alpha = 0.05$ and the adjusted threshold $\alpha = \alpha_{5:98}$.**

| $\alpha = 0.05$ | **Success rate** [# of the 169 A/B tests] \| **False-positive rate** [# of the 98 A/A tests] | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Statistic: | Average | | 0.5-Quantile (Q-0.5) / Median | | | Q-0.05 | Q-0.25 | Q-0.75 | Q-0.95 | Entropy | St. Dev. | |
| Stat.test: | T-test | BS | MW test | TW test | LR test | BS (Bootstrap) | | | | | | |
| S | 18 \| 2 | 23 \| 2 | 15 \| 5 | 21 \| 3 | 24 \| 2 | 0 \| 0 | 0 \| 0 | 0 \| 0 | 0 \| 0 | 8 \| 0 | 23 \| 3 | 27 \| 8 |
| log(ATpA) | 20 \| 3 | **28** \| 3 | 24 \| 3 | 24 \| 3 | 20 \| 4 | 17 \| 2 | 13 \| 2 | 17 \| 1 | 19 \| 5 | 19 \| 12 | 16 \| 3 | 16 \| 2 |
| Q | 34 \| 3 | 41 \| 4 | 35 \| 4 | 39 \| 5 | 38 \| 2 | 2 \| 0 | 0 \| 0 | 0 \| 0 | 1 \| 0 | 12 \| 0 | 37 \| 2 | 2 \| 5 |
| C | 52 \| 3 | 58 \| 3 | 46 \| 3 | 54 \| 3 | 57 \| 2 | 1 \| 0 | 1 \| 0 | 0 \| 0 | 8 \| 0 | 31 \| 1 | 57 \| 2 | 6 \| 1 |
| CpQ | 80 \| 9 | 81 \| 9 | 96 \| 10 | 94 \| 8 | 90 \| 8 | 62 \| 3 | 13 \| 2 | **70** \| 2 | 54 \| 4 | 7 \| 0 | 46 \| 3 | 6 \| 4 |
| logsumPT | 27 \| 2 | 27 \| 3 | 30 \| 3 | 34 \| 1 | 36 \| 1 | 18 \| 1 | 0 \| 0 | 13 \| 5 | 25 \| 3 | 24 \| 2 | 32 \| 2 | 39 \| 6 |
| $\alpha = \alpha_{5:98}$ | **Success rate** [# of the 169 A/B tests] \| **False-positive rate** [# of the 98 A/A tests] | | | | | | | | | | | |
| S | 37 \| 5 | **41** \| 5 | 19 \| 5 | 35 \| 5 | 35 \| 5 | 2 \| 5 | 0 \| 5 | 0 \| 5 | 9 \| 5 | 13 \| 5 | 33 \| 5 | 14 \| 5 |
| log(ATpA) | 27 \| 5 | 32 \| 5 | 33 \| 5 | 29 \| 5 | 24 \| 5 | 23 \| 5 | 20 \| 5 | 19 \| 5 | 19 \| 5 | 13 \| 5 | 29 \| 5 | 19 \| 5 |
| Q | 39 \| 5 | 43 \| 5 | 35 \| 5 | 39 \| 5 | 53 \| 5 | 4 \| 5 | 0 \| 5 | 1 \| 5 | 6 \| 5 | 26 \| 5 | 41 \| 5 | 1 \| 5 |
| C | 60 \| 5 | 65 \| 5 | 53 \| 5 | 61 \| 5 | 67 \| 5 | 11 \| 5 | 5 \| 5 | 3 \| 5 | 24 \| 5 | 41 \| 5 | 63 \| 5 | 14 \| 5 |
| CpQ | 79 \| 5 | 80 \| 5 | 85 \| 5 | **88** \| 5 | 83 \| 5 | 66 \| 5 | 18 \| 5 | 76 \| 5 | 62 \| 5 | 9 \| 5 | 54 \| 5 | 6 \| 5 |
| logsumPT | 36 \| 5 | 37 \| 5 | 34 \| 5 | 43 \| 5 | 47 \| 5 | 31 \| 5 | 11 \| 5 | 10 \| 5 | 33 \| 5 | 34 \| 5 | 37 \| 5 | 31 \| 5 |


[7] A. Deng, Y. Xu, R. Kohavi, and T. Walker. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *WSDM'2013*, 2013.

[8] B. T. Dinçer, C. Macdonald, and I. Ounis. Hypothesis testing for the risk-sensitive evaluation of retrieval systems. In *SIGIR'2014*, pages 23–32, 2014.

[9] A. Drutsa. Sign-aware periodicity metrics of user engagement for online search quality evaluation. In *SIGIR'2015*, 2015.

[10] A. Drutsa, G. Gusev, and P. Serdyukov. Engagement periodicity in search engine usage: Analysis and its application to search quality evaluation. In *WSDM'2015*, pages 27–36, 2015.

[11] A. Drutsa, G. Gusev, and P. Serdyukov. Future user engagement prediction and its application to improve the sensitivity of online experiments. In *WWW'2015*, pages 256–266, 2015.

[12] G. Dupret and M. Lalmas. Absence time and user engagement: evaluating ranking functions. In *WSDM'2013*, pages 173–182, 2013.

[13] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[14] F. E. Harrell and C. Davis. A new distribution-free quantile estimator. *Biometrika*, 69(3):635–640, 1982.

[15] B. J. Jansen, A. Spink, and V. Kathuria. How to define searching sessions on web search engines. In *Advances in Web Mining and Web Usage Analysis*, pages 92–109. Springer, 2007.

[16] H. Kashima. Risk-sensitive learning via expected shortfall minimization. In *SDM*, pages 529–533. SIAM, 2006.

[17] L. Keele, C. McConnaughy, and I. White. Statistical inference for experiments. *Unpublished manuscript*, 2008.

[18] R. Kohavi, T. Crook, R. Longbotham, B. Frasca, R. Henne, J. L. Ferres, and T. Melamed. Online experimentation at microsoft. *Data Mining Case Studies*, page 11, 2009.

[19] R. Kohavi, A. Deng, B. Frasca, R. Longbotham, T. Walker, and Y. Xu. Trustworthy online controlled experiments: Five puzzling outcomes explained. In *KDD'2012*, pages 786–794, 2012.

[20] R. Kohavi, A. Deng, B. Frasca, T. Walker, Y. Xu, and N. Pohlmann. Online controlled experiments at large scale. In *KDD'2013*, pages 1168–1176, 2013.

[21] R. Kohavi, A. Deng, R. Longbotham, and Y. Xu. Seven rules of thumb for web site experimenters. In *KDD'2014*, 2014.

[22] R. Kohavi, R. Longbotham, D. Sommerfield, and R. M. Henne. Controlled experiments on the web: survey and practical guide. *Data Min. Knowl. Discov.*, 18(1):140–181, 2009.

[23] J. Lehmann, M. Lalmas, G. Dupret, and R. Baeza-Yates. Online multitasking and user engagement. In *CIKM'2013*, pages 519–528, 2013.

[24] H. Mausser and D. Rosen. Beyond var: From measuring risk to managing risk. In *CIFEr'1999*, pages 163–178. IEEE, 1999.

[25] K. Nikolaev, A. Drutsa, E. Gladkikh, A. Ulianov, G. Gusev, and P. Serdyukov. Extreme states distribution decomposition method for search engine online evaluation. In *KDD'2015*, 2015.

[26] E. T. Peterson. *Web analytics demystified: a marketer's guide to understanding how your web site affects your business*. Ingram, 2004.

[27] K. Rodden, H. Hutchinson, and X. Fu. Measuring the user experience on a large scale: user-centered metrics for web applications. In *CHI'2010*, pages 2395–2398, 2010.

[28] T. Sakai. Evaluating evaluation metrics based on the bootstrap. In *SIGIR'2006*, pages 525–532, 2006.

[29] T. Sakai. Statistical reform in information retrieval? In *ACM SIGIR Forum*, volume 48, pages 3–12, 2014.

[30] C. E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[31] M. D. Smucker, J. Allan, and B. Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *CIKM'2007*, pages 623–632, 2007.

[32] M. D. Smucker, J. Allan, and B. Carterette. Agreement among statistical significance tests for information retrieval evaluation at varying sample sizes. In *SIGIR'2009*, pages 630–631, 2009.

[33] Y. Song, X. Shi, and X. Fu. Evaluating and predicting user engagement change with degraded search relevance. In *WWW'2013*, pages 1213–1224, 2013.

[34] G. P. Suciu, S. Lemeshow, and M. Moeschberger. Statistical tests of the equality of survival curves: reconsidering the options. *Handbook of Statistics*, 23:251–262, 2004.

[35] D. Tang, A. Agarwal, D. O'Brien, and M. Meyer. Overlapping experiment infrastructure: More, better, faster experimentation. In *KDD'2010*, pages 17–26, 2010.

[36] J. Urbano, M. Marrero, and D. Martín. A comparison of the optimality of statistical significance tests for information retrieval evaluation. In *SIGIR'2013*, pages 925–928, 2013.

[37] L. Wang, P. N. Bennett, and K. Collins-Thompson. Robust ranking models via risk-sensitive optimization. In *SIGIR'2012*, pages 761–770, 2012.

[38] W. Webber, A. Moffat, and J. Zobel. Statistical power in retrieval experimentation. In *CIKM'2008*, pages 571–580, 2008.