Statistical Analysis

Contact: Christina Choi (cchoi@mozilla.com), Joseph Kelly (Harvard)

After discussions with engineers and developers, we concluded that an increase in page rendering time is the primary indicator of a regression. Working with 30 days data for Tp5 (looking at page rendering time for top 100 websites), we recommended the best way to detect whether the sample means of the previous push and the new push were different.

First, we implemented a two independent sample t-test with unequal variance (Welch's t-test). Given the data, the assumption of equal variance between two population was too strong and unrealistic; hence, we adjusted for unequal variance to be safe. The null hypothesis would be that the means of the two populations are equal, and the alternative hypothesis would be that the mean of the new push is greater than the mean of the previous push (occurrence of a regression). From our sample data, we compute the standard error, degrees of freedom, test statistic, and p-value associated with the test statistic. Under the null hypothesis, if sample finding is a rare event, we reject the null hypothesis. To determine a rare event, we compare the p-value to the significance level (alpha), where we reject the null hypothesis when the p-value falls under the significance level. The significance level is a value which we pre-specify, which indicates the probability that the rare observation is due to chance; in other words, the probability of rejecting the null hypothesis when the null hypothesis is actually true (false positive). When null hypothesis is rejected, we can conclude that there is evidence that the new push has increased in mean rendering time. If we reject the null hypothesis, we should "flag" this push as a possible regression that a sheriff should look into. However, since the process is automated, we wrote our code such that each push is compared to the last "good' push where a "good" push is indicated by failing to reject the null hypothesis.

Since every webpage has very different page characteristics and rendering speed, we recommended that the t-test was implemented for each webpage. By doing so, we would have one hundred t-test results. Moving to page-centric testing led to multiple comparison problem, where increase in the number of hypothesis tests also increases our chances of rejecting the null hypothesis as multiple tests inflates significance level. This leads to increasing false positives, where we claim existence of a true difference in the mean when the difference is observed by chance. To correct for such inflation, we needed to implement a multiple testing correction.

One of the methods we explored to counter the problem of multiple comparisons was the Bonferroni method, where we control for the familywise error rate (FWER). This is a very conservative method where we control for the statistical significance level by alpha/n. By doing this, the threshold of a rare event becomes extremely conservative (alpha becomes extremely small and we need extremely small p-value to detect a difference). A more powerful and less restrictive procedure than Bonferroni is the Holm-Bonferroni method. This procedure gives us more power to detect regressions whilst still controlling the FWER. The procedure ranks the unadjusted p-values from the largest to smallest, and then we compare the smallest of the p-values and check to see whether it is less than (alpha)/(rank of p-value).

Then, we continue performing sequential hypothesis testing until we fail to reject null hypothesis and that would be the cutoff. Holm-Bonferroni correction performed better than Bonferroni correction. Despite the simplicity of methods where we control for familywise error rate, we were unable to obvious distinction in the presence of a regression and these methods were too conservative for our purpose.

We then looked into another method to counter the problem of multiple hypothesis testing using the false discovery rate: Benjamin-Hochberg procedure. This procedure controls for the expected proportion of false positives among all significant hypotheses (q-value). After exploring this method, we discovered that False discovery rate approach performed more powerfully in detecting truly significant results. To perform this method, we needed to order the p-values from smallest to largest, and assign indicator variable i=1 to the smallest, the next as i=2, and so on until i=m where m is the largest p-value. Then, individual p-value from a hypothesis test is compared to (i/m)*q where q is the chosen false discovery rate. After locating the largest p-value where p < (i/m)*q held true, we reject the corresponding hypothesis test as well as the corresponding tests for all smaller p-values. With Tp5 data, the False discovery rate procedure provided a good balance between false positive and false negative discoveries, and this method has been recommended.

After implementation of the t-test and FDR procedure, we hoped to observe noticeable difference between regressions and non-regressions; however, it was still not as distinguishable as we wanted. We noticed that there were large natural variations between consecutive pushes. Thus when comparing a new push to the current push we were often detecting changes attributable to noise rather than a true change in the mean page load time. We managed to bring down this natural variation by making changes to the experimental design, but these changes were not enough to detect small regressions. To counter this problem, we decided to implemented the exponential smoothing algorithm before performing the FDR procedure. This algorithm assigns exponentially decreasing weights to past observations and fits a trend line by incorporating all the "good" pushes. As a result, the null hypothesis would be that the mean of the trend line and the mean of the new push are equal. By smoothing we were able to reduce the noise present in the data and the testing procedure improved dramatically in detecting known regressions.