Outlier page visit times within sessions

Mikhail Popov

September 22, 2015

Background

In our ongoing research efforts to develop a measurement of user satisfaction with Search, we had an idea to detect when users had a page visit that deviated from the rest of their page visits in the same search session. The motivating hypothetical was:

1. User searches for a thing.
2. User opens N pages.
3. N-1 page visit times are short (e.g. $\leq 30s$).
4. 1 page visit time is long (e.g. $\geq 2min$).
5. User has probably found a thing they were looking for!

We used the data from the survival analysis earlier.

Results

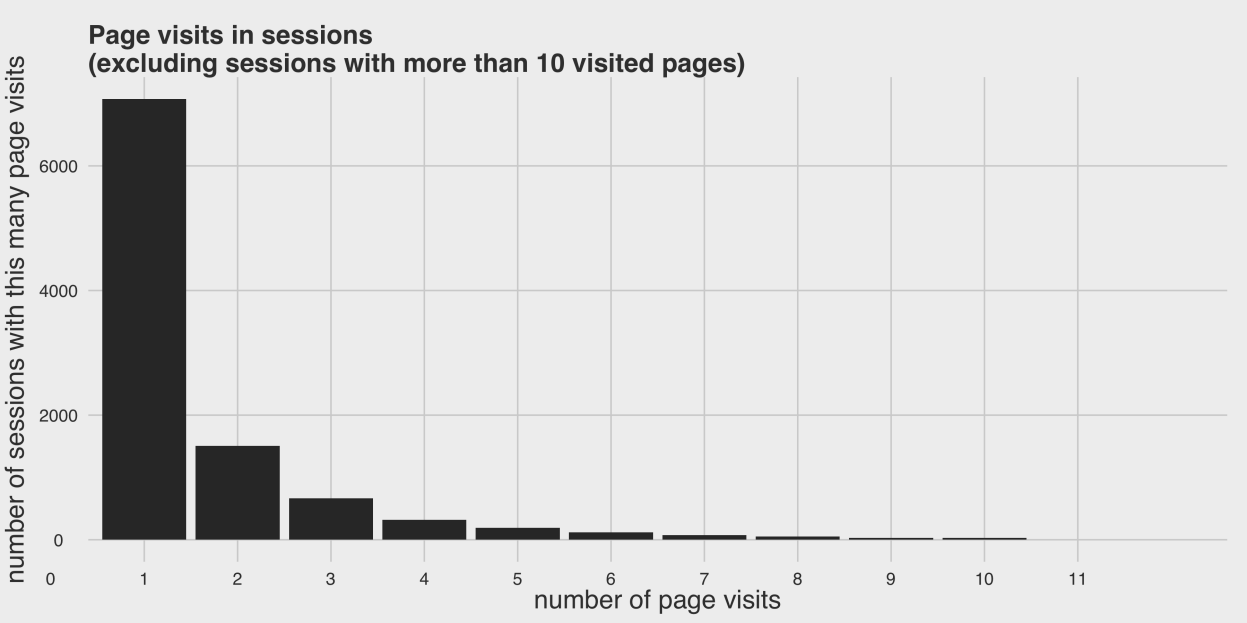

Figure 1: The *VAST* majority are 1-page sessions.

We can't really detect outliers in 1-page sessions, and since those account for the majority of our sessions, there is no reason to pursue this idea. At best, we can compute the median and the median absolute deviation (MAD) from the page visit times and then designate page visit times outside of median $\pm\, 2 \times MAD$ range as outliers.

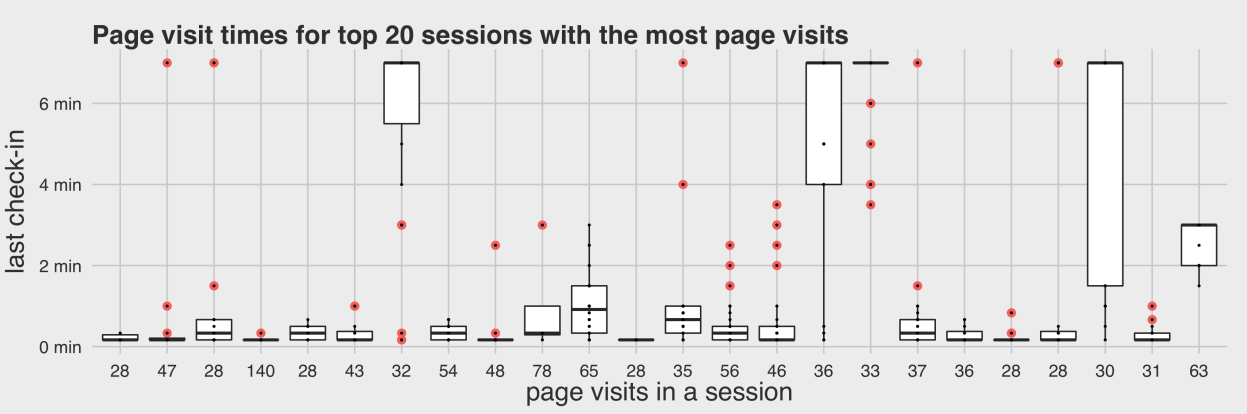

Figure 2: In *extreme sessions* with large numbers of visited pages – 140 in one case – we are able to detect page visit times that are outliers (red).

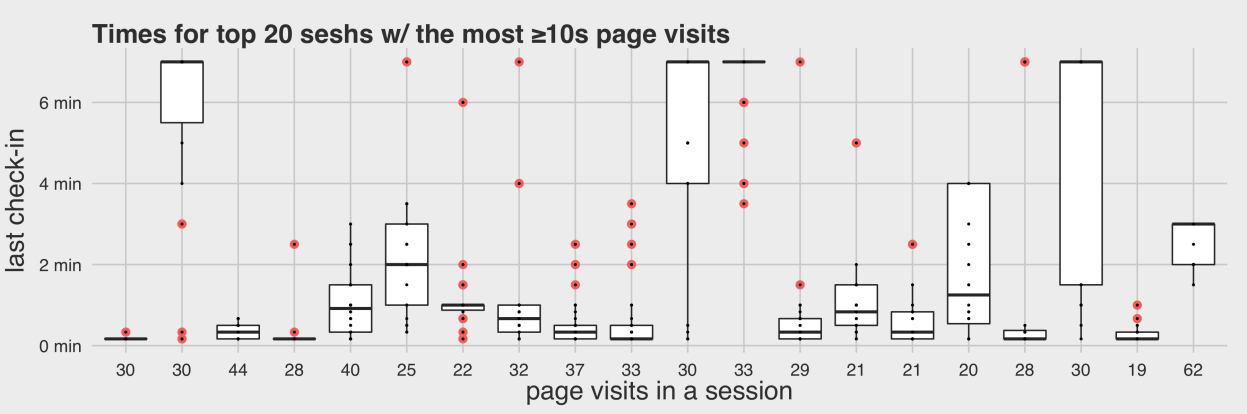

Figure 3: Top 20 sessions (by number of page visits) *AFTER* excluding page visits without check-ins – meaning the user closed the page within the first 10 seconds of opening it.

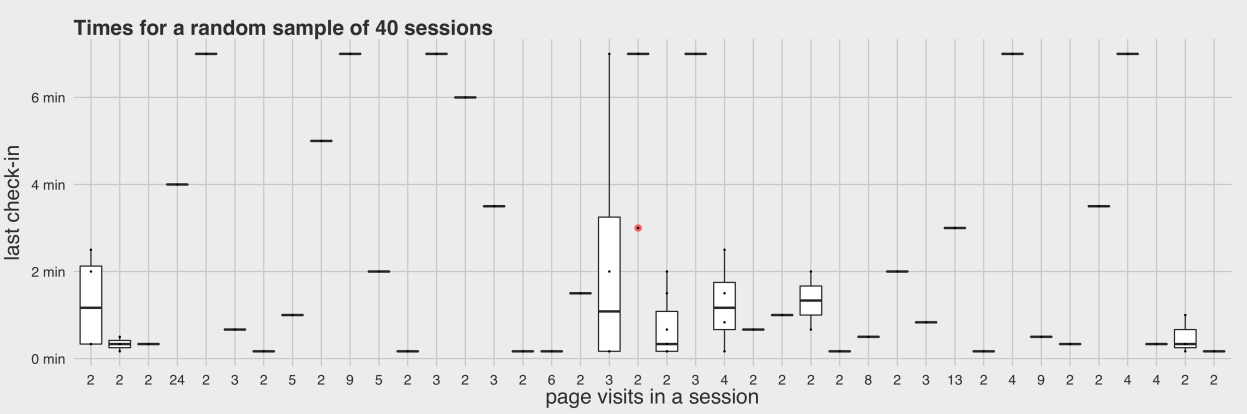

Figure 4: In general, sessions just don't have enough data points to detect outliers.

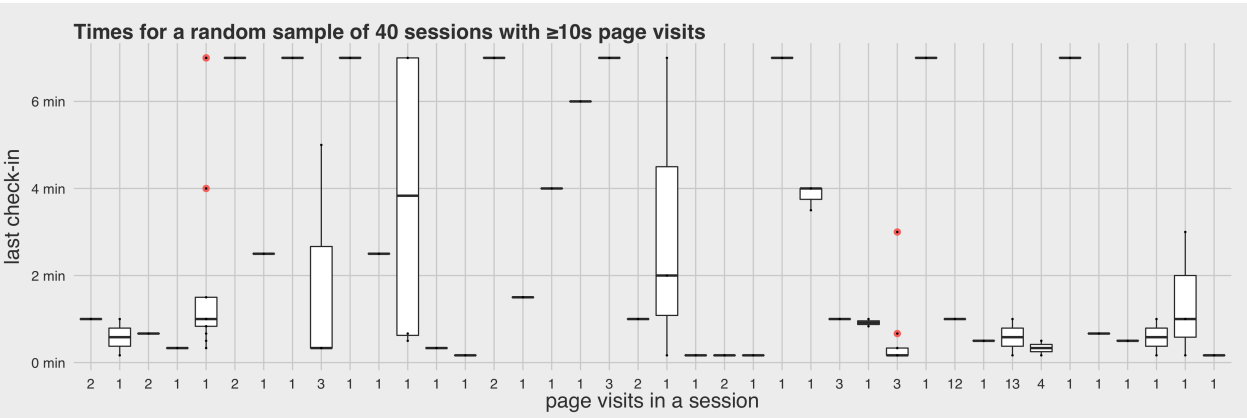

Figure 5: Even after we limit ourselves to page visits with check-ins, the random sample of sessions doesn't inspire confidence.