OCF Lab Session Analysis Part 1

Shicheng Huang

April 8, 2018

Introduction

I am a volunteer staff for the Open Computing Facility (OCF) at the University of California, Berkeley, where we provide free computer access to all students. Additionally, using the OCF desktop computers, students can print maximum of 10 pages per day, capping 100 pages each semester.

As a staff who spents average 7 hours per day in the lab, I often see people waiting for a computer. I am curious if I can give a decent estimate how long they have to wait. The first idea that comes into my mind is to survival analysis, estimate the wait time by estimating the probability of a session, out of all sessinos, will end within a short period of time.

First I have to estimate how long a person will stay in a desktop, given he/she has already spent T minutes with the desktop (i.e P(a session's additional duration S minutes|session's duration is t minutes))

Session Dataset

The dataset we use is the lab session data this semester. Below is a snippet of the session data. The field "host" represents each desktop. The field "duration" measure the duration of a session by minutes.

```
##       id                        host               start                 end
## 1 353281 outbreak.ocf.berkeley.edu 2018-04-08 10:18:42 2018-04-08 10:31:42
## 2 353278  cyclone.ocf.berkeley.edu 2018-04-08 09:56:52 2018-04-08 10:19:02
## 3 353272    venom.ocf.berkeley.edu 2018-04-08 08:48:18 2018-04-08 10:06:05
## 4 353275     acid.ocf.berkeley.edu 2018-04-08 09:18:38 2018-04-08 09:21:31
## 5 353274     acid.ocf.berkeley.edu 2018-04-08 09:10:47 2018-04-08 09:16:07
## 6 353271 sinkhole.ocf.berkeley.edu 2018-04-08 03:22:52 2018-04-08 04:47:01
##   duration
## 1 00:13:00
## 2 00:22:10
## 3 01:17:47
## 4 00:02:53
## 5 00:05:20
## 6 01:24:09
```

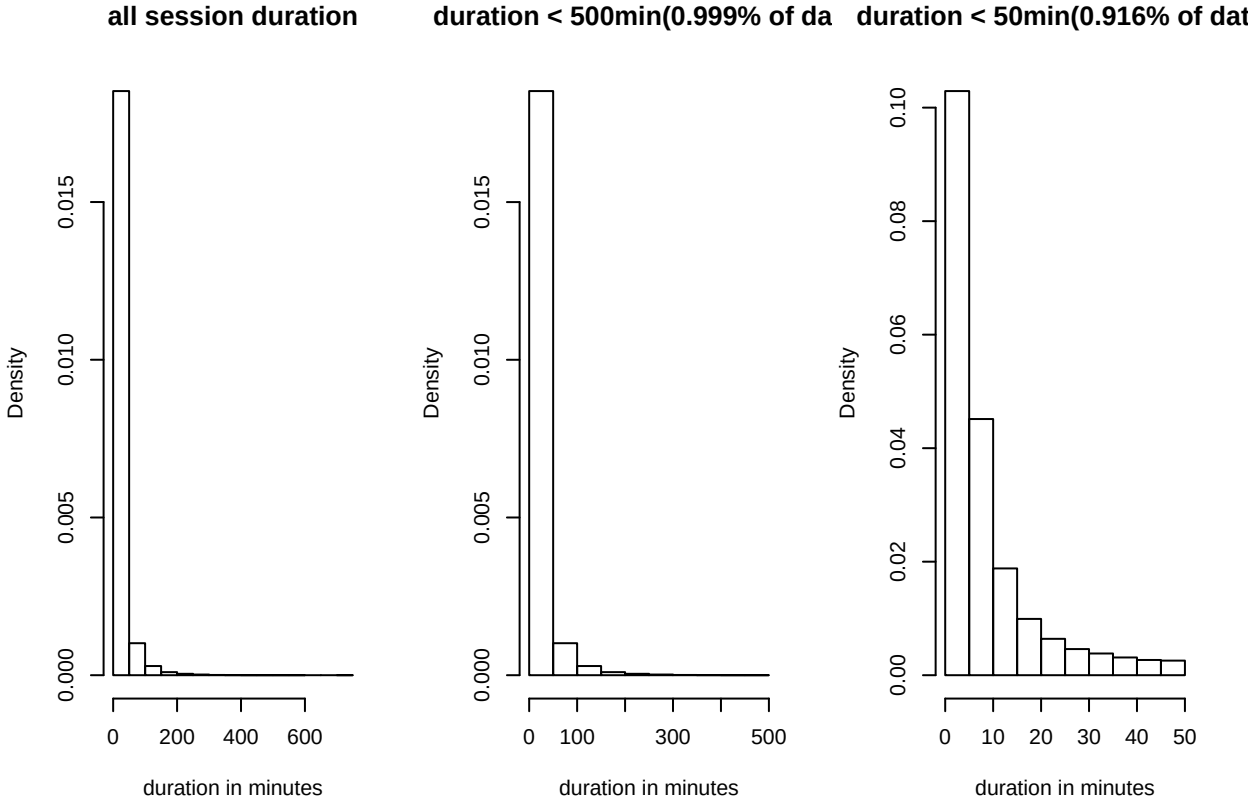
Basic Data processing

Here are the procedures I use to clean the data:

1. Because the lab volunteer staff often uses the desktops much longer than regular users, who mostly come to the lab to print, I exclude all sessions from the volunteer staff.
2. I filter out sessions that have 0 or negative durations. This is mostly a data engineering issue because it is physically very difficult and rare that some user logins and logouts within 1-2 seconds to have a 0 session duration.

3. Filter out sessions from host "blizzard.ocf.berkeley.edu" and "rruption.ocf.berkeley" because they are the fron desk desktop and desktop specific for volunteer staff to help student organizations with hosting websites.

Below are histograms of the session data. Because the raw histogram is extremely skewed, I make two histograms which filter out session that are longer than 500 minutes and 100 minutes respectively.



While taking a "closer" look, it seems like lots of sessinos are under 20 minutes. Below is a table of session duration quantiles. We can see 75% of the sessions are under 15 minutes. This makes sense because (unfortunately) most of the users come to the computer lab just to print.

```
##      quantiles num_sessions
## 5%         1.3         1652
## 10%        1.8         3304
## 15%        2.1         4956
## 20%        2.5         6607
## 25%        2.8         8259
## 30%        3.2         9911
## 35%        3.6        11562
## 40%        4.1        13214
## 45%        4.7        14866
## 50%        5.3        16517
## 55%        6.2        18169
## 60%        7.3        19821
## 65%        8.7        21473
## 70%       10.6        23124
## 75%       13.3        24776
## 80%       17.8        26428
## 85%       25.4        28079
## 90%       39.4        29731
```

```
## 95%       65.1        31383
## 100%     722.0        33034
```

There is a huge jump of session duration from the 95% quantile to the maximum. To find what are truely outliers, I also make a table showing the quantiles between 95% and 100% with 1% increment.

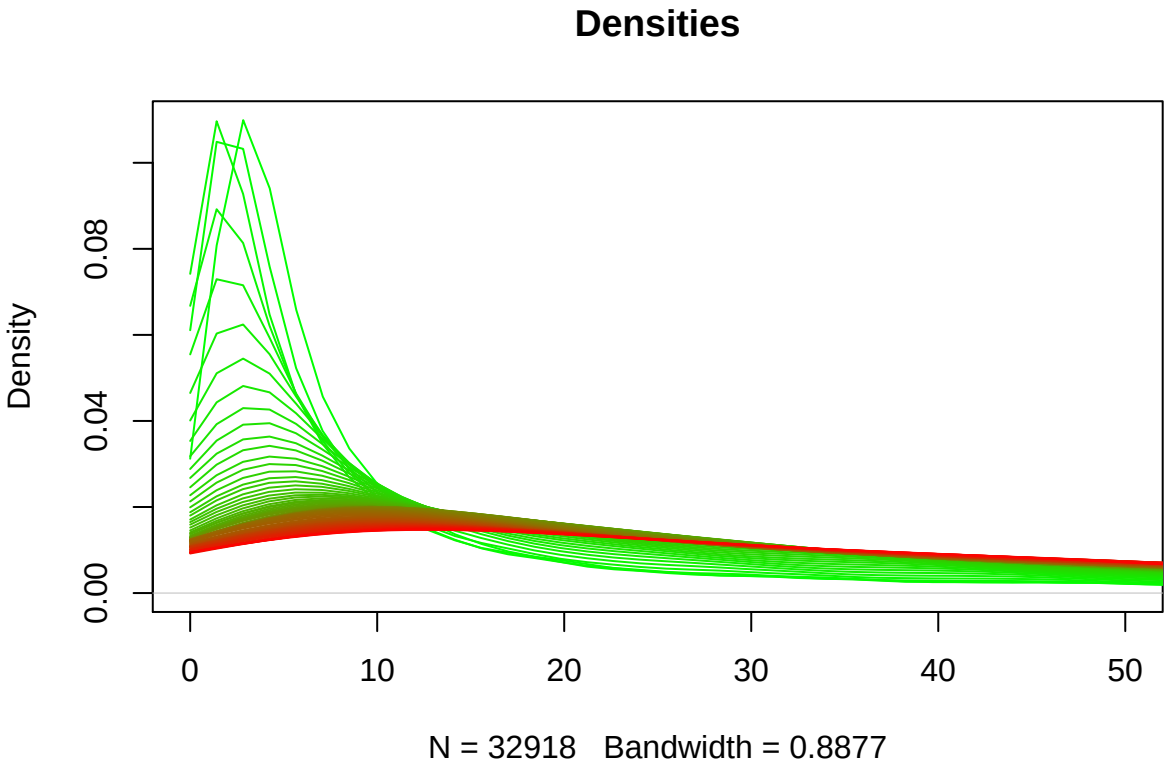
```
##      quantiles
## 95%       65.1
## 96%       74.9
## 97%       88.2
## 98%      107.7
## 99%      144.6
## 100%     722.0
```

Basic Survival P(additional session duration | session duration & non-staff)

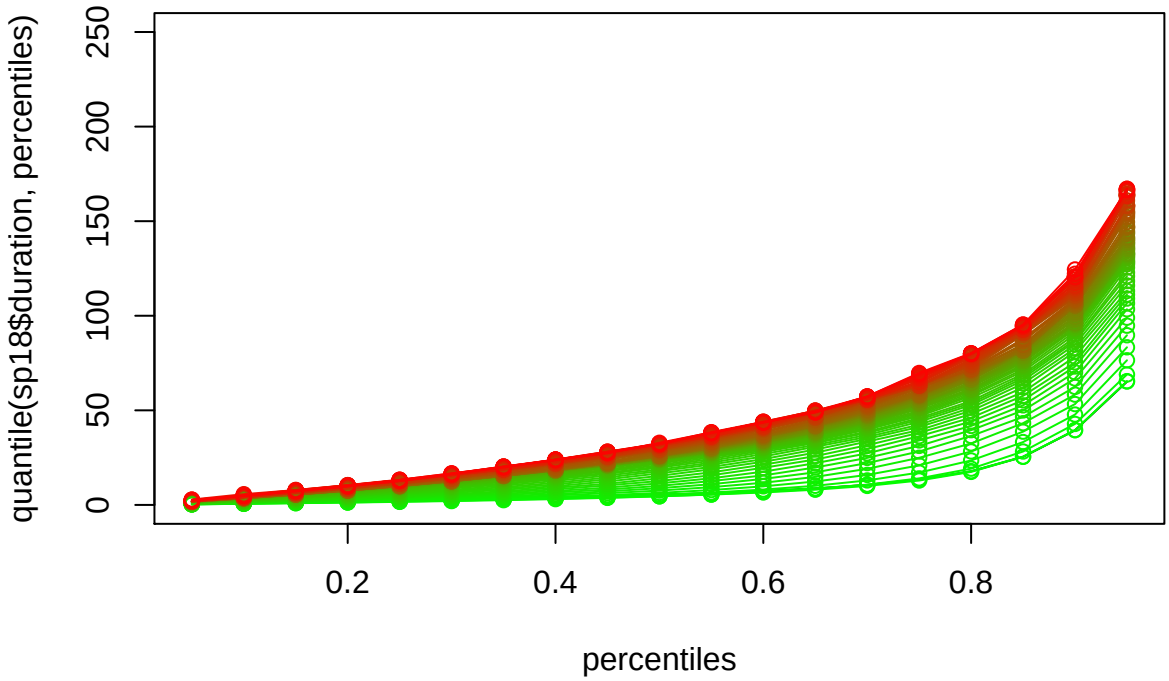
Imagine you come into the lab and wants to use the particular computer at the corner, but there is already a person sitting there. So you ponder when he is going to leave. One way to make an educated guess is to check how long he has been on the computer.

We define **current session duration** the time he has been on the computer, and the **remaining session duration** how long you have to wait (i.e the time between you start waiting and he leaves the computer).

Below graph shows the distributions of remaining session time given the current session duration. The red scale is porportional to current session duration. In other words, if the person has been using the computer for a long time, the distribution of the remaining session time will look very red.

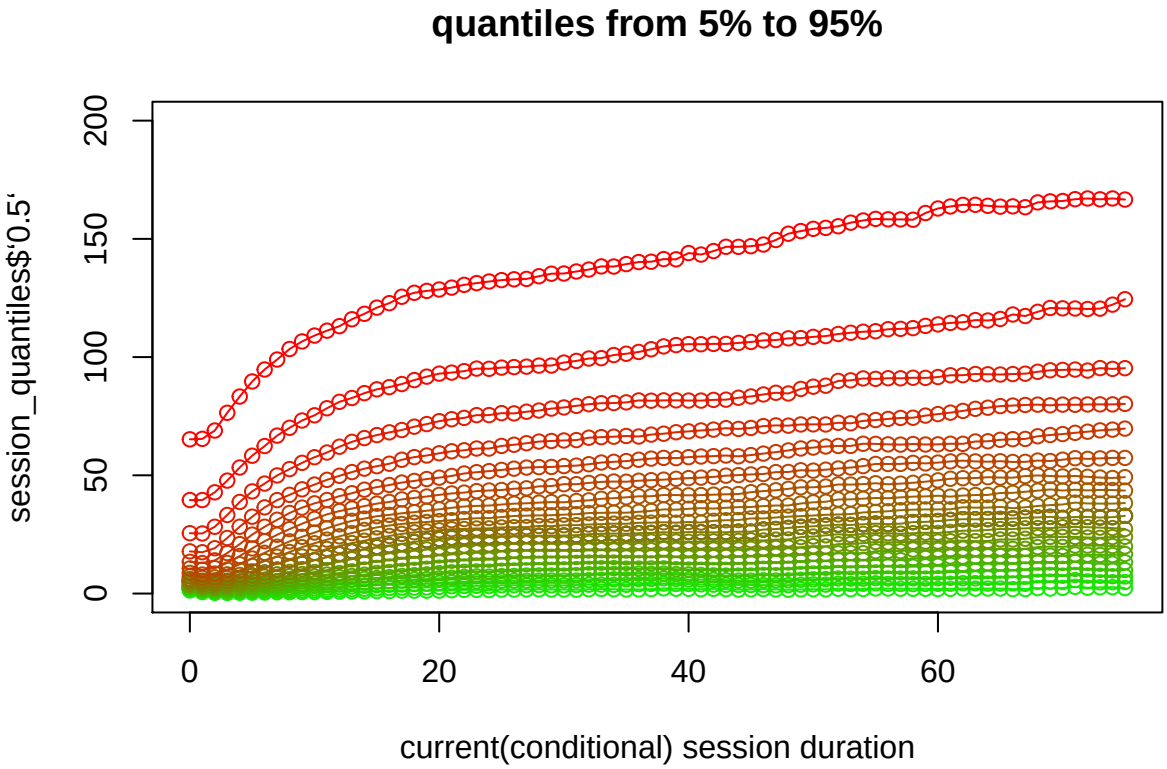


Another way to visualize a distribution is to see its different quantiles. The figure below shows the remaining duration of a session colored by the current session duration.

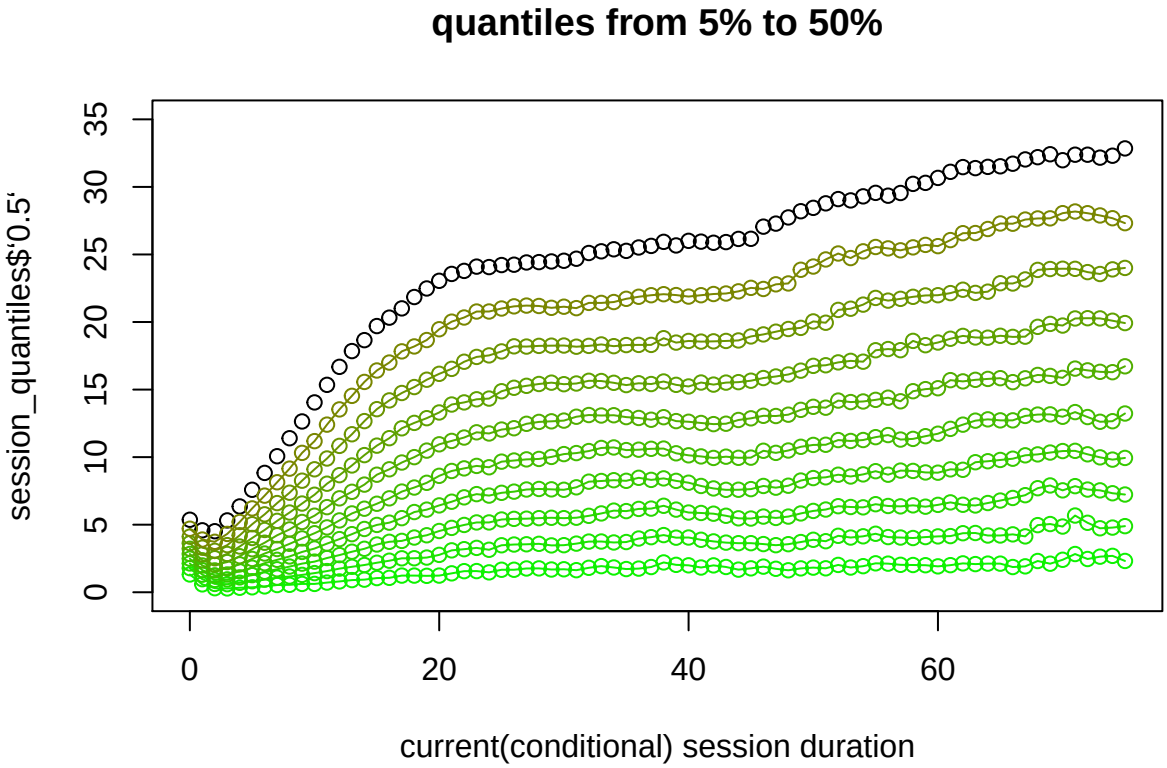


As the current session duration increases, indicated by the increase "redness" of the line, different quantiles of the remaining session time increases.

Below graph looks at how fast the quantiles are increasing as the given session time increases.



As the current session duration increases, the first quantiles of remaining session duration increases rapidly then slows down after the current session duration exceeds 20. To avoid overlapping lines, I make a graph that only shows the quantiles below 50%.



From the both graphs, we can learn that, when a usr first started his session, there is 50% chance that he will leave within 5-7 minutes (see the left most points when x = 0); however, if a user has been using a desktop for 20minutes or plus(x > 20), there is 50% chance that he will leave within the next 25 minutes or so, otherwise he will stay probably an additional 25minutes to 150minutes.

Now we can go back to our situation, if the person using the corner computer has been using it for a long time, it is best for your to find another computer or politely ask the person to leave.

Future Directions

In part 1, I have introduced you the problem context and learn that session duration data is extremely skew with some unreasonable outliers (1500+ minutes). Therefore, it is better to look at different quantiles of data. And we find that knowing how long a session has been (i.e the current session time) can help us infer how long the session will last from the current time (i.e remaining sesion time), but this heuristic's effectiveness will decrease after current session time exceeds 20 minutes or so.

In the next part, I will examine the effectiveness of inferring the remaining session time based on the information if the user has printed anything during the session.