OCF Lab Session Analysis Part 1: $P(T < s + t \mid T > t)$

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, we also let students print maximum of 10 pages per day and 100 pages per semester.

As a staff who spends average 7 hours per day in the lab, I often see people waiting for a computer. I wonder if I can have a decent estimate of when people have to wait for a computer. To break down the question, I first try to estimate the wait time for a single computer, since I sometimes have to wait for the particular computer at the corner of the lab. In this post, I will explore how does a desktop session remaining duration distribution changes conditioned on the current session duration. We define **current session duration** the time a student has been on a computer, and the **remaining session duration** how long it will take for him to leave to computer.

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

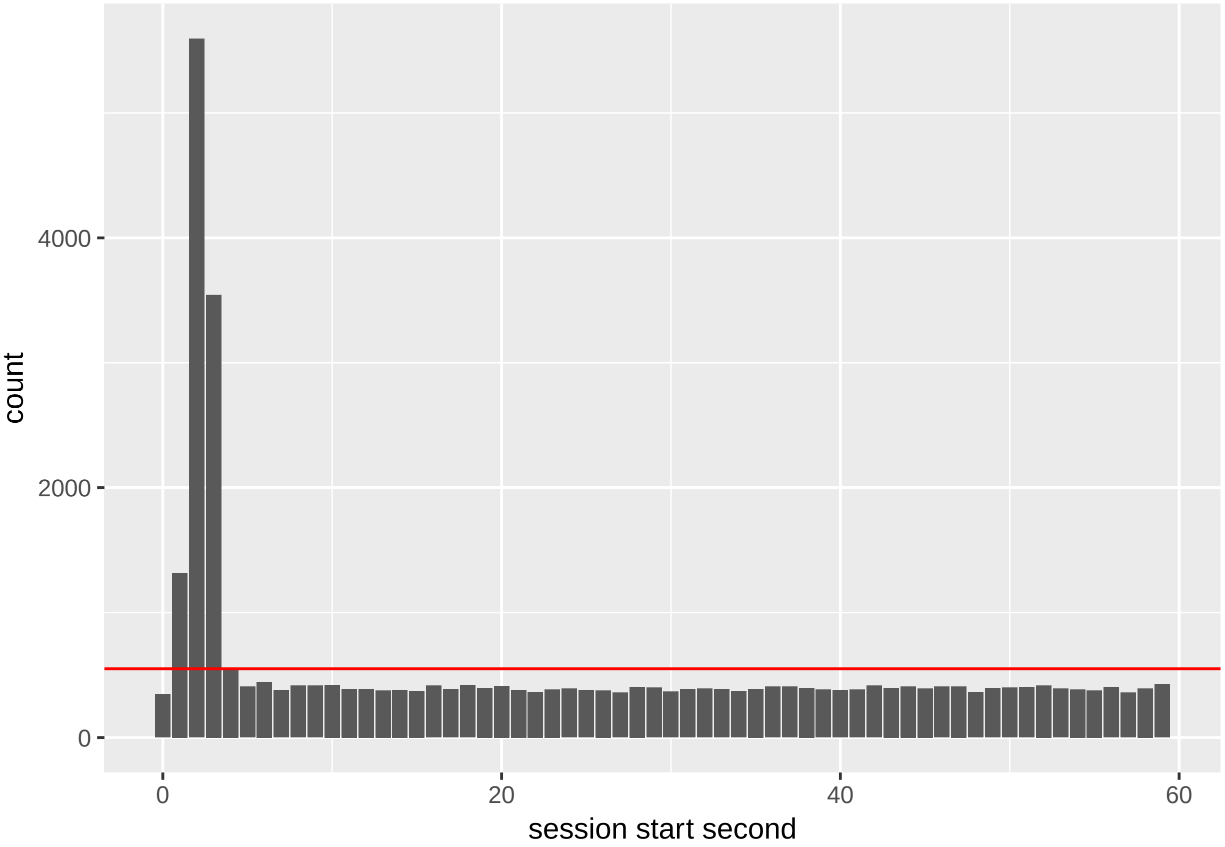
Here are the procedures I use to clean the data:

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.

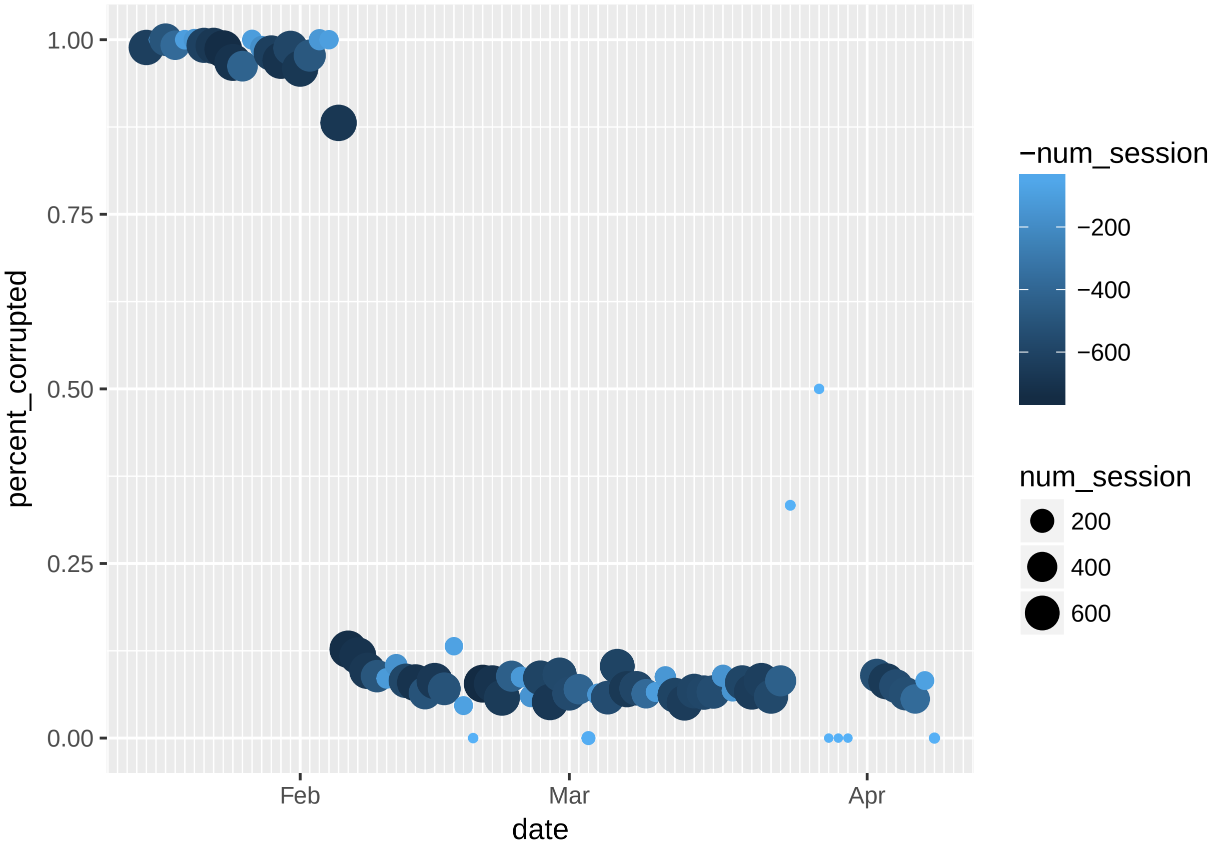
3. Filter out sessions from host "blizzard.ocf.berkeley.edu" and "eruption.ocf.berkeley" because they are the front desk desktop and desktop specific for volunteer staff to help student organizations with hosting websites.

Data Adjustments

Recall that there used to be a bug in our session tracking infrastructure that we could only record a session's start time at the beginning of the minute (know more about the bug from one of my previous post). As a result, lots of session appear to "start" around the beginning of the minute but they actually started the minute before. See below figure about the distribution of session start seconds, the red line is 1/60, the ideal proportion if all session start and end are uniformly random.



To find out the time interval when sessions data that are corrupted, I make a plot shows the percentage of sessions with start seconds < 4 against time. The darker and bigger the dot is, the more sessions are there in the day.



We can see the percentage is abnormally high until early February. The "peak" in the end of March is Spring break. As the table below show, after Feb 6th, the session tracking system goes back to normal again.

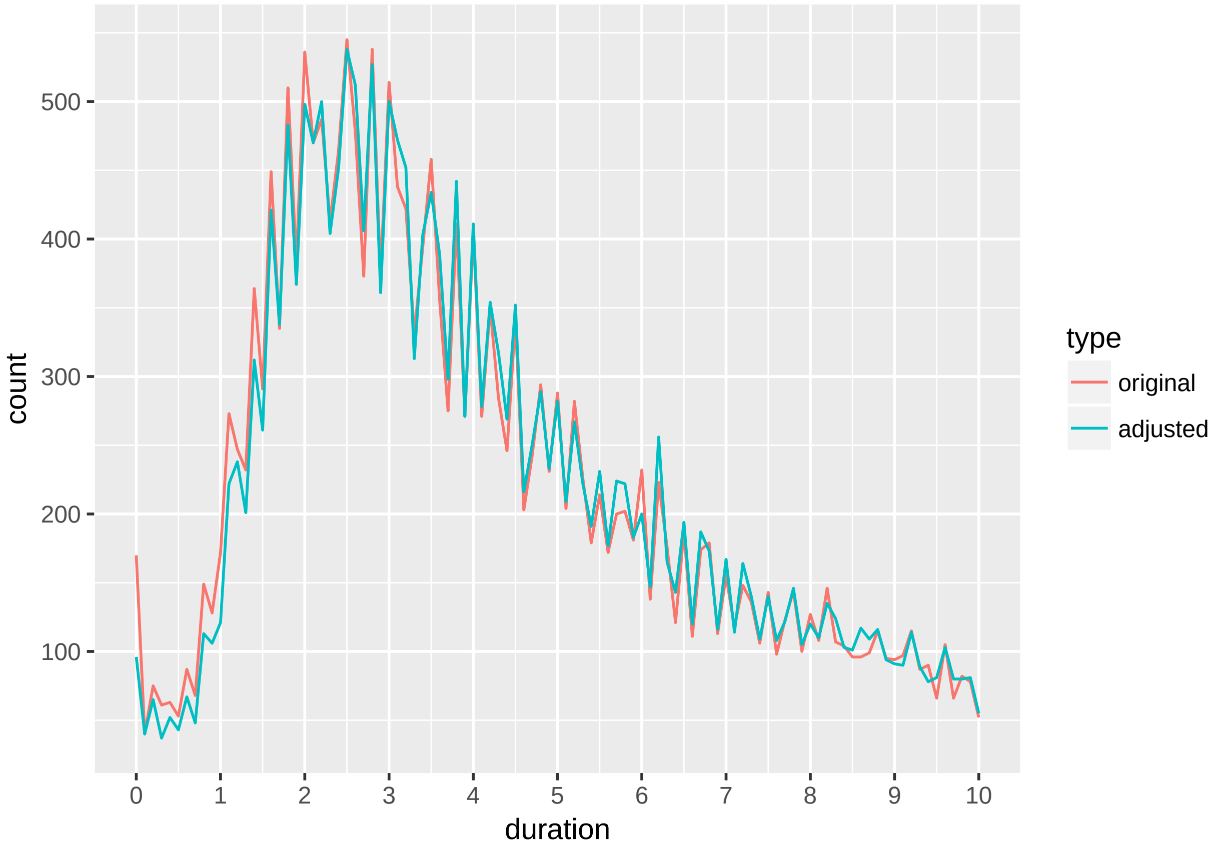
```
##         date percent_corrupted num_session
## 1 2018-02-01         0.9585799         676
## 2 2018-02-02         0.9772257         483
## 3 2018-02-03         1.0000000         131
## 4 2018-02-04         1.0000000          93
## 5 2018-02-05         0.8808824         680
## 6 2018-02-06         0.1272230         731
```

Thus, for all the session before 2018-02-05, I will adjust the duration by adding a random variable that is uniformly distributed from the set $\{0, 1, 2, \ldots 55 + S_{session\,start\,second}\}$.
Because if the session's recorded start time is X given our tracking system was malfunctioning, the real session start time could be from 0 to X or 5 to 55 from the previous minute. So the real session duration should be anywhere between 0 to $X + 1 + 55$ seconds longer.

Let's have a rough look at the difference before and after the adjustment through both graph and summary statistics.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.933   5.500  15.360  13.530 722.900
```

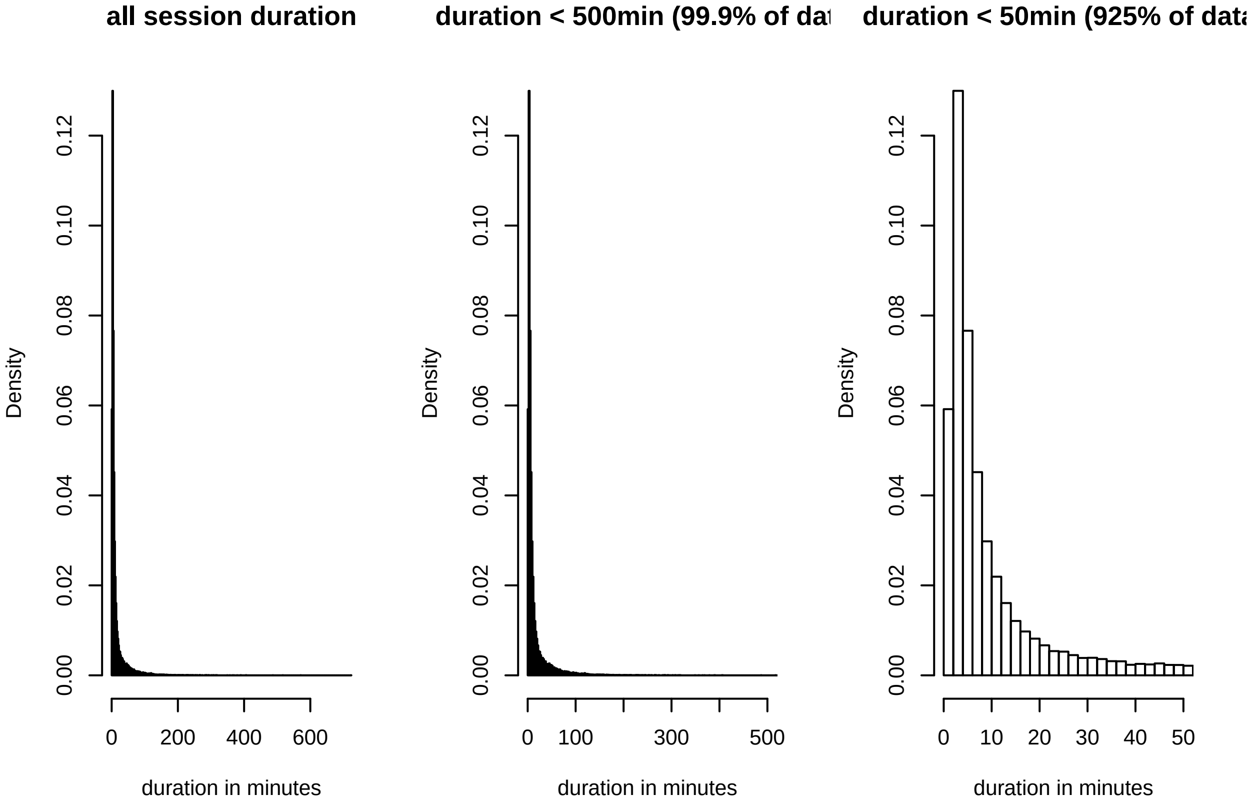
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.817   5.333  15.230  13.350 722.000
```



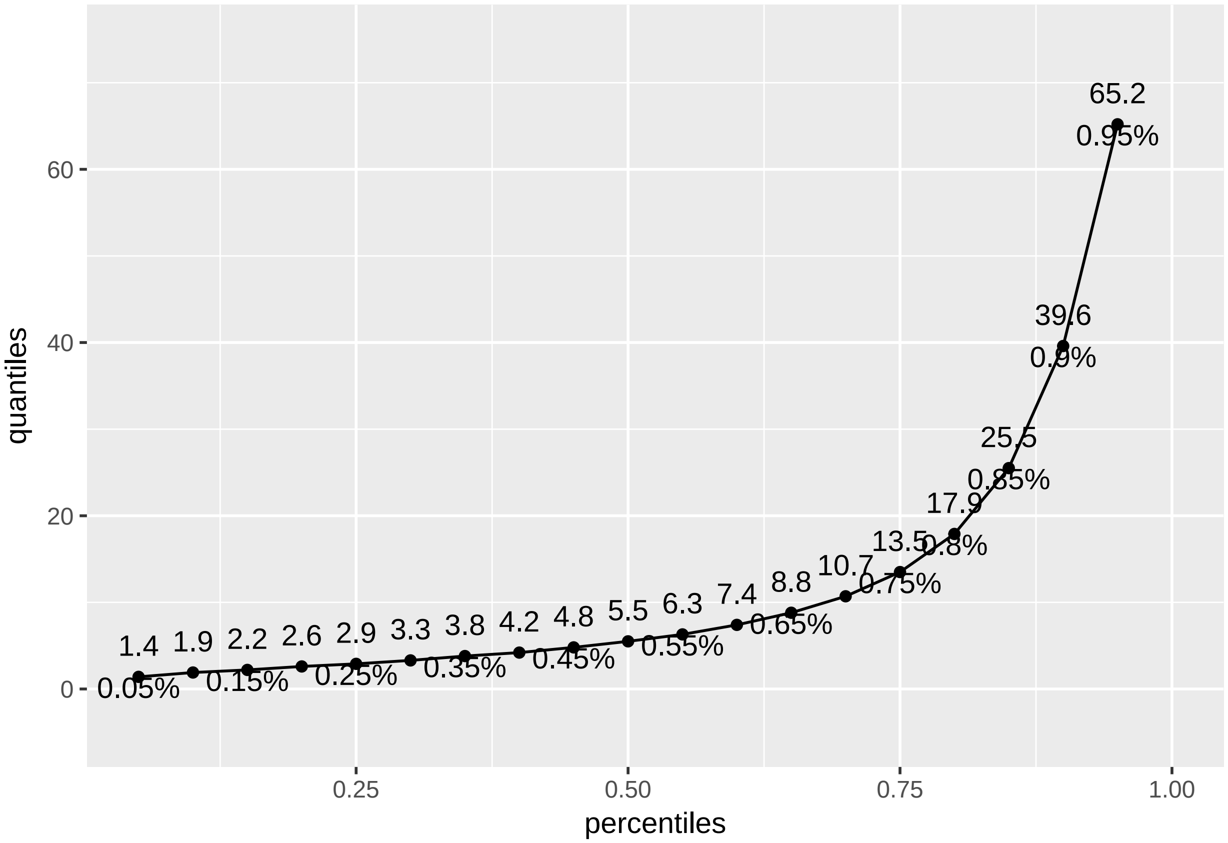
There isn't much visible difference but I think it is still important to take good care of the data inaccuracy issue.

Session Analysis

Lets first look at the distribution of the session duration. Because the raw histogram is extremely skewed, I make two other histograms with x axis limit (0, 500) and (0, 50) respectively.



While taking a closer look, it seems like lots of sessions are under 100 minutes. To visualize a skewed distribution, we can also see its different quantiles.



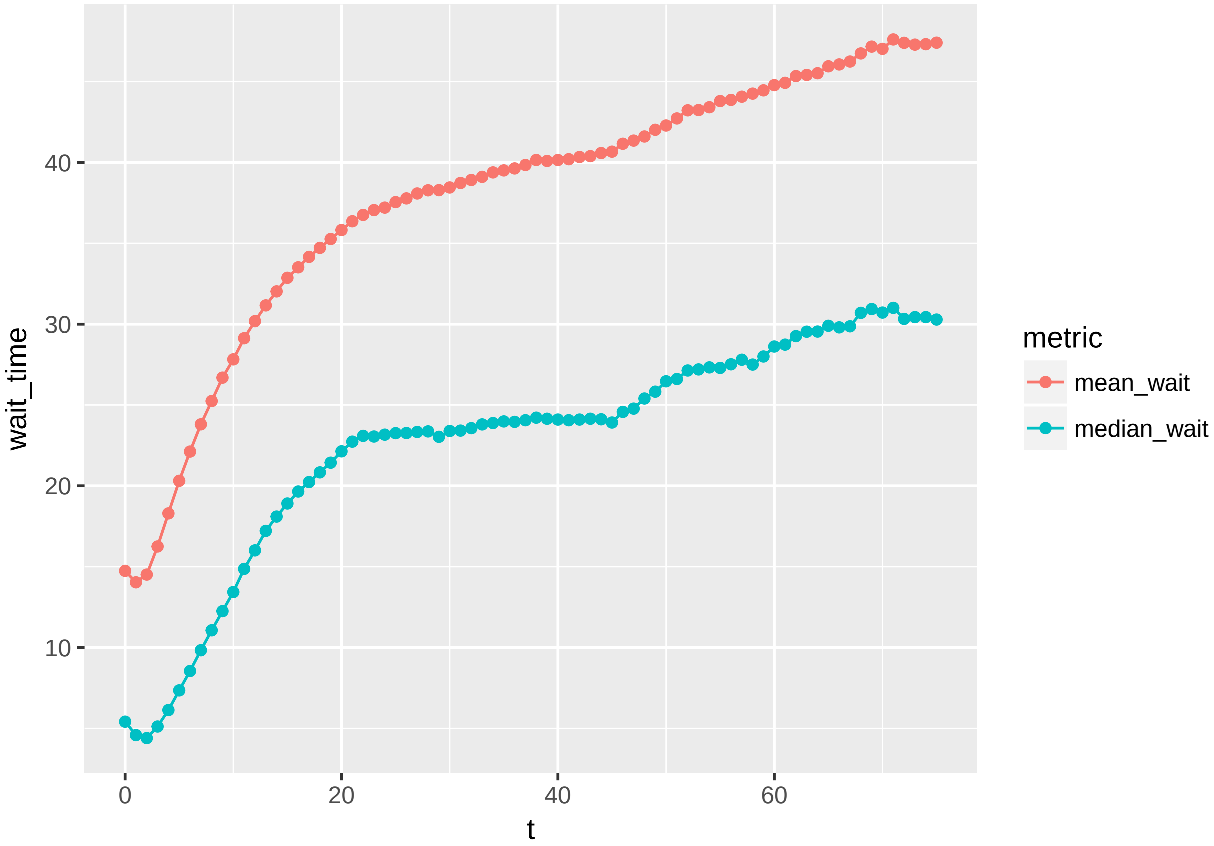
We can see 75% of the sessions are under 15 minutes. And 95% of the sessions are under 65 min. Maybe most of the users come to the lab just to print (I will investigate further in part 2). There is a huge jump of session duration from the 95% quantile to the the 100% quantile (722min). To find what are truely outliers, I also make a table showing the quantiles between 95% and 100% with 1% increment.

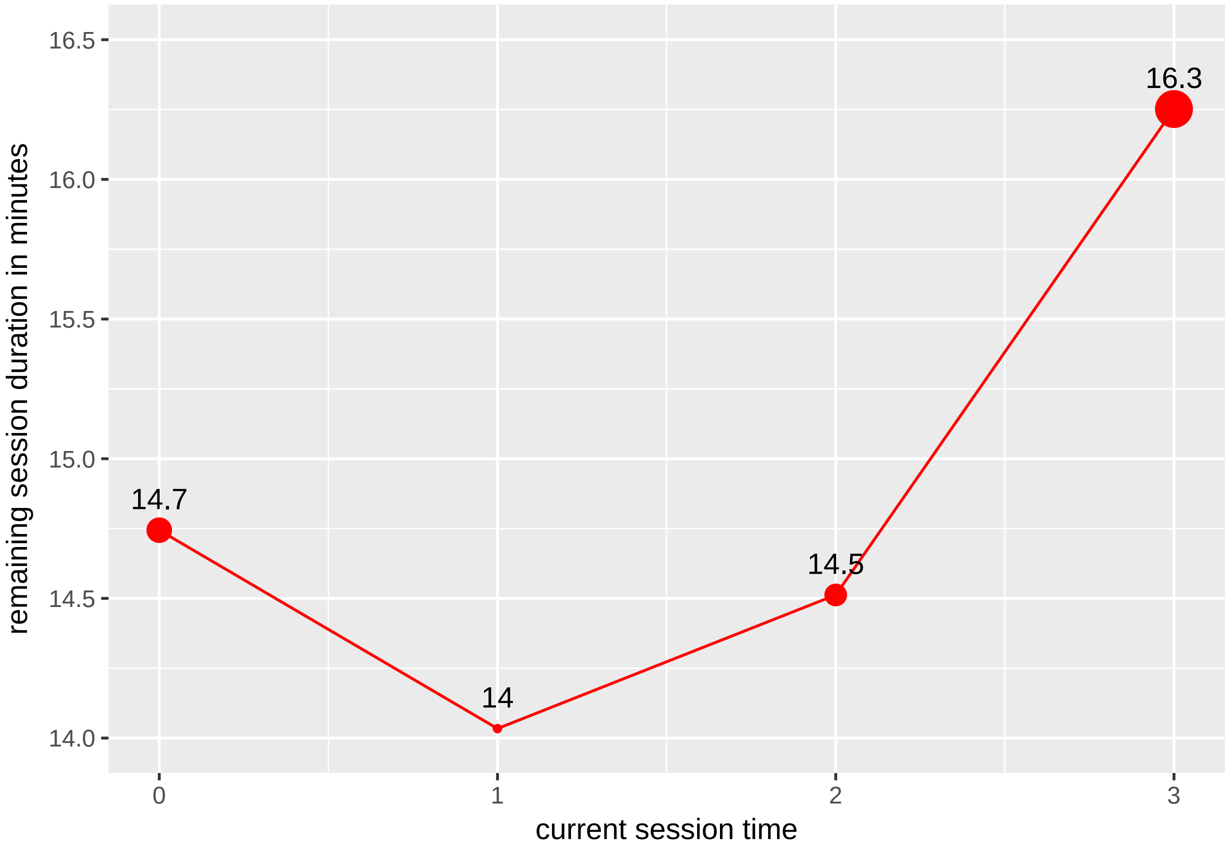
```
##      quantiles
## 95%       65.2
## 96%       75.1
## 97%       88.2
## 98%      107.7
## 99%      144.7
## 100%     722.9
```

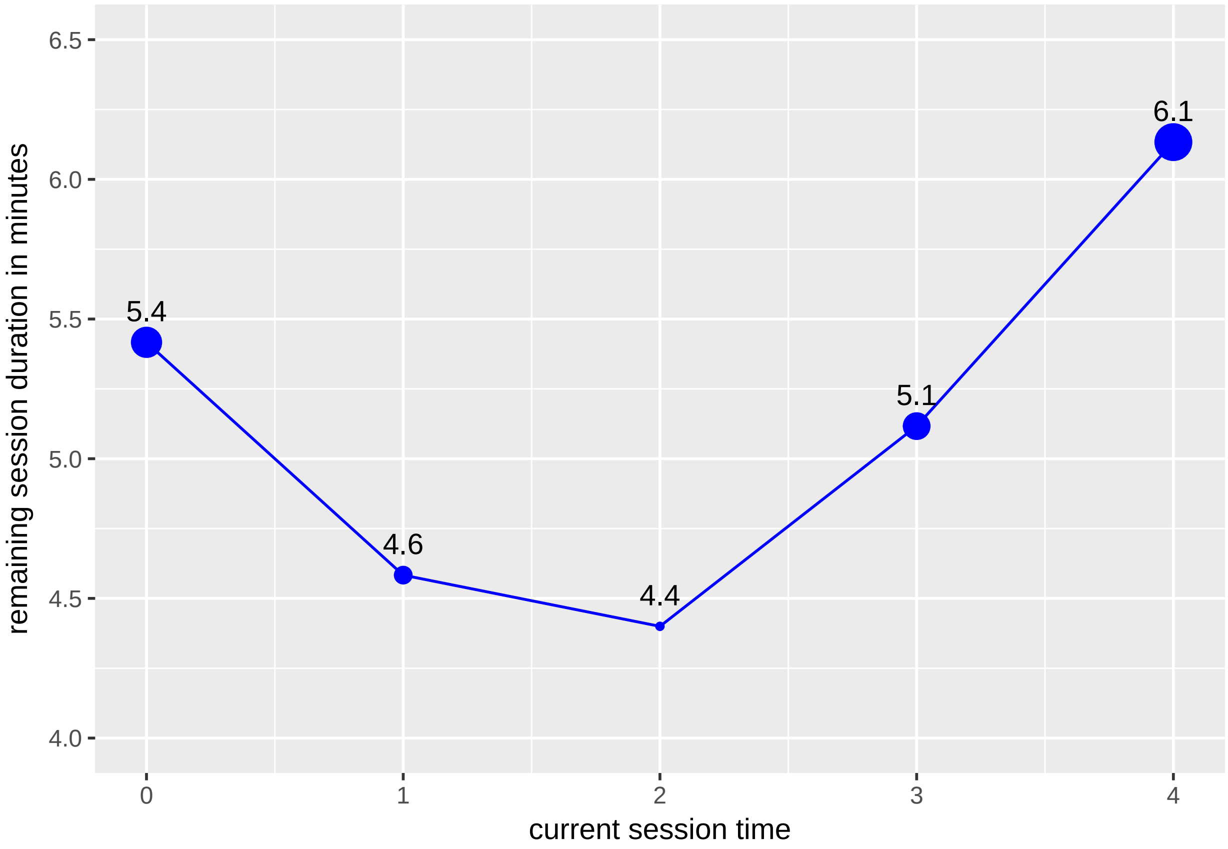
We can see, at least 99% of the sessions are "normal session" which people will leave within 3 hours.

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

Below graph shows the distributions of remaining session time given the current session duration. If the person has been using the computer for a long time, the distribution of the **remaining session time** will look very red.


wait_time
40
30
20
10
0
20
40
60
t
metric
mean_wait
median_wait


remaining session duration in minutes
16.5
16.0
15.5
15.0
14.5
14.0
14.7
14
14.5
16.3
0
1
2
3
current session time


remaining session duration in minutes
6.5
6.0
5.5
5.0
4.5
4.0
5.4
4.6
4.4
5.1
6.1
0
1
2
3
4
current session time

since many duration are cluster at the 2 - 3 min mark, we can anticipate a session will last only around 2-3min or so. As a result, remaining session time (both mean and median) is the lowest when the current session time is 2-3 min. After that, they are monotonically increasing.

Future Directions

In part 1, we learn that a majority of the sessions are short but its distribution is extremely right skew with some unreasonable outliers (700+ minutes). Therefore, it is better to look at different quantiles instead. 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 session time), but this heuristic's effectiveness will decrease after current session time exceeds 20 minutes or so.

In the next part, I will examine additional variables **day-of-the-week** and **the computer used**. I do expect sessions during the weekends will be longer because it seems like more people come to the lab just to study instead of printing; and students may favor computers differently because of their locations in the lab.