# MSAN 603 HW 3 - Funnel Analysis

Arda Aysu, Lawrence Barrett, Andre Duarte, Tim Zhou

1. As $\lambda \to \infty$, the number of people with higher survival times drops off. This can be seen in the graphs below.

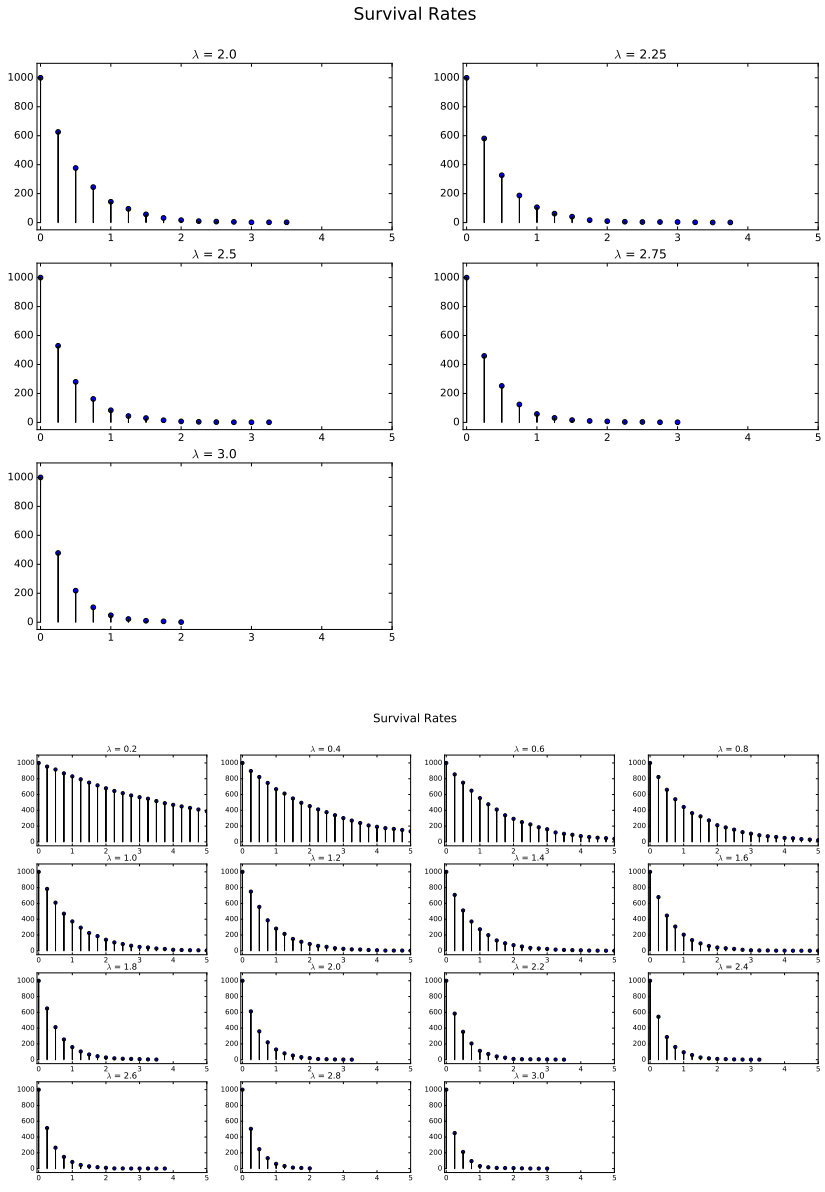



Figure 1: Survival Rate Plots for 1a and 1b

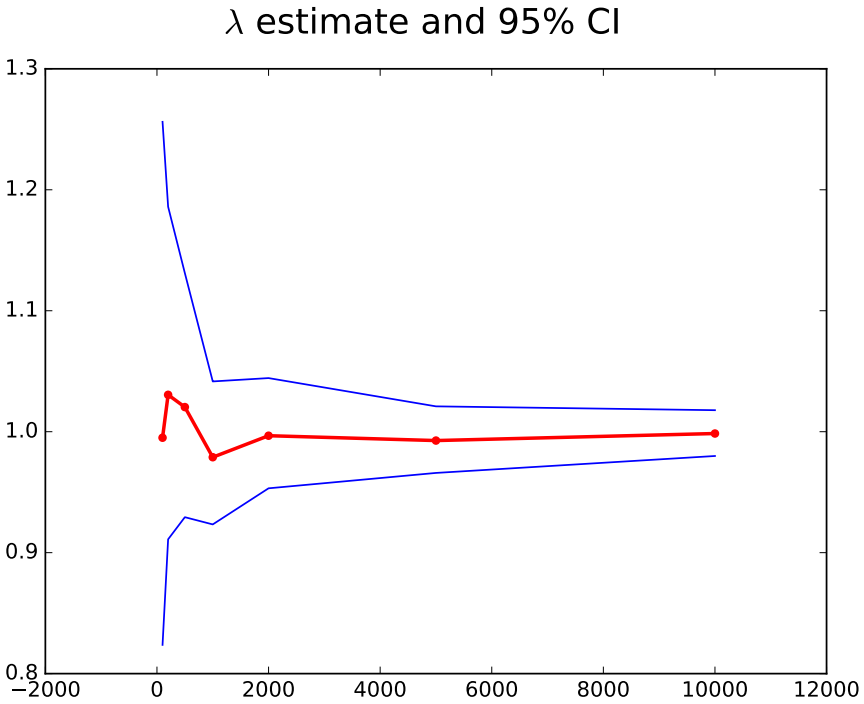
2. Please see attached python code.

3. (a) $\tilde{\lambda} = \dfrac{1}{\overline{X}}$ is an unbiased estimator because $\mathbb{E}\left[\tilde{\lambda}\right] = \mathbb{E}\left[\dfrac{1}{\overline{X}}\right] = \lambda$.

   (b) On 1000 users with 500 bootstraps, $\hat{\lambda} \approx 0.989$.

   (c) On 1000 users with 500 bootstraps, the 95% CI is $(0.929, 1.056)$.

   (d) A plot of the estimates and the 95% CI is shown below. Note that as $n_{\text{users}} \to \infty$, the CI tightens and the estimate approaches the expected value.




4. (a) The analytical estimate for $\lambda$ assumes that the exact quitting times $x_i$ of the users are known. With event data, we only have knowledge of the time of their last logged event (breakpoint), which might not equal $x_i$. For example, a user may have reached their final breakpoint after 3 seconds, but actually quit after 3.8 seconds. We would not have any knowledge of the 3.8 second time, which is their $x_i$.

   (b) Let $t_i, i = 1, \ldots, f$ be the times of the breakpoints. Then $\forall i \in \{1, \ldots, t\}, BP_{U_i} \in \{t_k\}_{k=1}^{f}$. We can then rewrite the likelihood function $L$ as

$$\begin{aligned} L &= \left\{\prod_{i=1}^{m_0} F\left(BP_1 \mid \lambda\right)\right\} \cdot \left\{\prod_{i=m_0+1}^{m_0+m_1} \left(F\left(BP_{U_i+1} \mid \lambda\right) - F\left(BP_{U_i} \mid \lambda\right)\right)\right\} \cdot \left\{\prod_{i=m_0+m_1+1}^{n} 1 - F\left(BP_t \mid \lambda\right)\right\} \\ &= \left\{\prod_{i=1}^{m_0} F\left(t_1 \mid \lambda\right)\right\} \cdot \left\{\prod_{i=m_0+1}^{m_0+m_1} \left(F\left(BP_{U_i+1} \mid \lambda\right) - F\left(BP_{U_i} \mid \lambda\right)\right)\right\} \cdot \left\{\prod_{i=m_0+m_1+1}^{n} 1 - F\left(t_f \mid \lambda\right)\right\} \\ &= \left[F\left(t_1 \mid \lambda\right)\right]^{m_0} \cdot \left\{\prod_{i=1}^{f-1} \left(F\left(t_{i+1} \mid \lambda\right) - F\left(t_i \mid \lambda\right)\right)^{c_i}\right\} \cdot \left[1 - F\left(t_f \mid \lambda\right)\right]^{m_2} \end{aligned}$$

   where $c_i$ is the number of users who quit between $t_i$ and $t_{i+1}$.

Taking the log likelihood gives

$$
\begin{aligned}
l &= \log\left[\left[F\left(t_1 \mid \lambda\right)\right]^{m_0} \cdot \left\{\prod_{i=1}^{f-1}\left(F\left(t_{i+1} \mid \lambda\right) - F\left(t_i \mid \lambda\right)\right)^{c_i}\right\} \cdot \left[1 - F\left(t_f \mid \lambda\right)\right]^{m_2}\right] \\
&= \log\left(\left(F\left(t_1 \mid \lambda\right)\right)^{m_0}\right) + \log\left(\prod_{i=1}^{f-1}\left(F\left(t_{i+1} \mid \lambda\right) - F\left(t_i \mid \lambda\right)\right)^{c_i}\right) + \log\left(\left(1 - F\left(t_f \mid \lambda\right)\right)^{m_2}\right) \\
&= m_0 \log\left(F\left(t_1 \mid \lambda\right)\right) + \left\{\sum_{i=1}^{f-1} \log\left(\left(F\left(t_{i+1} \mid \lambda\right) - F\left(t_i \mid \lambda\right)\right)^{c_i}\right)\right\} + m_2 \log\left(1 - F\left(t_f \mid \lambda\right)\right) \\
&= m_0 \log\left(F\left(t_1 \mid \lambda\right)\right) + \left\{\sum_{i=1}^{f-1} c_i \log\left(\left[F\left(t_{i+1} \mid \lambda\right) - F\left(t_i \mid \lambda\right)\right]\right)\right\} + m_2 \log\left(1 - F\left(t_f \mid \lambda\right)\right)
\end{aligned}
$$

$$
l = m_0 \log\left(1 - e^{-\lambda t_1}\right) + \sum_{i=1}^{f-1} c_i \log\left(e^{-\lambda t_i} - e^{-\lambda t_{i+1}}\right) - m_2 \lambda t_f
$$

(c) Please see attached python code.

(d) Please see attached python code.

5. (a) Code is attached.

The average difference between the MLE estimate of $\lambda$ using simulated quit times and the MLE estimate of $\lambda$ using less accurate event information is shown below for each breakpoint:

| **Break Points** | **Mean Difference (Bias)** |
|---|---|
| $[0.25, 0.75]$ | 0.07 |
| $[0.25, \quad 3]$ | 0.06 |
| $[0.25, \quad 10]$ | 0.14 |

Increasing the time between breakpoints results in an increase in bias, since larger gaps between breakpoints allow for more variation in the possible true quit times of users in those intervals.

This includes the final breakpoint - if the middle breakpoint(s) are chosen such that there could be many users who survive past the final breakpoint, this could also increase bias. This is shown in the second breakpoint choice in the table having lower bias than the first, even though the interval is wider.

(b) For funnel analysis, if event data is being used, breakpoints should be evenly distributed and similar in size over the expected lifetime of users, in order to minimize the variation within breakpoint intervals. This will help minimize bias in the estimate of $\lambda$. Ideally, you would want as many breakpoints as possible to minimize the size of each interval, but this is not usually feasible.