Lab 1, Definitions and Estimation

Dave Harrington

May 2018

The lab files come in two versions: a PDF with the problem statements and an Rmd file that produces the PDF. In most cases, you can work in the Rmd file and enter in your solutions. For the purely algebraic questions, you may either use LaTeX commands to enter your solutions in the Rmd file or write out the answers on paper.

The solution files (a PDF with the solutions, and the Rmd file to produce them) are contained in a separate folder under each lab. Your learning experience with the labs will be more effective if you do not look at the solutions until after you finish working on the questions.

Problem 1: The exponential distribution

Suppose T has an exponential distribution with rate parameter λ; i.e., T has density function

$$f(t) = \lambda e^{\lambda t}, \;\; t \geq 0.$$

a) What are the median, mean, standard deviation, and 20^{th} and 80^{th} percentiles for T, expressed as a function of λ?

b) Show that the cumulative hazard function for an exponential distribution with rate λ is given by the straight line $y = \lambda t, \;\; t \geq 0$.

c) Verify the *memoryless* property of exponential random variables,

$$\forall a, b > 0 : P(T > a + b | T > a) = P(T > b).$$

Problem 1 Solution.

If you are not familiar with LaTeX, it is perfectly fine to write the solution on paper. That applies to all problems in the labs with algebraic solutions.

a)

b)

c)

Problem 2: The Weibull distribution

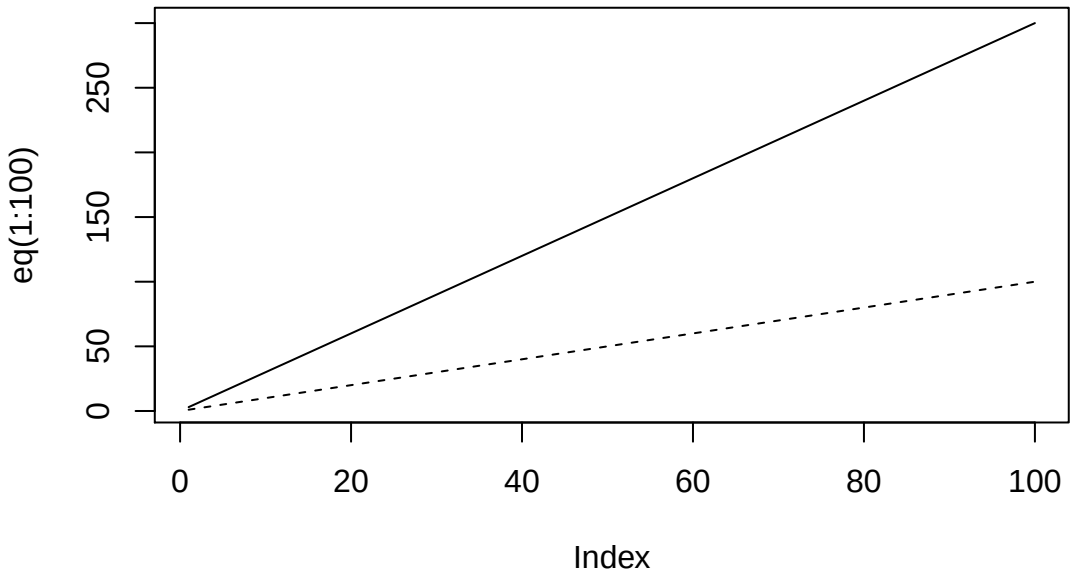
Suppose T has a Weibull distribution with survival function

$$S(t) = e^{-\lambda t^{\gamma}}.$$

a) What are the density, hazard, and cumulative hazard functions for T, expressed as a function of the parameters λ and γ?

b) To plot a function in R, first define the function then call the `plot()` command. Additional functions can then be added to the plot using the `lines` command. The `lines` command must be the next command after `plot()`. For example, the following R chunk produces a plot of two functions of x, $3x$ and x, for values of x from 1 to 100. The function $3x$, named `eq`, is plotted with a solid line; the function x, named `eq2`, is plotted with a dashed line.

```
eq = function(x){3*x}
eq2 = function(x){x}
plot(eq(1:100), type='l', lty = 1)
lines(eq2(1:100), type='l', lty = 2)
```



Make three separate plots, each showing three functions. On the first plot, plot the hazard functions for a Weibull distribution with scale parameter $\lambda = 0.5$ and shape parameters $\gamma = 0.5,\ 1,\ 1.4$; use the range $0 \leq t \leq 5$. On the second plot, plot the survival functions for the same combination of parameters. On the third plot, plot the density functions for the same combinations of parameters.

The code chunk below shows how to create the plot with the hazard functions. Add additional chunks to plot the survival functions and the density functions.

Describe one or two interesting features you observe from the plots of the hazard and survival functions.

```
#define shape and scale parameters
scale = 0.5
shape.1 = 0.5
shape.2 = 1.0
```

```r
shape.3 = 1.4

#define function wb.hazard
wb.hazard = function(lambda, gamma, t){gamma*lambda*t^(gamma - 1)}

#plot first function
plot(wb.hazard(scale, shape.1, (0:100)/20), #specify 0 < t < 5 as (0:100)/20
     type ='l', lty = 1, xlab = "time", ylab = "hazard",
     main = "Weibull hazard functions, scale = 0.5")

#plot additional functions
lines(wb.hazard(scale, shape.2, (0:100)/20), type = 'l', lty = 2)
lines(wb.hazard(scale, shape.3, (0:100)/20), type = 'l', lty = 3)

#add legend
legend(x = "topright", lty = 1:3,
       legend = c("gamma = 0.5", "gamma = 1.0", "gamma = 1.4"))
```

Problem 2 Solution.

a)

b)

Problem 3: Simulating a clinical trial

This exercise begins with the construction of a simulated dataset from a clinical trial with potentially censored failure time observations. The simulation builds the clinical trial dataset using the concept of staggered entry and finite follow-up shown in the graphic in the slide labeled 'Structure of event time data' in Unit 1.

The result of the simulation will be dataset of 300 participants, each with a treatment assignment, a follow-up time X, and a failure indicator δ.

The simulated trial has the following characteristics:

- Participants will enter at a constant rate (i.e., uniform entry) beginning on January 1, 2019; enrollment will end 8 years later on December 31, 2022. Assume that enrollment time is measured in months. The dates are not as important as the length of the enrollment period. Assume that the entry times will be uniformly distributed over the enrollment period.
- Data will be locked and analyzed when the last patient enrolled has been followed for 1 year.
- Assume that the trial will enroll a total of 300 participants.
- Assume this is a simple trial with only one treatment, and that patients in the trial will have a median survival of 18 months.
- Assume that the only form of censoring will be administrative censoring; censoring will happen only for participants whose death has not been observed by the end of the data collection period (i.e., at the time the data are locked).

The most efficient way to simulate the dataset is to, for each case, sample a uniformly distributed entry time and failure time, then use these to calculate the other items needed.

The simulation uses the following R functions. See the R help files for further information about syntax for the function arguments.

- `runif`: simulates uniform random variables
- `rbinom`: simulates binomial random variables
- `rexp`: simulates exponential random variables
- `pmin(a,b)`: finds the component pairwise minimum of two vectors `a` and `b`

When doing a simulation in R, it advisable to set the seed for the random number generator so that the results are reproducible. The code chunk below sets the seed with 14052018, the date for the first day of this course.

a) Run the simulation. Use the techniques discussed in the unit on estimation to check that the simulation behaves as expected. Since the parameters of the survival distribution in the simulation are known, there are several things that could be inspected, such as summary statistics for the failure times, summary statistics for the failure time distribution estimated from the censored data, as well as graphical comparisons of the underlying distributions and those estimated from the censored data.

b) Show that the failure time distribution estimated from the censored data and the true distribution are different than the distribution of X, the follow-up times. It is best to do this

with survival curves. When plotting a survival curve where there is no censoring, the `event` variable can simply be omitted. Why are the medians not useful in this case for showing that the distribution of `obs.time` differs from the other two distributions?

c) Show that the failure time distribution estimated from the censored data and the true distribution are different than the distribution of the follow-up times from only uncensored cases, i.e., the cases where $\delta = 1$. (Hint: The `subset()` command may be useful.)

```
###SIMULATION###

#clear the workspace
rm(list = ls())

#initialize parameters
total.enrollment = 300
enrollment.period = 96
followup.period = 12
max.obs.time = enrollment.period + followup.period
median.failure.time = 18

exp.rate = log(2)/median.failure.time
set.seed(14052018)   #set seed

#simulate entry times and failure times
entry.time = runif(total.enrollment, min = 0, max = enrollment.period)
failure.time = rexp(n = total.enrollment, rate = log(2)/median.failure.time)

#define longest possible follow-up time for each case
potential.obs.time = max.obs.time - entry.time

#calculate failure indicator and observation time
failed = (potential.obs.time > failure.time)   #failure indicator, delta
obs.time = pmin(potential.obs.time, failure.time)

###ANALYSIS###

#inspect distribution of failure times


#plot km estimator of survival


#compare km estimator with true survival function


#summary statistics and survival curve for X = obs.time


#summary statistics and survival curve for only uncensored cases
```

Problem 3 Solution.

a)

b)

c)

Problem 4: Calculating the Kaplan-Meier estimator

Listed below are values of survival time in years for 6 males and 6 females from a small study. Right-censored times are denoted with "+" as a superscript.

Group 0: 1.2, 3.4, 5.0^+, 5.1, 6.1, 7.1

Group 1: 0.4, 1.2, 4.3, 4.9, 5.0, 5.1^+

a) Let $\widehat{S}(t)$ and $\tilde{S}(t)$ denote the Kaplan-Meier estimator and the Fleming-Harrington estimator of the survivorship function of the failure time combining Groups 0 and 1. Calculate $\widehat{S}(1.2)$ and $\tilde{S}(1.2)$ by hand.

   The Fleming-Harrington estimator of the survival function at a time t is $\exp(-\widehat{\Lambda}(t))$, where $\widehat{\Lambda}$ is the Nelson-Aalen cumulative hazard estimator.

b) Figure 1 shows the Kaplan-Meier estimates separately for the two groups. Suppose that we know that the largest observation in the female group is a censored observation. Which group, 0 or 1, represents the female group?

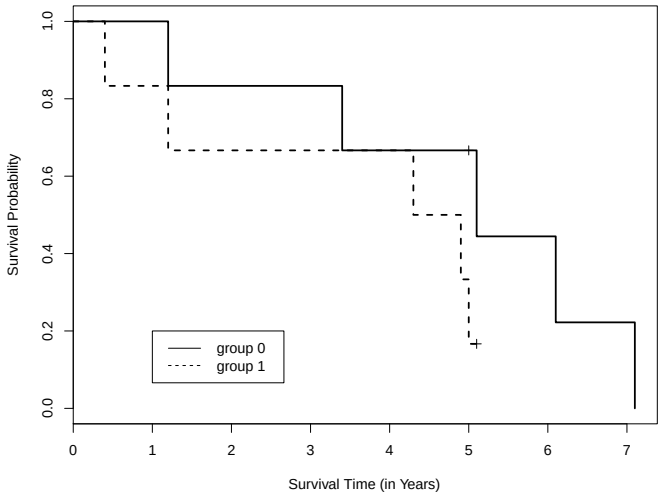


Figure 1: Kaplan-Meier estimates of survival times, by gender

c) The table below provides the Kaplan-Meier estimates for the survival function ($\widehat{S}(t)$), estimated failure time function $\widehat{F}(t)$, and the estimated standard errors for $\widehat{S}(t)$ using the Greenwood Formula for Group 1. Calculate the values $\widehat{S}(5.1)$ and $\widehat{F}(5.1)$.

Product-Limit Survival Estimates

years	$\widehat{S}(t)$	$\widehat{F}(t)$	$\widehat{s.e.}(\widehat{S}(t))$
0.00	1.0000	0	0
0.40	0.8333	0.1667	0.1521
1.20	0.6667	0.3333	0.1925
4.30	0.5000	0.5000	0.2041
4.90	0.3333	0.6667	0.1925
5.00	0.1667	0.8333	0.1521
5.10^{+}	????	????	.

d) Use the above table to calculate the estimated 75^{th} percentile of the survival time in Group 1. Provide a point estimate and a 95% confidence interval for $\widehat{S}(t)$ at that time, using the "log-log" approach. You can use the following result directly:

$$\text{Var}\left(\log(-\log(\widehat{S}(t)))\right) = \left(\frac{1}{\widehat{S}(t)\log(\widehat{S}(t))}\right)^2 \text{Var}\left(\widehat{S}(t)\right)$$

Problem 4 Solution.

a)

b)

c)

d)

Problem 5: Estimating the length of stay in a nursing home

The National Center for Health Services Research in the United States studied 36 for-profit nursing homes to assess the effects of different financial incentives on length of stay. "Treated" nursing homes received higher daily allowances for Medicaid patients and bonuses for improving a patient's health and sending them home. The study included 1,601 patients admitted between May 1, 1981 and April 30, 1982. The data appear in Morris, et al., *Case Studies in Biometry*, Chapter 12.

Data from this study are contained in the dataset `nursing.home` in the package `eventtimedata`. Refer to the package documentation for information on the variable definitions and coding.

a) Using the Kaplan-Meier estimator, provide a graph of the estimated survival function for length of stay for each of the two groups defined by the intervention variable `rx`. Length of stay is the variable `stay`, and the indicator for discharge is the numeric variable `cens`. Does the treatment appear to have changed the length of stay? Support your answer with some summary statistics; a later lab will ask for a significance test.

b) Do you notice anything strange about the appearance of the survival curves in the two groups?

c) Plot the cumulative hazard function for each group.

d) The 5-number summary for a distribution consists of the minimum and the maximum, the first and third quartile, and the median. As often happens with censored data, standard terms may need to be carefully redefined to accommodate censoring.

   Provide two 5-number summaries for the control group with these data. For the first summary, use just the observation times stored in `stay`, ignoring censoring; the `summary( )` command can be used. For the second 5-number summary, the elements should be the estimated 1st quartile, 3rd quartile, and median from the Kaplan-Meier and the largest and smallest observed times to discharge for censored cases. The quantiles from a KM fit in R can be calculated using the `quantile` function. The command `quantile(km, 0.5)` returns the median with its confidence interval for a KM fit named `km`.

e) Most manuscripts of randomized trials with event-time data report a 'median follow-up'. There are several definitions of median follow-up in use. The first uses the median of all observed follow-up times. This is the variable X in the slides and the variable `stay` in this dataset. The second reports the median of the follow-up times among participants who have not had an observed failure.

   Calculate each of these for the control group in the nursing home data. How might each of these statistics be useful?

Here is the beginning of a code chunk to get you started.

```
library(survival)
library(eventtimedata)
data(nursing.home)

nrshome.km = survfit(Surv(stay, cens) ~ rx, data = nursing.home)
quantile(nrshome.km, 0.5)

control.grp = (nursing.home$rx == "Control")
```

```
not.failed = (nursing.home$cens == 0)
```

Problem 5 Solution.

a)

b)

c)

d)

e)

Problem 6: The exponential distribution, again

Suppose T has an exponential distribution with rate parameter λ. Let $(t_1, t_2, \ldots, t_n)$ be a random sample of n observations from T.

a) The likelihood function for a set of observations from a distribution with density $f(t, \lambda)$ is

$$L(\lambda) = \prod_{i=1}^{n} f(t_i, \lambda).$$

Calculate the likelihood function for λ as a function of the data $(t_1, t_2, \ldots, t_n)$.

b) The maximum likelihood estimator (MLE) for a parameter λ is the value $\widehat{\lambda}$ that maximizes the likelihood function. What is the maximum likelihood estimator for λ based on the data $(t_1, t_2, \ldots, t_n)$? The maximum likelihood estimator is usually calculated by maximizing the log of the likelihood function

$$\ell(\lambda) = \log(L(\lambda)).$$

c) Suppose the observations from the exponential are subject to independent censoring. For each case, let

- U_i denote the censoring variable,
- T_i denote the underlying, possibly censored event time,
- X_i denote the observed follow-up time, and
- $\delta_i = I(T_i \leq U_i)$ be the failure indicator.

Assume that the data for each case consist of (x_i, δ_i), $i = 1, \ldots n$.

When data are censored in a parametric model, the likelihood for the data is given by

$$\prod_{i=1}^{n} [f(t_i)]^{\delta_i} [S(t_i)]^{1-\delta_i}$$

Show that the maximum likelihood estimator for λ based on the censored sample is

$$\widehat{\lambda} = \frac{\sum \delta_i}{\sum X_i}$$

d) In words, what does the $\sum X_i$ represent? Why do epidemiologists call $\widehat{\lambda}$ the *occurence/exposure* rate?

Problem 6 Solution.

a)

b)

c)

d)

Problem 7: Fitting an exponential distribution

This problem uses the code from the simulated dataset from a clinical trial with a total of 300 randomized patients (Problem 3).

a) Estimate and plot the survival distribution for all patients and plot the estimate for median survival with confidence intervals.

b) Use the formula for the MLE for an exponential distribution to estimate the rate parameter, then estimate the median survival using the formula derived earlier in the lab for the median of an exponential distribution.

c) On a single plot, plot the Nelson-Aalen cumulative hazard estimator and the estimated cumulative hazard from the exponential model. What do you notice?

Problem 7 Solution.

a)

b)

c)