STAT 417: Survival Analysis Methods

Computing Assignment 1: 1-8-25 (Due Thursday 1-9-25 at 11:59pm)

Name: Andrew Kerr

You will be using Minitab software for this assignment so please be sure to download it or have access to it. Minitab needs to be installed before the .mwx files required for this activity can be opened. Some Minitab instructions to answer the questions are provided at the end of the assignment. Try to figure out how to manipulate the data in the worksheet on your own.

1. **Chocolate Chips.** Melting times from a past chip activity are provided in the Minitab file MELT TIMES W025. Open the file to perform data manipulation and calculations using the Minitab software. We will assume that the times are a random sample of all chocolate chip melting times. The observed melting times (in seconds) are in the Melt Time column and column C contains the values of the censoring status variable (0 = right censored; 1 = complete).

(a) Briefly explain why an observed melting time of 85 seconds is right censored.

An observation of 85 seconds is right censored because 85 seconds is the end time for our study. Any times recorded at or after 85 seconds occur after the study concludes, so we do not know when the event time occurs for these subjects, only that it is after 85 seconds.

(b) Compute the (estimated) average melting time using the values in Melt Time. Minitab instructions for computing descriptive statistics are provided at the end of the assignment. Do you think this value over-estimates or under-estimates the true (population) average chip melting time? Briefly explain.

Average: **68.4146 seconds**

This value **under-estimates** the true average chip melting time because we have right censoring. This means that our large times are being censored to 85 seconds, lower than what their actual values are, which in turn lowers our average.

(c) Would it be more appropriate to add or subtract 5 seconds to the right censored times? Adjust the right censored times, accordingly, by adding or subtracting 5 seconds to the incomplete times, and recompute the sample average melting time. Verify your answer to the second part of Part (b).

It would be more appropriate to **add 5 seconds** to right censored times because right censoring lowers these times, so we should add time back to correct for this rather than remove more time.

New Average: **69.5122 seconds**

The new average is larger than the original average, as predicted in (b)!

(d) Even though you appropriately adjusted the times upward or downward, the correct amount of time to adjust each of the incomplete observations is unknown. Now remove the right censored chip melting times, recompute the average melting time, and report it. How does removing the incomplete observations affect the mean chip melting time found in Part (b)?

Truncated Average: **63.75 seconds**

Removing the uncompleted observations lowers the average chip melting time because we are removing the largest record values from the data.

(e) Summarize the effects of 1) treating the right censored data as complete and 2) removing right censored observations from the data set on the average chip melting time.

1) Under-estimates the true average chip melting time, and we do not know by how much since we do not know the difference between the censored and true values.

2) Greatly lowers the average chip melting time since we remove the largest observed values. Here our data is fully complete, but would restrict what we can generalize our results to since we are truncating our data.

2. **Age at First Bike Ride.** Recall the survey question that asked you to indicate your age when you learned to ride a bike. If the student could not remember her/his exact age, the student provided an age at which they knew how to ride a bike, but greater than the actual age that they had learned to ride a bike. These ages are provided in the Minitab file Bike Ages W2025 along with a left censoring status variable. The age (in years) when the student learned to ride a bike is in the Age column and column C contains the values of the censoring status variable (0 = left censored; 1 = complete).

(a) Describe the beginning of time, time metric, and time-to-event random variable.

Beginning of time: Birth (Age 0)

time metric: Years (age)

time-to-event RV: T = Age student first rode a bike

(b) Briefly explain what it means for an age to be left-censored (in the context of the problem).

Age is left-censored in this study when the student is unsure of what age they first learned how to ride a bike, but know that it is before some specified age.

(c) Compute the (estimated) average age when students first learned to ride a bike using the values in Age. Do you think this value over-estimates or under-estimates the true mean age at which students first learned to ride a bike? Briefly explain.

Estimated Average: **7.35714 years old**

This estimation will likely be an **overestimate** because the students who were unsure of what age they first rode a bike at will put a later age when they are certain they knew how to ride a bike by, which is an increase from this students true age, in turn inflating the estimated average.

(d) Would it be more appropriate to add or subtract 2 years to the left-censored observations? Adjust the left-censored ages, accordingly, by adding or subtracting 2 years to the incomplete times, and recompute the sample average age. Verify your answer to the second part of Part (c).

It would be more appropriate to **subtract 2** years from these left-censored observations.

New Estimated Average: **6.11905 years old**

The new average is smaller than the original average, as predicted in (c)!

(e) Remove the left-censored observations (ages) from the data set and recalculate the mean age of the sample. How does this value compare to your answer in Part (c)?

Truncated Average: **7.125 years old**

Removing the left-censored observations lowers the average age. This could be because students who were unsure of what age they first rode a bike imputed larger ages when they were certain they could ride a bike, thus inflating the estimated average.

(f) Describe the fundamental difference between the incomplete observations in Problems (1) and (2). (State something more meaningful than “Problem (1) has right censored data and Problem (2) has left censored data” :) )

The fundamental difference between the incomplete observations in Problems (1) and (2) lies in how these observations affected the data. In Problem (1), these observations biased the data downward since we are not able to observe the larger chip melting times. On the other hand, Problem (2) biases the data upwards because the event time is earlier than reported, inflating the average.

In short, **Problem (1) leads to an underestimation** while **Problem (2) leads to an overestimation**!

Minitab Instructions for Computing Descriptive Statistics:

1. Open Minitab software

2. On the top menu, follow: Stat → Basic Statistics → Display Descriptive Statistics...

3. Double click “Simple” box

4. Enter variable into the “Variables:” box.

5. Click “OK”