Advice on Statistical Methods for Bradford Hull's Research

Huashi Li (lihuashi1@arizona.edu), Alex Salce (asalce@arizona.edu), Taryn Laird (tarynl@arizona.edu)

2024-09-26

Executive Summary

Bradford Hull is a 5th year PhD Student in Molecular and Cellular Biology. He is currently researching lifespans of Caenorhabditis elegans worms and the effects of different stressors for worms living in different media. Part of his research goal is to perform statistical analyses to determine whether any of these stressors have significant effects on the lifespans of the worms, both within media (compared to a control) and between media for the same stressors. He is primarily interested in comparisons of the same stressors between media for his research.

By the time of our consultation, Bradford had already performed experiments and collected data for both agar and liquid media, using copper and DTT (dithiothreitol) as stressors. Due to some challenges unique to each media, he was not able to perform the same procedure to capture data for each media.

Bradford was able to provide his data in Excel for our review. The agar media experiments were performed in a manner suitable for the framework of Kaplan-Meier and log rank test methods for lifespan analysis, which he has already successfully performed using the `survival` package in R. However, his data collection for the liquid media experiments required a different procedure that we determined was not suitable for the Kaplan-Meier analysis framework, and the desired statistical analyses would require a different approach.

Detailed Summary

1.Background Bradford is seeking statistical consultation to help him determine how to adequately approach a statistical analysis for data which he has already captured from several experiments. His research intent is to study the effects of different stressors on the lifespans of worms living in different media, and to compare how the same stressors in different media can affect the lifespans of worms when prepared in different media. Ultimately this research should aid in generalizing protocol for research using Caenorhabditis elegans as test subjects.

During our consultation, he was able to describe in detail the procedures that were carried out. First, agar experimental units were made by adding $N = 40$ worms to 9 total agar culture dishes. Stressors, copper and DTT (dithiothreitol), were each added to three of the cultures so that in total there were three of each stressor, as well as 3 control cultures in agar media. The $N = 40$ and three replicates were chosen based on a power analysis from within the MCB department, but no further details were provided.

Next, liquid experimental units were made by adding approximately $N \approx 5000$ worms to each of three replicate liquid reservoirs per stressor; in total there 9 liquid culture reservoirs were prepared.

It is important to note that all worms were originally drawn from the same population batch of eggs, and that all experimental units were prepared at the same time (same day).

Ultimately, Bradford's goals for analysis are the following (in his words, taken from a follow up email statement).

- "...determine if the liquid data can be fit into a K-M curve (and therefore be able to use the log-rank test on it)"
- "...determine if liquid copper and DTT are different from liquid control"

- "determine if lifespan reductions for both copper and DTT from goal 2 are different from lifespan reductions in the agar data"

The team agreed that our next steps would be the following.

- Analyze whether the Kaplan-Meier (K-M) procedure is appropriate
  - Within media by stressor
  - Between media by analogous stressor
- Determine what conditions need to be met for a valid K-M analysis, and if those conditions cannot be met, what approach do we recommend for analysis?

2.Methods The study began the day the experimental units were prepared, and all worms were alive at the start time step. At every time step (usually one or two days), all experimental units had data collected. However, the procedure for data collection was not the same between experimental units in agar media and units in liquid media.

Agar At each timestep, each agar culture sample was able to be observed in its entirety, that is all of the original $N = 40$ worms in each sample could be observed, and the number of dead worms were counted. In this sense, individual worms were tracked throughout each timestep of the study, which is an important condition for Kaplan-Meier lifespan analysis. It is worth noting that each of the three replicates are in separate plates, and no worms were able to escape. Each experimental unit's data was recorded until all worms in their respective units were deceased.

Liquid Due to challenges with measurement/observation of the worms in liquid media, the experimental procedure did not allow for the observation of the same worms. At each time step, one sample of approximately 150μL was drawn from each separate replicate and raw numbers of dead/alive worms were assessed using special lab equipment. Each sample captured approximately 90 worms, and after data collection these worms were discarded. Each experimental unit's data was recorded until there were two consecutive measurements of all-dead worms.

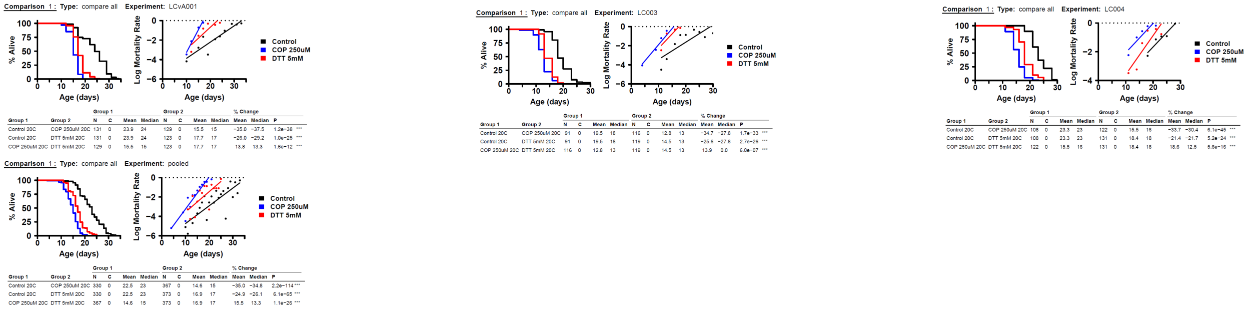
3.Results and Evaluation Three separate full experiments using the 9 experimental units (three treatments and three replicates) were performed between January and June 2024. A new experiment wouldn't be started until a previous was completed, so there was no overlap between studies. Each experiment lasted approximately a month.

Agar The experimental procedure for agar media captures time-to-event data (the event in question being death) for all individual worms throughout the length of the study. Observations were of the same individual worms in each experimental unit at every time step. Time steps were generally once every two days, excluding weekends, but in all cases all units were sampled at each time step.

All experimental units were sampled until every worm was dead, and no worms escaped any of the experimental units, so there is no censoring of any of the subjects. This means that we have adequate conditions for a Logrank test, and we are able to use a simple Empirical CDF to visualize the survival functions. A Kaplan-Meier curve would only technically be necessary to communicate any censored data, so it is not necessary here (although visually very similar).

The analyses that Bradford performed of the data using the `survival` package in R are valid for comparisons between treatments within the agar media, however we should note that the assumptions should be confirmed by Bradford for validity.

- All replicates have identical treatment conditions.
- Survival times are independent between subjects.
- Pooled analyses must maintain these same assumptions.



Comparison 1 : Type: compare all Experiment: LCvA001

Group 1	Group 2	Group 1 N	Group 1 C	Group 1 Mean	Group 1 Median	Group 2 N	Group 2 C	Group 2 Mean	Group 2 Median	% Change Mean	% Change Median	P	
Control 20C	COP 250uM 20C	131	0	23.9	24	129	0	15.5	15	−35.0	−37.5	1.2e−38	***
Control 20C	DTT 5mM 20C	131	0	23.9	24	123	0	17.7	17	−26.0	−29.2	1.0e−25	***
COP 250uM 20C	DTT 5mM 20C	129	0	15.5	15	123	0	17.7	17	13.8	13.3	1.6e−12	***

Comparison 1 : Type: compare all Experiment: LC003

Group 1	Group 2	Group 1 N	Group 1 C	Group 1 Mean	Group 1 Median	Group 2 N	Group 2 C	Group 2 Mean	Group 2 Median	% Change Mean	% Change Median	P	
Control 20C	COP 250uM 20C	91	0	19.5	18	116	0	12.8	13	−34.7	−27.8	1.7e−33	***
Control 20C	DTT 5mM 20C	91	0	19.5	18	119	0	14.5	13	−25.6	−27.8	2.7e−26	***
COP 250uM 20C	DTT 5mM 20C	116	0	12.8	13	119	0	14.5	13	13.9	0.0	6.0e−07	***

Comparison 1 : Type: compare all Experiment: LC004

Group 1	Group 2	Group 1 N	Group 1 C	Group 1 Mean	Group 1 Median	Group 2 N	Group 2 C	Group 2 Mean	Group 2 Median	% Change Mean	% Change Median	P	
Control 20C	COP 250uM 20C	108	0	23.3	23	122	0	15.5	16	−33.7	−30.4	6.1e−45	***
Control 20C	DTT 5mM 20C	108	0	23.3	23	131	0	18.4	18	−21.4	−21.7	5.2e−24	***
COP 250uM 20C	DTT 5mM 20C	122	0	15.5	16	131	0	18.4	18	18.6	12.5	5.6e−16	***

Comparison 1 : Type: compare all Experiment: pooled

Group 1	Group 2	Group 1 N	Group 1 C	Group 1 Mean	Group 1 Median	Group 2 N	Group 2 C	Group 2 Mean	Group 2 Median	% Change Mean	% Change Median	P	
Control 20C	COP 250uM 20C	330	0	22.5	23	367	0	14.6	15	−35.0	−34.8	2.2e−114	***
Control 20C	DTT 5mM 20C	330	0	22.5	23	373	0	16.9	17	−24.9	−26.1	6.1e−65	***
COP 250uM 20C	DTT 5mM 20C	367	0	14.6	15	373	0	16.9	17	15.5	13.3	1.1e−26	***

Figure 1: Agar survival analyses

Liquid The experimental procedure for the liquid media does not capture time-to-event data as would be required for a similar analysis as the agar media. This method uses a random sample draw from the experimental unit population without replacement at each time step, and can only observe the sampled worms at that time step.

K-M and log rank test analyses require that experimental subjects all be observed from the beginning of the study until either a death or a censoring event. We can actually safely assume that worms that were alive at a sample have been alive for the length of the experiment, however we cannot say whether or not any of the sampled dead worms were alive at the previous time step. So, we cannot estimate the survival curve in the same manner.

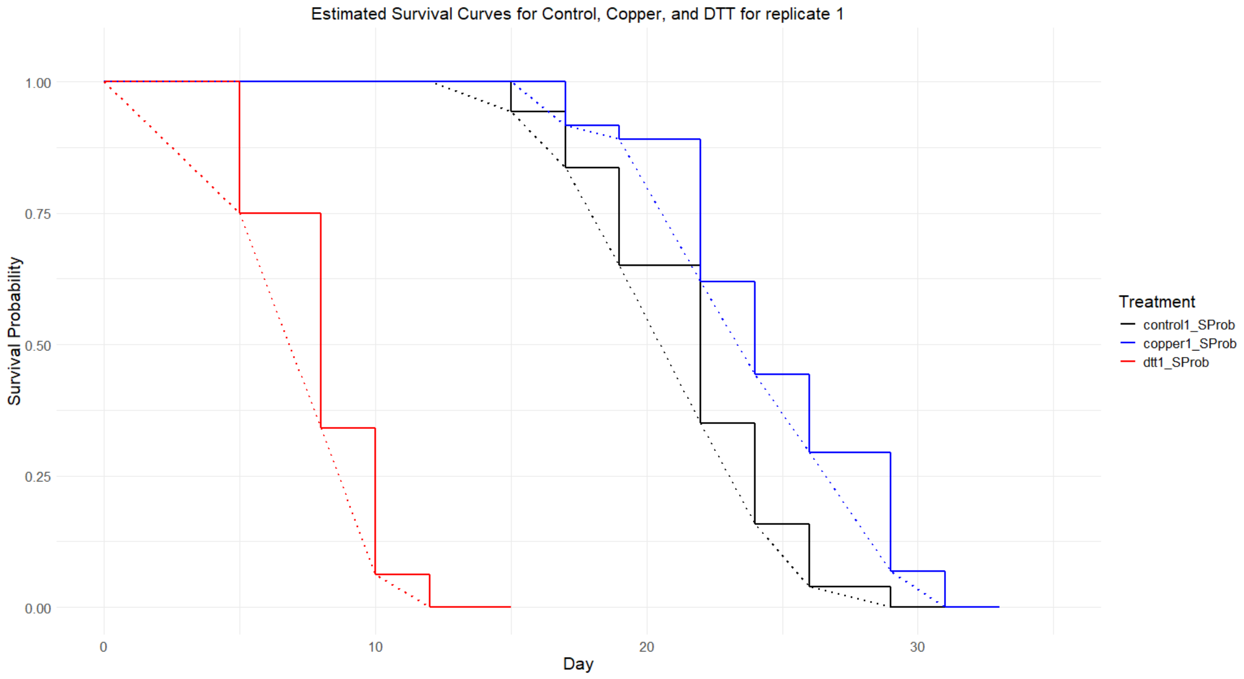
There are also some considerations that we should understand about the experimental procedure. We had concerns of worms possibly decaying/dissolving before they could be sampled, if they died before they were sampled. Per Bradford, "... There was little to no degradation of dead worms and worms only eat bacteria, so once they died they just floated around and were able to be sampled until the end of the experiment".

For the liquid media, the experiment does not meet the assumptions for Kaplan-Meier analysis, so it is not an appropriate method.

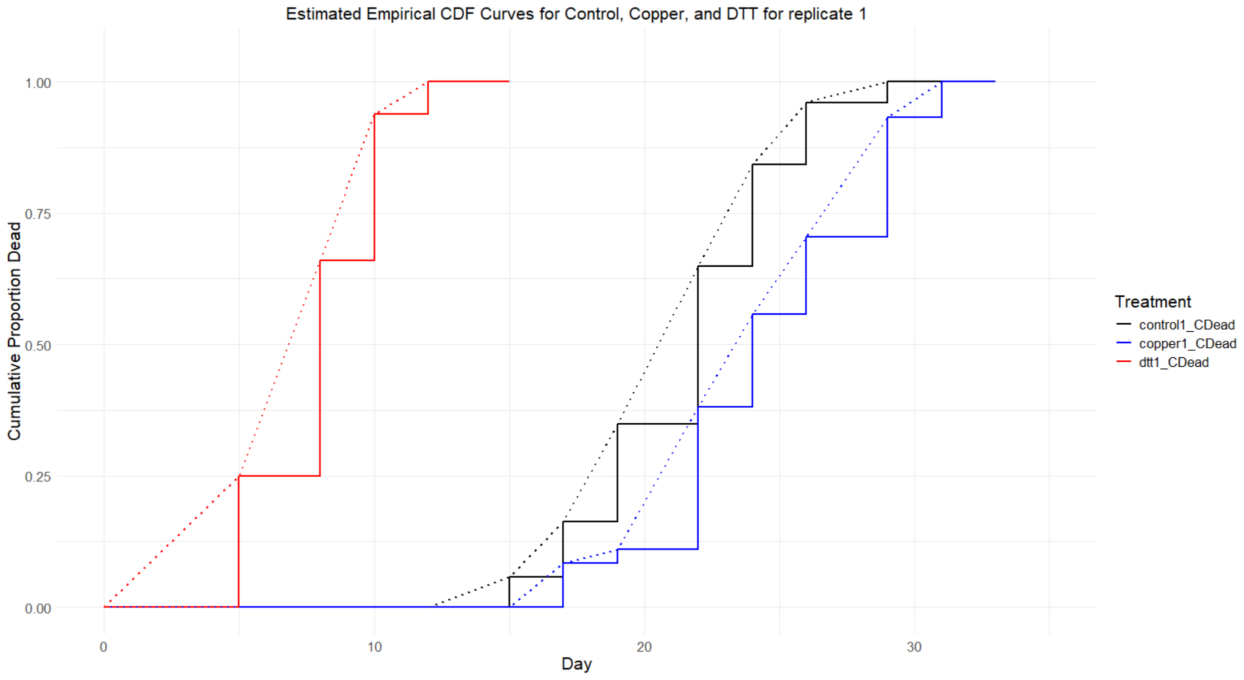
Recommendations

1.Estimated survival curves, Estimated empirical CDF for liquid media For the liquid media replication 1 of experiment 1, we calculated the survival probabilities and plotted the estimated survival curves and empirical CDF(See R code in appendix).
The survival ratio was calculated as the ratio of alive worms to the total worms for each treatment at each time point.
Cumulative survival probabilities were derived as the product of survival ratios over time for each treatment.
The cumulative proportion of dead worms was then computed as 1 minus the survival probability for each treatment.

Survival curves Control (Black Line): The survival probability for the control group shows a gradual decline over time. There is a noticeable decrease in survival starting around Day 22, with worms surviving until approximately Day 30.
Copper (Blue Line): The copper treatment group has a longer survival time.
DTT (Red Line): The DTT treatment group has the shortest survival time. The survival probability drops very quickly, and by around Day 10, the majority of worms are dead.



Empirical Cumulative Distribution Function (ECDF) Control (Black Line): The cumulative proportion of dead worms increases steadily over time, with a sharp increase around Day 22.
Copper (Blue Line): The copper group shows an increase in the cumulative proportion dead starting around Day 22. Like the control group, the copper group experiences a sharp increase in death around this point.
DTT (Red Line): The DTT treatment group has a steep increase in the cumulative proportion of dead worms early in the experiment, reaching close to 100% mortality by Day 10.



2.Median lifespan comparisons We can estimate the median lifespan time for each treatment from the survival curves by identifying the time point where the survival probability crosses 0.5. The median time is around 7 days for DTT, 21 days for control, and 23 days for copper in replicate 1 of experiment 1, so the median time follows the order: DTT < Control < Copper.

Comparing the median time between liquid and agar: In agar media, based on the survival curves, the median time is Copper < DTT < Control, which is different from the order in liquid media.

Recommended further study

Run a similar analysis as in replicate 1, and plot the survival curves and empirical CDF for replicates 2 and 3, as well as for experiments 2 and 3.

Appendix

1.Kaplan-Meier assumptions – Survival time: The survival time must be precisely measured for each individual in the population.
– Censoring: Censoring must be non-informative, meaning it's unrelated to the outcome of interest.
– Survival probabilities: Survival probabilities should be the same for people recruited early or late in the study.
– Event occurrence: The time of the event must be available.

```
# Load library
library(ggplot2)
library(tidyr)
library(dplyr)
# Read in the experiment 1 data file
data <- read.csv("C:/Bradford Hull/Liquid_experiment1.csv", header = TRUE)
# Calculate survival ratios, survival probabilities, and cumulative proportion of dead for replicate 1
data <- data %>%
        mutate(
               control1_SRatio = Control1_Alive / (Control1_Alive + Control1_Dead),
               control1_SProb  = cumprod(control1_SRatio),
               control1_CDead  = 1 - control1_SProb,

               copper1_SRatio  = Copper1_Alive / (Copper1_Alive + Copper1_Dead),
               copper1_SProb   = cumprod(copper1_SRatio),
               copper1_CDead   = 1 - copper1_SProb,

               dtt1_SRatio     = DTT1_Alive / (DTT1_Alive + DTT1_Dead),
               dtt1_SProb      = cumprod(dtt1_SRatio),
               dtt1_CDead      = 1 - dtt1_SProb)
# Add a row for Day 0 where all survival probabilities are 1 and all cumulative proportion of dead is 0
day_0 <- data.frame(
                    Day = 0,
                    control1_SProb = 1,
                    copper1_SProb = 1,
                    dtt1_SProb = 1,

                    control1_CDead = 0,
                    copper1_CDead = 0,
                    dtt1_CDead = 0)
# Add this row to the original data
data <- bind_rows(day_0, data) %>%
        arrange(Day)
# Data for plotting survival curves
data_SProb <- data %>%
              select(Day, control1_SProb, copper1_SProb, dtt1_SProb) %>%
              gather(key = "Treatment", value = "Survival_Prob", -Day)
# Plot the survival curves using ggplot2
survival_curves <- ggplot(data_SProb, aes(x = Day, y = Survival_Prob, color = Treatment)) +
                 geom_step(size = 1) +
                 geom_line(size = 1, linetype = "dotted") +
                 labs(title = "Estimated Survival Curves for Control, Copper, and DTT for replicate 1",
                 x = "Day",
                 y = "Survival Probability",
                 color = "Treatment") +
                 theme_minimal() +
                 theme(plot.title = element_text(hjust = 0.5, size = 12)) +
                 ylim(0, 1.05) +
                 xlim(0, 35) +
                 theme(text = element_text(size = 12)) +
                 scale_color_manual(values = c("control1_SProb" = "black",
                                               "copper1_SProb" = "blue",
                                               "dtt1_SProb" = "red"))
# Empirical CDF
# Data for plotting ECDF
data_ECDF <- data %>%
             select(Day, control1_CDead, copper1_CDead, dtt1_CDead) %>%
             gather(key = "Treatment", value = "Cumulative_Dead", -Day)
# Plot the ECDF curves using ggplot2
ECDF_curves<-ggplot(data_ECDF, aes(x = Day, y = Cumulative_Dead, color = Treatment)) +
             geom_step(size = 1) +
             geom_line(size = 1, linetype = "dotted") +
             labs(title = "Estimated Empirical CDF Curves for Control, Copper, and DTT for replicate 1"
             x = "Day",
             y = "Cumulative Proportion Dead",
             color = "Treatment") +
             theme_minimal() +
             theme(plot.title = element_text(hjust = 0.5,size = 12)) +
             ylim(0, 1.05) +
             xlim(0, 35) +
             theme(text = element_text(size = 12)) +
             scale_color_manual(values = c("control1_CDead" = "black",
                                           "copper1_CDead" = "blue",
                                           "dtt1_CDead" = "red"))
```

2. R code