Miniproject 4 – ANOVA, ANCOVA and Survival Analysis

StudentID 100225776 Applied Statistics

Question 1

We will use ANAOVA and ANCOVA to analyse the yield of cocoa plants grouped by height and by genotype to discover relationships in the data.

Part A – Visual Interpretation of Data

For this part I have generated two boxplots (with associated Stripcharts), one tracking the crop yield grouped by Heightgroup, the other with the yield grouped by the Genotype. In both, the data has been given a random horizontal jitter to make it easier to see samples with similar yields.

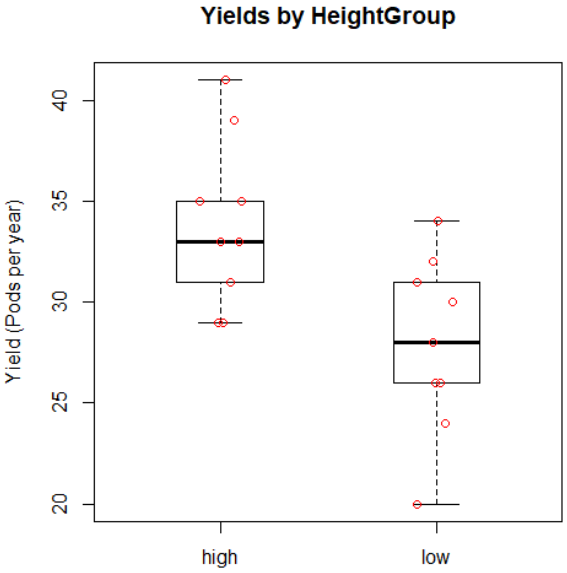

Figure 1- Crop yields by Heightgroup

Figure 1 shows data grouped by Heightgroup. From this we can clearly see that on average the taller plants produce a greater yield of cocoa pods, approximately five more pods per plant per year or approximately a 20% greater yield over the shorter plants, on average. The taller plants have a distribution with a longer tail towards the upper bound of the range, suggesting this may not be a normal distribution. The shorter plants are distributed more normally however there is a single value at the lowest bound of this group which appears significantly different from the rest of the set however this may be due to the relatively small sample size of the results and is still within the bounds of the set. Both data sets have the majority of their results sitting on or within the interquartile ranges of their groups, and the most productive plant from the high group has over twice the yield of the least productive plant in the low group. The plots suggest there is a clear difference in the yield of a plant based on what height group that plant sits within.

Figure 2 shows the same data as Figure 1, but grouped by genotype instead. We can see from this grouping that there is now one confirmed outlier, in group AA. This outlier could be a false reading (incorrect count of pods) which is highly possible as no other tree supposedly produced as high of a yield as that one did, and the value of 31 would place it in an acceptable position on both plots. It could also be a sign of a heavily tailed distribution. If the latter is true, we should be careful in assuming the distribution is normal and test for it. Aside from the outlier value, we have two values which are extreme but within the bounds of the set. The spread within each group is quite tight, with almost every value appearing within the interquartile range, though this is likely due to only having a few samples per group. The spread for group AA is very tight, implying that a plant with the AA genotype will almost certainly produce a yield in the 30-35 pod-per-year range, however the outlier value in this group pushes the mean to close to the third quartile. Despite there being only a

few samples per group, it is reasonable to say there is a difference between the yield of the plant based on its genotype.

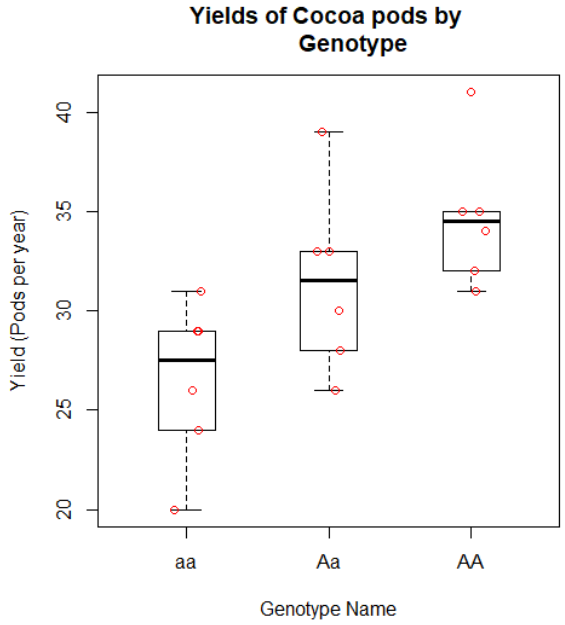

Figure 2 - Crop yields by Genotype

Figure 3 shows an interaction plot between genotypes for high plants compared to genotypes for low plants. As the lines begin to converge at Genotype AA, we can say that there is only a small interaction between Genotype and Yield. Figure 4 better shows this lack of interaction, where the plots for genotypes AA and aa in parallel show no interaction. Genotype Aa at an angle shows some slight interaction. In general, however, I would say this interaction is not statistically significant as the lines in both cases are mostly parallel and do not meet or cross at any point. This is surprising to me as I initially expected (from figure 1 and 2) that since the AA genotype produced taller trees, and taller trees produced a larger yield (and vice versa for the aa genotype), that there would be an interaction between the two.

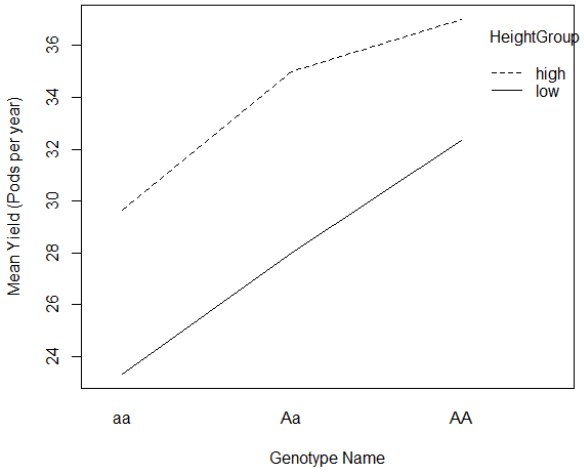

Figure 3 - Interaction Plot between Genotype and height group

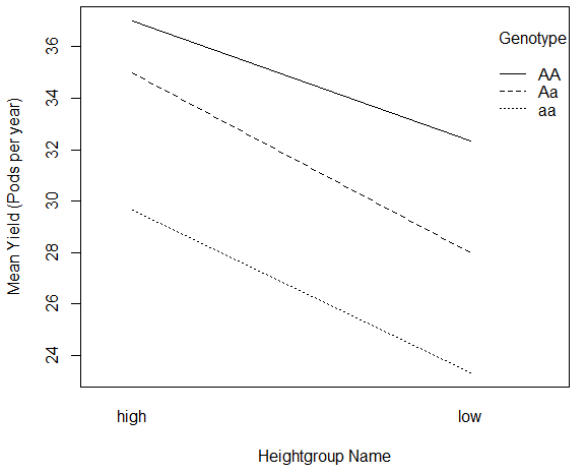

Figure 4 – Interaction between height group and Genotype

Part B – Two-Way ANOVA Interpretation

The two-way ANOVA will be used to test the below significant effects:

1. H_0 = The means of our observations grouped by Genotype are the same.
2. H_0 = The means of our observations grouped by Heightgroup are the same.
3. H_0 = There is no interaction between Genotype and Heightgroup.
4. H_0 = There are no significant effects present.

In addition, we also look to see if
To begin, we must first check for the data balance of observations across the groups. Since the variables Genotype and Heightgroup are not numerical, they are automatically treated as factors, so we can simply draw a table from them.

See Appendix A – R Commands – Lines 37-39.

Spread of observations per group:

	Genotype		
Heightgroup	*aa*	*Aa*	*AA*
High	3	3	3
Low	3	3	3

From these results we can see that the data are balanced with three observations within each group, so we can proceed with the two-way ANOVA.
We will use the ANOVA model

$$X_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}$$

to run a two-way ANOVA on our given data set, where α_i are the effects of Heightgroup on Yield, β_j are the effects of Genotype on Yield and $(\alpha\beta)_{ij}$ are the interaction effects between Heightgroup and Genotype. We create contrasts to ensure parameter estimates are unique, which is achieved using the contrast sum function (`contr.sum`) in R. We then run a linear model on our data set to produce the two-way ANOVA.

See Appendix A – R Commands – Lines 50-54.

Two-way ANOVA of Yield against Heightgroup and Genotype:

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Heightgroup	1	162.000	162.000	23.7073	0.0003856
Genotype	2	203.444	101.722	14.8862	0.0005620
Heightgroup: Genotype	2	4.333	2.167	0.3171	0.7341961
Residuals	12	82.000	6.833		

These results tell us that Yield is dependent on Heightgroup (P = 0.0003856) and that Yield is also dependant on Genotype (P = 0.000562). This allows us to reject our first and second null hypotheses.

We can also see that there is no significant interaction between Heightgroup and Genotype (p = 0.7342), as predicted from our interaction plots. This allows us to accept the third null hypothesis that there is no interaction between Genotype and Heightgroup. To test our fourth null hypothesis, we use the `summary()` function in R:

See Appendix A – R Commands – Lines 50-82.

Summary of ANOVA significant effects:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	30.8889	0.6161	50.133	2.6e-15
Heightgroup	3.0000	0.6161	4.869	0.000386
Genotype	-4.3889	0.8714	-5.037	0.000291
Genotype 2	0.6111	0.8714	0.701	0.496471
Heightgroup: **Genotype**	0.1667	0.8714	0.191	0.851510
Heightgroup: **Genotype 2**	0.5000	0.8714	0.574	0.576687

RSE: 2.614 on 12 df. **Multiple R^2**: 0.8185. **Adjusted R^2**: 0.7429. **F-statistic**: 10.82 on 5 and 12 DF. **P-value**: 0.0004069.

From the above results we can see that there is high significance in the results (P = 0.0004069), which is below the significance level of 0.05 which means that our fourth null hypothesis should be rejected, and we accept that significant effects are present.

We can find parameter estimates from running `results$coefficients`. However, then the coefficients would need to be calculated manually. The easier way of computing this is with `tapply()`, which will compare against all sample means and find $\bar{x}_{ij}$. From the ANOVA significant effects table, we can see that the overall average is 30.8889 and the sum of squares is 162. Once the parameter estimates have been found, we need to find the confidence interval. For this test, our confidence level is 0.95 as our chosen significance level is 0.05. We can see from the above table that the Residual Standard Error (RSE) is 2.614 on 12 degrees of freedom. Since each of the three Genotype groups contains 9 observations each, the standard error of each mean is:

$$\frac{MS(E)}{\sqrt{9}} = \frac{6.833}{3} = 2.2778$$

The variance is twice larger and so standard deviation is:

$$\frac{MS(E)}{\sqrt{4.5}} = \frac{6.8333}{2.1213} = 3.2213$$

The critical value from t-distribution with 12 df is now needs to be found, which is then multiplied by $\frac{\text{RSE}}{\sqrt{4.5}}$ to find the Margin of Error. The confidence intervals for each sample mean will be $\pm$ that.

See Appendix A – R Commands – Lines 85-120.

Summary of parameter estimates comparing to sample means:

	Genotype		
HeightGroup	*aa*	*Aa*	*AA*
high	29.66667	35	37
low	23.33333	28	32.33333

$$Margin\ of\ Error = CV \times \frac{\text{RSE}}{\sqrt{4.5}} = 2.1963$$

95% Confidence Intervals from combining the above findings:

	Genotype		
HeightGroup	*aa*	*Aa*	*AA*
high	27.47039 to 31.86295	32.80372 to 37.19628	34.80372 to 39.19628
low	21.13705 to 25.52961	25.80372 to 30.19628	30.13705 to 34.52961

Since the value 0 does not appear in all the above ranges we can reject the null hypothesis and the results are deemed significant.

We should now test the ANOVA assumptions.

Independence is the assumption that all experiments were carried out independent from each other. In this case, it is expected that each tree was recorded separately and correctly.

Normality can be tested informally by plotting standardised residuals against fitted values, or more formally with the use of a QQ plot. I will perform the latter to ensure we have a robust result and discuss the results below.

See Appendix A – R Commands – Lines 126-129.

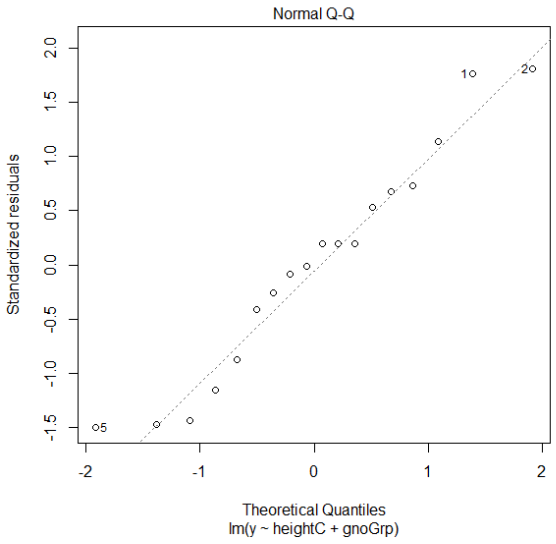

Figure 5 - QQ Plot to test for normality

Figure 5 shows the theoretical quartiles plotted against the standardized residuals. If both sets of quantiles came from the same (normal) set, we expect to see the points follow a straight line. R draws the line of best fit to show how close the quantiles fit our data, and as you can see, our data is quite a close fit. There are slight curves towards the ends of the line indicating that the distribution is likely tailed and has more extreme values than we would expect from a normal distribution.

Finally, **homogeneity of variances** can be tested for using the Lavene Test. Our null hypothesis here is that we assume all groups being tested have equal population variances. I will perform the test in R by running a one-way ANOVA on the absolute residuals (since we are using a two-way ANOVA) and discuss the results below.

See Appendix A – R Commands – Lines 131-152.

Summary of Lavene Test:

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	0.58302	0.06090	9.573	5.73e-07
Heightgroup	-0.16635	0.06090	-2.731	0.0182
Genotype	-0.16927	0.08613	-1.965	0.0729
Genotype 2	-0.27643	0.08613	-3.210	0.0075
Heightgroup: **Genotype**	-0.16417	0.08613	-1.906	0.0809
Heightgroup: **Genotype 2**	0.11847	0.08613	1.376	0.1941

RSE: 0.2584 on 12 df. **Multiple R^2**: 0.763. **Adjusted R^2**: 0.6642. **F-statistic**: 7.725 on 5 and 12 DF. **P-value**: 0.001851.

The results of our Lavene Test show a significant P-value of 0.001851. This means that there are statistically significant variances per group as it is below our significance level of 0.05. This means that we reject the null hypothesis that all groups have equal population variances and accept that the differences in sample variances are unlikely to have occurred based on random sampling from a population with equal variances. The assumptions can be considered reasonable for this data set.

Part C – ANCOVA Interpretation

Run an ANCOVA with Genotype factor and Height covariate and look for significant effects with interpretation on the results.

First, we need to ensure that the data is suitable for running a one-way ANCOVA. To do this, we plot the response, *yield*, against our covariate, *height*. This is shown in the results in Figure 6, below.

See Appendix A – R Commands – Lines 129-139.

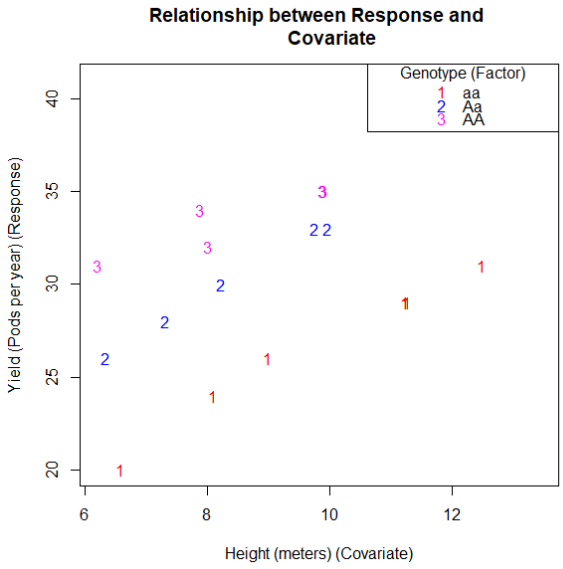

Figure 6 - Relationship of Response to Covariate

As we can see from Figure 6, there is a strong increasing relationship between the covariate and the response within each factor group, and the slopes are similar. This gives us an ideal situation for using a one-way ANCOVA.

Now we know we can run an ANCOVA, we must check the values of $\bar{x}_i$ and $\bar{y}_i$ to ensure they are in the correct range. We then centre the covariate on its mean and run a linear model on the response *yield* against the terms (*height* + *Genotype*) where height is centred, and genotype is grouped.

See Appendix A – R Commands – Lines 141-158.

Summary of ANCOVA Significant Effects:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	30.88889	0.19415	159.102	< 2e-16
Height (centered)	1.72409	0.09189	18.762	2.55e-11
Genotype	-5.14749	0.27752	-18.548	2.98e-11
Genotype 2	0.88984	0.27496	3.236	0.00597

RSE: 0.8237 on 14 df. **Multiple R^2**: 0.979. **Adjusted R^2**: 0.9745. **F-statistic**: 217.3 on 3 and 14 DF. **P-value**: 5.653e-12.

ANCOVA of Yield against Height Covariate and Genotype Factor:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Height (centered)	1	175.444	175.444	258.59	2.020e-10
Genotype	2	266.835	133.417	196.65	5.669e-11
Residuals	14	9.498	0.678		

This shows that both the Height and Genotype have high statistical significance and Genotype 2 has a lesser significance but still statistically significant, which allows us to reject the null hypothesis that height and genotype do not have a significant effect on yield. We can see the overall p-value is

5.653e-12, which is highly statistically significant and is below the significance level of 0.05 which means that our null hypothesis should be rejected, and we accept that significant effects are present. We can tell from these details that the ANCOVA results fit much better than those results found running a two-way ANOVA, meaning the ANCOVA is a better model for this data set.

Part D – Analysis of *Height* as a suitable ANCOVA Covariate

From the terms we were given in this data set, we had the choice of setting the covariate as either height or age. We know that the covariate must be a variable that is related to the dependant variable, in this case, yield, and not have a correlation with the factor, genotype. From what we have seen, there is a correlation of height to yield and no correlation of height to genotype, which makes height a good covariate candidate. Regarding age, although we haven't tested for it, we can expect that the age of a tree generally would not have a correlation with its yield beyond the first few years of growth, and we can't be sure that the age of a tree would have a correlation with its genotype (could certain genotypes live longer? This could be something we could test for separately). These two reasons would mean that age would not be as good for a covariate as height is.

Part E – Analysis of ANOVA vs ANCOVA

We can compare the results of an ANOVA and ANCOVA by finding the relative efficiency (e) between them. For this we divide the MS(E) of one method by the other. In our case, the MS(E) of our ANOVA was 6.833 and the ANCOVA had a MS(E) of 0.678.

$$e(\mathrm{T}_1, \mathrm{T}_2) = \frac{MS(E)_{\mathrm{ANOVA}}}{MS(E)_{\mathrm{ANCOVA}}} = \frac{6.833}{0.678} = 10.0782$$

We can say from this result that the ANCOVA is over ten times as efficient at predicting a relation between terms and response. This means we can get the same prediction with an ANCOVA on this data set by using use ten times less data compared to the prediction made from the ANOVA.

The suitability of a model can generally be summed up in the R^2 value given from summary. This is the statistical measure of the variance for a dependant variable explained by an independent variable, AKA the coefficient of determination:

$$\mathrm{R}^2 = 1 - \frac{\mathrm{SSR}}{\mathrm{SST}}$$

Therefore, the larger the R^2 value, the more of the variance is explained by the model. In our case, the ANOVA has an R^2 value of 81.85% versus the ANCOVA's 97.9%. This shows that the ANCOVA model can predict almost all the variance in our data set from the data given.

Question 2

We will use the hazard function of the lifetime of a lightbulb to find further details about it.

Part A – The Survival Function and Probability Density Function

The Hazard Function gives the probability that the event of interest (in this case, the failure of the bulb) will occur at time *t* if the measured variable survives to time *t*. The Hazard Function is denoted with the equation:

$$h(t) = \frac{f(t)}{S(t)}$$

Where f(t) is the Probability Density Function (pdf) and S(t) is the Survival Function.

The pdf is a function which gives the likelihood that the event of interest will occur between two sample points. It is denoted as the following.

$$f(t) = -S'(t)$$

The Survival Function gives us the probability that the event of interest has not occurred by the current time, given as duration *t*. The difference in time between the beginning point (here when the bulb is first turned on at t = 0) and the end point (failure of the bulb, denoted T) is known as Survival Time. Therefore, the Survival Function can be shown as

$$S(t) = P(T > t)$$

We can use the above to define the Hazard Function $h(t)$. Because S(0) = 1 (the probability that the event hasn't occurred by t = 0 is certain, we can define h(t) as the integrated hazard rate H(t).

$$H(t) = \int_0^t h(u)du$$

Which allows us to express the pdf as the following.

$$f(t) = h(t)e^{-H(t)}$$

If we simplify our given hazard function for the lifetime of the lightbulb to h(t) = *p*, we can define the Survival Function as the following.

$$S(t) = e^{-pt}$$

Therefore, the Survival Function for our lightbulb example can now be given as the following.

$$S(t) = \begin{cases} e^{-\alpha t^2}, & 0 \le t \le 1, \\ e^{-\alpha t}, & \text{otherwise} \end{cases}$$

This also allows us to determine the pdf in a similar way. Where h(t) = *p*, the pdf can be expressed as the following.

$$f(t) = pe^{-pt}$$

Therefore, the Probability Density Function for our lightbulb can be given as the following.

$$f(t) = \begin{cases} \alpha\, t\, e^{-\alpha t^2}, & 0 \le t \le 1, \\ \alpha\, e^{-\alpha t}, & \text{otherwise} \end{cases}$$

Part B – Determine the median lifetime of the bulb

We know that the median is the middle point in a set of values. In this case it would mean that 50% of the failures of the lightbulb lie each side of the median point. We know that for our example, the Cumulative Distribution Function (cdf) F(t) is given as the following.

$$F(t) = P(T \leq t) \; = \; 1 - P(T > t)$$

In a probability distribution with cdf of F(t), the median can be defined as *m* where

$$\int_{(-\infty,m]} \mathrm{dF(t)} \; \geq \frac{1}{2} \text{ and } \int_{[m,\infty)} \mathrm{dF(t)} \; \geq \frac{1}{2}$$

If a randomly chosen value (X) is distributed according to F within a given set, the following is true.

$$\mathrm{P(X} \; \leq \; m) \; \geq \frac{1}{2} \text{ and } \mathrm{P(X} \; \geq \; m) \; \geq \frac{1}{2}$$

We are looking for the largest value of t where $\mathrm{S}(m) \; \leq \; 0.5$, also denoted $t_{0.5}$, which can be computed as the solution to.

$$\mathrm{S(t)} \; = \; 1 \; - \; 0.5 \; = \; 0.5$$

We understand that the hazard function is a constant, α, therefore using the value we found for S(t) in part 1, we find the following.

$$\text{let } \mathrm{S}(\mathrm{t}_{0.5}) \; = \; e^{-\alpha t_{0.5}} \; = \; 0.5$$

Therefore, the median survival time is the following.

$$\mathrm{t}_{0.5} \; = \; \frac{\log 2}{\alpha}$$

Which can be rewritten to a function of α

$$\mathrm{f}(\alpha) \; = \; \frac{\log 2}{\mathrm{t}_{0.5}}$$

Part C – Determine the Probability of Survival > 2 months

To do this, we can recall our survival function again from part a.

$$S(t) = \begin{cases} e^{-\alpha t^2}, & 0 \; \leq \; t \; \leq \; 1, \\ e^{-\alpha t}, & \text{otherwise} \end{cases}$$

In this case, we are measuring the probability that the bulb will last longer than two months, therefore where t = 2. Here we can substitute t to get the following function.

$$\mathrm{S(2)} \; = \; e^{-\alpha 2}$$

Question 3

We will construct life tables from a given data set and use Kaplan Mier estimation to compare two clinical trials.

Part A – Construct Life Tables from Given Data

First, we will load the data into R and break into groups of six groups of four weeks each. I have created the six groups by registering a new column in the data frame called `Month`, which is then populated with the quotient of [the given week number minus one], plus one. This makes it easier to then calculate the values for tables A and B (corresponding to protocols A and B).

Regarding life table A, to generate the value for **ninit**, I count how many times protocol *A* is featured. For **nlost**, I first record in a vector called **lost** the month numbers where the status of each record was *0* and protocol *A* was used. I then take a factor of this vector, to ensure that any months that had zero records are included. I then create a table of the factored vector in order to count the number of entries in each month, then extract the frequencies of this table using `as.numeric`. This avoids us having to use a For Loop to do the count.

The process for finding **nevent** is the same as for nlost, except we are looking for status *1* rather than *0*.

Finally, I put these values into the function `lifetab()`in order to produce the life table. The process for getting the values for life table B is the same as the above, except we look for protocol value *B* instead. This gave us the following results:

See Appendix A – R Commands – Lines 212-274.

Life Table for treatments under Protocol A:

	nsubs	nlost	nrisk	nevent	surv	pdf	hazard	se.surv	se.pdf	se.hazard
0-1	21	0	21	1	1	0.047619	0.048780	0	0.046471	0.048766
1-2	20	1	19.5	3	0.952381	0.146520	0.166667	0.046471	0.078142	0.095890
2-3	16	1	15.5	3	0.805861	0.155973	0.214286	0.087186	0.082610	0.123006
3-4	12	1	11.5	0	0.649888	0	0	0.107160	NA	NA
4-5	11	2	10	1	0.649888	0.064989	0.105263	0.107160	0.062578	0.105117
5-6	8	4	6	4	0.584899	NA	NA	0.114467	NA	NA

Life Table for treatments under Protocol B:

	nsubs	nlost	nrisk	nevent	surv	pdf	hazard	se.surv	se.pdf	se.hazard
0-1	21	1	20.5	0	1	0	0	0	NaN	NaN
1-2	20	0	20.0	6	1	0.300000	0.352941	0	0.102469	0.141826
2-3	14	1	13.5	7	0.700000	0.362962	0.700000	0.102469	0.109016	0.247840
3-4	6	0	6.0	1	0.337037	0.056172	0.181818	0.107218	0.054303	0.181065
4-5	5	0	5.0	2	0.280864	0.112345	0.500000	0.103017	0.074057	0.342326
5-6	3	3	1.5	0	0.168518	NA	NA	0.087218	NA	NA

From this data we can plot the survival function S(t), probability density function f(t) and hazard function h(t) for each treatment, as seen in Figure 7.

See Appendix A – R Commands – Lines 277-304.

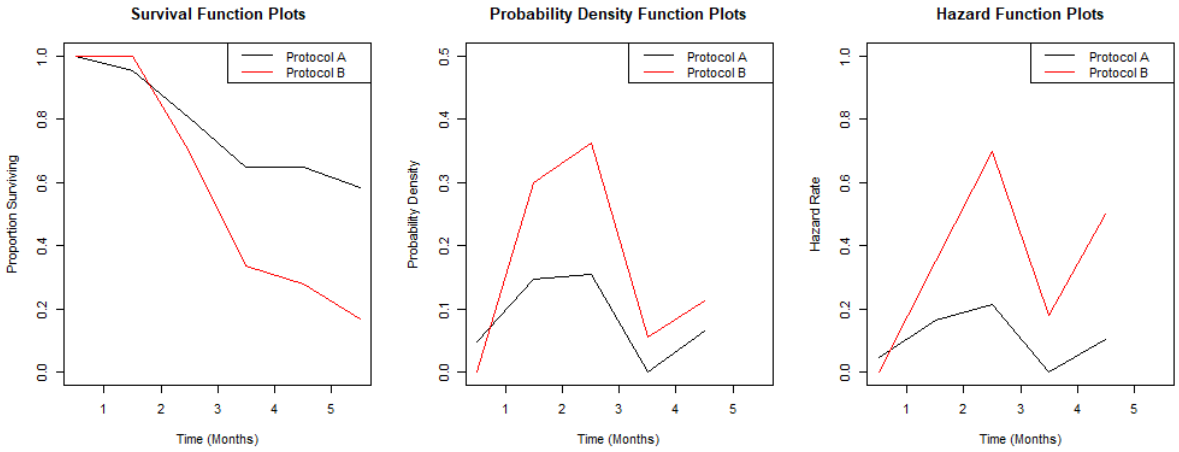

Figure 7 - Plots of Survival, Probability Density and Hazard Functions

We can also create survival plots for the above data. This is similar to the Survival Function plot in Figure 7 but with added confidence intervals. With survival plots, we have three choices in what happens to censored entries. We can either treat those entries as deaths, we can remove them, or we can take the proper approach by including censored events as tick marks on the plot where they occur in time. In Figure 8, I show a comparison between these three approaches with each plot showing a combination of both protocols.

See Appendix A – R Commands – Lines 305-320.

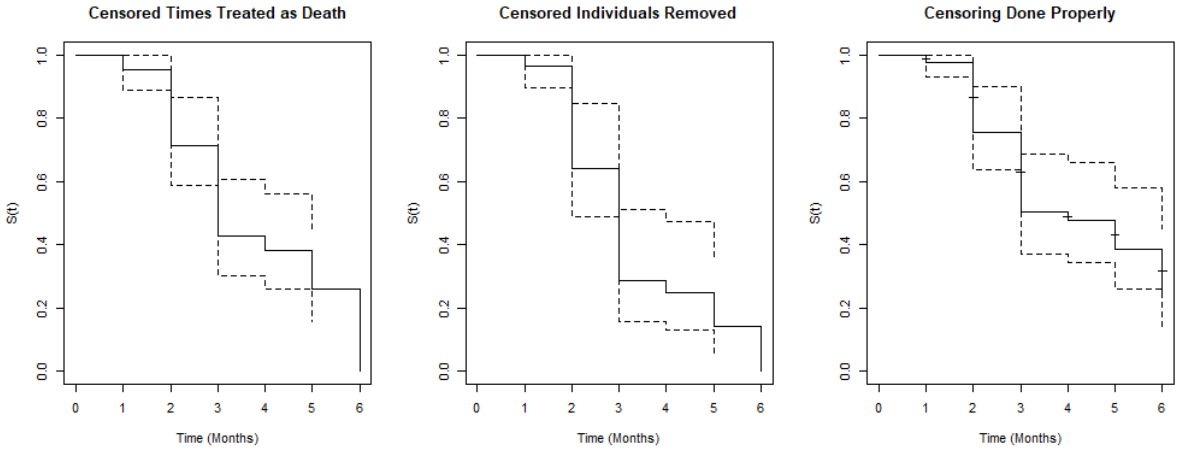

Figure 8 - Survival plots demonstrating three approaches to displaying censored events

Part B – Kaplan-Meier Estimation

Kaplan-Meier estimation is used to accurately account for censored data in survival plots. In Figure 8, the third plot we created used this method to display the survival rate where each tick on the plot represented a censored entry. The assumptions of Kaplan-Meier estimation are that censoring is not related to the death of a patient, patients that come into the study late have the same survival probability as those that came in early and that events happened at the times specified. In this part, we will use the original weekly data to produce survival plots rather than the monthly data as used in part A. These plots are displayed together in Figure 9.

See Appendix A – R Commands – Lines 322-338.

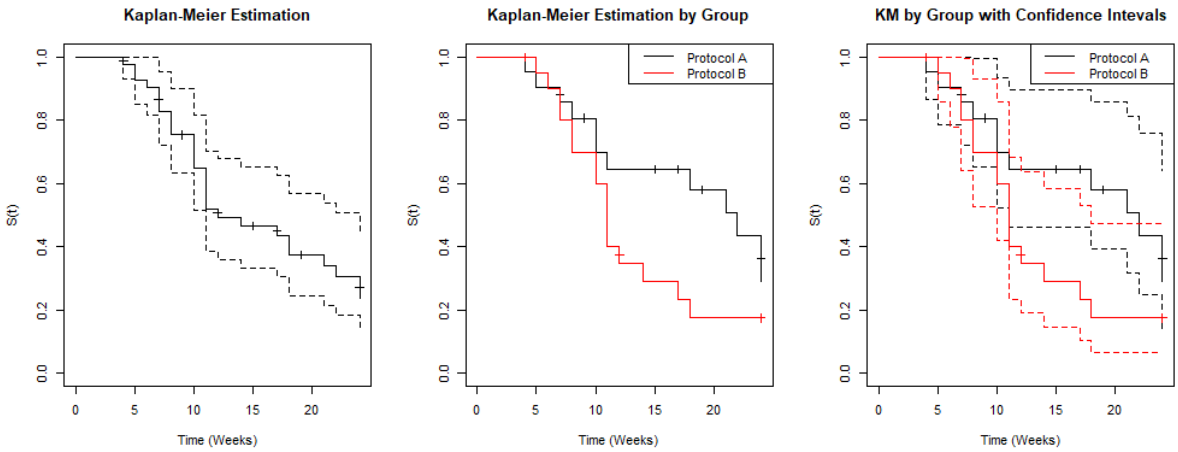

Figure 9 - Kaplan-Meier Estimation plots demonstrating three approaches to displaying groups and CI

Part C – Analysis of Data and Approaches

Although there are not enough samples in our data set when grouped by month to give an accurate reading, we can still infer a reasonable amount of information from the life tables shown in Part A and the plots in Figure 7.

The **surv** column and Survival Function Plots tell us that patients that were under treatment protocol A had a higher chance of surviving after month two, but those on protocol B had a slightly higher chance of survival before that point. By the conclusion of the study, over three times as many patients on protocol A had survived, compared to protocol B.

The **pdf** column and Probability Density Function Plots tell us that the greatest number of patients died during month three on both protocols, but that there were twice as many deaths from patients on protocol B during that period compared to protocol A. Protocol A had the fewest deaths between months three and four, which also saw a decline in deaths for patients on protocol A however the fewest deaths on protocol A occurred within the first month. Both groups saw an uptick of deaths during month five.

The **hazard** column and Hazard Function Plots give us similar information to the previous two columns and plots. After the initial month, deaths climbed for patients on both protocols before hitting a peak during month three, then dropping for month four, before climbing again towards the end of the survey period. Once again, we can see that the rate of death under protocol B is significantly higher than on protocol A.

The survival plots shown in Figure 8 explore how censored entries can be difficult to deal with correctly without inducing bias into the results. Both treating censored entries as deaths and removing censored entries produce plots that underestimate the true survival time figures. Only the final plot, the Kaplan-Meier approach, can reasonably deal with censored entries.

Part B used the same data but grouped weekly rather than monthly, which are detailed in Figure 9. This grouping produced much more granulated survival plots compared to those featured in Figure 8, producing smother lines.

Separating the plot into two groups, one for each protocol, gives a much more detailed look at survival rates over time compared to any of the previously produced data. We can see that survival

times for both protocols are very similar up to around week ten, when they diverge by a large degree so that by around week seventeen there is well over twice the likelihood of a patient on protocol A to survive compared to a patient on protocol B. We can also see that there are far more cases of censored data of patients on protocol A compared to those on protocol B, which could be the reason the survival rate for protocol A is much higher than for B. Protocol B remarkably reports zero deaths or censored entries in most of the final quarter of measurements, which tells us that even though more patients on protocol B died early on, those that were left after around week 17 were likely to continue living past the end of the study. At the same time we can also see that there is a large drop in survival rates in the final week for protocol A, looking at the raw data we can see that this is due to patients on protocol A both dying and being censored on this week, censored likely due to them still being alive but no longer being recorded due to the study period ending.

The third plot, which includes confidence intervals for both protocols, gives a lot of information but in my opinion is a little too messy to easily understand. It would probably be better to also have separate plots for each protocol when dealing with confidence intervals and use the two grouped plots in Figure 9 as reference for how effective each treatment is in comparison to the other. The plot we produced, however, is useful to see just how much overlap there is between the lower bounds of Protocol A and the upper bounds of Protocol B, which shows that it's possible one approach is as effective as the other, if the survival rates of Protocol A have been enhanced and the rates of protocol B have been suppressed in this study, which could be true based on the number of censored cases for protocol A. We may require more sampling and sampling for a longer period to see if this is the case, however.

It is clear from the plots that plotting by week rather than month provides us with a more useful interpretation, that Kaplan-Meier estimation is far more accurate than treating censored data than any other way and that splitting the Kaplan-Meier plot into groups gives us a much better look at how survival rates differ based on protocol, and thereby which protocol is more effective at keeping patients alive for longer. Therefore, I would say approach B is better in general for assessing survival rates.

Part D – Log-Rank Test of Treatments

A log-rank test can be used to investigate the significance in the differences of treatments to see if one approach does increase survival rates compared to another. It copes well with censored data. Our null hypothesis here is H_0 = no difference in hazard rates between groups. The assumptions of the Log-Rank test are the same assumptions that we use for the Kaplan-Meier estimation. In R, the log-rank test can be performed with the `survdiff( )` function with 0 as the *rho* parameter.

See Appendix A – R Commands – Line 341.

Log-Rank Test Results on Weekly Data:

	n	Observed	Expected	$\frac{(O-E)^2}{E}$	$\frac{(O-E)^2}{V}$
Protocol A	21	12	15.9	0.956	2.46
Protocol B	21	16	12.1	1.256	2.46

Chisq = 2.5 on 1 degrees of freedom, **p** = 0.1.

We are given a P value of 0.1, which is below our α of 0.05, indicating that there likely is not a significant difference between the two groups. This was hinted at with our Kaplan-Meier estimation in Figure 9 where the both protocols cross twice and remain close to each other until week ten, which can often lead to a weak test result. There were also far more censored cases for protocol A compared to protocol B, meaning there was a deviation from the Log-Rank (and Kaplan-Meier) assumptions. There were also deaths at the end of the measurement period for protocol A and many survivals for protocol B that brought the survival rates of the two protocols back to almost the same value at the end of the study.
The difference between the Observed and Expected variables tells us that for there to be a significant difference between the two groups, that there needed to be approximately four more events observed in protocol A group, four fewer observations in protocol B group or a mixture of both over the measurement period.
The fourth column gives us the difference between the observed and expected values, the smaller this value, the closer our data fits what is expected from treatments that are significantly different. If we add these two figures up, we get the X^2 value of 2.212. By using a Chi Square table, we find that the value we need to reject the null hypothesis with a p-value of 0.05 on 1 degrees of freedom must be greater than $\chi^2 = 3.84$. An X^2 value below this shows that we do not have enough difference between the two values to be statistically relevant. Indeed, the **Chisq** value we are given is 2.5 which, again, is too low to reject the null hypothesis.

With this explored, we must therefore accept the null hypothesis that there is no difference between the groups, with the caveat that we may need more data to be certain.

Appendix A – R Commands

```
## Question 1 ######################################################################
# We will use ANCOVAs to analyse the yield of cocoa plants grouped by height
# and also grouped by genotype to discover relationships in the data.
#
# First we load the data:
library(readr)
yields <- read_csv("CocoaYield.csv", col_names = TRUE)
#
# Part a - We will visualize the data and interpret what is seen.
# For this we will use boxplots. First on Yield grouped by HeightGroup:
boxplot(Yield ~ HeightGroup, data = yields, main = "Yields of Cocoa pods by
        HeightGroup", xlab = "Heightgroup Name",
        ylab = "Yield (Pods per year)", boxwex = 0.40)
stripchart(Yield ~ HeightGroup, data = yields, add = TRUE, vertical = TRUE,
           method = "jitter", pch = 1, col = "red")
# Saved as Fig1.png
# Then Yield grouped by Genotype:
boxplot(Yield ~ Genotype, data = yields, main = "Yields of Cocoa pods by
        Genotype", xlab = "Genotype Name", ylab = "Yield (Pods per year)",
        boxwex = 0.40)
stripchart(Yield ~ Genotype, data = yields, add = TRUE, vertical = TRUE,
           method = "jitter", pch = 1, col = "red")
# Saved as Fig2.png
# We'll do two interaction plots too to look for corolation:
interaction.plot(yields$Genotype, yields$HeightGroup, yields$Yield,
                 trace.label = "HeightGroup",
                 xlab = "Genotype Name", ylab = "Mean Yield (Pods per year)")
# Saved as Fig3.png
interaction.plot(yields$HeightGroup, yields$Genotype, yields$Yield,
                 trace.label = "Genotype",
                 xlab = "Heightgroup Name",
                 ylab = "Mean Yield (Pods per year)")
# Saved as Fig4.png
# See report for analysis.
#
# Part B - We will run a two way ANOVA on the above data.
#
# First we require factors:
y        <- yields$Yield
height   <- yields$Height
genotype <- yields$Genotype

hgtGrp   <- factor(yields$HeightGroup)
gnoGrp   <- factor(yields$Genotype)
# Now we check for data balance:
table(hgtGrp, gnoGrp)
# RESULTS
#        gnoGrp
# hgtGrp aa Aa AA
# high    3  3  3
# low     3  3  3
# Data are balanced with 3 observations per group.
# We now create contrasts
contrasts(hgtGrp) <- contr.sum
contrasts(gnoGrp) <- contr.sum
# And run the two-way ANOVA:
results <- lm(y ~ hgtGrp*gnoGrp, data = yields)
anova(results)
```

```
# RESULTS
#               Df  Sum Sq Mean Sq F value    Pr(>F)
# hgtGrp         1 162.000 162.000 23.7073 0.0003856 ***
# gnoGrp         2 203.444 101.722 14.8862 0.0005620 ***
# hgtGrp:gnoGrp  2   4.333   2.167  0.3171 0.7341961
# Residuals     12  82.000   6.833
# See report for analysis.
# We must also run a summary to check for significant effects:
sm <- summary(results)
sm
# RESULTS
# Residuals:
#    Min     1Q Median     3Q    Max
# -3.333 -2.000 -0.500  1.583  4.000
#
# Coefficients:
#                  Estimate Std. Error t value Pr(>|t|)
# (Intercept)      30.8889     0.6161  50.133  2.6e-15 ***
# hgtGrp1           3.0000     0.6161   4.869 0.000386 ***
# gnoGrp1          -4.3889     0.8714  -5.037 0.000291 ***
# gnoGrp2           0.6111     0.8714   0.701 0.496471
# hgtGrp1:gnoGrp1   0.1667     0.8714   0.191 0.851510
# hgtGrp1:gnoGrp2   0.5000     0.8714   0.574 0.576687
#
# Residual standard error: 2.614 on 12 degrees of freedom
# Multiple R-squared:  0.8185, Adjusted R-squared:  0.7429
# F-statistic: 10.82 on 5 and 12 DF,  p-value: 0.0004069
# See Report for Analysis.
#
# Next, we need to find the sample means:
Xbarij <- tapply(y, yields[,3:4],mean)
Xbarij
# RESULTS:
#                     Genotype
# HeightGroup       aa Aa       AA
# high        29.66667 35 37.00000
# low         23.33333 28 32.33333
#
# Store the RSE, MS(E), SE and SD:
RSE <- sm$sigma
MSE <- anova(results)['Residuals', 'Mean Sq']
SE  <- MSE / sqrt(9)
SD  <- MSE / sqrt(4.5)
# Use these to find the critical value from t-distribution with 12 df then
# find the Margin of Error from that:
CV <- qt(.95,12)
ME <- CV * (RSE / sqrt(4.5))
ME
# RESULTS:
# [1] 2.196281
#
# Finally, add and subtract the ME to each sample mean to find
# the bounds for 95% CI:
Xbarij - ME
Xbarij + ME
# RESULTS:
# Min                     Genotype
# HeightGroup       aa       Aa       AA
# high        27.47039 32.80372 34.80372
# low         21.13705 25.80372 30.13705
#
```

```
# Max                     Genotype
# HeightGroup       aa       Aa       AA
# high        31.86295 37.19628 39.19628
# low         25.52961 30.19628 34.52961
# See Report for Analysis
#
# Test for normality with QQ plot:
plot(results, which = 2)
# Saved as fig5.png
# See report for analysis.
#
# Test for Homogeneity using the Lavene Test:
yields$absres <- abs(results$residuals)
tmp <- lm(absres ~ hgtGrp*gnoGrp, data = yields)
summary(tmp)

# RESULTS:
# Residuals:
#    Min       1Q   Median       3Q      Max
# -0.40790 -0.12780  0.03445  0.19538  0.26062
#
# Coefficients:
#                   Estimate Std. Error t value Pr(>|t|)
#  (Intercept)       0.58302    0.06090   9.573 5.73e-07 ***
#  hgtGrp1          -0.16635    0.06090  -2.731   0.0182 *
#  gnoGrp1          -0.16927    0.08613  -1.965   0.0729 .
#  gnoGrp2          -0.27643    0.08613  -3.210   0.0075 **
#  hgtGrp1:gnoGrp1 -0.16417    0.08613  -1.906   0.0809 .
#  hgtGrp1:gnoGrp2  0.11847    0.08613   1.376   0.1941
#
# Residual standard error: 0.2584 on 12 degrees of freedom
# Multiple R-squared:  0.763,  Adjusted R-squared:  0.6642
# F-statistic: 7.725 on 5 and 12 DF,  p-value: 0.001851
# See report for analysis.
#
# Part C - Run an ANCOVA with Genotype factor and Height covariate, report
# on significant effects with interpretration.
#
# First, check the suitability of ANCOVA by plotting yield against height:
plot(y ~ height, type = "n", main = "Relationship between Response and
     Covariate", xlab = "Height (meters) (Covariate)",
     ylab = "Yield (Pods per year) (Response)")
points(height[gnoGrp=="aa"], y[gnoGrp=="aa"], pch="1", col=2)
points(height[gnoGrp=="Aa"], y[gnoGrp=="Aa"], pch="2", col=4)
points(height[gnoGrp=="AA"], y[gnoGrp=="AA"], pch="3", col=6)
legend( x = "topright", title = "Genotype (Factor)",
        legend = c("aa", "Aa", "AA"), pch = c("1","2","3"), col = c(2,4,6))
# Saved as Fig5.png
# See report for analysis.
#
# Factors and contrasts have previously been set, we will just check xbari and
# ybari:
Xbari <- tapply(height, gnoGrp, mean)
ybari <- tapply(y, gnoGrp, mean)
Xbari
ybari
# RESULTS:
# > Xbari
# aa       Aa       AA
# 9.775000 9.173333 9.056667
# > ybari
```

```
# aa       Aa       AA
# 26.50000 31.50000 34.66667
#
# Now center the covariate and run the ANCOVA:
heightC <- height - mean(height)
results <- lm(y ~ heightC + gnoGrp) # covariate first.
summary(results)
# RESULTS:
# Residuals:
# Min       1Q   Median       3Q      Max
# -1.12065 -0.53749  0.05905  0.47209  1.34477
#
# Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
#  (Intercept) 30.88889    0.19415 159.102  < 2e-16 ***
#  heightC      1.72409    0.09189  18.762 2.55e-11 ***
#  gnoGrp1     -5.14749    0.27752 -18.548 2.98e-11 ***
#  gnoGrp2      0.88984    0.27496   3.236  0.00597 **
#
# Residual standard error: 0.8237 on 14 degrees of freedom
# Multiple R-squared:  0.979,  Adjusted R-squared:  0.9745
# F-statistic: 217.3 on 3 and 14 DF,  p-value: 5.653e-12
anova(results)
# RESULTS:
#           Df  Sum Sq Mean Sq F value    Pr(>F)
# heightC    1 175.444 175.444  258.59 2.020e-10 ***
# gnoGrp     2 266.835 133.417  196.65 5.669e-11 ***
# Residuals 14   9.498   0.678
# See report for analysis.
################################################################################

# Question 3 ###################################################################
# We will construct life tables from a given data set and use Kaplan Mier
# estimation to compare two clinical trials.
#
# Part A - Prepare data and produce life tables and plots.
# First, load the package and data:
library(KMsurv)
cancer_data <- read_csv("CancSurv.csv", col_names = TRUE)
# Use the Time variable to assign each entry into one of 6 'Month' groups:
cancer_data$Month <- c((cancer_data$Time - 1) %/% 4 + 1)
# Assign variables for Life Table A and create it:
tis   <- c(0:6) # 6 groups plus 1
ninit <- sum(cancer_data$Protocol == "A")
lost  <- cancer_data$Month[cancer_data$Status == 0 &
                           cancer_data$Protocol == "A"]
nlost <- as.numeric(table(factor(lost, levels = 1:6)))
event <- cancer_data$Month[cancer_data$Status == 1 &
                           cancer_data$Protocol == "A"]
nevent <- as.numeric(table(factor(event, levels = 1:6)))
lifetableA <- lifetab(tis, ninit, nlost, nevent)
lifetableA
# RESULTS:
#     nsubs nlost nrisk nevent      surv        pdf     hazard    se.surv
# 0-1    21     0  21.0      1 1.0000000 0.04761905 0.04878049 0.00000000
# 1-2    20     1  19.5      3 0.9523810 0.14652015 0.16666667 0.04647143
# 2-3    16     1  15.5      3 0.8058608 0.15597306 0.21428571 0.08718565
# 3-4    12     1  11.5      0 0.6498877 0.00000000 0.00000000 0.10716023
# 4-5    11     2  10.0      1 0.6498877 0.06498877 0.10526316 0.10716023
# 5-6     8     4   6.0      4 0.5848990         NA         NA 0.11446690
#
```

```
#     se.pdf  se.hazard
# 0.04647143 0.04876598
# 0.07814240 0.09589035
# 0.08261011 0.12300575
# NaN        NaN
# 0.06257811 0.10511726
# NA         NA
#
# Same for Life Table B:
ninit <- sum(cancer_data$Protocol == "B")
lost  <- cancer_data$Month[cancer_data$Status == 0 &
                           cancer_data$Protocol == "B"]
nlost <- as.numeric(table(factor(lost, levels = 1:6)))
event <- cancer_data$Month[cancer_data$Status == 1 &
                           cancer_data$Protocol == "B"]
nevent <- as.numeric(table(factor(event, levels = 1:6)))
lifetableB <- lifetab(tis, ninit, nlost, nevent)
lifetableB
# RESULTS:
#     nsubs nlost nrisk nevent      surv        pdf    hazard    se.surv
# 0-1    21     1  20.5      0 1.0000000 0.00000000 0.0000000 0.00000000
# 1-2    20     0  20.0      6 1.0000000 0.30000000 0.3529412 0.00000000
# 2-3    14     1  13.5      7 0.7000000 0.36296296 0.7000000 0.10246951
# 3-4     6     0   6.0      1 0.3370370 0.05617284 0.1818182 0.10721839
# 4-5     5     0   5.0      2 0.2808642 0.11234568 0.5000000 0.10301783
# 5-6     3     3   1.5      0 0.1685185         NA        NA 0.08721828
#
#     se.pdf se.hazard
# NaN        NaN
# 0.10246951 0.1418263
# 0.10901684 0.2478407
# 0.05430301 0.1810653
# 0.07405736 0.3423266
# NA         NA
# See report for analysis.
# Create plots of functions from the above data:
x  <- 0.5+c(0:5)
y  <- seq(0, 1, by = 0.2)
y2 <- seq(0, 0.5, by = 0.1)
par(mfrow=c(1,3))

plot(x, y, type = "n", xlab = "Time (Months)", ylab = "Proportion Surviving",
     main = "Survival Function Plots")
lines(x, lifetableA[,5], type = "l", col = 1)
lines(x, lifetableB[,5], type = "l", col = 2)
legend( x = "topright",lty = 1, col = c(1,2), y.intersp = 2.5,
        legend = c("Protocol A", "Protocol B"))

plot(x, y2, type = "n", xlab = "Time (Months)", ylab = "Probability Density",
     main = "Probability Density Function Plots")
lines(x, lifetableA[,6], type = "l", col = 1)
lines(x, lifetableB[,6], type = "l", col = 2)
legend( x = "topright",lty = 1, col = c(1,2), y.intersp = 2.5,
        legend = c("Protocol A", "Protocol B"))

plot(x, y, type = "n", xlab = "Time (Months)", ylab = "Hazard Rate",
     main = "Hazard Function Plots")
lines(x, lifetableA[,7], type = "l", col = 1)
lines(x, lifetableB[,7], type = "l", col = 2)
legend( x = "topright",lty = 1, col = c(1,2), y.intersp = 2.5,
        legend = c("Protocol A", "Protocol B"))
```

```
# Saved as Fig7.png
#
# Create Survival plots from the original data:
library(survival)

fit <- survfit(Surv(Month, Status*0+1) ~ 1, data = cancer_data)
plot(fit, main = "Censored Times Treated as Death",
     xlab = "Time (Months)", ylab = "S(t)")

cancer_data2 <- subset(cancer_data, Status == 1)
fit <- survfit(Surv(Month, Status) ~ 1, data = cancer_data2)
plot(fit, main = "Censored Individuals Removed",
     xlab = "Time (Months)", ylab = "S(t)")

fit <- survfit(Surv(Month, Status) ~ 1, data = cancer_data)
plot(fit, main = "Censoring Done Properly",
     xlab = "Time (Months)", ylab = "S(t)", mark.time = TRUE)
# Saved as Fig8.png
#
# Part B - Produce Kaplan Meier plots from weekly data:
fit <- survfit(Surv(Time, Status) ~ 1, data = cancer_data)
plot(fit, main = "Kaplan-Meier Estimation",
     xlab = "Time (Weeks)", ylab = "S(t)", mark.time = TRUE)

fit <- survfit(Surv(Time, Status) ~ Protocol, data = cancer_data)
plot(fit, main = "Kaplan-Meier Estimation by Group", conf.int = FALSE,
     xlab = "Time (Weeks)", ylab = "S(t)", mark.time = TRUE, col = c(1,2))
legend( x = "topright",lty = 1, col = c(1,2), y.intersp = 2.5,
        legend = c("Protocol A", "Protocol B"))

fit <- survfit(Surv(Time, Status) ~ Protocol, data = cancer_data)
plot(fit, main = "KM by Group with Confidence Intevals", conf.int = TRUE,
     xlab = "Time (Weeks)", ylab = "S(t)", mark.time = TRUE, col = c(1,2))
legend( x = "topright",lty = 1, col = c(1,2), y.intersp = 2.5,
        legend = c("Protocol A", "Protocol B"))
# Saved as Fig9.png
#
# Part D - Do a log-rank test to investigate the significance of differences:
survdiff(Surv(Time, Status) ~ Protocol, data = cancer_data, rho = 0)
# RESULTS:
#             N Observed Expected (O-E)^2/E (O-E)^2/V
# Protocol=A 21       12     15.9     0.956      2.46
# Protocol=B 21       16     12.1     1.256      2.46
#
# Chisq= 2.5  on 1 degrees of freedom, p= 0.1
# See report for analysis.
################################################################################
```