COLUMBIA UNIVERSITY — Prof. Marco Avella
Statistical Inference and Modeling — TA: Casey Bradshaw, Leon Fernandes
Graded Homework 2 — Spring 2020

You are free to present your solutions to the exercises below in *groups of at most three*. This homework is due on March 7. Buena suerte!

Exercise 1

The data set **transplant.txt** consists of data regarding allotransplant and autotransplant. The first column is the time to death or relapse (in months), the second one shows the types of transplant (1=allogeneic, 2=autologous), and the last column is the survival indicator (0=alive without relapse, 1=dead or relapse). Provide a minimum amount of `R` output to justify your answers.

1. Do you think that it is reasonable to assume that the right censoring in this data set is random? Why?

2. Plot the Kaplan-Meier estimators of the survival curves of the two transplant groups. Does the plot seem to indicate a difference between these groups? Which type of transplant seems to be more efficient?

3. Fit an exponential model by using the function `survreg`. Do the signs of the fitted parameters agree with your intuition from the last point?

4. Using point 3, do we observe a significant difference between the two groups based on the likelihood ratio statistic? Does this conclusion depend on parametric model assumptions?

5. Do a visual model check using the fitted exponential models and the Kaplan-Meier fits. Comment briefly what you observe.

6. Try fitting a Weibull model now. Does this model fit the data better? Does this new model provide evidence against the relevance of an exponential model?

Solution:

1. Given the little information available, assuming random censoring seems like a natural and convenient assumption. We do not see an obvious pattern in the censoring. Perhaps some auxiliary information such as weight, age, sex, etc. could explain part of the censoring...

2. The Kaplan-Meier estimates seem to suggest that treatment 2 is more efficient early on but treatment 1 is more efficient later.

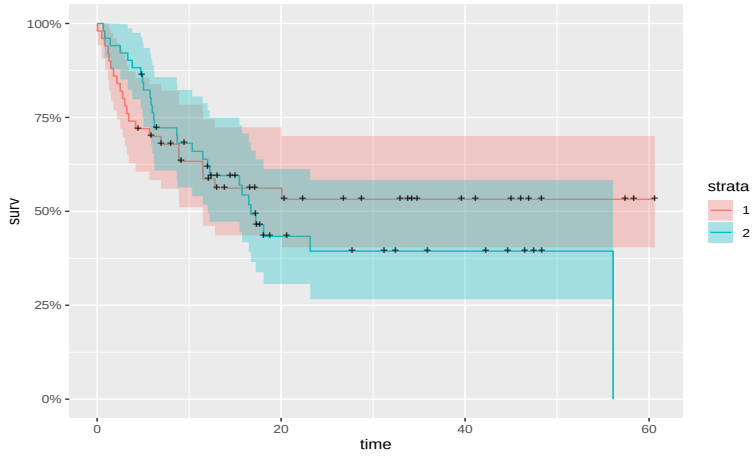

Figure 1: Kaplan-Meier estimates of the two treatments

3. The estimated coefficients do show a significant difference between the two groups. The sign of the coefficient also matches what we see in the figure above. Indeed, writing the survival function of an exponential random variable as $S(x) = e^{-\lambda x}$, the estimated coefficients for the two arms are $\hat{\lambda}_1 = e^{-3.742}$ and $\hat{\lambda}_2 = e^{-(3.742-0.325)}$.

```
Call:
survreg(formula = Surv(t, d) ~ type, data = ncog, dist = "exp")
             Value Std. Error     z      p
(Intercept)  3.742      0.213 17.55 <2e-16
type2       -0.325      0.285 -1.14   0.25

Scale fixed at 1

Exponential distribution
Loglik(model)= -228   Loglik(intercept only)= -228.6
Chisq= 1.31 on 1 degrees of freedom, p= 0.25
Number of Newton-Raphson Iterations: 5
n= 101
```

4. The output given above also gives a p-value of 0.25 using the likelihood ratio statistic which is approximately χ^2_1. This provides evidence to not reject H_0 where H_0: "there is no difference between the treatments". This also agrees with the insignificant coefficient for type 2 treatment in the output.
   These conclusions depend on model assumptions (exponential parametric model) and asymptotic approximation to the distribution of the likelihood ratio statistic.

5. The visual check is not very satisfactory for the fitted survival function for both treatments. After 40 months, both fitted curves depart from the point-wise confidence intervals of the Kaplan-Meier counterpart.

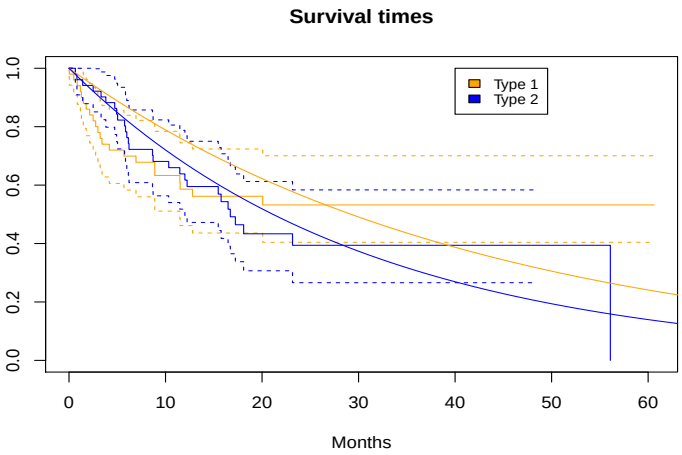

Figure 2: Checking the exponential fits

6.
```
Call:
survreg(formula = Surv(t, d) ~ type, data = ncog)
              Value Std. Error     z      p
(Intercept)  3.968      0.330 12.02 <2e-16
type2       -0.374      0.420 -0.89 0.3742
Log(scale)   0.388      0.123  3.15 0.0016

Scale= 1.47

Weibull distribution
Loglik(model)= -222   Loglik(intercept only)= -222.4
Chisq= 0.8 on 1 degrees of freedom, p= 0.37
Number of Newton-Raphson Iterations: 5
n= 101
```

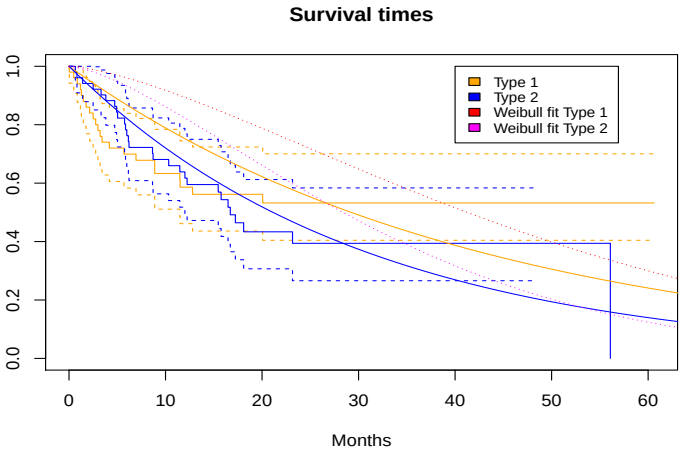

Figure 3: Checking the Weibull fits

We see that the Weibull model does not seem to improve the fit of the exponential model. The coefficient for type 2 treatment is again not significant and visually there is also no improvement in the estimated Weibull survival functions.

Exercise 2

The purpose of this exercise is to explore the numerical performance of the following approaches to missing data:

(a) Complete case analysis.

(b) Available case analysis.

(c) Mean imputation

(d) Mean imputation with the bootstrap.

(e) The EM-algorithm

We will start by looking at the Student Score Data[1] which has both a small sample size and missing observations. Let λ_1 denote the largest eigenvalue of the population covariance matrix Σ and $\hat{\lambda}_1$ its estimated value using the sample covariance[2]. In the absence of missing data, it can be shown that

$$\sqrt{n}(\log \hat{\lambda}_1 - \log \lambda_1) \xrightarrow[n\to\infty]{\mathcal{D}} N(0, 2).$$

Solutions to 1–3 below should be presented in *no more than two pages*. Provide a minimum amount of `R` output to justify your answers.

1. Estimate the covariance matrix of the 5 variables in the Student Score Data using methods (a)–(e) and comment your results.

2. How would you construct an asymptotic confidence interval leveraging the asymptotic normality of $\hat{\lambda}_1$? Use this result to construct confidence intervals for λ_1 using the estimated covariances from point 1. Comment your findings.

3. Note that the student score data is just a subset of the mathmarks data[3] with artificially generated missing entries. Use the full data to compute the sample covariance and give a confidence interval for λ_1. Compare this with your findings in the previous two questions.

[1] Table 1 in Efron (1994), *Journal of the American Statistical Association.*

[2] These quantities are interesting for multivariate analysis and unsupervised learning. For example, in a principal component analysis, the largest empirical eigenvalue is the variance explained by the first principal component

[3] Available in the `R` package "SMPracticals".

4. Can you derive the form of the EM-algorithm for an i.i.d. normal sample with missing data?

   Remember that with partially observed vectors $X_i = (X_{io}^T, X_{im}^T)^T$ the EM-algorithm simplifies to the iterations:

$$\mu^{(k+1)} : \sum_{i=1}^{n} (\hat{X}_i - \mu) = 0$$

$$\Sigma^{(k+1)} : \sum_{i=1}^{n} \left(\Sigma - (\hat{X}_i - \mu)(\hat{X}_i - \mu)^T - \mathbf{C}_i^{(k)} \right) = 0$$

   with $\hat{X}_{io} = X_{io}$ and

$$\hat{X}_{im} = \mu_{im}^{(k)} + \Sigma_{imo}^{(k)} (\Sigma_{ioo}^{(k)})^{-1} (X_{io} - \mu_{io}^{(k)})$$

   and $C_{ijk}^{(k)} = 0$ if X_{ij} or X_{ik} are observed and

$$C_{imm}^{(k)} = \Sigma_{imm}^{(k)} - \Sigma_{imo}^{(k)} (\Sigma_{ioo}^{(k)})^{-1} \Sigma_{iom}^{(k)}$$

Solution:

1. We see large differences in the estimated covariance matrices due to the small sample size and the large number of missing entries for variables 1 and 5.

   (a) Complete case analysis:

```
> cov(M,use="complete")
  x1       x2    x3       x4       x5
x1 458.7 207.9000 135.4 180.8000 330.9000
x2 207.9 337.3667 168.4 150.4667 372.5667
x3 135.4 168.4000 160.4 110.0000 282.4000
x4 180.8 150.4667 110.0 109.4667 211.8667
x5 330.9 372.5667 282.4 211.8667 565.7667
```

   (b) Available case analysis:

```
> cov(M,use="pairwise")
      x1        x2        x3        x4       x5
x1 219.42424  95.15152  73.37879  93.66667 330.9000
x2  95.15152 210.35968 127.02767 132.52569 236.5758
x3  73.37879 127.02767 154.62451 131.91107 204.6364
x4  93.66667 132.52569 131.91107 160.17391 173.1212
x5 330.90000 236.57576 204.63636 173.12121 365.1515
```

   (c) Mean imputation:

```
> cov(mean.impute(M))
      x1        x2        x3        x4        x5
x1 109.71212  47.57576  36.68939  46.83333  75.02273
x2  47.57576 210.35968 127.02767 132.52569 118.28788
x3  36.68939 127.02767 154.62451 131.91107 102.31818
x4  46.83333 132.52569 131.91107 160.17391  86.56061
x5  75.02273 118.28788 102.31818  86.56061 182.57576
```

   (d) Mean imputation with the bootstrap:

```
> bootimpS
          [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 105.69342  47.92343  35.15371  45.13967  70.80995
[2,]  47.92343 203.81663 121.87264 127.80139 112.52102
[3,]  35.15371 121.87264 146.73067 125.48260  95.27575
[4,]  45.13967 127.80139 125.48260 153.20956  80.89462
[5,]  70.80995 112.52102  95.27575  80.89462 172.42492
```

(e) The EM-algorithm:

```
> Mls(M)$sig
        x1        x2        x3       x4       x5
[1,] 322.13065  46.29316  94.09942 115.2921 196.2932
[2,]  46.29316 201.21361 121.50473 126.7637 167.3158
[3,]  94.09942 121.50473 147.90170 126.1758 194.0407
[4,] 115.29212 126.76371 126.17580 153.2098 136.8706
[5,] 196.29320 167.31583 194.04074 136.8706 336.3849
```

2. Taking the largest eigenvalue of the different estimated matrices gives different estimated eigenvalues. One can also construct a confidence interval for λ_1 using these different estimators of λ_1 to be

$$\left(\hat{\lambda} \pm 1.96\hat{\lambda}\sqrt{2/n}\right)$$

Applying naively this principle leads to the confidence intervals

(a) Complete case analysis: $(530.3, 2062.5)$

(b) Available case analysis: $(378.0, 1470.1)$

(c) Mean imputation: $(226.1, 879.4)$

(d) Mean imputation with the bootstrap: $(265.5, 885.8)$

(e) The EM-algorithm: $(323.8, 1259.3)$

Complete case analysis is useless for this data set because it only has 1 or 2 complete cases. Complete case analysis is somehow more satisfactory but the interval is very different from the one obtained with the EM algorithm. Mean imputation with bootstrap gives the narrowest interval, followed by mean imputation - but is to be taken with caution since it underestimates the variance. With such small samples the validity of a normal approximation is always questionable.

3. The confidence interval obtained using the full data set and the normal approximation is $(484.0, 890.0)$. This interval should be in principle more accurate as $\hat{\lambda}_1$ should have smaller bias and smaller variance. These two features will translate to an interval that is both centered closer to the true population λ_1 and narrower.

The bootstrap method and CCA provided confidence intervals that don't cover the interval $(484.0, 890.0)$ obtained with all the data. Furthermore, if we use the same data but with all the complete observations corresponding to the ones analyzed in the previous points, then we get the interval $(245.0, 450.5)$. Only mean imputation covers this interval.

4. For the EM algorithm, we need to solve

$$\frac{\partial}{\partial \boldsymbol{\mu}} \mathbb{E}\left[\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma} \big| X_1, ..., X_n) \Big| X_{1o}, X_{2o}, ..., X_{no}; \mu^{(k)}, \Sigma^{(k)}\right] = 0_{p\times 1}$$
$$\frac{\partial}{\partial \boldsymbol{\Sigma}} \mathbb{E}\left[\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma} \big| X_1, ..., X_n) \Big| X_{1o}, X_{2o}, ..., X_{no}; \mu^{(k)}, \Sigma^{(k)}\right] = 0_{p\times p}$$

Note $\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma} \big| X_1, ..., X_n) = \sum_{i=1}^n \ell_i(\boldsymbol{\mu}, \boldsymbol{\Sigma} \big| X_i)$. Therefore we need to calculate for each i,

$$\mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\mu}} \ell_i(\boldsymbol{\mu}, \boldsymbol{\Sigma} \big| X_i) \Big| X_{io}; \mu^{(k)}, \Sigma^{(k)}\right]$$
$$\mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\Sigma}} \ell_i(\boldsymbol{\mu}, \boldsymbol{\Sigma} \big| X_i) \Big| X_{io}; \mu^{(k)}, \Sigma^{(k)}\right]$$

As derived in class, the log likelihood of the ith observation is given by

$$\begin{aligned} \ell_i(\boldsymbol{\mu}, \boldsymbol{\Sigma} | X_i) &= -\tfrac{p}{2}\log 2\pi - \tfrac{1}{2}\log\det\boldsymbol{\Sigma} - \tfrac{1}{2}(X_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(X_i - \boldsymbol{\mu}) \\ &= -\tfrac{p}{2}\log 2\pi - \tfrac{1}{2}\log\det\boldsymbol{\Sigma} - \tfrac{1}{2}\mathtt{tr}\left((X_i - \boldsymbol{\mu})(X_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\right) \end{aligned}$$

Thus,

$$\frac{\partial}{\partial \boldsymbol{\mu}} \ell_i(\boldsymbol{\mu}, \boldsymbol{\Sigma} | X_i) = -\boldsymbol{\Sigma}^{-1}(X_i - \boldsymbol{\mu}) \quad (1)$$
$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\Sigma}} \ell_i(\boldsymbol{\mu}, \boldsymbol{\Sigma} | X_i) &= -\tfrac{1}{2}\boldsymbol{\Sigma}^{-1} + \tfrac{1}{2}\boldsymbol{\Sigma}^{-1}(X_i - \boldsymbol{\mu})(X_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1} \\ &= \tfrac{1}{2}\boldsymbol{\Sigma}^{-1}((X_i - \boldsymbol{\mu})(X_i - \boldsymbol{\mu})^T - \boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1} \quad (2) \end{aligned}$$

Note that $\mathbb{E}[X_{im}|X_{io}; \mu^{(k)}, \Sigma^{(k)}] = \hat{X}_{im}^{(k)}$, so that $\mathbb{E}[X_i|X_{io}; \mu^{(k)}, \Sigma^{(k)}] = \hat{X}_i^{(k)}$. Taking the conditional expectation of (1) summing over all i and equalizing to zero shows that

$$\sum_{i=1}^n \boldsymbol{\Sigma}^{-1}(\hat{X}_i^{(k)} - \boldsymbol{\mu}) = 0$$

and hence $\mu^{(k+1)}$ is given by solving:

$$\mu^{(k+1)} : \sum_{i=1}^n (\hat{X}_i^{(k)} - \boldsymbol{\mu}) = 0$$

In order to derive the second estimating equation consider first the following calculations:

$$\mathbb{E}[(X_i - \boldsymbol{\mu})(X_i - \boldsymbol{\mu})^T \big| X_{io}; \mu^{(k)}, \Sigma^{(k)}]$$
$$= \mathbb{E}\left[\begin{bmatrix} (X_{io} - \boldsymbol{\mu}_{io})(X_{io} - \boldsymbol{\mu}_{io})^T & (X_{io} - \boldsymbol{\mu}_{io})(X_{im} - \boldsymbol{\mu}_{im})^T \\ (X_{im} - \boldsymbol{\mu}_{im})(X_{im} - \boldsymbol{\mu}_{io})^T & (X_{im} - \boldsymbol{\mu}_{im})(X_{im} - \boldsymbol{\mu}_{im})^T \end{bmatrix} \Big| X_{io}; \mu^{(k)}, \Sigma^{(k)} \right]$$

with

$$\mathbb{E}[(X_{io} - \boldsymbol{\mu}_{io})(X_{io} - \boldsymbol{\mu}_{io})^T \big| X_{io}; \mu^{(k)}, \Sigma^{(k)}] = (\hat{X}_{io}^{(k)} - \boldsymbol{\mu}_{io})(\hat{X}_{io}^{(k)} - \boldsymbol{\mu}_{io})^T \tag{3}$$
$$\mathbb{E}[(X_{io} - \boldsymbol{\mu}_{io})(X_{im} - \boldsymbol{\mu}_{im})^T \big| X_{io}; \mu^{(k)}, \Sigma^{(k)}] = (\hat{X}_{io}^{(k)} - \boldsymbol{\mu}_{io})(\hat{X}_{im}^{(k)} - \boldsymbol{\mu}_{im})^T \tag{4}$$

and

$$\begin{aligned} \mathbb{E}[(X_{im} - \boldsymbol{\mu}_{im})(X_{im} - \boldsymbol{\mu}_{im})^T \big| X_{io}; \mu^{(k)}, \Sigma^{(k)}] &= \mathbb{E}[X_{im} X_{im}^T \big| X_{io}; \mu^{(k)}, \Sigma^{(k)}] \\ &\quad - \boldsymbol{\mu}_{im}(\hat{X}_{im}^{(k)})^T - \hat{X}_{im}^{(k)} \boldsymbol{\mu}_{im}^T + \boldsymbol{\mu}_{im} \boldsymbol{\mu}_{im}^T \\ &= (\hat{X}_{im}^{(k)} - \boldsymbol{\mu}_{im})(\hat{X}_{im}^{(k)} - \boldsymbol{\mu}_{im})^T \\ &\quad + C_{imm}^{(k)} \end{aligned} \tag{5}$$

where the last inequality comes from using the variance of the conditional normal distribution:

$$\begin{aligned} \mathbb{E}[X_{im} X_{im}^T \big| X_{io}; \mu^{(k)}, \Sigma^{(k)}] &= \left(\Sigma_{imm}^{(k)} - \Sigma_{imo}^{(k)} (\Sigma_{ioo}^{(k)})^{-1} \Sigma_{iom}^{(k)} \right) \\ &\quad + \mathbb{E}\left[X_{im} \big| X_{io}; \mu^{(k)}, \Sigma^{(k)} \right] \mathbb{E}\left[X_{im}^T \big| X_{io}; \mu^{(k)}, \Sigma^{(k)} \right] \\ &= C_{imm}^{(k)} + \hat{X}_{im} \hat{X}_{im}^T \end{aligned}$$

Therefore, using (3), (4) and (5),

$$\mathbb{E}[(X_i - \boldsymbol{\mu})(X_i - \boldsymbol{\mu})^T \big| X_{io}; \mu^{(k)}, \Sigma^{(k)}] = (\hat{X}_i^{(k)} - \boldsymbol{\mu})(\hat{X}_i^{(k)} - \boldsymbol{\mu})^T + C_i^{(k)}$$

Now taking the conditional expectation over (2) equalizing it to zero we get

$$\sum_{i=1}^{n} \boldsymbol{\Sigma}^{-1} \left((\hat{X}_i^{(k)} - \boldsymbol{\mu})(\hat{X}_i^{(k)} - \boldsymbol{\mu})^T + C_i^{(k)} - \boldsymbol{\Sigma} \right) \boldsymbol{\Sigma}^{-1} = 0_{p \times p}.$$

We therefore get Σ^{k+1} by solving

$$\Sigma^{(k+1)} : \sum_{i=1}^{n} \left(\boldsymbol{\Sigma} - (\hat{X}_i^{(k)} - \boldsymbol{\mu})(\hat{X}_i^{(k)} - \boldsymbol{\mu})^T - C_i^{(k)} \right) = 0$$

Exercise 3

The data set **CentralPark.csv** consists of precipitation data from weather station at Central Park, New York. The data was collected from the National Oceanic and Atmospheric Administration (NOAA). The variable PRCP shows the observed amount of rain at time t in mm. Consider a first order Markov Chain model with a two dimensional state space corresponding to the states $\{0, 1\} = \{\text{“rainy day”}, \text{“no rain”}\}$, where we define a rain day as one with a PRCP of at least 1.5 mm. Suppose the estimated transition probability matrices

$$\mathbf{T} = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix}$$

obtained using data collected above.

1. Interpret the meaning of a_i
2. What's the long-term probability of observing a rainy day in Central Park (use a_i to express the result).
3. Can you estimate a_i for the month of July using the historical Central Park data?
4. Are the probability laws of “$X_{t+1}|X_t = 1$” and “$1-X_{t+1}|X_t = 0$” significantly different in Central Park?
5. Does a higher order chain improve the fit of the data?

Solution:

1. Let's flip the states to mean $(0, 1) = (\text{“no rain”}, \text{“rainy day”})$ to avoid confusing notation. Then,

   a_1 is the probability that there's no rain tomorrow given that there was no rain today.
   a_2 is the probability that it rains tomorrow given that there was no rain today.
   a_3 is the probability that there's no rain tomorrow given that there was rain today.
   a_4 is the probability that it rains tomorrow given that there was rain today.

2. The equation defining the stationary distribution of p_1 is

   $$a_2(1 - p_1) + a_4 p_1 = p_1 \iff p_1 = \frac{a_2}{1 + a_2 - a_4}$$

   Therefore the estimated long-term probability of observing a rainy day in Central Park is $\frac{a_2}{1+a_2-a_4}$.

3. See R code. We get the estimates $\hat{a}_1 = 77.63\%$, $\hat{a}_2 = 22.37\%$, $\hat{a}_3 = 70.29\%$ and $\hat{a_4} = 29.71\%$.

4. Note $\mathbb{P}(X_{t+1} = 1|X_t = 1) = p_{00} = a_4$ and $\mathbb{P}(1 - X_{t+1} = 1|X_t = 0) = \mathbb{P}(X_{t+1} = 0|X_t = 0) = p_{11} = a_1$.

   This amounts to testing $H_0 : a_1 = a_4$. Using the asymptotic normality of $(\hat{p}_{00}, \hat{p}_{11})$ and the asymptotic independence of their components, we have the following approximation under the null hypothesis

   $$\hat{p}_{00} - \hat{p}_{11} \approx N(0, p_{00}(1 - p_{00})/n_{0\cdot} + p_{11}(1 - p_{11})/n_{1\cdot}).$$

   This can be used as a test statistic for H_0. We reject H_0 (see R code).

5. The likelihood ratio statistic gives strong evidence against the first order model (see R code).

Exercise 4

Consider the following model for a DNA sequencing experiment. Assume that the probability of obtaining one of the four bases A,C,G,T are $p_A = 1-\theta$, $p_C = \theta-\theta^2$, $p_G = \theta^2 - \theta^3$ and $p_T = \theta^3$, where $0 \leq \theta \leq 1$. Further assume that we have n independent realizations where n_A, n_C, n_G, n_T are the observed occurrences for each base and $n_A + n_C + n_G + n_T = n$.

1. Give the joint distribution of (N_A, N_C, N_G, N_T)

2. Show that the MLE of θ is

   $$\hat{\theta} = \frac{N_C + 2N_G + 3N_T}{N_A + 2N_C + 3N_G + 3N_T}$$

3. Find the asymptotic distribution of $\hat{\theta}$

4. Find constants a_A, a_C, a_G, a_T such that $T = a_A N_A + a_C N_C + a_G N_G + a_T N_T$ is unbiased for θ.

5. Find the variance of T and five the asymptotic relative efficiency between T and $\hat{\theta}$

6. Compare the two estimators discussed above with the MLE that does not assume that p_A, p_C p_G and p_T depend on a common unknown parameter θ.

7. Using the last point, propose a test statistic for the null hypothesis H_0 : $\mathbf{p} = \mathbf{p}(\theta)$, where $\mathbf{p} = (p_A, p_C, p_G, p_T)^T$ and $\mathbf{p}(\theta) = (1-\theta, \theta-\theta^2, \theta^2-\theta^3, \theta^3)^T$ for some fixed value $\theta \in (0, 1)$.

Solutions:

1. For n_A, n_C, n_G, n_T such that $n_A + n_C + n_G + n_T = n$,

$$\begin{aligned} f_{N_A,N_C,N_G,N_T}(n_A, n_C, n_G, n_T; \theta) &= \mathbb{P}\left(N_A = n_A, N_C = n_C, N_G = n_G, N_T = n_t\right) \\ &= \frac{n!}{n_A! n_C! n_G! n_T!}(1-\theta)^{n_A}(\theta - \theta^2)^{n_C}(\theta^2 - \theta^3)^{n_G}(\theta^3)^{n_T} \end{aligned}$$

2. Since

$$\frac{\partial}{\partial \theta} \log f_{N_A,N_C,N_G,N_T}(n_A, n_C, n_G, n_T; \theta) = n_A \frac{-1}{1-\theta} + n_C \frac{(1-2\theta)}{\theta - \theta^2} + n_G \frac{2\theta - 3\theta^2}{\theta^2 - \theta^3} + n_T \frac{3}{\theta}$$

the MLE is obtained by setting the above equation to zero. Multiplying by $\theta(1-\theta)$ shows

$$\left(n_A + 2n_C + 3n_G + 3n_T\right)\theta = n_C + 2n_G + 3n_T$$

Thus the MLE of θ is

$$\hat{\theta} = \frac{N_C + 2N_G + 3N_T}{N_A + 2N_C + 3N_G + 3N_T}$$

3.

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow[n\to\infty]{\mathcal{D}} N\left(0, \frac{\theta(1-\theta)}{1+\theta+\theta^2}\right)$$

4. We want to have

$$\mathbb{E}[T] = n\left(a_A(1-\theta) + a_C(\theta - \theta^2) + a_G(\theta^2 - \theta^3) + a_T\theta^3\right) = \theta$$

We can therefore pick $a_A = 0$ and $a_C = a_G = a_T = 1/n$.

5. Note that with the above choices of constants a_A, a_C, a_G, a_T we have that $T = \frac{n - N_A}{n}$ and hence $nT \sim Binomial(n, \theta)$. By CLT therefore,

$$\sqrt{n}(T - \theta) \xrightarrow[n\to\infty]{\mathcal{D}} N\left(0, \theta(1-\theta)\right)$$

The asymptotic relative efficiency of $\hat{\theta}$ over T is

$$\frac{\frac{\theta(1-\theta)}{1+\theta+\theta^2}}{\theta(1-\theta)} = \frac{1}{1+\theta+\theta^2}$$

6. Without any dependence on θ, the MLE is $(\hat{p}_A, \hat{p}_C, \hat{p}_G, \hat{p}_T) = \left(\frac{N_A}{n}, \frac{N_C}{n}, \frac{N_G}{n}, \frac{N_T}{n}\right)$ while the MLE using part 2 is $\left(1-\hat{\theta}, \hat{\theta}(1-\hat{\theta}), \hat{\theta}^2(1-\hat{\theta}), \hat{\theta}^3\right)$. Assuming $a_A = 0$ and $a_C = a_G = a_T = 1/n$, $T = (N_A + N_G + N_T)/n$ which is an estimator of $1 - p_A$ corresponding exactly to $1 - \hat{p}_A$.

7. Let $\mathbf{p}(\theta) = (1-\theta, \theta-\theta^2, \theta^2-\theta^3, \theta^3)^T$. Using the asymptotic normality of $\hat{\mathbf{p}} = (\hat{p}_A, \hat{p}_C, \hat{p}_G, \hat{p}_T)^T$ we can construct the Wald statistic for a fixed value of θ as

$$W_n = n\left(\hat{\mathbf{p}} - \mathbf{p}(\theta)\right)^T \hat{\mathbf{V}}^{-1}\left(\hat{\mathbf{p}} - \mathbf{p}(\theta)\right)$$

where

$$\hat{\mathbf{V}}_{ij} = \begin{cases} \frac{\hat{p}_i(1-\hat{p}_i)}{n} & \text{if } i = j \in \{A, C, G, T\} \\ -\frac{\hat{p}_i\hat{p}_j}{n} & \text{if } i \neq j \in \{A, C, G, T\} \end{cases}$$

Under $H_0 : \mathbf{p} = \mathbf{p}(\theta)$ we have that $W_n \xrightarrow[n\to\infty]{\mathcal{D}} \chi^2_4$.

Exercise 5 (Optional bonus question)

Consider once again the data set **transplant.txt** from exercise 1. In this exercise you will use the **log rank test** to check the null hypothesis that the survival function of both groups is the same. Treatment A denotes the allogenic transplant and Treatment B denote the autologous transplant.
Denote by $\tau^{(1)} < \tau^{(2)} < \cdots < \tau^{(K)}$ the K distinct times of observed deaths/relapses. For $1 \leq k \leq K$, let $n_A^{(k)}$ and $n_B^{(k)}$ be the number of patients at risk undergoing treatment A and treatment B, respectively, at time $\tau^{(k)}$. Also, $n_d^{(k)}$ be the total number of deaths/relapses observed at time $\tau^{(k)}$ across both treatment groups A and B.
Thus, at time $\tau^{(k)}$ there were $n_A^{(k)} + n_B^{(k)}$ patients, and if $y^{(k)}$ is the number of deaths/relapses under treatment A, then $n_d^{(k)} - y^{(k)}$ is the number of deaths/relapses under treatment B.

1. Show that the conditional distribution of $y^{(k)}$ given $(n_A^{(k)}, n_B^{(k)}, n_d^{(k)})$ is a *HyperGeometric*$(n_A^{(k)} + n_B^{(k)}, n_A^{(k)}, n_d^{(k)})$ random variable under H_0 via the following steps:

   Hint: To calculate $P(y^{(k)} = m, n_A^{(k)} = m_A, n_B^{(k)} = m_B, n_d^{(k)} = m_d)$ *note that at time* $\tau^{(k)}$*, out of the* m_A *at risk from group A,* m *will die/relapse and out of the* m_B *at risk from group B,* $m_d - m$ *will die/relapse. Suppose* $i_1, ..., i_{m_A}$ *are the individuals at risk from group A and* $j_1, ..., j_{m_B}$ *are the individuals at risk from group B. Show that the probability of any* m *out of* $\{i_1, ..., i_{m_A}\}$ *and any* $m_d - m$ *out of* $\{j_1, ..., j_{m_B}\}$ *dying/relapsing does not depend on the value of* m*.*

2. Verify that the conditional mean and variance of $y^{(k)}$ given $(n_A^{(k)}, n_B^{(k)}, n_d^{(k)})$ under H_0 are given by

$$E^{(k)} = \frac{n_A^{(k)} n_d^{(k)}}{n^{(k)}}$$

$$V^{(k)} = \frac{n_A^{(k)} n_B^{(k)} n_d^{(k)} n_s^{(k)}}{(n^{(k)})^2 (n^{(k)} - 1)}$$

where $n^{(k)} = n_A^{(k)} + n_B^{(k)}$ and $n_s^{(k)} = n^{(k)} - n_d^{(k)}$.

3. For each $1 \leq k \leq K$ prove that under H_0, $\text{var}[y^{(k)} - E^{(k)}] = \mathbb{E}[V^{(k)}]$.
Hint: Recall the conditional variance formula for random variables X, Y

$$\text{Var}[X] = \text{Var}[\mathbb{E}[X|Y]] + \mathbb{E}[\text{Var}[X|Y]]$$

4. Show under H_0 that

$$\text{Var}\left[\sum_{k=1}^{K}(y^{(k)} - E^{(k)})\right] = \sum_{k=1}^{K}\text{Var}\left[y^{(k)} - E^{(k)}\right] = \sum_{k=1}^{K}\mathbb{E}[V^{(k)}]$$

5. To test H_0 we use the log rank test statistic:

$$Z = \frac{\sum_{k=1}^{K}\left(y^{(k)} - E^{(k)}\right)}{\left(\sum_{k=1}^{K}V^{(k)}\right)^{1/2}}$$

By using central limit theorem arguments it can be shown that under H_0, Z is asymptotically a standard normal distribution.

Using this fact test the null hypothesis that the two survival functions obtained in Exercise 1.2 above are identical.

Solution:

1. Let $\{i_1^d, ..., i_m^d\} \subset \{i_1, ..., i_{m_A}\}$ and $\{j_1^d, ..., j_{m_d-m}^d\} \subset \{j_1, ..., j_{m_B}\}$ be the individuals that die/relapse at time $\tau^{(k)}$ from the ones at risk.

Note that the event $\{Y_i = \tau^{(k)}, \delta_i = 1\}$ is the event that patient i died/relapsed at time $\tau^{(k)}$. Under H_0, the probability of this event is independent of which treatment group i belongs to.

Also $\{Y_i > \tau^{(k)}\}$ is the event that patient i is at risk at $\tau^{(k)}$. Similarly, under H_0, the probability of this event is also independent of which treatment group i belongs to.

Therefore, the event E that at time $\tau^{(k)}$, $i_1, ..., i_{m_A}, j_1, ..., j_{m_B}$ are at risk and $i_1^d, ..., i_m^d, j_1^d, ..., j_{m_d-m}^d$ are the individuals that died/relapsed, given that

patients in $\{i_1, ..., i_{m_A}, j_1, ..., j_{m_B}\}^c$ are not at risk, can be expressed as

$$E = \left(\bigcap_{l=1}^{m} \left\{Y_{i_l^d} = \tau^{(k)}, \delta_{i_l^d} = 1\right\}\right) \qquad (i_1^d, ..., i_m^d \text{ died/relapsed})$$
$$\bigcap \left(\bigcap_{l=1}^{m_d - m} \left\{Y_{j_l^d} = \tau^{(k)}, \delta_{j_l^d} = 1\right\}\right) \qquad (j_1^d, ..., j_{m_d-m}^d \text{ died/relapsed})$$
$$\bigcap \left(\bigcap_{l \in \{i_1,...,i_{m_A}\} \setminus \{i_1^d,...,i_m^d\}} \left\{Y_l > \tau^{(k)}\right\}\right) \qquad (\text{elements of } \{i_1, ..., i_{m_A}\} \text{ at risk})$$
$$\bigcap \left(\bigcap_{l \in \{j_1,...,j_{m_B}\} \setminus \{j_1^d,...,j_{m_d-m}^d\}} \left\{Y_l > \tau^{(k)}\right\}\right) \qquad (\text{elements of } \{j_1, ..., j_{m_B}\} \text{ at risk})$$

The probability of E under H_0 decomposes into a product using the IID assumption of patients,

$$\begin{aligned}\mathbb{P}(E) &= \mathbb{P}(Y_1 = \tau^{(k)}, \delta_1 = 1)^m \cdot \mathbb{P}(Y_1 = \tau^{(k)}, \delta_1 = 1)^{m_d - m} \\ &\quad \cdot \mathbb{P}(Y_1 > \tau^{(k)})^{m_A - m} \cdot \mathbb{P}(Y_1 > \tau^{(k)})^{m_B - (m_d - m)} \\ &= \mathbb{P}(Y_1 = \tau^{(k)}, \delta_1 = 1)^{m_d} \cdot \mathbb{P}(Y_1 > \tau^{(k)})^{m_A + m_B - m_d} = c\end{aligned}$$

Thus $\mathbb{P}(E) = c$ is independent of m. Summing over all indices we see,

$$\begin{aligned}&\mathbb{P}(y^{(k)} = m, n_A^{(k)} = m_A, n_B^{(k)} = m_B, n_d^{(k)} = m_d) \\ &= \sum_{\{i_1,...,i_{m_A}\} \subset \{1,...,N_A\}} \sum_{\{i_1,...,i_{m_B}\} \subset \{1,...,N_B\}} \sum_{\{i_1^d,...,i_m^d\} \subset \{i_1,...,i_{m_A}\}} \sum_{\{j_1^d,...,j_{m_d-m}^d\} \subset \{j_1,...,j_{m_B}\}} c \\ &= \binom{N_A}{m_A}\binom{N_B}{m_B}\binom{m_A}{m}\binom{m_B}{m_d - m} c = \binom{m_A}{m}\binom{m_B}{m_d - m} c'\end{aligned}$$

where N_A and N_B are the total number of patients we started out with for each group. c' is also independent of m. Hence,

$$\mathbb{P}\left(y^{(k)} = m | n_A^{(k)} = m_A, n_B^{(k)} = m_B, n_d^{(k)} = m_d\right) \propto \binom{m_A}{m}\binom{m_B}{m_d - m}$$

which proves that under H_0,

$$y^{(k)} | n_A^{(k)}, n_B^{(k)}, n_d^{(k)} \sim \mathit{HyperGeometric}(n_A^{(k)} + n_B^{(k)}, n_A^{(k)}, n_d^{(k)})$$

2. Follows from the mean and variance of a hypergeometric random variable.

3. Note, $\text{Var}\left[\mathbb{E}\left[y^{(k)} | n_A^{(k)}, n_B^{(k)}, n_d^{(k)}\right] - E^{(k)}\right] = \text{Var}[0] = 0$. Thus by the conditional variance formula,

$$\text{Var}[y^{(k)} - E^{(k)}] = \mathbb{E}\left[V^{(k)}\right]$$

4. The second equality follows from part 3 above. The first equality follows since for $k < l$, $\text{Cov}\left[y^{(k)} - E^{(k)}, y^{(l)} - E^{(l)}\right] = 0$, because

$$\begin{aligned}
\mathbb{E}\left[(y^{(k)} - E^{(k)})(y^{(l)} - E^{(l)})\right] &= \mathbb{E}\left[\mathbb{E}\left[(y^{(k)} - E^{(k)})(y^{(l)} - E^{(l)})|n_A^{(l)}, n_B^{(l)}, n_d^{(l)}\right]\right] \\
&= \mathbb{E}\left[(y^{(k)} - E^{(k)})\mathbb{E}\left[(y^{(l)} - E^{(l)})|n_A^{(l)}, n_B^{(l)}, n_d^{(l)}\right]\right] \\
&= 0
\end{aligned}$$

   and $\mathbb{E}[y^{(k)} - E^{(k)}] = \mathbb{E}[y^{(l)} - E^{(l)}] = 0$.

5. See R code. We get a very large p-value and hence the data shows that the survival functions are the same.