1. Let random variable $X$ follows a Poisson distribution with density function

$$p(X = x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!},$$

where $x = 0, 1, 2, \ldots$. Suppose that we are interested in estimating $\theta = e^{-3\lambda}$ based on a single observation.

(a) Show that $T = (-2)^X$ is an unbiased estimator of $\theta$.
(b) Is $T$ the only unbiased estimator of $\theta$? Justify your answer.
(c) Would you use $\hat{\theta} = T$ in practice if you have a single observation? Briefly explain your reasons.
(d) Suppose you have a random sample $X_1, X_2, \ldots, X_n$ of size $n$ from the above Poisson distribution, could you find a sensible estimate of $\theta$. Is your estimate unique? Justify your answer.
(e) With the aforementioned random sample $X_1, X_2, \ldots, X_n$ please propose a 95% confidence interval estimate for $\theta$.

2. Bacterial growth on a petri dish is modeled as follows: At time $t = 0$ one bacterium is placed on the dish. Each bacterium, independently of other bacteria, divides into two bacteria after an exponential($\lambda$) waiting time. Here, $\lambda$ denotes the scale parameter (that is, the mean is $1/\lambda$). Let $X(t)$ denote the number of bacteria at time $t > 0$. Clearly, $X(0) = 1$. Assume the bacteria do not die.

(a) Show that the time $S_n$ until the number of bacteria becomes $1+n$ has a cumulative distribution function given by

$$(1 - e^{-\lambda t})^n, \quad 0 < t < \infty.$$  

(b) Show that $X(t)$ is distributed as a geometric variable with mean $e^{\lambda t}$. 

1
(c) Given that \(X(30) = 11\), (i) what is the conditional distribution of the age of the youngest bacterium at time \(t = 30\)? (ii) What is the conditional distribution of the birth time of the next bacterium to be born after \(t = 30\).

3. Suppose that we have two independent random samples: \(X_1, X_2, \ldots, X_m\) from \(N(\mu, \theta_1)\) and \(Y_1, Y_2, \ldots, Y_n\) from \(N(\mu, \theta_2)\) respectively.

   (a) Assuming the variances \(\theta_1, \theta_2\) known, derive the MVUE of the common mean \(\mu\).

   (b) Assuming the variances \(\theta_1, \theta_2\) known, derive the likelihood ratio test for \(H_0: \mu = \mu_0\) versus \(H_1: \mu \neq \mu_0\). Is this test uniformly most powerful?

   (c) If the variances \(\theta_1, \theta_2\) are unknown, derive the MLEs of \(\mu, \theta_1, \theta_2\).

   (d) Suppose that the variances \(\theta_1, \theta_2\) are unknown. Consider a plug-in estimator of \(\mu\) obtained by replacing \(\theta_i\) by the sample variance \(S_i^2\) (for \(i = 1, 2\)) in the MVUE of \(\mu\) in part (a). Is the plug-in estimator unbiased for \(\mu\)?

4. In a medical study, investigators are interested in the time from diagnosis to death. There are 30 patients enrolled in the study and they are followed for about 5 years. Some of the patients die of the condition before the end of follow up. The remaining patients are still alive at the end of follow up and hence their death time is right censored. Denote the observations as \((t_i, \delta_i, x_i)\), where \(t_i = \min(t_i^*, C_i)\) is the observed time, \(t_i^*\) is the death time, \(C_i\) is the censoring time, \(\delta_i\) is the death indicator (\(\delta_i = 1\) if \(t_i^* \leq C_i\), and \(\delta_i = 0\) otherwise), \(x_i\) is the covariate age. Note that \(t_i^*, C_i\) are the underlying death and censoring times which may not be observed.

   Now if it is of interest to study the age at enrollment as a risk factor of the time to death, one will naturally consider using Cox proportional hazards model.

   (a) Write the proportional hazards model with age at enrollment as the only covariate. Briefly explain the assumptions on right censoring.

   (b) Write the (log) full likelihood of observations \((t_i, \delta_i, x_i), i = 1, \ldots, 30\). The Breslow’s estimate of baseline hazard at an observed death time is:

   \[
   \hat{\lambda}_0(t_i) = \frac{\delta_i}{\sum_{l \in R_i} \exp(x_l \beta)};
   \]

   where \(R_i\) is the at-risk set at \(t_i\). Assuming that the baseline hazard function taking value 0 except at the observed death time, derive the (log) partial likelihood by plugging in the Breslow estimate into the full likelihood.

   (c) Write the (log) partial likelihood to be maximized for the Cox model (even if you couldn’t derive it in part b). Prove that the partial likelihood has a unique maximum.

   (d) Now consider only one binary covariate \(x\) taking values 0 and 1. One can make comparison between the 0 and 1 group by testing \(H_0: \beta = 0\) where beta is
the regression coefficient of the binary covariate. Derive the score function (first derivative of log partial likelihood function) under the null hypotheses. One can also make comparison between group 0 and 1 using the logrank test. Write the numerator of logrank statistic for comparison between the groups defined by $x$. What is the relationship between these two expressions assuming there is no tie in the observed times? Show all your work.

5. Consider a regression model

$$Y_i = \beta X_i + \epsilon_i,$$

where $\epsilon_i \sim N(0, 1)$ are i.i.d. for $i = 1, 2, \ldots, n$.

(a) Assume that $X_i$’s are fixed.

i. Show that the least squares estimator $\hat{\beta}_n$ of $\beta$ based on $n$ data points is unbiased.

ii. What is the variance of $\hat{\beta}_n$?

iii. Show that $\hat{\beta}_n$ is the uniformly minimum variance unbiased estimator of $\beta$.

(b) Suppose that $X_i \sim N(0, 1)$ ($i = 1, 2, \ldots, n$) and $X_i$’s and $\epsilon_i$’s are independent.

i. What is the variance of $\hat{\beta}_n$?

ii. Is $\hat{\beta}_n$ the UMVUE of $\beta$? Prove your conclusion.

iii. Derive the asymptotic distribution of $\sqrt{n}(\hat{\beta}_n - \beta)$.

6. Let $Y_i = (Y_{ij1}, \ldots, Y_{ijm_i})$ be independent vectors of $n$ patients, $i = 1, \ldots, n; j = 1, \ldots, m_i$. Also let $\mu_{ij} = E(Y_{ij}) = g(\beta_0^* + \beta_1^* t_i)$, $g$ is an inverse link function, $t_i = 0$ if no treatment and 1 if treatment. For a continuous outcome, the identity link is used; for a repeated binary outcome, a probit link can be used, $g = \Phi$, where $\Phi$ is the cumulative distribution function of a standard normal variable, $\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-s^2/2} ds$. For either type of outcomes, we can model the intra-patient correlation by using random effects models: $E(Y_{ij}|U_i) = \mu_{ij} + U_i$ for a repeated continuous outcome and $E(Y_{ij}|U_i) = \Phi(\mu_{ij} + U_i)$ for a repeated binary outcome, where $\mu_{ij} = g(\beta_0^* + \beta_1^* t_i)$ and $U_i \sim N(0, \sigma^2)$ are i.i.d.

(a) Find the following:

i. $E(Y_{ij})$ for a repeated continuous outcome;

ii. if $X_i$’s are independent, $X_i \sim N(u_i, \sigma_i^2), i = 1, 2$, find the distribution of $X_1 - X_2$.

iii. $E(Y_{ij})$ for a repeated binary outcome under the above probit model.

(b) Show what the conditional ($\beta_1$) and unconditional ($\beta_1^*$) treatment effects are for the repeated continuous outcome.

(c) Show what the conditional ($\beta_1$) and unconditional ($\beta_1^*$) treatment effects are for a repeated binary outcome.