1. Consider a statistic \( \hat{\theta}_n \) that is known to be asymptotically normally distributed. In particular, suppose that

\[
Z_n = \sqrt{n} \left( \frac{\hat{\theta}_n - \theta}{\sqrt{V(\theta)}} \right) \rightarrow_d N(0,1)
\]

where \( \text{Var}[\hat{\theta}_n] = V(\theta)/n \) may depend upon the value of \( \theta \). In this case, a hypothesis test having one-sided type I error \( \alpha/2 \) might be based on a critical function which rejects \( H_0 : \theta \leq \theta_0 \) in favor of alternative hypothesis \( H_1 : \theta > \theta_0 \) when

\[
Z_n = \sqrt{n} \left( \frac{\hat{\theta}_n - \theta_0}{\sqrt{V(\theta)}} \right) > z_{1-\alpha/2} = \Phi^{-1}(1 - \alpha/2),
\]

where \( \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du \). The power function for this hypothesis test is then

\[
Pwr(\theta) = \Pr \left[ \sqrt{n} \left( \frac{\hat{\theta}_n - \theta_0}{\sqrt{V(\theta)}} \right) > z_{1-\alpha/2} \mid \theta \right].
\]

(a) Suppose we wish to design a study to test \( H_0 : \theta \leq \theta_0 \) vs. \( H_1 : \theta > \theta_0 \) with one-sided type I error \( \alpha/2 \), and we wish to attain power \( \beta \) at \( \theta = \theta_1 \). Derive a general formula for the required sample size that will lead to these operating characteristics.

Solution: We require

\[
\beta = \Pr \left[ \sqrt{n} \left( \frac{\hat{\theta}_n - \theta_0}{\sqrt{V(\theta_0)}} \right) > z_{1-\alpha/2} \mid \theta = \theta_1 \right]
\]

\[
= \Pr \left[ \hat{\theta}_n > \frac{z_{1-\alpha/2} \sqrt{V(\theta_0)}}{\sqrt{n}} \mid \theta = \theta_1 \right]
\]

\[
= \Pr \left[ \sqrt{n} \frac{\hat{\theta}_n - \theta_1}{\sqrt{V(\theta_1)}} > \frac{z_{1-\alpha/2} \sqrt{V(\theta_0)}}{\sqrt{V(\theta_1)}} + \sqrt{n} (\theta_0 - \theta_1) \mid \theta = \theta_1 \right]
\]

\[
= 1 - \Phi \left( \frac{z_{1-\alpha/2} \sqrt{V(\theta_0)}}{\sqrt{V(\theta_1)}} + \sqrt{n} (\theta_0 - \theta_1) \right)
\]

\[
\Rightarrow \Phi^{-1}(1 - \beta) = -\Phi^{-1}(\beta) = \frac{z_{1-\alpha/2} \sqrt{V(\theta_0)}}{\sqrt{V(\theta_1)}} + \sqrt{n} (\theta_0 - \theta_1)
\]

\[
\Rightarrow \sqrt{n} (\theta_0 - \theta_1) = z_{1-\alpha/2} \sqrt{V(\theta_0)} + z_{\beta} \sqrt{V(\theta_1)}
\]

\[
\Rightarrow n = \left[ \frac{z_{1-\alpha/2} \sqrt{V(\theta_0)} + z_{\beta} \sqrt{V(\theta_1)}}{(\theta_0 - \theta_1)} \right]^2
\]

(b) Now consider a randomized clinical trial where patients are randomized to group A vs. group B and are followed for a censored time-to-event endpoint. Further suppose that patients are randomized to
group $A$ with probability $\pi$ and to group $B$ with probability $1 - \pi$. Finally suppose the trial will compare groups via a proportional hazards model with a single dichotomous covariate $X$ that is an indicator of assignment to group $B$, so that

$$\lambda_i(t) = \lambda_0(t)e^{\theta X_i}.$$  

Suppose we wish to test the null hypothesis $H_0 : \theta = 0$ (ie. no difference in the hazard functions between the two groups) and would like to design a study to have power $\beta$ for detecting a hazard ratio of $e^{\theta_1}$. In this case, it can be shown that

$$Z_D = \sqrt{D \hat{\theta}_D - \theta \sqrt{V(\theta)}} \rightarrow d N(0, 1)$$

where $D$ is the total number of observed deaths and $V(\theta) = \frac{1}{\text{var}[X]} = \frac{1}{\pi(1-\pi)}$. Use this fact and your work above to derive a formula for the number of events that would be required to obtain power $\beta$ for detecting a hazard ratio of $e^{\theta_1}$.

**Solution:** From the work above, we have

$$D = \left( \frac{z_{1-\alpha/2} + z_\beta}{\pi(1-\pi)\theta_1^2} \right)^2$$

(c) Above, you solved for the total number of required events. In general, it necessary to know the expected number of patients required for study planning purposes. That is, we must translate the number of events to the expect number of patients. To do this, note that:

$$N = \frac{D}{\pi \Pr_0[\text{Event}] + (1 - \pi) \Pr_1[\text{Event}]}$$

where $D$ is the total number of required events. Assuming that patients uniformly enroll in the study over the first $T_E$ years and the study lasts for $T_L > T_E$ years with no additional drop out, show that

$$\Pr_i[\text{Event}] = 1 - \frac{1}{T_E} \int_0^{T_E} S_i(T_L - t) dt$$

where $S_i(\cdot)$ denotes the assumed survival distribution group $i$, $i = 0, 1$.

**Solution:**

$$\Pr_i[\text{Event}] = \int_0^{T_A} \Pr[\text{Event & Entry at } t] dt$$

$$= \int_0^{T_A} \Pr[\text{Event | Entry at } t] \Pr[\text{Entry at } t] dt$$

$$= 1 - \int_0^{T_A} \Pr[\text{No Event | Entry at } t] \Pr[\text{Entry at } t] dt$$

$$= 1 - \frac{1}{T_A} \int_0^{T_A} \Pr[\text{No Event | Entry at } t] dt \quad \text{(unif acc)}$$

$$= 1 - \frac{1}{T_A} \int_0^{T_A} S_i(T_L - t) dt$$

(d) Finally, suppose that you are part of a study team that wishes to determine the sample size for a randomized, placebo-controlled clinical trial. A one-sided (unweighted) log-rank test (equivalent to a Cox proportional hazards model with a single binary covariate) with significance level .025 will be performed. It is desired that the study attains 90% power to detect the alternative that the new
treatment reduces the hazard function by 25% (i.e., a hazard ratio of 0.75). Patients will be uniformly enrolled for 3 years and followed for an additional year (i.e., the study will last for a total of 4 years). Based on historical data, it is expected that survival times in the untreated group will be exponentially distributed and that the 5-year survival probability will be 60%. Suppose randomization will be 1:1 and that loss-to-followup will be negligible in the study. Determine how many events and how many patients will be required in the study under these assumptions.

**Solution:** First note from part (b), the study will require a total of

\[ D = \frac{(z_{1-\alpha/2} + z_\beta)^2}{\pi(1 - \pi)\theta^2} = \frac{4(1.96 + 1.28)^2}{[\log(.75)]^2} = 507.84 \rightarrow 508 \]

Now to convert this to the number of required patients, note that the survival function for an exponential distribution is given by \( S(t) = e^{-\lambda t} \). Thus, by assumption the hazard in for the placebo arm is \( \lambda_0 = -\log(.6)/5 \), and under the hypothesized treatment effect the hazard for the treatment arm is \( \lambda_1 = -.75 \log(.6)/5 \). From this and the accrual followup assumptions, the probability of an event in group \( i \) is

\[
\Pr_i[\text{Event}] = 1 - \frac{1}{3} \int_0^4 e^{-\lambda_i (t-4)} \, dt \\
= 1 - \frac{1}{3\lambda_i} (e^{-\lambda_i} - e^{-4\lambda_i})
\]

and

\[ N = \frac{507.84}{0.5 \Pr_0[\text{Event}] + 0.5 \Pr_1[\text{Event}]} . \]

Performing the calculations in R, we have

\[
\begin{align*}
> & \text{lambda0} <- -\log(.6)/5 \\
> & \text{lambda1} <- .75*\text{lambda0} \\
> & c(\text{lambda0}, \text{lambda1}) \\
> & [1] \ 0.102165 \ 0.076624 \\
> & \text{Pevent.0} <- 1 - (1/(3*\text{lambda0}))*((\exp(-\text{lambda0}) - \exp(-4*\text{lambda0}))) \\
> & \text{Pevent.1} <- 1 - (1/(3*\text{lambda1}))*((\exp(-\text{lambda1}) - \exp(-4*\text{lambda1}))) \\
> & c(\text{Pevent.0}, \text{Pevent.1}) \\
> & [1] \ 0.22237 \ 0.17251 \\
> & D <- 4*(\text{qnorm}(.975)+\text{qnorm}(.9))^2 / (\log(.75))^2 \\
> & N <- D / (.5*\text{Pevent.0} + .5*\text{Pevent.1}) \\
> & c(D,N) \\
> & [1] \ 507.84 \ 2572.16
\end{align*}
\]

Thus the study would be required to enroll approximately \( N = 2573 \) subjects over 3 years.

2. Consider data \((y_i, \delta_i), i = 1, ..., n\), where \( \delta_i = 1 \) if \( y_i \) is the observed failure time, and \( \delta_i = 0 \) if \( y_i \) is a censored time. Suppose we wish to model these data using a Bayesian framework, assuming an exponential distribution of rate \( \lambda > 0 \), so that if \( T_i \) is the true survival time for subject \( i \) then \( T_i \sim \text{Exp}(\lambda) \) where \( f(t_i; \lambda) = \lambda e^{-\lambda t_i} \), for \( t_i > 0 \), and zero otherwise. Further assume that censoring and failure times are independent.
(a) In a Bayesian framework, inference is based upon the posterior distribution of the model parameters, conditional upon the observed data. Generally speaking, for unknown parameter vector $\hat{\theta}$ and data $\vec{x}$, the posterior distribution can be obtained via Bayes Rule as

$$
\pi_{\hat{\theta}|\vec{x}}(\hat{\theta}|\vec{x}) = \frac{f_{\vec{x}|\hat{\theta}}(\vec{x}|\theta)\pi(\theta)}{\int f_{\vec{x}|\hat{\theta}}(\vec{x}|\theta)\pi(\theta)\,d\theta},
$$

where $f_{\vec{x}|\hat{\theta}}(\vec{x}|\theta)$ denotes the joint distribution of $\vec{X}$ (i.e., the likelihood) and $\pi(\theta)$ denotes the prior distribution for $\theta$.

For a general set of survival data $(\vec{y}, \vec{\delta})$, assume a Gamma($a, b$) prior distribution for $\lambda$. Show that the posterior distribution $\lambda|\vec{y}, \vec{\delta} \sim$ Gamma($a', b'$), and identify $a'$, $b'$. (Hint: Recall the form of the $i$-th subject’s contribution to the likelihood under independent censoring.)

Solution: For the $i$th individual we have

$$
L_i(\lambda) = f(y_i, \delta_i|\lambda) = [f(y_i; \lambda)]^{\delta_i}[1 - F(y_i; \lambda)]^{1-\delta_i} = (\lambda^{\delta_i}e^{-\lambda y_i^{\delta_i}})(e^{-\lambda y_i(1-\delta_i)}) = \lambda^{\delta_i}e^{-\lambda y_i},
$$

so that $f(\vec{y}, \vec{\delta}|\lambda) = \lambda^{\sum_{i=1}^n \delta_i}e^{-\lambda \sum_{i=1}^n y_i}$. Now, let $Y = \sum_{i=1}^n y_i$ and $D = \sum_{i=1}^n \delta_i$. Then,

$$
f(\vec{y}, \vec{\delta}|\lambda) = \lambda^{D}e^{-\lambda Y}
$$

and the prior density for $\lambda$ is given by

$$
g(\lambda) = \frac{1}{\Gamma(a)(1/b)^a} \lambda^{a-1}e^{-\lambda b}.
$$

So for the posterior distribution we have

$$
p(\lambda|\vec{y}, \vec{\delta}) = \frac{f(\vec{y}, \vec{\delta}|\lambda)g(\lambda)}{\int_0^\infty f(\vec{y}, \vec{\delta}|\lambda)g(\lambda)d\lambda} = \frac{1}{\Gamma(a)(1/b)^a} \lambda^{D+a-1}e^{-\lambda(Y+b)}
$$

$$
= \frac{1}{\Gamma(a + D)(Y + b)^{-(a+b)}} \lambda^{D+a-1}e^{-\lambda(Y+b)}
$$

Thus $\lambda|\vec{y}, \vec{\delta} \sim$ Gamma($a', b'$), where $a' = a + \sum_{i=1}^n \delta_i$, and $b' = b + \sum_{i=1}^n y_i$.

(b) Table 1 shows data reported on the times, in weeks, of remission (symptom-free condition) of leukemia patients. The patients were paired and one patient randomly received the drug 6-mercaptopurine (6-MP) while the other acted as a control. For now, we will assume we have independence within and between treatment groups. Suppose we wish to model the data from each group using exponential distributions with $\lambda_k, k = 1, 2$ being assigned the same gamma prior Gamma($a = 3, b = 45$). For the data in Table 1 obtain the posterior distributions $\lambda_k|\vec{y}_k, \vec{\delta}_k, k = 1, 2$. 

Solution: From Table 1, note that \( Y_1 = \sum_{i=1}^{21} y_{1i} = 359, Y_2 = \sum_{i=1}^{21} y_{2i} = 182, D_1 = \sum_{i=1}^{21} \delta_{1i} = 9, \) and \( D_2 = \sum_{i=1}^{21} \delta_{2i} = 21. \) So, by part (a) we have

\[
\lambda_1 | \vec{y}_1, \vec{\delta}_1 \sim \text{Gamma}(3 + 9, 45 + 359) = \text{Gamma}(12, 404)
\]

\[
\lambda_2 | \vec{y}_2, \vec{\delta}_2 \sim \text{Gamma}(3 + 21, 45 + 182) = \text{Gamma}(24, 227)
\]

(c) Under squared error loss, the Bayes estimator is posterior mean of a parameter. Using the data from Table 1, find the Bayes estimator for \( \lambda_1 \) and \( \lambda_2 \) assuming squared error loss.

Solution: The Bayes estimator under squared error loss is given by the posterior mean. In this case, we have

\[
\hat{\lambda}_1 = \frac{12}{404} = 0.029703
\]

\[
\hat{\lambda}_2 = \frac{24}{227} = 0.10573
\]

(d) Using the data from Table 1, provide a 95% credible interval for \( \lambda_1 \) and \( \lambda_2 \) formed by taking the lower 2.5-percentile and upper 97.5-percentile of the posterior distribution. (Hint: Use R to obtain this.)

Solution: Below are the limits of a quantile interval as computed in R:

```r
## 95% CI for lambda_1
> c(qgamma(.025, shape=12, rate=404), qgamma(.975, shape=12, rate=404))
[1] 0.015348 0.048718

## 95% CI for lambda_2
> c(qgamma(.025, shape=24, rate=227), qgamma(.975, shape=24, rate=227))
[1] 0.067741 0.152032
```

(e) The posterior predictive distribution for a new observation \( z \) is obtained by integrating the likelihood for \( z \) (conditional upon the unknown parameters) over the posterior of the unknown parameters. More specifically, in the context of the current problem, \( p(z | \vec{y}_k, \vec{\delta}_k) = \int_0^\infty f(z | \lambda_k)p(\lambda | \vec{y}_k, \vec{\delta}_k)d\lambda_k. \) Analytically obtain the predictive distribution for the time of remission which we would expect for a patient receiving each of the drugs. Using R, draw a random sample of size 1000 from each posterior predictive distribution and compare the resulting histograms.

Solution: For general \( a, b \) and \( k = 1, 2 \), we have

\[
f(z | \lambda_k) = \lambda e^{-\lambda z}
\]

\[
p(\lambda | \vec{y}_k, \vec{\delta}_k) = \frac{1}{\Gamma(a + D_k)(Y_k + b)^{-(a+D_k)}}\lambda^{D_k + a - 1}e^{-\lambda(Y_k+b)}
\]

So that the predictive distribution is given by

\[
p(z | \vec{y}_k, \vec{\delta}_k) = \int_0^\infty f(z | \lambda_k)p(\lambda | \vec{y}_k, \vec{\delta}_k)d\lambda_k
\]

\[
= \frac{1}{\Gamma(a + D_k)(Y_k + b)^{-(a+D_k)}}\int_0^\infty \lambda^{D_k + a}e^{-\lambda(Y_k+b+z)}d\lambda_k
\]

\[
= \frac{1}{\Gamma(a + D_k)(Y_k + b)^{-(a+D_k)}}\Gamma(a + D_k + 1)(Y_k + b + z)^{-(a+D_k+1)}
\]

\[
= \frac{(a + D_k)(Y_k + b + z)^{-(a+D_k+1)}}{(Y_k + b)^{-(a+D_k)}}
\]
Thus in the present scenario with $a = 3, b = 45, T_1 = 359, T_2 = 182, D_1 = 9, D_2 = 21$, we have

$$p(z|\tilde{y}_1, \delta_1) = 12(404^{12})(z + 404)^{-13}$$
$$p(z|\tilde{y}_2, \delta_2) = 24(227^{24})(z + 227)^{-25}.$$ 

A histogram of the posterior predictive survival times for each treatment arm is given in Figure 1 (code to generate this is given below). We can see that the predictive survival times for the 6-MP arm are longer than those for the placebo arm. On very nice feature of the Bayesian framework is that we can automatically estimate functional of the posterior predictive distribution off of these samples. For example, the posterior predictive probability of surviving past 20 months in the 6-MP and placebo arms is estimated to be 0.5564 and 0.1349, respectively (again, code is given below).

```r
##
##### 6-MP Data
##
> y1 <- c(6,6,6,6,7,9,10,10,11,13,16,17,19,20,22,23,25,32,32,34,35)
> d1 <- c(0,1,1,1,0,0,1,0,1,1,0,0,1,1,0,0,0,0,0,0)
> y2 <- c(1,1,2,2,3,4,4,5,5,8,8,8,8,11,11,12,12,15,17,22,23)
> d2 <- rep(1, 21)
> Y1 <- sum( y1 )
> D1 <- sum( d1 )
> Y2 <- sum( y2 )
> D2 <- sum( d2 )

##
##### Draws from the posterior predictive distn of lambda1 and lambda2
##
> nsim <- 10000
> lambda1 <- rgamma( nsim, 3+D1 ) / (45+Y1)
> lambda2 <- rgamma( nsim, 3+D2 ) / (45+Y2)

> pred.t1 <- rexp( nsim, lambda1 )
> pred.t2 <- rexp( nsim, lambda2 )
> par(mfrow=c(2,1))
> hist( pred.t1, nclass=90, xlim=c(0,150), ylim=c(0,5100),
> main="Treated with 6-MP", xlab="Predictived Survival Time" )
> hist( pred.t2, nclass=20, xlim=c(0,150), ylim=c(0,5100),
> main="Treated with Placebo", xlab="Predictived Survival Time" )

##
##### Predictive prob of T > 20
##
> sum( pred.t1 > 20 ) / nsim
[1] 0.5564
> sum( pred.t2 > 20 ) / nsim
[1] 0.1349
```
Figure 1: Histogram of posterior predictive survival times for each treatment arm.