Problem Set 6: Written Part

Your Name Goes Here

Details

How to Write Up

The written part of this assignment can be either typeset using latex or hand written.

Grading

5% of your grade on this assignment is for turning in something legible. This means it should be organized, and any Rmd files should knit to pdf without issue.

An additional 15% of your grade is for completion. A quick pass will be made to ensure that you've made a reasonable attempt at all problems.

Across both the written part and the R part, in the range of 1 to 3 problems will be graded more carefully for correctness. In grading these problems, an emphasis will be placed on full explanations of your thought process. You don't need to write more than a few sentences for any given problem, but you should write complete sentences! Understanding and explaining the reasons behind what you are doing is at least as important as solving the problems correctly.

Solutions to all problems will be provided.

Collaboration

You are allowed to work with others on this assignment, but you must complete and submit your own write up. You should not copy large blocks of code or written text from another student.

Sources

You may refer to our text, Wikipedia, and other online sources. All sources you refer to must be cited.

Problem I: Confidence intervals for exponential distribution

obtain interval estimates from the Bayesian framework and by maximum likelihood. We will use a different parameterization of the exponential than we used on the midterm, so don't worry if it seems like your results aren't matching up. *You will need to do the written part before you can do some parts of the R part for this question.*

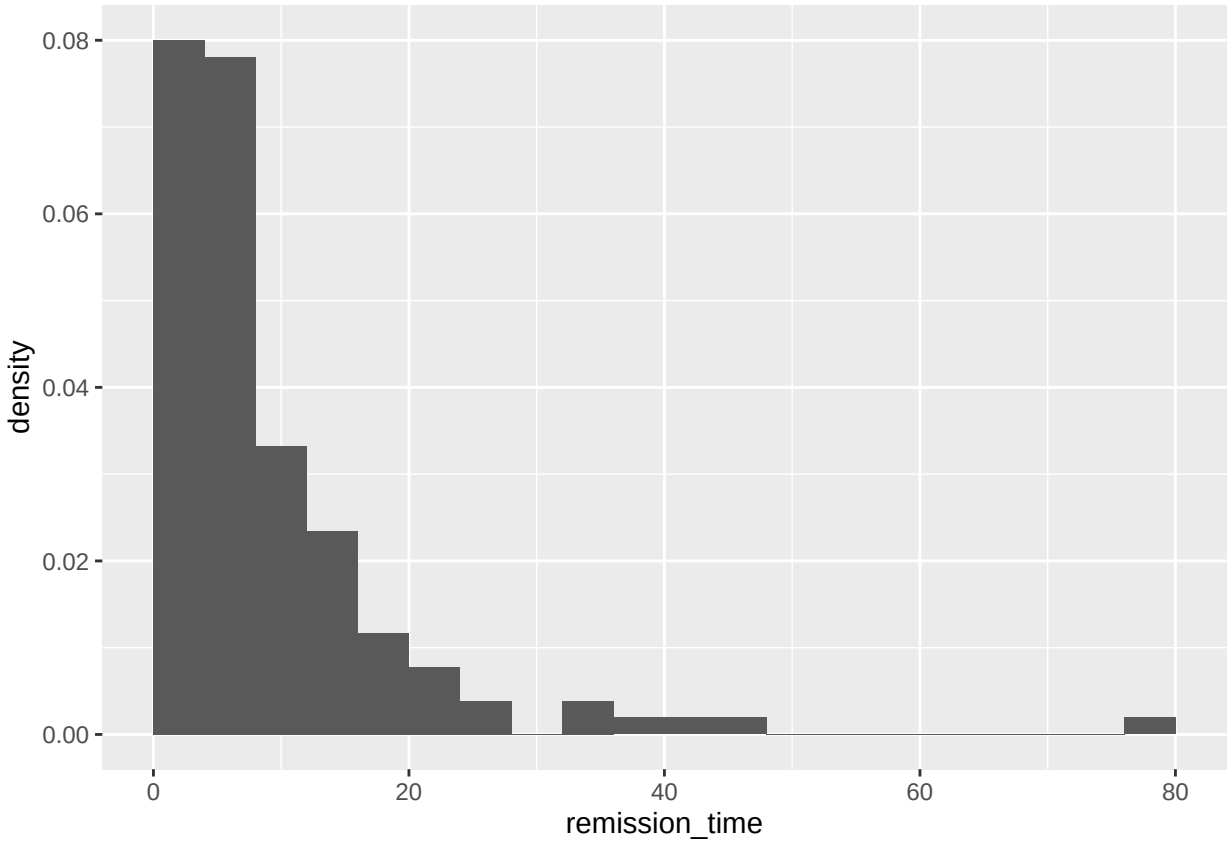
The following R code reads in remission times in months for 128 bladder cancer patients and makes an initial plot of the data. These data were reported in Lee ET, Wang JW (2003) Statistical methods for survival data analysis.

```
library(readr)
library(ggplot2)

cancer_remissions <- read_csv("http://www.evanlray.com/data/misc/bladder_cancer/bladder_cancer.csv")
```

```
## Parsed with column specification:
## cols(
##   remission_time = col_double()
## )
```

```
ggplot(data = cancer_remissions, mapping = aes(x = remission_time)) +
  geom_histogram(mapping = aes(y = ..density..), binwidth = 4, boundary = 0)
```



Set up and foundational facts

Let's use an exponential model for these data: $X_1, \ldots, X_n \sim \text{Exp}(\lambda)$, where the X_i are independent.

Here are some facts about the exponential distribution (please use this parameterization of the exponential distribution for this problem):

If $X \sim Exp(\lambda)$ then the density function is given by $f(x|\lambda) = \lambda e^{-\lambda x}$

The mean and variance are given by $E(X) = \frac{1}{\lambda}$ and $Var(X) = \frac{1}{\lambda^2}$.

Based on a sample $x_1, \ldots, x_n$, the log-likelihood function is:

$$L(\lambda|x_1, \ldots, x_n) = n\log(\lambda) - \lambda \sum_{i=1}^{n} x_i$$

The first derivative of the log-likelihood function is:

$$\frac{d}{d\lambda} L(\lambda|x_1, \ldots, x_n) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$$

The second derivative of the log-likelihood function is:

$$\frac{d^2}{d\lambda^2} L(\lambda|x_1, \ldots, x_n) = -\frac{n}{\lambda^2}$$

The maximum likelihood estimator can be shown to be: $\hat{\lambda}_{MLE} = 1/\bar{X}$.

It is also possible to conduct a Bayesian analysis using a prior for λ that is Gamma(α, β), which has pdf $f_\Lambda(\lambda|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$. This is a conjugate prior, and the posterior distribution is $\lambda|x_1, \ldots, x_n, \alpha, \beta \sim \text{Gamma}(\alpha + n, \beta + \sum_{i=1}^{n} x_i)$.

Overview

So far, we have at our disposal 5 (!) different approaches for obtaining an interval estimate for the unknown model parameter λ of the Exponential model:

1. A frequentist confidence interval based on an exact distribution for a relevant statistic;
2. A frequentist confidence interval based on a large-sample normal approximation to the sampling distribution of the maximum likelihood estimator.
3. A Bayesian credible interval based on quantiles of the exact Gamma posterior distribution.
4. A Bayesian credible interval based on a large-sample normal approximation to the posterior distribution.
5. A Bayesian credible interval based on a MCMC sample drawn from the exact posterior distribution.

We will also develop two bootstrap confidence interval methods, but those are not featured on this problem set. We will explore bootstrap interval estimation for the parameter λ, and compare with the approaches developed here, on the next problem set.

1. Exact confidence interval

It can be shown that $2n\lambda\bar{X} \sim \chi^2_{2n}$, where χ^2_{2n} denotes a Chi-squared distribution with $2n$ degrees of freedom. Use this fact to derive an expression for an exact $(1-\alpha)100\%$ confidence interval for λ in terms of the quantiles $\chi^2_{2n}(\frac{\alpha}{2})$ and $\chi^2_{2n}(1-\frac{\alpha}{2})$.

2. Large-sample approximate confidence interval

Find an expression for a frequentist confidence interval for λ based on a large-sample normal approximation to the sampling distribution of the maximum likelihood estimator. In problem set 5, we have show that for a large sample size n, it is approximately the case that $\hat{\Lambda}^{MLE} \sim \text{Normal}\left(\lambda, \frac{1}{I(\hat{\lambda}_{MLE})}\right)$ where $I(\widehat{\lambda}_{MLE}) = \frac{n}{\hat{\lambda}^2_{MLE}} = \frac{n}{(1/\bar{x})^2} = n\bar{x}^2$. The approximation to use in deriving the confidence interval is therefore $\hat{\Lambda}^{MLE} \sim \text{Normal}\left(\lambda, \frac{1}{n\bar{x}^2}\right)$

3. No work to do for the exact Bayesian credible interval in the written part

4. Large-sample approximate credible interval

Find an expression for a Bayesian credible interval for λ based on a large-sample normal approximation to the posterior distribution of Λ.

In problem set 5, we have shown that for a large sample size n, it is approximately the case that $\Lambda | X_1 = x_1, \ldots, X_n = x_n \sim \text{Normal}\left(\frac{1}{\bar{x}}, \frac{1}{n\bar{x}^2}\right)$

5. No work to do for the MCMC-based Bayesian credible interval in the written part

Problem II: Likelihood ratio tests for exponential distribution

Suppose that $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Exponential}(\lambda)$. As a reminder, their pdf is $f(x_i|\lambda) = \lambda e^{-\lambda x_i}$.

Suppose we want to conduct a test of the simple hypotheses $H_0 : \lambda = 10$ vs $H_A : \lambda = 4$.

1. Find the form of the likelihood ratio statistic for this test. It is not important for this part whether or not your answer is a random variable.

2. Show that you could calculate the p-value for the test based on a comparison of $\bar{X}$ and $\bar{x}$.

3. As stated above, it can be shown that $2n\lambda\bar{X} \sim \chi^2_{2n}$, where χ^2_{2n} denotes a Chi-squared distribution with $2n$ degrees of freedom. Based on this information, how could you calculate the p-value for the test using the `pchisq` function in R?

4. Is a small or a large value of $\bar{x}$ stronger evidence against the null hypothesis? Justify your answer in a sentence or two.

Problem III: Facts about likelihood ratio tests

True or False? No need to justify your answer.

1. The likelihood ratio statistic W is always less than or equal to 1.

2. The p-value of a test is the probability that the null hypothesis is correct

3. In testing a simple versus simple hypothesis via the likelihood ratio, the p-value equals the likelihood ratio.