STAT 510: Homework 10

David Dalpiaz

Due: Monday, December 7, 11:59 PM

General Directions

This assignment is worth 10 points with the potential to obtain one buffer point. For each exercise, you may obtain a score of 0, 0.5, or 1.

- To obtain a score of **1**, your answer must be correct, contain valid supporting work, and be reasonably formatted up to and including boxing your answer when possible.
- A score of **0.5** will be given to solutions which show reasonabe effort, but contain errors. (A score of **1** may be granted to some solutions containing errors if they are extremely minor.)
- A score of **0** will be given to a blank solution or a solution that shows no reasonable progress towards the correct solution. Note that if you do not indicate a page for a problem on Gradescope, it will be considered blank.

Please submit your assignment to Gradescope by the due date listed above. You may submit up to 48 hours late with a two point late penalty. After that, no late work will be accepted.

Any grade disputes must be petitioned through Gradescope within one week of receiving a grade.

Please attempt to submit your work as a single PDF and complete the process of indicating which problem is on which page. You may need to merge together PDF files from various sources and scans. We will keep track of best practice for submitting to Gradescope in this Piazza thread.

Homework assignments are meant to be learning experiences. You may discuss the exercises with other students, but you must write the solutions on your own. Directly sharing or copying any part of a homework solution is an infraction of the University's rules on academic integrity. Any violation will be punished as severely as possible.

For this, and all homework assignments, you may use any computational tools that you wish, such as a statistical computing enviroment or integral solver. The course staff is most familiar with `R`, so we will be able to best support `R` users, but you may use any software that you like.

Graded Exercises

Exercise 1 (Normal Means MLE)

Consider $X_1, \ldots, X_k \sim N(\theta_i, \sigma_i^2)$. That is, each observation is drawn independently from a normal distribution with potentially different means and variances. Assume the variances are known.

Define $\theta = (\theta_1, \ldots, \theta_k)$.

- Find the MLE for θ, $\hat{\theta}$.
- Find $\mathbb{E}\left[\hat{\theta}\right]$.
- Find $\mathbb{V}\left[\hat{\theta}\right]$. Also note what this result simplifies to when $\sigma_1 = \ldots = \sigma_k = 1$.

Exercise 2 (Estimating a Variance with One Observation)

Consider a single observation, $X \sim N(0, \sigma^2)$.

- Find an unbiased estimator of σ^2.
- Find the MLE of σ.

Exercise 3 (Inverse Guassian MLE)

Let $X_1, \ldots, X_n$ be a random sample from the inverse Gaussian distribution

$$f(x; \mu, \lambda) = \left(\frac{\lambda}{2\pi x^3}\right)^{\frac{1}{2}} \exp\left(-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right), \quad x > 0.$$

Find the MLE of μ and λ.

Exercise 4 (A Regression MLE)

Consider $Y_1, \ldots, Y_n$ such that

$$Y_i = \beta x_i + \epsilon_i$$

where

- the x_i are fixed, known constants
- $\epsilon_i \sim N(0, \sigma^2)$
- σ^2 is unknown.

Find the MLE of β as well as its mean and variance.

Exercise 5 (Beta-Geometric Model)

Assume:

- Likelihood: $X_1, \ldots, X_n \sim \text{Geometric}(p)$
- Prior: $p \sim \text{Beta}(\alpha, \beta)$

Find the posterior mean of $p \mid X_1, \ldots, X_n$, that is, the Bayes estimator of p under squared error loss.

Exercise 6 (A Simple LRT)

Suppose $X_1, \ldots, X_n \sim N(\mu, \sigma^2 = 2)$. Derive the likelihood ratio test for

$$H_0 : \mu = 10 \quad \text{versus} \quad H_1 : \mu \neq 10.$$

Use the data stored below in `norm_data` to carry out the test by calculating an approximate p-value using the large sample properties of the likelihood ratio test statistic.

```
set.seed(42)
norm_data = rnorm(n = 100, mean = 10.4, sd = sqrt(2))
```

Exercise 7 (A LRT for Two Proportions)

Suppose $X_1, \ldots, X_{n_x} \sim \text{Bernoulli}(p_x)$ and $Y_1, \ldots, Y_{n_y} \sim \text{Bernoulli}(p_y)$. Derive the likelihood ratio test for

$$H_0 : p_x = p_y \quad \text{versus} \quad H_1 : p_x \neq p_y.$$

Assuming $n_x = 20$, $n_y = 30$, and $p_x = p_y = 0.3$, repeatedly simulate from this setup and for each simulation:

- Calculate the likelihood ratio test statistic.
- Calculate the value of the usual "textbook" test statistic where $\hat{p}$ is the pooled estimate of the proportion.

$$z = \frac{\hat{p}_x - \hat{p}_y}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}}$$

Using the results of these simulations:

- Plot a histogram of the calculated likelihood ratio test statistics and overlay the approximate distribution of the test statistic under the null hypothesis.
- Create a scatter plot of the likelihood ratio versus the textbook test statistics. What do you notice?

Exercise 8 (An ANOVA Adjacent LRT)

Suppose

- $X_1, \ldots, X_{n_x} \sim N(\mu_x, \sigma_x^2)$.
- $Y_1, \ldots, Y_{n_y} \sim N(\mu_y, \sigma_y^2)$.
- $Z_1, \ldots, Z_{n_z} \sim N(\mu_z, \sigma_z^2)$.

Derive the likelihood ratio test for $H_0 : \sigma_x^2 = \sigma_y^2 = \sigma_z^2$ versus an alternative that allows for at least one unequal variance.

Use the data stored below in the vectors `x`, `y`, and `z` to carry out the test by calculating an approximate p-value using the large sample properties of the likelihood ratio test statistic. (Note that this data is **not** tidy, but is instead stored in a format that is easy to understand.) Does the result match your expectation?

```
set.seed(42)
x = rnorm(n = 50, mean = -5, sd = 1)
y = rnorm(n = 60, mean =  0, sd = 1)
z = rnorm(n = 70, mean =  5, sd = 1)
```

Exercise 9 (Free Points)

Draw a smiley face!

Exercise 10 (Free Points)

Draw a smiley face!

Exercise 11 (Free Points)

Draw a smiley face!

Exercise 12 (Bayes Risk in the Beta-Bernoulli Model)

Suppose $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ and $p \sim \text{Beta}(\alpha, \beta)$. Using squared error loss, find the Bayes estimator and the Bayes risk.

Exercise 13 (The James-Stein Estimator)

Consider $X_1, \ldots, X_k \sim N(\theta_i, 1)$. Define $\theta = (\theta_1, \ldots, \theta_k)$. Consider the loss

$$L\left(\theta, \hat{\theta}\right) = \sum_{i=1}^{k} (\theta - \hat{\theta})^2.$$

where $\hat{\theta}$ is some estimator of θ.

Use simulation to compare the risk of the MLE to the James-Stein estimator. Consider at least three simulation setups:

- $k = 2$
- a relative "large" k and a dense θ vector
- a relative "large" k and a sparse θ vector

You are free to further specify k and θ as you wish. You are also free to add additional setups. Summarize your findings.