Midterm 2 review problems

Note: no practice problems on p-values or LRT are provided because those topics were covered on most recent problem set!

1. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Unif}[\theta, \theta+1]$, where $\theta > 0$ is the unknown parameter. Suppose we'd like to test the following hypotheses:

$$H_0 : \theta = 0 \qquad \text{vs.} \qquad H_1 : \theta > 0$$

   Our rejection rule is to reject H_0 if $Y_n > 1$ *or* $Y_1 > k$, for some constant k. Here, $Y_n = \max\{X_1, \ldots, X_n\}$ and $Y_1 = \min\{X_1, \ldots, X_n\}$.

   Suppose we'd like our test to have size α. What value k should we choose?

2. Let $X_1, \ldots, X_n$ be a sample from $\text{Unif}[\theta - 0.5, \theta + 0.5]$ with θ unknown, and let $X = \sum_{i=1}^n X_i$.

   (a) Is the random variable $V = X - n\theta$ pivotal? Why or why not?

   (b) Find a function $r(v, \mathbf{x})$ for which $r(V, \mathbf{X}) = \theta$.

   (c) Suppose $Y_1, \ldots, Y_n$ are iid $\text{Unif}[-0.5, 0.5]$, let $Y = \sum Y_i$, and let F be the CDF for Y. Use parts (a) and (b) to find a formula for a γ-level confidence interval for θ in terms of F. *Note: Y is not a named distribution that we've previously studied.*

   (d) Use R to approximate $F^{-1}(0.025)$ and $F^{-1}(0.975)$ by simulating 10,000 samples from $\text{Unif}[-0.5, 0.5]$.

   (e) Suppose $X = 25$ and $n = 50$. Find the endpoints of the observed 0.95-level confidence interval for θ.

3. Let $X_1, \ldots, X_n | \theta \overset{\text{iid}}{\sim} \text{Unif}[0, \theta]$, where $\theta > 0$ unknown. On a previous homework, you obtained an interval estimator for θ based on the pivot $U = \frac{Y}{\theta} \in [0, 1]$, where $Y = \max\{X_1, \ldots, X_n\}$. On that homework, you found the following:

$$F_U(u) = \begin{cases} 0 & \text{if } u < 0 \\ u^n & \text{if } 0 \le u \le 1 \\ 1 & \text{if } u > 1 \end{cases}$$

   I previously asked you to find a γ-coefficient confidence interval whose bounds are strictly greater than Y.

   (a) Now, I'd like you construct a different γ-coefficient interval estimator for θ of the following form: $[Y, B(\mathbf{X})]$. Find $B(\mathbf{X})$.

   (b) For fixed γ, comment on if and how the length of your interval in (a) changes as $n \to \infty$. Is this behavior good or bad, and why?

4. Suppose we have a single $X|\theta \sim \text{Unif}[0, \theta]$. Consider a test of the following two simple hypotheses:

$$H_0 : \theta = 1 \qquad \text{vs.} \qquad H_1 : \theta = 2$$

(i.e. $\Omega = \{1, 2\}$).

   (a) Find a test that is level $\alpha_0 = 0$.

   (b) Find the power of your test from (a). *Note: we want the* power*, not power function.*

5. A two stage clinical trial is planned to understand the efficacy of a certain protocol treatment. Let p be the proportion of patients in the treatment who respond to the treatment. Consider testing the following hypotheses:

$$H_0 : p = 0.10 \qquad \text{vs.} \qquad H_1 : p > 0.10$$

The two stage trial proceeds as follows:

- At the first stage, 10 patients are recruited and given the treatment. If 2 or more of these 10 patients are found to respond to the treatment protocol, then H_0 is rejected and the study is terminated.
- If the study was not terminated, we proceed to a second stage: an additional 10 patients are recruited and treated in the second stage. If a total of 4 or more patients among all 20 total patients respond to the treatment, then H_0 is rejected. No matter the results at this stage, the study is terminated.

For example: if 5 patients in the first stage are found to respond to the treatment, H_0 is rejected and the study is finished. However, if only 1 patient was responded in the first stage, we recruit another 10 patients. Suppose 1 additional person in stage two responded to the treatment. Since $2 < 4$, we fail to reject H_0. The study is terminated.

   (a) Find the probability that H_0 is rejected at the first stage, assuming the null is true. State your answer exactly (using PMFs), and then use `R` to obtain an approximate probability.

   (b) Find the overall probability that H_0 is rejected, assuming the null is true. State your answer exactly (using PMFs), and then use `R` to obtain an approximate probability.

6. Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ where both μ and σ^2 are unknown. How should c be chosen such that the interval $(-\infty, \bar{X} + c]$ is a 95% confidence interval for μ?

7. Suppose that an iid sample of $n = 15$ observations from a Normal distribution with unknown mean μ and variance σ^2 yields $\bar{X} = 10$ and $s^2 = 25$. Find 90% confidence intervals for μ and σ^2.

8. An investigator is planning a study. They have data that are assumed to be Normally distributed with unknown mean μ and known variance $\sigma^2 = 2.5$. The null hypothesis is $H_0 : \mu = 3$ and alternate hypothesis is $H_1 : \mu \neq 3$. The investigator has budget to obtain $n = 10$ samples.

   (a) Provide and also sketch the distribution of the sampling distribution of $\bar{X}$ under the null hypothesis. Be sure to label the axes and add some x-axis values.

   (b) For a level 0.05 test, for what values of the test statistic $r(\mathbf{X}) = \bar{X}$ will the investigator reject the null hypothesis?

9. Two college students collected data on the price of hardcover textbooks from two disciplinary areas: Mathematics and the Natural Sciences, and the Social Sciences. The data can be loaded into R by running the following code (Don't worry about interpreting what the code itself is doing).

```
bookprices <- read.csv("https://people.carleton.edu/~kstclair/data/BookPrices.csv")
books_ss <- subset(bookprices, Area == "Social Sciences")$Price
books_mns <- subset(bookprices, Area == "Math & Science")$Price
```

   In particular, the vector `books_ss` contains a list of prices for Social Science texts, and the vector `books_mns` contains a list of prices for Math and Science texts. Let $\bar{x}_{ss}$ denote the sample mean price of social science texts and let $\bar{x}_{mns}$ denote the sample mean price of Math and Science texts.

   (a) Compute $\bar{x}_{ss}$ and $\bar{x}_{mns}$. Then compute the ratio $\frac{\bar{x}_{ss}}{\bar{x}_{mns}}$.

   (b) Use bootstrapping to simulate 10^4 sample means from the sample of Social Science textbooks, and 10^4 sample means from the sample of Math and Natural Sciences textbooks. Visualize the approximate bootstrap distributions using histograms.

   (c) Use the bootstrap statistics in the previous part to create 10^4 bootstrap statistics for the ratio of mean prices (social science / math and natural science). Create a histogram of the approximate bootstrap distribution.

   (d) Create a 95% *empirical* bootstrap interval for the ratio of the means. What does this interval suggest about the true ratio?

   (e) Use your approximate bootstrap distribution to estimate the standard deviation and the bias of $\dfrac{\bar{x}_{ss}}{\bar{x}_{mns}}$ as an estimator for the true ratio of mean prices. Approximately what proportion of the mean squared error of $\dfrac{\bar{x}_{ss}}{\bar{x}_{mns}}$ is due to bias?