MAT/STA 336 **Midterm 2 Review** **Spring 2023**

The first midterm exam will be a take-home exam which will be made available in the Midterm 2 folder under the documents section of PWeb at 9am on Wednesday, April 26th and due at 11:59pm (uploaded to Gradescope) on Wednesda, May 3rd.

Content. The exam will be lightly cumulative, but with emphasis on the material covered since the first midterm. In particular, it will focus on Chapter 8 (8.4, 8.5), Chapter 9 (9.1, 9.5, 9.6, 9.7) and Section 12.6 of DeGroot and Schervish's *Probability and Statistics.* There will be some questions that ask you to use R.

Format. The exam is intended to take 3 hours to complete, although you may take up to 5 hours to complete it. These 5 hours do not need to be consecutive. You should monitor your own time, and record on the test your estimate for the total amount of time you actively worked on the exam.

Your solutions to the exam should be neatly neatly written or typed. If you scan a handwritten assignment, be sure to review the legibility of your scan on Gradescope after you submit.

Resources. You may use any notes you've taken for this class, your work on any previous homework or daily assignments, lecture notes I've posted on the course website, the recorded lecture video from Monday 2/20 and DeGroot and Schervish's *Probability and Statistics* textbook, as well as Blitzstein's *Introduction to Probability* textbook.

For problems asking you to do analysis or perform computations using R, you may use either a local installation of R or the Grinnell R Studio server, and you may reference any of the R help files (available by typing `?functionname` in the console).

You may not use any other resources other than those listed above. If you have questions about whether a resource can be used, you are welcome to message me.

Preparation. The best preparation you can do for the exam is to organize your notes and/or homework to make finding information and examples as quick and efficient as possible. Beyond that, you should attempt to accurately assess what topics you have mastered and which you need to practice more. A good starting point is to review the list of objectives on each daily assignment. Another way to prepare is to create your own study guide with summaries of the important concepts, along with example problems you've designed and solved. Exam problems will be comparable in difficulty to those exhibited in class and assigned for homework. Some exam questions may be similar to problems you have seen before, while others will require you to synthesize your knowledge in new ways.

On the exam, you may be asked to do the following:

- Rephrase a key definition and/or theorem in your own words.
- Determine whether a given statement is true or false.
- Interpret or explain a statistics concept in everyday language.
- Sketch the proof of an important result discussed in class.
- Perform calculations using relevant techniques from the course.
- Provide a short, rigorous proof of a novel statement or result.
- Create and analyze a statistical model for a particular phenomenon.
- Use R to simulate a random phenomenon.

For extra practice, several additional review problems are printed below. Solutions to these problems can be found on the exams page of the course website. While these questions are representative of the typical scope and difficulty of individual exam questions, this review is not comprehensive, nor does it necessarily represent the total amount of time available for the exam.

Practice Problems.

(1) Suppose $X_1, \ldots, X_n$ are a sample from $N(\theta, \sigma^2)$, where θ is unknown but σ^2 is known.

(a) Construct the shortest-length 0.95 confidence interval for θ.

(b) Suppose that θ has a prior distribution $N(\mu, \nu^2)$. Find the shortest length 0.95 posterior credible interval for θ.

(c) Show that as $\nu^2 \to \infty$, the interval in part (b) converges to the interval in part (a).

(d) Explain what this suggests about the relationship between the frequentist confidence interval and the bayesian credible interval, using the concept of *improper priors*.

Solution. (a) We'll use the MLE $\bar{X}$ as our estimator for θ. Since the sampling distribution for $\bar{X}$ is $N(\mu, \sigma^2/n)$, which is symmetric, then the shortest length confidence interval is the one that assigns equal area to each tail. Let c be the 0.975 quantile of the standard Normal distribution. By Exercise 8.5.1, a 0.95-level confidence interval is of the form

$$\left(\bar{X} - c\frac{\sigma}{\sqrt{n}}, \bar{X} + c\frac{\sigma}{\sqrt{n}}\right)$$

(b) By Theorem 7.3.3, the Normal distribution $N(\mu, \nu^2)$ is the conjugate prior for θ, when samples are taken from $N(\theta, \sigma^2)$ with known σ^2. Moreover, the posterior distribution of $\theta|\mathbf{X}$ is $N(\mu_1, \nu_1^2)$ with

$$\mu_1 = \frac{\sigma^2\mu + n\nu^2\bar{x}}{\sigma^2 + n\nu^2} \qquad \nu_1^2 = \frac{\sigma^2\nu^2}{\sigma^2 + n\nu^2}$$

The 0.95 credible interval of smallest length is the one assigning equal area to the two tails. Let c_1 and c_2 be the 0.05 and 0.975 quantiles of the $N(\mu_1, \nu_1^2)$ distribution, giving a 0.9 credible interval of (c_1, c_2).

(c) By Theorem 7.3.1 and 7.3.2, $\mu_1 \to \bar{x}$ and $\nu_1^2 \to \sigma^2/n$, as $\nu^2 \to \infty$ (this can also be verified directly form the formula in the previous part). Therefore, the posterior distribution $N(\mu_1, \nu_1^2)$ converges to $N(\bar{x}, \frac{\sigma^2}{n})$ as $\nu^2 \to \infty$. In this case, c_1 and c_2 converge to c_1' and c_2', the 0.025 and 0.975 quantiles of $N(\bar{x}, \frac{\sigma^2}{n})$. Now, let F be the CDF of $N(\bar{x}, \frac{\sigma^2}{n})$ and let Φ be the CDF of $N(0,1)$. Then by the location-sclae properties of the Normal distribution

$$F(x) = \Phi\left(\frac{x - \bar{x}}{\sigma/\sqrt{n}}\right)$$

Hence, if c_1' and c_2' are the 0.025 and 0.975 quantile of Φ, then solving $c_i' = \frac{x_i - \bar{x}}{\sigma/\sqrt{n}}$ for x_i are the 0.025 and 0.975 quantiles for F:

$$x_1' = \bar{x} + c_1\frac{\sigma}{\sqrt{n}} \qquad x_2' = \bar{x} + c_2\frac{\sigma}{\sqrt{n}}$$

as desired.

(d) The previous problem shows that we can obtain the frequentist confidence interval as the limit of the Bayesian posterior credible interval, using an improper prior. Therefore, we can think about the frequentist interval as a credible interval, but with a prior reflecting maximal amount of ignorance.

(2) Let $X_1, \ldots, X_n$ be a sample from $\text{Unif}(\theta - 0.5, \theta + 0.5)$ with θ unknown, and let $X = \sum X_i$

(a) Explain why the random variable $V = X - n\theta$ is pivotal.

(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.

Solution. (a) By expanding $X - n\theta$ as a sum, we see that

$$X - n\theta = \sum (X_i - \theta)$$

and each $(X_i - \theta)$ is distributed as $\text{Unif}(-0.5, 0.5)$. Hence, as $X - n\theta$ is a sum of variables, none of whose distributions depend on θ, then the distribution of $X - n\theta$ also does not depend on θ.

(b) Let $x = \sum x_i$ and define $r(v, \mathbf{x}) = \frac{x-v}{n}$. Then

$$r(V, \mathbf{X}) = \frac{X - V}{n} = \frac{X - (X - n\theta)}{n} = \theta$$

(c) Let $c_1 = F^{-1}\left(\frac{1-\gamma}{2}\right)$ and $c_2 = F^{-1}\left(\frac{1+\gamma}{2}\right)$. Then, as r is a decreasing function of v (and hence, reverses inequalities),

$$\gamma = P(c_1 < V < c_2) = P(r(c_1, \mathbf{X}) > r(V, \mathbf{X}) > r(c_2, \mathbf{X})) = P\left(\frac{X - c_2}{n} < \theta < \frac{X - c_1}{n}\right)$$

showing that $A = \frac{X-c_2}{n}$ and $B = \frac{X-c_1}{n}$ are a γ-level confidence interval for θ.

(d) The following code runs 10,000 simulations and computes the 0.025 and 0.975 quantiles of the distribution (when $n = 50$).

```
set.seed(1002)
trials <- 10000
y <- rep(0,trials)
for (i in 1:trials){
 y[i] = sum(runif(50, -0.5, 0.5))
 }
 quantile(y, c(0.025, 0.975))
 ## -4.016697  4.002452
```

which we round to -4 and 4.

(e) With $X = 25$ and $n = 50$, the .95-level confidence interval is

$$\left(\frac{25-4}{50}, \frac{25+4}{50}\right) \quad = \quad (0.42, 0.58)$$

(3) 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% bootstrap percentile 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 $\frac{\bar{x}_{ss}}{\bar{x}_{mns}}$ as an estimator for the true ratio of mean prices. Approximately what proportion of the mean squared error of $\frac{\bar{x}_{ss}}{\bar{x}_{mns}}$ is due to bias?

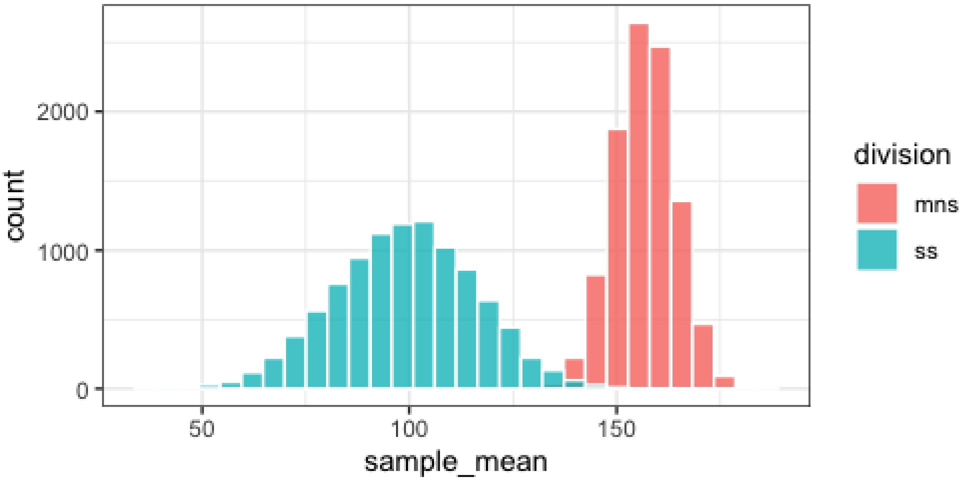
Solution. (a) The following code computes $\bar{x}_{ss}$ and $\bar{x}_{mns}$, as well as $\frac{\bar{x}_{ss}}{\bar{x}_{mns}}$:

```
xbar_ss <- mean(books_ss)
xbar_mns <- mean(books_mns)
ratio <- xbar_ss/xbar_mns
## 98.99  156.73   0.6315
```

(b) The following code creates the 10^4 bootstrap statistics from each sample:

```
trials <- 10^4
boot_ss <- rep(0, trials)
boot_mns <- rep(0, trials)
for (i in 1:trials){
  boot_ss[i] <- mean(sample(books_ss, size = length(books_ss), replace = T))
  boot_mns[i] <- mean(sample(books_mns, size = length(books_mns), replace = T))
  }
```

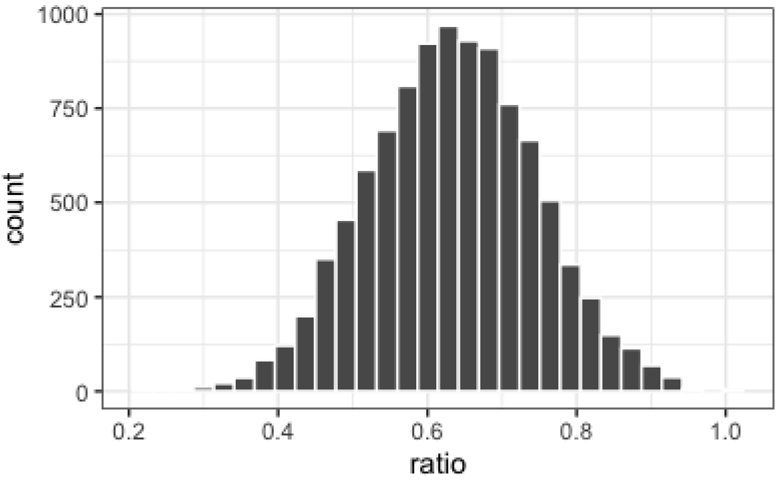
Histograms of the two bootstrap distributions are shown below:



(c) To create a bootstrap ratio statistics, we take the ratios of the bootstrap means for each (independent) trial:

```
boot_ratio <- boot_ss / boot_mns
```

The histogram of the distribution is shown below:



(d) The 95% bootstrap percentile interval is obtain from the 0.025 and 0.975 quantiles of the bootstrap distribution:

```
quantile(boot_ratio, c(0.025, 0.975))
## 0.4179 0.8560
```

Since this interval gives a range of plausible values for the true ratio, we estimate that social science texts cost between 42% and 86% less than mathematical and natural science texts.

(e) The bias estimate is the difference between the mean of the bootstrap statistic and the observed statistic,

```
mean(boot_ratio) - ratio
# 0.0025
```

while the standard deviation estimate is the standard deviation of the bootstrap statistic:

```
sd(boot_ratio)
# 0.11
```

The estimated mean squared error is the sum of squares of the bias and standard deviation estimates:

```
(mean(boot_ratio) - ratio)^2 + (sd(boot_ratio))^2
## 0.01
```

Of this 0.01 MSE, only $\frac{\text{bias}^2}{\text{MSE}} = 0.0005$ is due to the bias.

(4) Suppose that a single observation X is obtained from $\text{Unif}(0, \theta)$ with θ unknown. Consider hypotheses

$$H_0 : \theta = 1 \qquad H_1 : \theta = 2$$

(a) Specify a test of the above hypotheses with size of 0, but also power of 0.

(b) Specify a test of the above hypotheses with size of 0, but with power strictly greater than 0.

(c) For the same hypotheses, consider a procedure which rejects when $X \geq 0.5$. What is the size and power of this test?

(d) For the same hypotheses, specify a size 0.05 test with power strictly greater than 0.5.

Solution. Throughout this problem, since both null and alternative hypotheses are simple, then the size of a test δ is the value of the power function $\pi(\theta|\delta)$ on $\theta = 1$, while the power of the test is the value of the power function $\pi(\theta|\delta)$ on $\theta = 1$.

(a) Let δ be the test that never rejects H_0. Hence $\pi(\theta = 1|\delta) = P(\text{reject} H_0|\theta = 1) = 0$ and similarly, $\pi(\theta = 2|\delta) = P(\text{reject} H_0|\theta = 2) = 0$. This shows that the test has both size and power of 0.

(b) Now, consider the test δ which rejects when $X > 1$. Then

$$\pi(\theta = 1|\delta) = P(\text{reject} H_0|\theta = 1) = P(X > 0.5|\theta = 1) = \frac{1}{2}$$

while

$$\pi(\theta = 2|\delta) = P(\text{reject} H_0|\theta = 2) = P(X > 0.5|\theta = 2) = \frac{1}{4}$$

showing that this test has size 1/2 and power 1/4.

(c) In this case,

$$\pi(\theta = 1|\delta) = P(\text{reject} H_0|\theta = 1) = P(X > 1|\theta = 1) = 0$$

while

$$\pi(\theta = 2|\delta) = P(\text{reject} H_0|\theta = 2) = P(X > 1|\theta = 2) = \frac{1}{2}$$

(d) Let δ be the test which rejects H_0 when $X \geq 0.95$. Then,

$$\pi(\theta = 1|\delta) = P(\text{reject} H_0|\theta = 1) = P(X > 0.95|\theta = 1) = 0.05$$

while

$$\pi(\theta = 2|\delta) = P(\text{reject} H_0|\theta = 2) = P(X > 0.95|\theta = 2) = \frac{1.05}{2} = 0.525$$

showing that this test has size 0.05 and power 0.525.

(5) Let $\mathbf{X}$ be a random sample from a distribution with unknown parameter θ, and suppose that for each value θ_0 of θ, and each number $0 \leq \alpha_0 \leq 1$, there exists a level α_0 test procedure δ_{θ_0} of the hypotheses

$$H_0 : \theta \geq \theta_0 \qquad H_1 : \theta < \theta_0$$

For each possible value $\mathbf{x}$ of the $\mathbf{X}$, define a set $\omega(\mathbf{x})$ by

$$\omega(\mathbf{x}) = \{\theta_0 : \delta_{\theta_0} \text{ rejects } H_0 \text{ when } \mathbf{x} \text{ is observed}\}$$

(a) Prove that $\omega(\mathbf{X})$ is a $1 - \alpha_0$ confidence set for θ.

(b) Suppose that $X_1, \ldots, X_n \sim N(\theta, 1)$, and consider procedures $\{\delta_{\theta_0}\}$ which reject H_0 when $\bar{X} \leq c_{\theta_0}$, where c_{θ_0} is chosen so that δ_{θ_0} is a size α_0 test. Give an explicit description of $\omega(\mathbf{x})$ as an interval; i.e. explain why $\omega(\mathbf{x})$ is an interval, rather than some other type of set, and specify the random variable(s) that determine the endpoints of the interval.

Solution. (a) For fixed value of θ, δ_θ is a α_0-level test, by construction. Then

$$\begin{aligned} P(\theta \in \omega(\mathbf{X})) =& P(\delta_\theta \text{ does not reject } H_0 \text{ when } \mathbf{X} \text{ is observed}) \\ =& 1 - P(\delta_\theta \text{ does reject } H_0 \text{ when } \mathbf{X} \text{ is observed}) \\ \geq 1 - \alpha_0 \end{aligned}$$

Hence, $\omega(\mathbf{X})$ is indeed a $1 - \alpha_0$ confidence set.

(b) We first calculate the power function

$$\pi(\theta|\delta_{\theta_0}) = P(\bar{X} \leq c_{\theta_0}|\theta) = P\left(\frac{\bar{X} - \theta}{1/\sqrt{n}} \leq \frac{c_{\theta_0} - \theta}{1/\sqrt{n}}|\theta\right) = \Phi\left(\frac{c_{\theta_0} - \theta}{1/\sqrt{n}}\right)$$

which is a decreasing function of θ on H_0. Hence, the size δ_{θ_0} is obtained for $\theta = \theta_0$. If δ_{θ_0} is to be an α_0-sized test, then

$$c_{\theta_0} = n^{-1/2}\Phi^{-1}(\alpha_0) + \theta_0$$

Observe now that if $\theta_0 < \theta_0'$, then $c_{\theta_0} < c_{\theta_0'}$. If $\delta_{\theta_0'}$ does not reject when $\mathbf{X} = \mathbf{x}$ is observed, then $\mathbf{x} > c_{\theta_0'}$. Thus, $\mathbf{x} > c_{\theta_0}$ and so δ_{θ_0} also does not reject when $\mathbf{X} = \mathbf{x}$ is observed. It follows that if $\theta_0' \in \omega(X)$, then $\theta_0 \in \omega(X)$, which shows that $\omega(\mathbf{X})$ is an interval.

On the other hand, if

$$\bar{X} = c_{\theta_0} = n^{-1/2}\Phi^{-1}(\alpha_0) + \theta_0$$

then δ_{θ_0} rejects H_0. As does every test with $\theta_0' < \theta_0$. Hence, $\bar{X} - n^{-1/2}\Phi^{-1}(\alpha_0)$ is the upper endpoint of the interval.