Stat 343 Final Review Problems

Final Structure and Coverage

I haven't written the final yet, but it will likely have 3 kinds of questions:

1) conceptual questions to make sure you understand what's going on.
2) some problems where you're asked to do some math.
3) a computational question.

Conceptual Topics

- All the things from the midterm
- Things about likelihood ratios, inverting confidence intervals/hypothesis tests, and p-values

For examples, see the multiple choice problems on the most recent homework assignment.

Here's one more.

Problem 1

If n = 1,000,000,000, is there an unbiased estimator with lower variance than the MLE? Justify your answer.

Solution: The answer I had in mind was no. For large n, if all the regularity conditions are satisfied, the variance of the MLE is the inverse of the Fisher information. The Cramer-Rao Lower Bound states that the variance of any unbiased estimator must be at least as large as the inverse of the Fisher information. Therefore, for large n, any unbiased estimator must have variance at least as large as the variance of the MLE.

There are two places where this argument is a little shakey, so that you could argue the answer is "maybe". First, the result about the variance of the MLE is really asymptotic - the variance of the MLE approaches the inverse of the Fisher information as the sample size goes to infinity. For any finite sample size, even 1,000,000,000, the variance of the MLE could be larger than the inverse of the Fisher information. Second, we could be working with a probability distribution where the regularity conditions are not satisfied, in which case the theorems do not apply.

Example Worked Problems

The midterm will have problems roughly similar in content to the examples below.

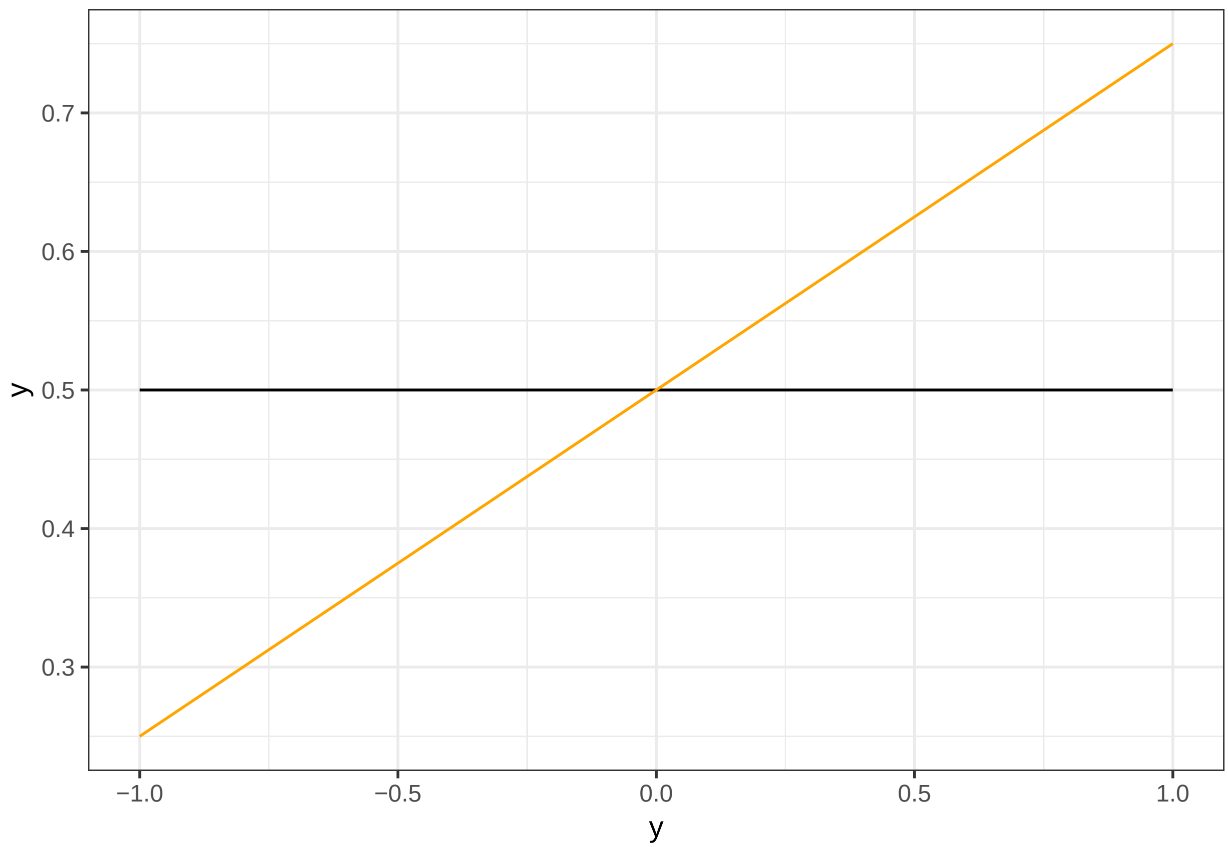
Problem 1

Suppose that $Y_1, Y_2, \dots, Y_n$ are i.i.d. samples from a distribution with pdf given by $f_Y(y|\theta) = \frac{1+\theta y}{2}$ for $-1 < y < 1$ and $-1 < \theta < 1$. An investigator is interested in testing $H_0 : \theta = 0$ vs. $H_A : \theta = 0.5$.

(a) Plot a picture of the pdfs of Y_1 under the null and alternative hypotheses. (One plot, two lines).

```
library(ggplot2)
f <- function(y, theta) {
  (1 + theta * y) / 2
}

ggplot(data = data.frame(y = c(-1, 1)), mapping = aes(x = y)) +
  stat_function(fun = f, args = list(theta = 0)) +
  stat_function(fun = f, args = list(theta = 0.5), color = "orange") +
  theme_bw()
```



(b) Consider a sample of size n. Derive a general structure of the rejection region of the most powerful test of size $\alpha = 0.05$ for testing $H_0 : \theta = 0$ vs. $H_A : \theta = 0.5$. By "general structure", I mean that you should find the test statistic and indicate in general what the rejection region of the test looks like, but you do not need to derive the exact critical value for the test.

The likelihood ratio statistic is

$$
\begin{aligned}
W &= \frac{\mathcal{L}(0|Y_1, \ldots, Y_n)}{\mathcal{L}(0.5|Y_1, \ldots, Y_n)} \\
&= \frac{0.5^n}{\prod_{i=1}^n \{(1 + 0.5 \cdot Y_i)/2\}}
\end{aligned}
$$

We will reject the null hypothesis if the observed value of this statistic is less than some critical value c:

$\frac{0.5^n}{\prod_{i=1}^n \{(1+0.5 \cdot y_i)/2\}} < c$

(c) Suppose now that $n = 1$. On your plot from part (a), show how you could calculate the p-value of the test if you observe $y = 0.5$.

The p-value is the probability of obtaining a test statistic at least as small as the observed value of the test statistic, given that the null hypothesis is true. In this case, that is calculated as

$$
\begin{aligned}
p - value &= P\left(\frac{0.5}{\{(1 + 0.5 \cdot Y)/2\}} \leq \frac{0.5}{\{(1 + 0.5 \cdot 0.5)/2\}} |\theta = 0\right) \\
&= P\left(1.25 \leq 1 + 0.5Y|\theta = 0\right) \\
&= P\left(Y \geq 0.5|\theta = 0\right) \\
&= \int_{0.5}^1 0.5 dy \\
&= 0.5(1 - 0.5) \\
&= 0.25
\end{aligned}
$$

(d) Still working with $n = 1$, find the rejection region of the test.

We will reject for values of y at least as large as the critical value y^* at which the p-value is 0.05.

$$
\begin{aligned}
0.05 &= P\left(Y \geq y^*|\theta = 0\right) \\
&= \int_{y^*}^1 0.5 dy \\
&= 0.5(1 - y^*)
\end{aligned}
$$

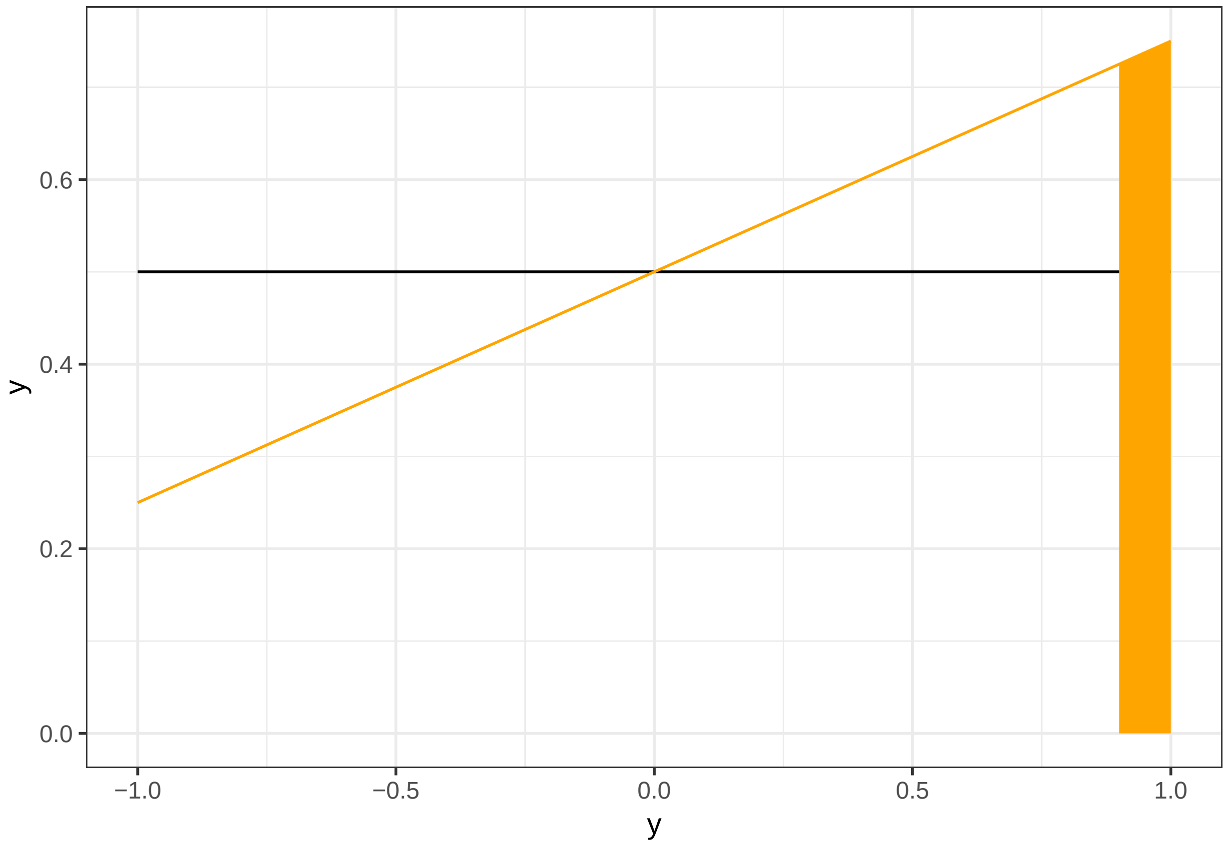
Solving for y^*, we obtain $y^* = 0.9$.

(e) Still working with $n = 1$, indicate on your picture the area corresponding to the power of the test.

```
library(ggplot2)
f <- function(y, theta) {
  (1 + theta * y) / 2
}

poly <- data.frame(
  y = c(0.9, 1, 1, 0.9, 0.9),
  f = c(0, 0, f(1, theta = 0.5), f(0.9, theta = 0.5), 0)
)

ggplot(data = data.frame(y = c(-1, 1)), mapping = aes(x = y)) +
  stat_function(fun = f, args = list(theta = 0)) +
  stat_function(fun = f, args = list(theta = 0.5), color = "orange") +
  geom_polygon(data = poly, mapping = aes(x = y, y = f), fill = "orange") +
  theme_bw()
```



Problem 2

In the written part of Lab 13, you showed that if $X \sim \text{Binomial}(n, \theta)$, then the maximum likelihood estimator $\hat{\theta} = X/n$ has expected value θ and variance $\theta(1-\theta)/n$. Regarding X as a sum of results of iid Bernoulli trials, find an approximation to the distribution of $\hat{\theta}$ if n is large. Use your approximation to find an approximate 95% confidence interval for θ. How could you use your confidence interval to conduct a test of $H_0 : \theta = 0.2$?

Let $X_i \sim \text{Bernoulli}(\theta)$ denote the outcome of trial number i, so that $X = \sum_{i=1}^n X_i$. By our theorem about the asymptotic approximate distribution of the MLE, $\hat{\theta} \sim \text{Normal}\left(\theta, \frac{1}{I(\hat{\theta})}\right)$, which can be written as $\hat{\theta} \sim \text{Normal}\left(\theta, \frac{\hat{\theta}(1-\hat{\theta})}{n}\right)$.

Denoting the qth quantile of the standard normal distribution by $z(q)$, it therefore follows that

$$
\begin{aligned}
& P\left(z(\alpha/2) \leq \frac{\hat{\theta} - \theta}{\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}} \leq z(1-\alpha/2)\right) \\
\Rightarrow & P\left(z(\alpha/2)\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} \leq \hat{\theta} - \theta \leq z(1-\alpha/2)\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}\right) \\
\Rightarrow & P\left(-\hat{\theta} + z(\alpha/2)\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} \leq -\theta \leq -\hat{\theta} + z(1-\alpha/2)\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}\right) \\
\Rightarrow & P\left(-\hat{\theta} + z(\alpha/2)\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} \leq -\theta \leq -\hat{\theta} + z(1-\alpha/2)\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}\right) \\
\Rightarrow & P\left(\hat{\theta} - z(\alpha/2)\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} \leq \theta \leq \hat{\theta} - z(1-\alpha/2)\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}\right)
\end{aligned}
$$

To conduct the hypothesis test at significance level α, we could use the above method to compute a $(1-\alpha)100\%$ confidence interval for θ. We would reject the null hypothesis if the confidence interval did not include 0.2.

Problem 3

Consider a situation in which we've observed $x_1 = 1$, $x_2 = 2$, and $x_3 = 4$. Calculate the bootstrap distribution of the median. Your answer should be in the form of a table listing possible values of the median and their probabilities. (Hint: there are 27 possibilities, and they are all equally likely). Use the distribution to find a 90% confidence interval for the median based on the bootstrap percentile method. Is it appropriate to use this approach? Why or why not?

The bootstrap distribution of the median consists of the medians of all possible samples of size 3 drawn with replacement from the set $\{1, 2, 4\}$. If you list those out, you will find that there are 27 such samples. For 7 of them, the median is 1: the one sample where 1 was selected three times, three samples where 1 was selected twice and 2 once, and three samples where 1 was selected twice and 4 was selected once. By a similar argument, the median is 4 in 7 of the samples. The median is 2 in the remaining 13 samples. Our bootstrap distribution of the median therefore is:

P(median $= 1) = 7/27$ P(median $= 2) = 13/27$ P(median $= 4) = 7/27$

Noting that $1.35/27 = 0.05$, the 5th percentile of the bootstrap distribution of the median is 1 and the 95th percentile of the bootstrap distribution of the median is 4. Therefore, a 90% bootstrap confidence interval for the median is [1, 4].

Using the bootstrap percentile method would be a terrible idea in this example, because our sample size is too small for the bootstrap distribution to approximate the sampling distribution well. The bootstrap percentile method in particular should not be used with small sample sizes.