Math 341 / 641 Fall 2024
Midterm Examination Two

Professor Adam Kapelner

November 19, 2024

Full Name ____

Code of Academic Integrity

Since the college is an academic community, its fundamental purpose is the pursuit of knowledge. Essential to the success of this educational mission is a commitment to the principles of academic integrity. Every member of the college community is responsible for upholding the highest standards of honesty at all times. Students, as members of the community, are also responsible for adhering to the principles and spirit of the following Code of Academic Integrity.

Activities that have the effect or intention of interfering with education, pursuit of knowledge, or fair evaluation of a student's performance are prohibited. Examples of such activities include but are not limited to the following definitions:

Cheating Using or attempting to use unauthorized assistance, material, or study aids in examinations or other academic work or preventing, or attempting to prevent, another from using authorized assistance, material, or study aids. Example: using an unauthorized cheat sheet in a quiz or exam, altering a graded exam and resubmitting it for a better grade, etc.

I acknowledge and agree to uphold this Code of Academic Integrity.

____ ____________________

signature date

Instructions

This exam is 110 minutes (variable time per question) and closed-book. You are allowed **one** page (front and back) of a "cheat sheet", blank scrap paper (provided by the proctor) and a graphing calculator (which is not your smartphone). Please read the questions carefully. Within each problem, I recommend considering the questions that are easy first and then circling back to evaluate the harder ones. No food is allowed, only drinks.

Problem 1 Consider $m = 1,000$ tests of independence of different single nucleotide polymorphisms in mice. From these tests, we result in m p-values. We wish to let the familywise error-rate (FWER) be 5%. Below are the first 150 p-vals rounded to 6 significant digits sorted from smallest to largest. The bracket numbers on the left are the index of the number to help you count them (e.g. the second line shows [11] which means the first number is the 11th smallest p-value). There are 10 p-vals per line.

```
  [1] 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000001 0.000001

 [11] 0.000005 0.000007 0.000007 0.000010 0.000012 0.000012 0.000019 0.000034 0.000038 0.000120

 [21] 0.000166 0.000190 0.000197 0.000420 0.000507 0.000937 0.004868 0.007945 0.008571 0.013348

 [31] 0.024456 0.024794 0.029519 0.029621 0.033902 0.036871 0.048581 0.053524 0.060118 0.089462

 [41] 0.091529 0.091766 0.103048 0.109418 0.112729 0.113030 0.114637 0.117140 0.126016 0.129215

 [51] 0.135389 0.136046 0.136753 0.141048 0.143251 0.149764 0.153472 0.154205 0.155544 0.157578

 [61] 0.158298 0.159050 0.161296 0.163764 0.164217 0.166613 0.166983 0.170020 0.171261 0.171923

 [71] 0.174144 0.178189 0.179773 0.180444 0.183544 0.185305 0.190947 0.190966 0.192081 0.196738

 [81] 0.196744 0.198668 0.198925 0.199524 0.203801 0.204903 0.206711 0.206848 0.209806 0.210644

 [91] 0.211662 0.217509 0.218852 0.219836 0.220203 0.227602 0.227983 0.231062 0.233572 0.234139

[101] 0.234714 0.235919 0.238669 0.242114 0.243362 0.244109 0.244461 0.244791 0.247373 0.250629

[111] 0.251423 0.251787 0.254297 0.256812 0.258325 0.259333 0.260674 0.262117 0.262640 0.265076

[121] 0.268175 0.270445 0.273475 0.275958 0.275983 0.278991 0.279362 0.279822 0.280783 0.281936

[131] 0.284556 0.287701 0.288057 0.288981 0.289011 0.290727 0.294255 0.295627 0.295918 0.296420

[141] 0.297899 0.298104 0.299034 0.300128 0.300867 0.300935 0.301045 0.301628 0.301719 0.302135
```

(a) [3 pt / 3 pts] Using Bonferroni control of FWER, how many rejections would be made?

The Bonferroni control threshold is FWER$/m = 0.05/1000 = 0.00005$. We just count the number of p-values in the above table from top left to the right and down as they're sorted. We tally $\boxed{19}$ tests.

(b) [3 pt / 6 pts] Using Dunn-Sidak control of FWER, how many rejections would be made?

The Dunn-Sidak control threshold is $1 - (1 - \text{FWER})^{1/m} = 1 - (1 - 0.05)^{1/1000} = 0.0000513$. Thus doesn't change the answer from the previous answer which tallied $\boxed{19}$.

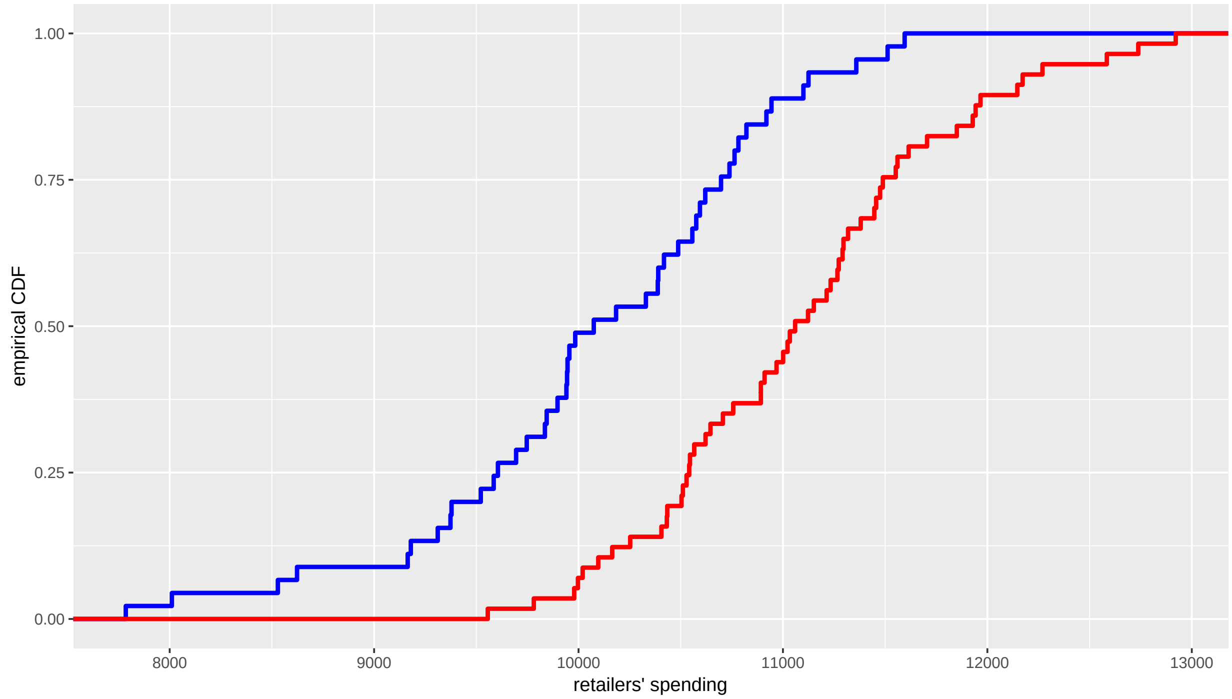
(c) [5 pt / 11 pts] Using Simes control of FWER, how many rejections would be made?

The Simes control threshold is the number of p-values below the first crossing of the "linear setup line" given by j/m (where j is the index of the test) with the line of the p-values. The line thus looks like 0.001, 0.002, 0.003, etc. Following that pattern, we see it crosses after the $\boxed{35}$th sorted p-value.

(d) [6 pt / 17 pts] If instead of controlling FWER at 5%, you are comfortable controlling FDR at 5%, how many expected Type I errors are made in this family of tests?

The FDR control is the same as the Simes control protocol. Thus since we had 35 rejections from the previous problem, we know by definition of FDR, the number of Type I errors is distributed as a Binomial(35, 0.05) whose expectation is $35 \cdot 0.05 = \boxed{1.75}$.

Problem 2 A company wants to analyze the purchasing behavior of large retailers in two different regions: Region 1 and Region 2. The company has data on the total amount spent by retailers in each region over the last year. The company is interested in proving that these retailers are different, so they can divide management resources to better focus on the retailers. In region 1 there are $n_1 = 45$ retailers and in region 2 there are $n_2 = 57$ retailers. Below is a plot of the empirical both $\hat{F}_1(x)$ and $\hat{F}_2(x)$ for retailers in region 1 and retailers in region 2 respectively.



(a) [3 pt / 20 pts] Estimate the value of $\hat{\hat{d}} := \sup_{x \in \mathbb{R}} \left\{ \left| \hat{F}_1(x) - \hat{F}_2(x) \right| \right\}$ as best as possible.

At $x \approx 10000$ the difference appears to be maximal at $\hat{\hat{d}} \approx \boxed{0.4}$.

(b) [5 pt / 25 pts] At $\alpha = 5\%$, make a decision on the suitability of H_0, i.e. that spending in the two regions are realized from the same DGP. Note that $F_K(1.36) = 95\%$ where F_K denotes the CDF of Kolmogorov's distribution.

The test statistic for the 2-sample Kolmogorov-Smirnov test is:

$$\hat{\hat{k}} := \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \hat{\hat{d}} = \sqrt{\frac{45 \cdot 57}{45 + 57}} 0.4 = 2.01$$

where the distance statistic $\hat{\hat{d}}$ was estimated from the plot in the previous question. Since this statistic is larger than the threshold of 1.36 which is given above, we $\boxed{\text{reject } H_0}$.

(c) [1 pt / 26 pts] The test in the previous question is... circle one...
exact / approximate.

Problem 3 Consider the following DGP:

$$X_1, \ldots, X_n \overset{iid}{\sim} \text{Laplace}(0, \theta) := \frac{1}{2\theta} e^{-|x|/\theta}$$

Here are some facts about this DGP:

$$\mathbb{E}[X] = 0, \quad \mathbb{V}\text{ar}[X] = 2\theta^2, \quad \mathbb{E}[|X|] = \theta$$

And below is some of the busy work of the problem done for you:

$$\begin{aligned}
\mathcal{L}(\theta; \boldsymbol{X}) &= \prod_{i=1}^{n} \frac{1}{2\theta} e^{-|X_i|/\theta} = \frac{1}{2^n} \frac{1}{\theta^n} e^{-\frac{1}{\theta}\sum_{i=1}^{n} |X_i|} \\
\ell(\theta; \boldsymbol{X}) &= -n \ln(2) - n \ln(\theta) - \frac{1}{\theta} \sum_{i=1}^{n} |X_i| \\
s(\theta; \boldsymbol{X}) = \ell'(\theta; \boldsymbol{X}) &= -\frac{n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^{n} |X_i| \\
0 &= -\frac{n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^{n} |X_i| \quad \Rightarrow \quad n = \frac{1}{\theta} \sum_{i=1}^{n} |X_i| \quad \Rightarrow \quad \hat{\theta}^{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} |X_i| \\
\ell''(\theta; \boldsymbol{X}) &= \frac{n}{\theta^2} - 2\frac{1}{\theta^3} \sum_{i=1}^{n} |X_i| \\
\mathbb{E}[-\ell''(\theta; \boldsymbol{X})] &= \mathbb{E}\left[-\frac{n}{\theta^2} + 2\frac{1}{\theta^3} \sum_{i=1}^{n} |X_i|\right] = 2\frac{1}{\theta^3} \mathbb{E}\left[\sum_{i=1}^{n} |X_i|\right] - \frac{n}{\theta^2} = n\left(\frac{2\theta}{\theta^3} - \frac{1}{\theta^2}\right) = \frac{n}{\theta^2}
\end{aligned}$$

(a) [4 pt / 30 pts] What is the asymptotic distribution of $\hat{\theta}^{\text{MLE}}$? Your answer can only include brand name variable notation, θ, n and fundamental constants. Merely writing the relevant result of the monster theorem will give you one point at most.

The relevant result of the monster theorem is:

$$\hat{\theta}^{\text{MLE}} \mathrel{\dot\sim} \mathcal{N}\left(\theta, \frac{I(\theta)^{-1}}{n}\right) = \mathcal{N}\left(\theta, \frac{1}{I_n(\theta)}\right)$$

We are given $I_n(\theta)$ in the last line from the problem statement. So we merely substitute:

$$\hat{\theta}^{\text{MLE}} \mathrel{\dot\sim} \mathcal{N}\left(\theta, \frac{\theta^2}{n}\right) = \mathcal{N}\left(\theta, \left(\frac{\theta}{\sqrt{n}}\right)^2\right)$$

For the rest of this problem, consider the following dataset:

$$\boldsymbol{x} = < 2.37, 3.71, 0.46, -0.44, 1.37, 9.09, -3.86, -1.69, -3.00, 0.46 >$$

(b) [6 pt / 36 pts] Compute the maximum likelihood estimate of θ to the nearest three decimal values.

The formula for $\hat{\hat{\theta}}^{\text{MLE}}$ was found for us in the problem statement. Using our calculator:

$$\begin{aligned}\hat{\hat{\theta}}^{\text{MLE}} &= \frac{1}{n}\sum_{i=1}^{n}|x_i| = \frac{2.37+3.71+0.46+0.44+1.37+9.09+3.86+1.69+3+0.46}{10} \\ &= \frac{26.45}{10} = 2.645\end{aligned}$$

If you did not answer the previous question, for the rest of the problem, use $\hat{\hat{\theta}}^{\text{MLE}} = 2.645$.

(c) [6 pt / 42 pts] Calculate a 95% confidence interval for θ to the nearest three decimals.

$$\begin{aligned}CI_{\theta,95\%} &= \left[\hat{\hat{\theta}}^{\text{MLE}} \pm 1.96\frac{1}{\sqrt{I_n\left(\hat{\hat{\theta}}^{\text{MLE}}\right)}}\right] \\ &= \left[\hat{\hat{\theta}}^{\text{MLE}} \pm 1.96\frac{\hat{\hat{\theta}}^{\text{MLE}}}{\sqrt{n}}\right] = \left[2.645 \pm 1.96\frac{2.645}{\sqrt{10}}\right] = [1.006, 4.284]\end{aligned}$$

(d) [6 pt / 48 pts] Test $H_0 : \theta = 2$ using the Wald Test at $\alpha = 5\%$. Your answer must include computing the proper retainment region.

The Wald test is the asymptotically valid Z test using the asymptotic normality of $\hat{\theta}^{\text{MLE}}$. Thus we can compute a retention region:

$$RET_{5\%} = \left[\theta_0 \pm 1.96\frac{1}{\sqrt{I_n(\theta_0)}}\right] = \left[\theta_0 \pm 1.96\frac{\theta_0}{\sqrt{n}}\right] = \left[2 \pm 1.96\frac{2}{\sqrt{10}}\right] = [0.76, 3.24]$$

Since $\hat{\hat{\theta}}^{\text{MLE}} = 2.645 \in RET_{5\%}$, we $\boxed{\text{fail to reject } H_0}$.

(e) [6 pt / 54 pts] Test $H_0 : \theta = 2$ using the Score Test at $\alpha = 5\%$. Your answer must include computing the proper test statistic and comparing it to the proper retainment region.

The Score test provides an asymptotically valid z test statistic:

$$\hat{\hat{z}} = \frac{s(\theta_0; \boldsymbol{X})}{\sqrt{I_n(\theta_0)}} = \frac{-\frac{n}{\theta_0} + \frac{1}{\theta_0^2}\sum_{i=1}^n |x_i|}{\sqrt{\frac{n}{\theta_0^2}}} = \frac{-\frac{10}{2} + \frac{1}{2^2}26.45}{\sqrt{\frac{10}{2^2}}} = 1.0198$$

Since $\hat{\hat{z}} = 1.02 \in RET_{5\%} = [\pm 1.96]$, we $\boxed{\text{fail to reject } H_0}$.

(f) [6 pt / 60 pts] Test $H_0 : \theta = 2$ using the Likelihood Ratio Test at $\alpha = 5\%$. Your answer must include computing the proper test statistic and comparing it to the proper retainment region.

The Score test provides an asymptotically valid chi-squared 1 test statistic:

$$\begin{aligned}
\hat{\hat{\Lambda}} &= 2\left(\ell(\hat{\hat{\theta}}^{\text{MLE}}; \boldsymbol{x}) - \ell(\theta_0; \boldsymbol{x})\right) \\
&= 2\left(\left(\cancel{-n\ln(2)} - n\ln(\theta) - \frac{1}{\theta}\sum_{i=1}^n |X_i|\right) - \left(\cancel{-n\ln(2)} - n\ln(\theta) - \frac{1}{\theta}\sum_{i=1}^n |X_i|\right)\right) \\
&= 2\left(\left(-n\ln\left(\hat{\hat{\theta}}^{\text{MLE}}\right) - \frac{1}{\hat{\hat{\theta}}^{\text{MLE}}}\sum_{i=1}^n |X_i|\right) - \left(-n\ln(\theta_0) - \frac{1}{\theta_0}\sum_{i=1}^n |X_i|\right)\right) \\
&= 2\left(-10\ln(2.645) - \frac{1}{2.645}26.45 + 10\ln(2) + \frac{1}{2}26.45\right) = 0.8595
\end{aligned}$$

By the equivalence of z-tests and chi-squared-1 tests, we know that the $RET_{5\%} = [0, 1.96^2] = [0, 3.84]$. Since $\hat{\hat{\Lambda}} = 0.8595 \in RET_{5\%}$, we $\boxed{\text{fail to reject } H_0}$.

(g) [4 pt / 64 pts]
The interval in (c) is... circle one... exact / approximate.
The decision in (d) is... circle one... exact / approximate.
The decision in (e) is... circle one... exact / approximate.
The decision in (f) is... circle one... exact / approximate.

Problem 4 Let's consider a scenario where we want to investigate whether there is an association between gender and preference for car models Honda vs Toyota among middle-aged suburbanites in Long Island. We collect the data in a contingency table, where the rows represent gender (Male vs Female), and the columns represent product preference (Honda vs Toyota). We ask an SRS of middle-aged suburbanites in Long Island which model they prefer and their gender. The contingency table of the data is below:

	Honda	Toyota
Male	17	53
Female	28	46

(a) [10 pt / 74 pts] At $\alpha = 5\%$, test the null hypothesis that gender and car model are independent. Your answer must include computing the proper test statistic and comparing it to the proper retainment region.

This is a chi-squared test. We first complete the table with empirical probabilities of being in each row and column:

	Honda	Toyota	Total (prob estimate)
Male	17	53	70 (.486)
Female	28	46	74 (.514)
Total (prob estimate)	45 (.313)	99 (.687)	144

This then allows us to construct expected values under the null of independent via multiplying the row-column probability estimates:

	Honda	Toyota
Male	21.875	48.125
Female	23.125	50.875

We now compute Pearson's test statistic:

$$\hat{\hat{\phi}} = \frac{(17-21.875)^2}{21.875} + \frac{(53-48.125)^2}{48.125} + \frac{(28-23.125)^2}{23.125} + \frac{(46-50.875)^2}{50.875} = 3.075$$

As there are two rows and two columns, the degrees of freedom on the chi-squared is 1. By the equivalence of z-tests and chi-squared-1 tests, we know that the $RET_{5\%} = [0, 1.96^2] = [0, 3.84]$ and thus we $\boxed{\text{fail to reject } H_0}$.

Problem 5 You have measurements of survival of $n = 11$ yeast cultures in a laboratory. Pertinent statistics are $\bar{x} = 40.3$ and $s = 114.6$. Let θ denote the mean survival of the yeast cultures.

(a) [6 pt / 80 pts] Assume the survivals above are iid normally distributed. Compute a $CI_{\theta,95\%}$ to two decimals. Here are a table of 97.5%iles of Student's T distribution under a variety of degrees of freedom (df).

df	1	2	3	4	5	6	7	8	9	10	11	12	13
$F^{-1}(.975)$	12.71	4.30	3.18	2.78	2.57	2.45	2.36	2.31	2.26	2.23	2.20	2.18	2.16

$$CI_{\theta,95\%} = \left[\bar{x} \pm t_{n-1,.975}\frac{s}{\sqrt{n}}\right] = \left[\bar{x} \pm t_{10,.975}\frac{s}{\sqrt{n}}\right] = \left[40.3 \pm 2.23\frac{114.6}{\sqrt{11}}\right] = [-36.75, 117.35]$$

(b) [1 pt / 81 pts] The interval above is... circle one... exact / approximate.

(c) [3 pt / 84 pts] What is conceptually wrong with the above CI?

It contains negative values. Survival must be positive.

(d) [8 pt / 92 pts] Let $\phi = \ln(\theta)$. Compute a $CI_{\phi,95\%}$ to two decimals.

We need to use delta method and the three theorems:

$$\frac{\hat{\phi} - \phi}{|g'(\hat{\theta})|\hat{SE}\left[\hat{\theta}\right]} \overset{d}{\to} \mathcal{N}(0,\, 1)$$

And when using the finite approximation:

$$\begin{aligned} CI_{\phi,95\%} &\approx \left[\hat{\hat{\phi}} \pm 1.96\left|g'(\hat{\hat{\theta}})\right|\hat{SE}\left[\hat{\theta}\right]\right] = \left[\ln(\bar{x}) \pm 1.96\frac{1}{|\bar{x}|}s\right] = \left[\ln(40.3) \pm 1.96\frac{114.6}{40.3}\right] \\ &= [-1.88, 9.27] \end{aligned}$$

Problem 6 Consider $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$ and the Laplace principle of indifference prior on the entire parameter space, $\Theta = (0, 1)$.

(a) [8 pt / 100 pts] Find the posterior distribution step-by-step from first principles. If it's a brand-name rv, mark it so and provide the values of its parameters.

$$
\begin{aligned}
f(\theta \mid \boldsymbol{X}) &= \frac{\mathbb{P}(\boldsymbol{X} \mid \theta) f(\theta)}{\mathbb{P}(\boldsymbol{X})} \propto \mathbb{P}(\boldsymbol{X} \mid \theta) f(\theta) \\
&= \left(\prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i}\right) \mathbb{1}_{\theta \in (0,1)} \\
&= \theta^{\sum x_i} (1-\theta)^{\sum 1 - x_i} \mathbb{1}_{\theta \in (0,1)} \\
&= \theta^{\sum x_i + 1 - 1} (1-\theta)^{n - \sum x_i + 1 - 1} \mathbb{1}_{\theta \in (0,1)} \\
&\propto \text{Beta}\left(1 + \sum x_i,\, 1 + n - \sum x_i\right)
\end{aligned}
$$