Math 341 / 641 Fall 2023
Midterm Examination Two Solutions

Professor Adam Kapelner

November 8, 2023

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 **two** 8.5" $\times$ 11" page (front and back) "cheat sheets", 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. Show as much partial work as you can and justify each step. No food is allowed, only drinks.

Problem 1 We revisit the data from the International Mouse Phenotyping Consortium (IMPC). As described in Karp et al. (2017), the IMPC coordinates a large study to functionally annotate every protein coding gene by exploring the impact of the gene knockout on the resulting phenotype for up to 234 traits of interest. There are $m = 172328$ gene-phenotype tests.

(a) [3 pt / 3 pts] If we want to control FWER at 1%, what is the Bonferroni threshold α for each of the m tests to three significant digits?

$$\alpha = \frac{FWER}{m} = \frac{0.01}{172328} = 5.80 \times 10^{-8}$$

(b) [3 pt / 6 pts] Write in English about the practical problem of employing the control in part (a).

This α is very small and thus we will have low power and rarely reject and thus we will not make many discoveries.

(c) [6 pt / 12 pts] Let $p_{(1)}, p_{(2)}, \ldots, p_{(m)}$ denote the sorted Fisher's p-value for all m tests. The FDR procedure at 1% expected false discoveries yields 4,579 tests that are rejected. Estimate the value of the maximum p-value of the set of the 4,579 rejections' p-values to three significant digits. Is this higher than the Bonferroni α? Yes/no.

The FDR procedure is identical to the Simes procedure where we order all tests by their p-value and reject the first a_* defined by $a_* := \max\left\{a \,:\, p_{(a)} \leq FWER\frac{a}{m}\right\}$. Here we are given that $a_* = 4579$ so we can just solve for $p_{(a_*)}$ to obtain the upper bound on the p-value:

$$p_{(a_*)} \leq FWER\frac{a_*}{m} = 0.01\frac{4579}{172328} = 0.000266$$

Yes, this is larger than the Bonferroni FWER control's α.

Problem 2 We wish to test differences between the distribution of the number of LIRR delays post-corona pandemic and the distribution of the number of LIRR delays pre-corona pandemic. We'll define population #1 to be pre-pandemic (March, 2020 and before) and population #2 to be post-pandemic (January, 2022 to present day). This is real monthly LIRR data going back to January, 2016 as found at catalog.data.gov. We will assume the monthly samples for both populations are iid and independent between the two populations as well. Some relevant summary statistics are below:

$$n_1 = 51,\ \bar{x}_1 = 49.8,\ s_1^2 = 104.83^2$$
$$n_2 = 21,\ \bar{x}_2 = 16.7,\ s_2^2 = 30.70^2$$

(a) [3 pt / 15 pts] To test difference in population means, would a 2-sample t-test be appropriate? Explain in English why or why not.

The exact or approximate t-test both require the two populations' DGP's to be iid normal. Here, we are dealing with number of delays which means $\mathcal{S}_X = \mathbb{N}_0$ for both populations. This means the two populations' DGP's are not normal and hence the t-test is **inappropriate**.

(b) [6 pt / 21 pts] Regardless of whether it is appropriate or not, use an F-test to attempt to prove the variances are unequal i.e. $H_a : \sigma_1^2 \neq \sigma_2^2$ at $\alpha = 5\%$. The relevant values you need are $F_W(0.502) = 2.5\%$ and $F_W(2.25) = 97.5\%$ where $W \sim F_{50,20}$. Indicate the RET region, the decision and write a concluding sentence.

$$\hat{\hat{r}} := \frac{s_1^2}{s_2^2} = \frac{104.83^2}{30.70^2} = 11.66 \notin \text{RET} = [0.512, 2.25] \quad \Rightarrow \quad \text{Reject } H_0$$

There is sufficient evidence to conclude the variance of the number of LIRR delays pre-pandemic and the variance of the number of LIRR delays post-pandemic are unequal.

(c) [8 pt / 29 pts] Regardless of whether it is appropriate or not, run a 2-sample t-test of a difference in means, i.e. $H_a : \theta_1 \neq \theta_2$, assuming equal variance at $\alpha = 5\%$. The relevant t-value is $t_{n_1+n_2-2,1-\alpha/2} = t_{70,97.5\%} = 1.99$. Indicate the decision and write a concluding sentence.

$$\begin{aligned} s_{\text{pooled}} &= \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} = \sqrt{\frac{50 \cdot 104.83^2 + 20 \cdot 30.70^2}{70}} = \sqrt{8118.8} = 90.10 \\ \hat{\hat{T}} &= \frac{\bar{x}_1 - \bar{x}_2}{s_{\text{pooled}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{49.8 - 16.7}{90.10\sqrt{\frac{1}{51} + \frac{1}{21}}} = \frac{33.1}{32.28} = 1.02 \in \text{RET} = [-1.99, 1.99] \quad \Rightarrow \quad \text{Retain } H_0 \end{aligned}$$

There is insufficient evidence to conclude the mean number of LIRR delays pre-pandemic and the mean number of LIRR delays post-pandemic are unequal.

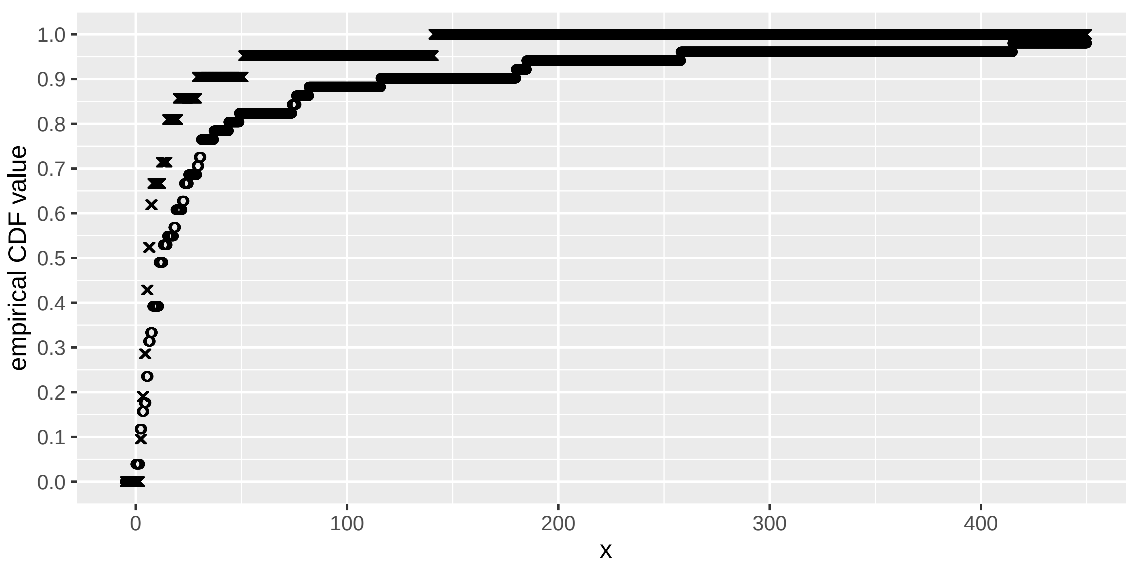
(d) [6 pt / 35 pts] Assume the answer to (a) was "no", run an asymptotically valid test of a difference in means, i.e. $H_a : \theta_1 \neq \theta_2$, at $\alpha = 5\%$. Calculate the test statistic, provide the RET region, indicate the decision and write a concluding sentence.

We run the Wald test whose test statistic is asymptotically normal.

$$\hat{\hat{Z}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{49.8 - 16.7}{\sqrt{\frac{104.83^2}{51} + \frac{30.70^2}{21}}} = \frac{33.1}{16.14} = 2.05 \notin \text{RET} = [-1.96, 1.96] \quad \Rightarrow \quad \text{Reject } H_0$$

There is sufficient evidence to conclude the mean number of LIRR delays pre-pandemic and the mean number of LIRR delays post-pandemic are unequal.

(e) [10 pt / 45 pts] This test can also be accomplished with a 2-sample Kolmogorov-Smirnov (KS) test albeit less-powerfully as the KS test looks for any difference in the two distributions (not only the mean). Below is a plot of $\hat{F}(x)$ for both populations. Population 1 is plotted as "o" and population 2 as "x".



Run the 2-sample KS test at $\alpha = 5\%$. Note that $F_K(1.359) = 95\%$ where $K \sim$ Kolmogorov dist. The alternative hypothesis is $H_a : \text{DGP}_1 \neq \text{DGP}_2$. Calculate the test statistic, provide the RET region, indicate and decision and write a concluding sentence.

We can estimate $\hat{\hat{D}}_{n_1,n_2}$ from the $\hat{F}_1(x)$ and $\hat{F}_2(x)$ plots above. At around $x = 60$ it appears as if the difference is slightly less than 3 y-axis gridlines where each has width 0.05. So let's estimate $\hat{\hat{D}}_{n_1,n_2} = 0.14$. We then calculate the test statistic:

$$\sqrt{\frac{n_1 n_2}{n_1 + n_2}} \hat{\hat{D}}_{n_1,n_2} = \sqrt{\frac{51 \cdot 21}{51 + 21}} 0.14 = 0.544 \in \text{RET} = [0, 1.359] \quad \Rightarrow \quad \text{Retain } H_0$$

There is insufficient evidence to conclude the distribution of the number of LIRR delays pre-pandemic and the distribution of the number of LIRR delays post-pandemic are unequal.

Problem 3 Assume the following DGP:

$$X_1, \ldots, X_n \overset{iid}{\sim} \text{Poisson}(\theta_1) := \frac{\theta^x e^{-\theta}}{x!} \mathbb{1}_{x \in \mathbb{N}_0}$$

Here are some facts about this DGP from the previous midterm:

$$\begin{aligned}
\mathcal{L}(\theta; X_1, \ldots, X_n) &= \frac{\theta^{\sum_{i=1}^n X_i} e^{-n\theta}}{\prod_{i=1}^n X_i!} \\
\ell(\theta; X_1, \ldots, X_n) &= \ln(\theta) \sum_{i=1}^n X_i - n\theta - \sum_{i=1}^n \ln(X_i!) \\
s(\theta; X_1, \ldots, X_n) &= \frac{\sum_{i=1}^n X_i}{\theta} - n = n\left(\frac{\bar{X}}{\theta} - 1\right) \\
I_n(\theta) &= \frac{n}{\theta} \quad \Rightarrow \quad I(\theta) = \frac{1}{\theta} \\
\hat{\theta}^{\text{MLE}} &= \bar{X} \text{ and it is the UMVUE, i.e. } \mathbb{V}\text{ar}\left[\bar{X}\right] = \text{CRLB} := \frac{1}{I_n(\theta)} = \frac{\theta}{n}
\end{aligned}$$

(a) [6 pt / 51 pts] Provide the score test statistic for testing $H_a : \theta \neq \theta_0$. The statistic must be a function of $X_1, \ldots, X_n, n, \theta_0$ only.

$$\hat{Z} \mid H_0 = \frac{s(\theta_0; X_1, \ldots, X_n)}{\sqrt{I_n(\theta_0)}} = \frac{n\left(\frac{\bar{X}}{\theta_0} - 1\right)}{\sqrt{\frac{n}{\theta_0}}}$$

(b) [6 pt / 57 pts] Provide the likelihood ratio test statistic $\hat{\Lambda} := 2\ln\left(\hat{LR}\right)$ for testing $H_a : \theta \neq \theta_0$. The statistic must be a function of $X_1, \ldots, X_n, n, \theta_0$ only.

$$\begin{aligned}
\hat{\Lambda} &:= 2\ln\left(\frac{\frac{\left(\hat{\theta}^{\text{MLE}}\right)^{\sum_{i=1}^n X_i} e^{-n\hat{\theta}^{\text{MLE}}}}{\prod_{i=1}^n X_i!}}{\frac{\theta_0^{\sum_{i=1}^n X_i} e^{-n\theta_0}}{\prod_{i=1}^n X_i!}}\right) \\
&= 2\ln\left(\frac{\left(\hat{\theta}^{\text{MLE}}\right)^{\sum_{i=1}^n X_i} e^{-n\hat{\theta}^{\text{MLE}}}}{\theta_0^{\sum_{i=1}^n X_i} e^{-n\theta_0}}\right) \\
&= 2\ln\left(\frac{\bar{X}^{n\bar{X}} e^{-n\bar{X}}}{\theta_0^{n\bar{X}} e^{-n\theta_0}}\right) \\
&= 2\left(n\bar{X}\ln\left(\frac{\bar{X}}{\theta_0}\right) - n(\bar{X} - \theta_0)\right)
\end{aligned}$$

Simplifying is not necessary to get full credit but it will make (d) easier to compute.

(c) [6 pt / 63 pts] Consider the reparameterization of $\phi = \sqrt{\theta}$. Provide an asymptotically normal test statistic for testing $H_a : \phi \neq \phi_0$. The statistic must be a function of $X_1, \ldots, X_n, n, \theta_0$ only.

The delta method theorem states that

$$\frac{\hat{\phi} - \phi_0}{|g'(\theta)|\mathrm{SE}[\hat{\theta}]} \xrightarrow{d} \mathcal{N}(0,\, 1)$$

We compute $|g'(\theta)| = |-\frac{1}{2}\theta^{-\frac{1}{2}}| = \frac{1}{2}\theta^{-\frac{1}{2}}$ and $\mathrm{SE}[\hat{\theta}] = \sqrt{\theta/n}$ as given in the problem description. Putting this all together and simplifying, the test statistic is

$$\hat{Z} \mid H_0 = \frac{\hat{\phi} - \phi_0}{\frac{1}{2}\theta^{-\frac{1}{2}}\sqrt{\theta/n}} = \frac{\hat{\phi} - \phi_0}{\frac{1}{\sqrt{4n}}}$$

For the rest of this problem, consider the prepandemic LIRR delay data where $n = 51$ and $\bar{x} = 49.8$.

(d) [6 pt / 69 pts] Use the likelihood ratio test to test if $H_a : \theta \neq 35$. Indicate the RET region and the decision.

$$\begin{aligned} \hat{\hat{\Lambda}} &= 2\left(n\bar{x}\ln\left(\frac{\bar{x}}{\theta_0}\right) - n(\bar{x} - \theta_0)\right) \\ &= 2\left(51 \cdot 49.8 \ln\left(\frac{49.8}{35}\right) - 51(49.8 - 35)\right) \\ &= 281.8 \notin \mathrm{RET} = [0, 3.84] \;\Rightarrow\; \text{Reject } H_0 \end{aligned}$$

(e) [6 pt / 75 pts] Create a 95% confidence interval estimate for $\phi = \sqrt{\theta}$.

$$\begin{aligned} \hat{\hat{CI}}_{\phi,95\%} &\approx \left[\hat{\hat{\phi}} \pm z_{1-\alpha/2}\, |g'(\theta)|\, \hat{\mathrm{SE}}[\hat{\theta}]\right] \\ &= \left[\sqrt{\hat{\hat{\theta}}} \pm z_{1-\alpha/2}\frac{1}{\sqrt{4n}}\right] \quad \text{where} \quad \sqrt{\hat{\hat{\theta}}} = \sqrt{\bar{x}} = \sqrt{49.8} = 7.06 \\ &= \left[7.06 \pm 1.96 \cdot \frac{1}{\sqrt{4 \cdot 51}}\right] \\ &= [6.92, 7.19] \end{aligned}$$

Problem 4 We seek to test if hair length and eye color of cats are dependent. Here is a sample of cats organized into a contingency table with rowsums and colsums supplied:

	Blue Eyes	Yellow Eyes	Green Eyes	Total
Short Hair	25	48	13	86
Long Hair	8	32	24	64
Total	33	80	37	150

Let $\theta_{i,j}$ denote the joint probability of having hair length of row i and eye color of column j. Let $\theta_{i.}$ denote the marginal probability of having hair length of row i. Let $\theta_{.j}$ denote the marginal probability of having eye color of column j.

Below are some 95%iles of different chi-squared distributions by degrees of freedom.

degrees of freedom	1	2	3	4	5	6	7	8	9	10
x where $F_{\chi^2}(x) = .95$	3.84	5.99	7.81	9.49	11.07	12.59	14.07	15.51	16.92	18.31

(a) [3 pt / 78 pts] Using the θ notation above, write the null and alternative hypotheses.

$$H_a : \exists i, j \quad \theta_{i,j} \neq \theta_{i.}\theta_{.j}$$
$$H_0 : \forall i, j \quad \theta_{i,j} = \theta_{i.}\theta_{.j}$$

(b) [10 pt / 88 pts] Run the test from part (a) at $\alpha = 5\%$. Indicate the decision.

We first calculate the expected counts below. The margins record the proportions in each row and column:

	Blue Eyes	Yellow Eyes	Green Eyes	
Short Hair	18.91	50.45	21.2	$(\theta_{1.} = 0.573)$
Long Hair	14.09	37.60	15.8	$(\theta_{2.} = 0.427)$
	$(\theta_{.1} = 0.22)$	$(\theta_{.2} = 0.587)$	$(\theta_{.2} = 0.247)$	

Now we calculate the chi-squared statistic cell by cell calculating $(O_{ij} - E_{ij})^2/E_{ij}$:

	Blue Eyes	Yellow Eyes	Green Eyes
Short Hair	1.96	0.12	3.17
Long Hair	2.63	0.83	4.26

We sum to obtain the statistic. The threshold comes from the $\chi^2_{(r-1)(c-1)=2}$ distribution:

$$\hat{\hat{\phi}} = 12.97 \notin \text{RET} = [0, 5.99] \;\; \Rightarrow \;\; \text{Reject } H_0$$

Problem 5 In class we modeled the maximum daily wind speed at JFK in the year 2013 and thus $n = 365$ for our dataset $\boldsymbol{x}$. We fit seven different models to the data by computing the MLE's of all their parameters and calculated their AIC's below from lowest to highest left to right.

distribution	gamma	logistic	normal	gumbel	weibull	frechet	gompertz	exponential
$k := \dim[\boldsymbol{\theta}]$	2	2	2	2	2	2	2	1
AIC	2262.6	2265.3	2289.4	2290.6	2300.2	2352.7	2404.5	2872.2
Normalized Akaike Weight	0.799	0.201	0.000	0.000	0.000	0.000	0.000	0.000

(a) [3 pt / 91 pts] Which DGP of the seven is the best fitting model and why?

The gamma model is the best fitting as it has the lowest AIC metric.

(b) [3 pt / 94 pts] Compute the corrected AIC for the best fitting model.

$$\begin{aligned} AICc_{m_*} &= -2\ell_{m_*} + 2k_{m_*}\frac{n}{n-k_{m_*}-1} = AIC - 2k_{m_*} + 2k_{m_*}\frac{n}{n-k_{m_*}-1} \\ &= 2262.6 - 2\cdot 2 + 2\cdot 2\frac{365}{365-2-1} \\ &= 2262.63 \end{aligned}$$

(c) [6 pt / 100 pts] Compute $\ell(\hat{\hat{\theta}}^{\text{MLE}}; \boldsymbol{x})$ where the DGP is $X_1, \ldots, X_n \overset{iid}{\sim} \text{Exp}(\theta)$.

Let $m = 7$ denote the exponential model.

$$AIC_7 = -2\ell_7 + 2k_7 \;\Rightarrow\; \ell_7 = -\frac{AIC_7 - 2k_7}{2} = -\frac{2872.2 - 2\cdot 1}{2} = -1435.1$$