Math 341 / 641 Fall 2024
Midterm Examination One

Professor Adam Kapelner

October 1, 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 Benford's Law represents the distribution of the leading digit in base-10 numbers within datasets across a wide variety of natural phenomenon such as street addresses, stock prices, population numbers, etc:

$$X \sim \text{Benford} := \log_{10}\left(\frac{x+1}{x}\right)\mathbb{1}_{x\in\{1,2,\ldots,9\}}, \quad \theta_0 := \mathbb{E}[X] = 3.44, \quad \sigma_0^2 := \mathbb{V}\text{ar}[X] = 6.06$$

For example, here are 15 numbers sampled from Benford's Law sorted: 11217 11426 12612 13368 15644 25311 27342 39332 41511 42632 44125 52322 78431 81673 82152 . Notice how the first digit (in black) is more likely to be a 1 or 2 vs an 8 or 9.

Let the parameter of interest be the mean value of the first digit of numeric entries (call it θ). Benford's Law is used by the Internal Revenue Service (IRS) to catch people committing tax fraud since numeric entries on tax forms is known to follow Benford's law.

(a) [2 pt / 2 pts] Circle one: the value of θ, then mean numeric value of the first digit for the fields in any individual's tax return is... known / unknown

(b) [2 pt / 4 pts] To catch someone cheating on their taxes, we need to prove beyond a reasonable doubt that this individual's θ is not the expectation value of Benford's Law. Thus the alternative hypothesis is $H_a : \theta \neq 3.44$. What is the null hypothesis for the parameter θ?

$$H_0 : \theta = 3.44$$

(c) [2 pt / 6 pts] The "1040 form" the IRS uses has 52 numeric entries. We collect the first digit from each of these fields. Let these digits be the data $x_1, x_2, \ldots, x_{52}$. What is the value of n? $n = 52$

(d) [2 pt / 8 pts] What estimator would you choose to estimate θ? $\hat{\theta} = \bar{X}$

(e) [2 pt / 10 pts] Regardless of what answer you put for the previous question, for the remainder of this problem use $\hat{\theta} = \bar{X}$. Circle one: the distribution of this estimator is... known exactly / unknown

(f) [5 pt / 15 pts] Compute $RET_{5\%}$, the retainment region at $\alpha = 5\%$. Remember, the full null hypothesis is that the data DGP is $X_1, \ldots, X_n \overset{iid}{\sim}$ Benford. Round the values to the nearest two decimals.

$$\text{RET}_{5\%} = \left[\theta_0 \pm z_{1-\frac{\alpha}{2}}\frac{\sigma_0}{\sqrt{n}}\right] = \left[3.44 \pm 1.96\frac{\sqrt{6.06}}{\sqrt{52}}\right] = [3.44 \pm 0.67] = [2.77, 4.11]$$

(g) [2 pt / 17 pts] Circle one: the RET above is... exact / approximate

(h) [6 pt / 23 pts] For Bob's tax return, $\bar{x} = 5.27$ for the 52 first digits of the numeric fields. Run the test and write a concluding sentence.

$\bar{x} \notin \text{RET}_{5\%}$ hence we reject H_0 and conclude that there is statistically significant evidence that the first digits of Bob's numeric values on his form 1040 do not adhere to Benford's Law, i.e., Bob is cheating the IRS.

(i) [3 pt / 26 pts] Circle one: you could've made a ... Type I error / Type II error

(j) [6 pt / 32 pts] Fisher's approximate p_{val} for this test equals $2\mathbb{P}(Z > z)$ where $Z \sim \mathcal{N}(0, 1)$. Compute the value of z to the nearest two decimals.

$$\begin{aligned} p_{val} &= 2\mathbb{P}\left(\hat{\theta} > \hat{\hat{\theta}} \mid H_0\right) = 2\mathbb{P}\left(\bar{X} > \bar{x} \mid H_0\right) = 2\mathbb{P}\left(\frac{\bar{X} - \theta_0}{\sigma_0/\sqrt{n}} > \frac{\bar{x} - \theta_0}{\sigma_0/\sqrt{n}}\right) \\ &= 2\mathbb{P}\left(Z > \frac{5.27 - 3.44}{\sqrt{6.06}/\sqrt{52}}\right) \quad \Rightarrow \quad z = 5.36 \end{aligned}$$

(k) [6 pt / 38 pts] Compute an approximate $CI_{\theta,95\%}$ to the nearest two decimals.

$$CI_{\theta,95\%} = \left[\bar{x} \pm z_{1-\frac{\alpha}{2}}\frac{\sigma_0}{\sqrt{n}}\right] = \left[5.27 \pm 1.96\frac{\sqrt{6.06}}{\sqrt{n}}\right] = [5.27 \pm 0.67] = [4.60, 5.94]$$

(l) [8 pt / 46 pts] When people commit fraud, they may fabricate numbers "completely randomly". This means the first digits of their numbers are drawn from the DGP:

$$X_1, \ldots, X_n \overset{iid}{\sim} U(\{1, 2, \ldots, 9\}) \text{ where } \theta := \mathbb{E}[X] = 5, \quad \sigma^2 := \mathbb{V}\text{ar}[X] = 6.67$$

How many digits n would you need to detect if someone was cheating by sampling numbers randomly (via the above DGP) with probability 90% if the null assumption is the same as previously, i.e. the DGP is $X_1, \ldots, X_n \overset{iid}{\sim}$ Benford. Note that $z_{10\%} = -1.28$. Assume $\alpha = 5\%$. Round the result to the nearest natural number.

As the "real" $\hat{\theta}$ distribution is centered to the right of the $\hat{\theta} \mid H_0$ distribution, we must equate the right bound of the retainment region (from part f) to the 10%ile of the distribution of $\hat{\theta}$ and solve for the sample size n:

$$\begin{aligned}
\theta_0 + z_{1-\frac{\alpha}{2}} \frac{\sigma_0}{\sqrt{n}} &= \theta + z_{10\%} \frac{\sigma}{\sqrt{n}} \\
3.44 + 1.96 \frac{\sqrt{6.06}}{\sqrt{n}} &= 5 - 1.28 \frac{\sqrt{6.67}}{\sqrt{n}} \\
1.96 \frac{\sqrt{6.06}}{\sqrt{n}} + 1.28 \frac{\sqrt{6.67}}{\sqrt{n}} &= 1.56 \\
\frac{8.13}{\sqrt{n}} &= 1.56 \\
n &= \left(\frac{8.13}{1.56} \right)^2 = 27.165 \quad \Rightarrow \quad n = 27
\end{aligned}$$

Problem 2

Consider the "inverse gamma" DGP, a famous rv we'll study later in class:

$$X_1, \ldots, X_n \overset{iid}{\sim} \text{InvGamma}(\theta_1, \theta_2) := \frac{\theta_2^{\theta_1}}{\Gamma(\theta_1)} x^{-\theta_1 - 1} e^{-\theta_2 / x} \mathbb{1}_{x > 0}$$

where $\Gamma(u)$ is called the "gamma" function which is $\mathbb{R} \to \mathbb{R}$ and ensures the Humpty-Dumpty identity. The mean and variance of this rv are given below:

$$\mathbb{E}[X] = \frac{\theta_2}{\theta_1 - 1}, \quad \mathbb{V}\text{ar}[X] = \frac{\theta_2^2}{(\theta_1 - 1)^2 (\theta_1 - 2)}$$

(a) [2 pt / 48 pts] How many parameters can be targets of inference in this DGP? 2

(b) [8 pt / 56 pts] Show that the method moments estimator for θ_1 is $\hat{\theta}_1^{MM} = \frac{2\hat{\mu}_2 - \hat{\mu}_1^2}{\hat{\mu}_2 - \hat{\mu}_1^2}$.

$$\begin{aligned}
\mu_1 &:= \mathbb{E}[X] = \frac{\theta_2}{\theta_1 - 1} \quad \Rightarrow \quad \theta_2 = \mu_1(\theta_1 - 1) \\
\mu_2 - \mu_1^2 &:= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \mathbb{V}\mathrm{ar}[X] = \frac{\theta_2^2}{(\theta_1 - 1)^2(\theta_1 - 2)} \\
&= \frac{(\mu_1(\theta_1 - 1))^2}{(\theta_1 - 1)^2(\theta_1 - 2)} \\
&= \frac{\mu_1^2}{\theta_1 - 2} \\
\Rightarrow \quad \theta_1 - 2 &= \frac{\mu_1^2}{\mu_2 - \mu_1^2} \\
\Rightarrow \quad \theta_1 &= \frac{\mu_1^2}{\mu_2 - \mu_1^2} + 2 = \frac{\mu_1^2}{\mu_2 - \mu_1^2} + 2\frac{\mu_2 - \mu_1^2}{\mu_2 - \mu_1^2} \quad \Rightarrow \quad \hat{\theta}_1^{MM} = \frac{2\hat{\mu}_2 - \hat{\mu}_1^2}{\hat{\mu}_2 - \hat{\mu}_1^2}
\end{aligned}$$

That completes the problem. But we'll need $\hat{\theta}_2^{MM}$ for the next problem, so we'll derive it here now by substituting $\hat{\theta}_1^{MM}$ for θ_1:

$$\begin{aligned}
\theta_2 &= \mu_1(\theta_1 - 1) \\
&= \mu_1\left(\frac{2\mu_2 - \mu_1^2}{\mu_2 - \mu_1^2} - 1\right) = \mu_1\left(\frac{2\mu_2 - \mu_1^2}{\mu_2 - \mu_1^2} - \frac{\mu_2 - \mu_1^2}{\mu_2 - \mu_1^2}\right) \\
\Rightarrow \quad \hat{\theta}_2^{MM} &= \frac{\hat{\mu}_1\hat{\mu}_2}{\hat{\mu}_2 - \hat{\mu}_1^2}
\end{aligned}$$

(c) [2 pt / 58 pts] Circle one: as n gets larger, the $\mathbb{M}\mathrm{SE}\left[\hat{\theta}_2^{MM}\right]$ ... decreases / increases

(d) [3 pt / 61 pts] Let $\boldsymbol{x} =< 0.918, 0.386, 0.395, 0.553, 1.643, 0.536 >$. Using the method of moments estimation technique, find the estimates for the two parameters $\hat{\hat{\theta}}_1^{MM}$ and $\hat{\hat{\theta}}_2^{MM}$. Round to three decimals.

We first estimate the moments and then substitute

$$\hat{\hat{\mu}}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i = 0.7385, \qquad \hat{\hat{\mu}}_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 = 0.7400$$

$$\hat{\hat{\theta}}_1^{MM} = \frac{2\hat{\hat{\mu}}_2 - \hat{\hat{\mu}}_1^2}{\hat{\hat{\mu}}_2 - \hat{\hat{\mu}}_1^2} = 4.802, \qquad \hat{\hat{\theta}}_2^{MM} = \frac{\hat{\hat{\mu}}_1\hat{\hat{\mu}}_2}{\hat{\hat{\mu}}_2 - \hat{\hat{\mu}}_1^2} = 2.807$$

(e) [8 pt / 69 pts] Assume we now know that $\theta_1 = 5$ going forward in this problem. Show that the maximum likelihood estimator for the second parameter for n draws from this inverse gamma DGP is $\hat{\theta}_2^{\text{MLE}} = 5n \left(\sum_{i=1}^n X_i^{-1}\right)^{-1}$.

$$\begin{aligned}
\mathcal{L} &= \prod_{i=1}^n \frac{\theta_2^5}{\Gamma(5)} X_i^{-5-1} e^{-\theta_2/X_i} = \frac{\theta_2^{5n}}{\Gamma(5)^n} e^{-\theta_2 \sum_{i=1}^n \frac{1}{X_i}} \prod_{i=1}^n X_i^{-6} \\
\ell &= 5n \ln(\theta_2) - n \ln(\Gamma(5)) - \theta_2 \sum_{i=1}^n \frac{1}{X_i} + \ln\left(\prod_{i=1}^n X_i^{-6}\right) \\
\ell' &= \frac{5n}{\theta_2} - \sum_{i=1}^n \frac{1}{X_i} \overset{set}{=} 0 \quad \Rightarrow \quad \frac{5n}{\theta_2} = \sum_{i=1}^n \frac{1}{X_i} \quad \Rightarrow \quad \hat{\theta}_2^{\text{MLE}} = \frac{5n}{\sum_{i=1}^n X_i^{-1}}
\end{aligned}$$

(f) [6 pt / 75 pts] Find the Cramer-Rao Lower Bound for any unbiased estimator for θ_2.

Using the log likelihood ℓ' from (e), we continue:

$$\begin{aligned}
\ell'' &= -\frac{5n}{\theta_2^2} \\
I(\theta_2) &= \mathbb{E}[-\ell''] = \mathbb{E}\left[-\left(-\frac{5n}{\theta_2^2}\right)\right] = \frac{5n}{\theta_2^2} \\
\mathbb{V}\text{ar}\left[\hat{\theta}_2\right] &\geq \frac{I(\theta_2)^{-1}}{n} = \frac{\left(\frac{5n}{\theta_2^2}\right)^{-1}}{n} = \frac{\theta_2^2}{25n^2}
\end{aligned}$$

Problem 3 We are trying to prove that less than 2% of all electronic devices a certain company manufactures are defective. Let θ denote the real proportion of defective devices.

(a) [2 pt / 77 pts] What is the null hypothesis? $H_0 : \theta \geq 0.02$

(b) [2 pt / 79 pts] We now sample and record if the device is defective or not. What sampling procedure should be employed to ensure the results can be believed for the entire manufacturing process? simple random sample (SRS)

(c) [3 pt / 82 pts] We sample $n = 300$ according to the procedure given by the correct answer for (b). What is the DGP this sample was realized from?

$$X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}\,(\theta)$$

(d) [5 pt / 87 pts] We choose to use the binomial exact test. Below is a table of the PDF of the Binomial $(300,\ 0.02)$ rv.

x	0	1	2	3	4	5	6	7	8	9	10
$p(x)$	0.002	0.014	0.044	0.088	0.134	0.162	0.162	0.139	0.104	0.069	0.041

Find the three smallest possible nonzero sizes of the binomial exact test each to the nearest three digits.

The three possible sizes are the three left tails which are the CDF values $F(0), F(1), F(2)$ which are $0.002, 0.016, 0.060$.

(e) [5 pt / 92 pts] Find $RET_{5\%}$, the retainment region at $\alpha = 5\%$.

Since $\alpha = 5\%$, we cannot use $\{2, 3, \ldots, 100\}$ as this would result in a size of 6%. Thus, we must use $\{1, 2, 3, \ldots, 100\}$ resulting in a size of 1.6% as it's the largest size $\leq \alpha$.

(f) [5 pt / 97 pts] Of the 300 sampled devices, 3 were defective. Run the test. No need to write a concluding sentence.

Since $3 \in RET_{5\%}$, we retain H_0.

(g) [3 pt / 100 pts] If you don't believe the result of this test but you believe the sampling was done correctly, what is a legitimate criticism of the experiment? The sample size was insufficiently large, i.e., the test was underpowered.