Math 341 / 641 Fall 2024
Final Examination

Professor Adam Kapelner

December 17, 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 120 minutes (variable time per question) and closed-book. You are allowed **three** pages (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 This problem concerns 2-sample tests and under which settings they are appropriate to employ. Use the following standard lecture notes' notation throughout:

n_1 and n_2 denotes the sample size from population #1 and #2 respectively,
$X_{1,1}, \ldots, X_{1,n_1}$ and $X_{2,1}, \ldots, X_{2,n_2}$ denote rv's in population #1 and #2 respectively,
$\bar{x}_1$ and $\bar{x}_2$ are the averages of the values in populations #1 and #2 respectively and
s_1^2 and s_2^2 are the sample variances of the values in populations #1 and #2 respectively.

You can assume both samples are drawn using an SRS.

(a) [4 pt / 4 pts] Provide a setting where you would use an **exact two sample Z** test. Make sure to indicate (1) the DGP of the samples from the two populations, (2) what is known about the parameter(s) if anything, (3) the null hypothesis and (4) the test statistic. You do not need to state the statistic's null distribution.

(1) $X_{1,1}, \ldots, X_{1,n_1} \overset{iid}{\sim} \mathcal{N}(\theta_1, \sigma_1^2), \quad X_{2,1}, \ldots, X_{2,n_2} \overset{iid}{\sim} \mathcal{N}(\theta_2, \sigma_2^2)$

(2) the values of σ_1^2, σ_2^2 are known

(3) $H_0 : \theta_1 - \theta_2 = 0$ or $H_0 : \theta_1 = \theta_2$

(4) $\dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

(b) [4 pt / 8 pts] Provide a setting where you would use an **exact two sample T** test. Make sure to indicate (1) the DGP of the samples from the two populations, (2) what is known about the parameter(s) if anything, (3) the null hypothesis and (4) the test statistic. You do not need to state the statistic's null distribution.

(1) $X_{1,1}, \ldots, X_{1,n_1} \overset{iid}{\sim} \mathcal{N}(\theta_1, \sigma_1^2), \quad X_{2,1}, \ldots, X_{2,n_2} \overset{iid}{\sim} \mathcal{N}(\theta_2, \sigma_2^2)$

(2) $\sigma_1^2 = \sigma_2^2$ but its value is unknown

(3) $H_0 : \theta_1 - \theta_2 = 0$ or $H_0 : \theta_1 = \theta_2$

(4) $\dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$

(c) [4 pt / 12 pts] Provide a setting where you would use an **approximate two sample Z** test. Make sure to indicate (1) the DGP of the samples from the two populations, (2) what is known about the parameter(s) if anything, (3) the null hypothesis and (4) the test statistic. You do not need to state the statistic's null distribution.

There are many acceptable answers to this question. Below are two:

(1) $X_{1,1}, \ldots, X_{1,n_1} \overset{iid}{\sim} \text{Bernoulli}(\theta_1), \quad X_{2,1}, \ldots, X_{2,n_2} \overset{iid}{\sim} \text{Bernoulli}(\theta_2)$

(2) N/A

(3) $H_0 : \theta_1 - \theta_2 = 0$ or $H_0 : \theta_1 = \theta_2$

(4) $\dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\bar{x}_1(1-\bar{x}_1)}{n_1} + \frac{\bar{x}_2(1-\bar{x}_2)}{n_2}}}$

(1) $X_{1,1}, \ldots, X_{1,n_1} \overset{iid}{\sim} f_1(\theta_1, \sigma_1^2), \quad X_{2,1}, \ldots, X_{2,n_2} \overset{iid}{\sim} f_2(\theta_2, \sigma_2^2)$
where f_1 and f_2 are either unknown or otherwise non-normal PDF's or PMF's

(2) finite means and variances for both population DGP's

(3) $H_0 : \theta_1 - \theta_2 = 0$ or $H_0 : \theta_1 = \theta_2$

(4) $\dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

(d) [4 pt / 16 pts] Provide a setting where you would use an **approximate two sample T** test. Make sure to indicate (1) the DGP of the samples from the two populations, (2) what is known about the parameter(s) if anything, (3) the null hypothesis and (4) the test statistic. You do not need to state the statistic's null distribution.

(1) $X_{1,1}, \ldots, X_{1,n_1} \overset{iid}{\sim} \mathcal{N}(\theta_1, \sigma_1^2), \quad X_{2,1}, \ldots, X_{2,n_2} \overset{iid}{\sim} \mathcal{N}(\theta_2, \sigma_2^2)$

(2) nothing is known

(3) $H_0 : \theta_1 - \theta_2 = 0$ or $H_0 : \theta_1 = \theta_2$

(4) $\dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

Note: this is the case where the test statistic is drawn from the Fisher-Behrens distribution which can be approximated by the Welch-Satterthwaite T distribution.

(e) [4 pt / 20 pts] Provide a setting where you would use an **exact two sample F** test. Make sure to indicate (1) the DGP of the samples from the two populations, (2) what is known about the parameter(s) if anything, (3) the null hypothesis and (4) the test statistic. You do not need to state the statistic's null distribution.

(1) $X_{1,1}, \ldots, X_{1,n_1} \overset{iid}{\sim} \mathcal{N}(\theta_1, \sigma_1^2), \quad X_{2,1}, \ldots, X_{2,n_2} \overset{iid}{\sim} \mathcal{N}(\theta_2, \sigma_2^2)$

(2) nothing is known

(3) $H_0 : \sigma_1^2/\sigma_2^2 = 1$ or $H_0 : \sigma_1^2 = \sigma_2^2$

(4) $\dfrac{s_1^2/n_1}{s_2^2/n_2}$

Problem 2 Consider the following data with sample size $n = 50$:

2.22	2.44	4.11	2.58	2.79	3.98	2.40	2.99	3.82	3.84
3.07	6.00	2.48	5.07	2.39	3.96	2.68	2.82	4.17	4.53
1.78	4.40	3.99	1.83	6.78	26.53	6.49	2.61	3.13	3.09
3.67	8.63	3.22	3.18	4.15	2.56	3.27	2.66	3.61	3.03
4.61	0.04	3.91	1.05	3.42	2.16	2.68	1.13	6.75	10.15

Useful summary statistics are as follows: $\bar{x} = 4.06$ and $s^2 = 13.75$. We do not know the model that generated this data so we wish to select one using the AIC metric. On the top of the following page is a table of models, their number of parameters denoted K and AIC metrics below. They are sorted in order from lowest AIC to highest AIC (but one is missing).

(a) [2 pt / 22 pts] Which is the best fitting iid rv model according to the AIC metric?

Generalized Student's T

iid rv model name (DGP)	K	AIC
Generalized Student's T	2	208.5
Cauchy	2	208.8
Gumbel	2	221.8
Generalized Logistic	3	223.7
Laplace	2	224.3
Gamma	2	229.7
Weibull	2	235.0
Logistic	2	237.0
Exponential	1	242.1
Gompertz	2	265.8
Normal	2	?
Frechet	2	291.4

(b) [8 pt / 30 pts] Compute the iid normal model's AIC to one decimal point.

First we note that $\hat{\hat{\theta}}^{MLE} = \bar{x} = 4.06$ and $\hat{\hat{\sigma}}^{2MLE} = \frac{n-1}{n}s^2 = \frac{49}{50} \cdot 13.75 = 13.475$. Using these quantities we can compute AIC from scratch:

$$\begin{aligned}
AIC &= -2\ell\left(\hat{\hat{\theta}}^{MLE}, \hat{\sigma}^{2MLE} ; \boldsymbol{x}\right) + 2k \\
&= -2\ln\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\hat{\hat{\sigma}}^{2MLE}}} e^{-1/\left(2\hat{\hat{\sigma}}^{2MLE}\right)}(x_i - \hat{\hat{\theta}}^{MLE})^2\right) + 2k \\
&= n\ln(2\pi) + n\ln\left(\hat{\sigma}^{2MLE}\right) + \left(\hat{\sigma}^{2MLE}\right)^{-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 + 2k \\
&= n\ln(2\pi) + n\ln\left(\hat{\sigma}^{2MLE}\right) + \frac{(n-1)s^2}{\hat{\hat{\sigma}}^{2MLE}} + 2k \\
&= 50\ln(6.283) + 50\ln(13.475) + \frac{49 \cdot 13.75}{13.475} + 2 \cdot 2 \\
&= 275.9
\end{aligned}$$

Problem 3 Every time your professor leaves his apartment, he plays a game. When the door swings shut, he holds his key in middair and tries to estimate where the keyhole is. As the door shuts, the key either goes in the keyhole, or it doesn't. Let θ denote the probability of succeeding in this worthwhile endeavor.

(a) [3 pt / 33 pts] Out of the past 17 tries, he hasn't managed to succeed. Using Laplace's prior of indifference, provide the three Bayesian point estimates for θ as exact as possible.

Under the uniform prior $f(\theta \,|\, \boldsymbol{x}) = \text{Beta}\left(\sum x_i + 1,\, n - \sum x_i + 1\right) = \text{Beta}\left(1,\, 18\right)$. Thus,

$$\hat{\theta}^{\text{MMSE}} = \frac{1}{19}, \quad \hat{\theta}^{\text{MMAE}} = \texttt{qbeta}(0.5, 1, 18), \quad \hat{\theta}^{\text{MAP}} = 0,$$

(b) [2 pt / 35 pts] Which of the point estimates above is unrealistic and why?

$\hat{\theta}^{\text{MAP}} = 0$ because it's unrealistic to estimate that this event is impossible.

He collects more data. Of the past 87 tries, he now manages 7 successes. Use this dataset and Laplace's prior of indifference for the remaining questions in this problem.

(c) [6 pt / 41 pts] Assume Laplace's prior of indifference, compute $\mathbb{B}\text{ias}\left[\hat{\theta}^{\text{MMSE}}\right]$.

$$\mathbb{B}\text{ias}\left[\hat{\theta}^{\text{MMSE}}\right] = \mathbb{E}\left[\frac{\sum X_i + 1}{n+2}\right] - \theta = \frac{\mathbb{E}\left[\sum X_i\right] + 1}{n+2} - \frac{(n+2)\theta}{n+2} = \frac{n\theta + 1 - n\theta - 2\theta}{n+2} = \frac{1-2\theta}{n+2}$$

(d) [4 pt / 45 pts] Provide a 95% CI for θ to three decimal places.

$$CI_{\theta, 95\%} = \left[\bar{x} + \pm 1.96 \cdot \sqrt{\frac{\bar{x}(1-\bar{x})}{n}}\right] = \left[\frac{7}{87} + \pm 1.96 \cdot \sqrt{\frac{\frac{7}{87} \cdot \frac{80}{87}}{87}}\right] = [0.023, 0.138]$$

(e) [2 pt / 47 pts] Is the CI in the previous problem within the parameter space Θ? Circle one... yes / no

(f) [4 pt / 51 pts] Find an expression for the ninety-eight percent CR for θ.

Under the uniform prior $f(\theta \,|\, \boldsymbol{x}) = \text{Beta}\left(\sum x_i + 1,\, n - \sum x_i + 1\right) = \text{Beta}\left(8,\, 81\right)$. Thus,

$$CR_{\theta, 98\%} = [\texttt{qbeta}(0.01, 8, 81),\ \texttt{qbeta}(0.99, 8, 81)]$$

(g) [4 pt / 55 pts] Find an expression for the Bayesian p-value for the test of $H_a : \theta > 10\%$.

$$p_{val} = \mathbb{P}\left(H_0 \mid \boldsymbol{x}\right) = \mathbb{P}\left(\theta \leq 0.1 \mid \boldsymbol{x}\right) = \texttt{pbeta}(0.1, 8, 81)$$

(h) [6 pt / 61 pts] For the next 100 attempts, provide a 95% interval of the number of future successes.

$$\begin{aligned} p(X_* \mid \boldsymbol{x}) &= \text{BetaBinomial}\left(n_*, \sum x_i + 1, n - \sum x_i + 1\right) \\ &= \text{BetaBinomial}\left(100, 8, 81\right) \\ PI_{X_*, 95\%} &= \left[\texttt{qbetabinom}(0.025, 100, 8, 81),\ \texttt{qbetabinom}(0.975, 100, 8, 81)\right] \end{aligned}$$

Problem 4 We have the following IQ scores measured on a small sample of $n = 5$ from a special subpopulation of interest:

```
90.6 102.8  87.5 123.9 104.9
```

Assume the mean IQ score by definition is 100 but we don't know the variance in this specific population, and that is our parameter of interest. Its MLE is calculated to be 165.77.

Also, we have a lot of previous data on subpopulations we can use to create an informative prior to make our inference better. Using this previous data, we fit a prior to be $\sigma^2 \sim \text{Invgamma}(37, 8325)$.

(a) [2 pt / 63 pts] Inference on σ^2 can only be considered generalizable and externally valid if this sample of $n = 5$ was collected how?

Via a simple random sample (SRS)

(b) [5 pt / 68 pts] The prior given in the problem header is ... circle as many as apply ...

objective / subjective / informative / uninformative / indifferent

(c) [4 pt / 72 pts] The prior given in the problem header is equivalent to seeing pseudosamples with what standard deviation (i.e., the square root of variance)?

$$\begin{aligned} \alpha &= 37 = \frac{n_0}{2} \quad \Rightarrow \quad n_0 = 74 \\ \beta &= 8325 = \frac{n_0 \sigma_0^2}{2} = 37\sigma_0^2 \quad \Rightarrow \quad \sigma = \sqrt{8325/37} = 15 \end{aligned}$$

(d) [4 pt / 76 pts] Find the probability the next IQ in this subpopulation is greater than the mean IQ *exactly*.

$$\begin{aligned} f(X_* \mid \boldsymbol{x}, \theta) &= T_{n_0+n}\left(\theta, \frac{n_0\sigma_0^2 + n\hat{\hat{\sigma}}^{2\,MLE}}{n_0+n}\right) = T_{74+5}\left(100, \frac{8325 + 5 \cdot 165.77}{74+5}\right) \\ &= T_{79}\left(100, 115.9\right) \\ \mathbb{P}\left(X_* > 100 \mid \boldsymbol{x}, \theta\right) &= 0.5 \end{aligned}$$

(e) [6 pt / 82 pts] Find an expression for the probability the next IQ in this subpopulation is greater than 130.

From the previous question, we found that $f(X_* \mid \boldsymbol{x}, \theta) = T_{79}\left(100, 115.9\right)$ thus

$$\mathbb{P}\left(X_* > 130 \mid \boldsymbol{x}, \theta\right) = 1 - \texttt{pt.scaled}(130, 79, 100, \sqrt{115.9})$$

Problem 5 Let

$$X_1, \ldots, X_n \overset{iid}{\sim} \text{Log}\mathcal{N}\left(\theta,\, \sigma^2\right) := \frac{1}{\sqrt{2\pi\sigma^2 x^2}} e^{-\frac{1}{2\sigma^2}(\ln(x)-\theta)^2} \mathbb{1}_{x>0} \text{ with } \sigma^2 \text{ known}$$

(a) [9 pt / 91 pts] Find Jeffrey's prior for θ if proper. If it is not proper, find what it is proportional to.

$$\begin{aligned}
\mathcal{L}(\theta\,;\,\boldsymbol{x}, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2 x_i^2}} e^{-\frac{1}{2\sigma^2}(\ln(x)-\theta)^2} \mathbb{1}_{x_i>0} \\
\ell(\theta\,;\,\boldsymbol{x}, \sigma^2) &= -\frac{n}{2}\ln\left(2\pi\sigma^2\right) - \sum_{i=1}^{n} \ln\left(x_i\right) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} (\ln\left(x_i\right) - \theta)^2\right) + \ln\left(\prod_{i=1}^{n} \mathbb{1}_{x_i>0}\right) \\
\ell'(\theta\,;\,\boldsymbol{x}, \sigma^2) &= -\frac{1}{2\sigma^2}\frac{d}{d\theta}\left[\sum_{i=1}^{n} \ln\left(x_i\right)^2 - 2\theta\sum_{i=1}^{n} \ln\left(x_i\right) + n\theta^2\right] \\
\ell'(\theta\,;\,\boldsymbol{x}, \sigma^2) &= \frac{\sum_{i=1}^{n} \ln\left(x_i\right)}{\sigma^2} - \frac{n\theta}{\sigma^2} \\
\ell''(\theta\,;\,\boldsymbol{x}, \sigma^2) &= -\frac{n}{\sigma^2} \\
I(\theta) &= \mathbb{E}\left[-\ell''(\theta\,;\,\boldsymbol{x}, \sigma^2)\right] = \frac{n}{\sigma^2} \\
f_J(\theta) &\propto \sqrt{I(\theta)} = \sqrt{\frac{n}{\sigma^2}} \propto 1
\end{aligned}$$

Problem 6 Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{U}(0, \theta) := \frac{1}{\theta}\mathbb{1}_{x \in (0,\theta)}$. The conjugate prior for this DGP is:

$$\theta \sim \text{ParetoI}(k, \lambda) := \frac{\lambda k^\lambda}{\theta^{\lambda+1}}\mathbb{1}_{\theta > k}.$$

(a) [9 pt / 100 pts] Show that the posterior for θ is ParetoI and find its parameters.

$$\begin{aligned}
f(\theta \mid \boldsymbol{x}) \propto f(\boldsymbol{x} \mid \theta) k(\theta) &\propto \left(\prod_{i=1}^n \frac{1}{\theta}\mathbb{1}_{x_i \in (0,\theta)}\right) \frac{1}{\theta^{\lambda+1}}\mathbb{1}_{\theta > k} \\
&\propto \frac{1}{\theta^n}\left(\prod_{i=1}^n \mathbb{1}_{x_i > 0}\mathbb{1}_{x_i < \theta}\right) \frac{1}{\theta^{\lambda+1}}\mathbb{1}_{\theta > k} \\
&\propto \frac{1}{\theta^n}\left(\prod_{i=1}^n \mathbb{1}_{x_i < \theta}\right) \frac{1}{\theta^{\lambda+1}}\mathbb{1}_{\theta > k} \\
&\propto \frac{1}{\theta^{n+\lambda+1}}\mathbb{1}_{\theta > x_1}\mathbb{1}_{\theta > x_2} \cdot \ldots \cdot \mathbb{1}_{\theta > x_n}\mathbb{1}_{\theta > k} \\
&\propto \frac{1}{\theta^{n+\lambda+1}}\mathbb{1}_{\theta > \max\{x_1, \ldots, x_n, k\}} \\
&\propto \text{ParetoI}(\max\{x_1, \ldots, x_n, k\}, n + \lambda)
\end{aligned}$$

Distribution of r.v.	Quantile Function	PMF / PDF function	CDF function	Sampling Function
beta	`qbeta`(p, α, β)	`d-`(x, α, β)	`p-`(x, α, β)	`r-`(α, β)
betabinomial	`qbetabinom`(p, n, α, β)	`d-`(x, n, α, β)	`p-`(x, n, α, β)	`r-`(n, α, β)
gamma	`qgamma`(p, α, β)	`d-`(x, α, β)	`p-`(x, α, β)	`r-`(α, β)
ext negative binomial	`qnbinom`(p, r, θ)	`d-`(x, r, θ)	`p-`(x, r, θ)	`r-`(r, θ)
normal	`qnorm`(p, θ, σ)	`d-`(x, θ, σ)	`p-`(x, θ, σ)	`r-`(θ, σ)
inversegamma	`qinvgamma`(p, α, β)	`d-`(x, α, β)	`p-`(x, α, β)	`r-`(α, β)
Student's T	`qt.scaled`(p, k, μ, σ)	`d-`(x, k, μ, σ)	`p-`(x, k, μ, σ)	`r-`(k, μ, σ)

Table 1: Functions from `R` (in alphabetical order) that can be used on this exam. The hyphen in colums 3, 4 and 5 is shorthand notation for the full text of the r.v. which can be found in column 2.