Final Examination: QUESTION BOOKLET

Prof. Will Fithian

Fall 2020

- The exam begins at 3:10pm and ends at 6:00pm. There is a grace period for turning in the exam until 6:10pm; after that, the exam accrues a 20-point penalty plus 20 points more for every additional 10 minutes of lateness. If you are unable to submit to Gradescope, take timestamped photos and send them to us by email as soon as you possibly can.
- Any communication with classmates or anyone else other than me during the exam, about any subject remotely related to statistics, is strictly forbidden. That includes statements like "Problem 2 is so hard!"
- The exam is open book, open notes, open lecture videos, and any general resources from the Internet (**not** any materials specifically related to this test, obviously). These are not standard problems so hunting around for the answers to them in textbooks is unlikely to be worth your time.
- **Some students are taking the exam later due to time zone issues. Do not post anything about the exam on Piazza until I post the solutions tomorrow afternoon.**
- All parts of all problems are worth 5 points. There are 20 total parts, for 100 total points.
- Be neat! If we can't read it, we can't grade it.
- You can treat any results from lecture or homework as "known," and use them in your work without rederiving them, but do make clear what result you're using.
- For a multi-part problem, you may treat results of previous parts as given (if you don't prove the result for part (a), you can still use it to solve part (b)).
- I have starred some parts which I believe are the most difficult, and which I expect most students won't necessarily be able to solve in the time allotted. They are not worth more points than the less difficult parts, so don't waste too much time on them until you're happy with your answers to the latter.
- Be careful to justify your reasoning and answers. We are primarily interested in your understanding of concepts, so show us what you know.
- You can ask questions by email to me, with [210A Exam] in the subject line, and I will respond as quickly as I can. But my answer to most questions is just "I am satisfied with the wording of the exam as written."
- Check your email every so often just in case I have to correct something.

Good luck!

1. One Poisson, two Poissons (30 points, 5 points / part).

Some useful facts / notation for this problem:

- For $\theta > 0$, the Poisson density for $X \sim \text{Pois}(\theta)$ is $\frac{\theta^x e^{-\theta}}{x!}$ on $x = 0, 1, \ldots$. The mean and variance are both θ.
- Let $P(n)$ denote the set of integers $0 \le i \le n$ with the same parity (odd/even) as n, i.e. for which $n - i$ is even:

$$P(n) = \{i \in 0, 1, \ldots, n : \ n - i \text{ is even}\},$$

so for example $P(10) = \{0, 2, 4, 6, 8, 10\}$ while $P(9) = \{1, 3, 5, 7, 9\}$.

Suppose we observe two independent random variables, with

$$X \sim \text{Pois}(\theta), \quad \text{and } Y \sim \text{Pois}(\theta^2),$$

where $\theta > 0$ is an unknown parameter.

(a) Show that the model is an exponential family and find its complete sufficient statistic.

(b) Give an explicit expression for the UMVU estimator of θ. Evaluate it when $X = Y = 2$ (give your answer as a fraction, or a decimal with at least 3 significant digits).

(c) Now suppose that you observe an i.i.d. sample of n pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ where each pair has the same distribution specified above. That is, $X_i \sim \text{Pois}(\theta)$ and $Y_i \sim \text{Pois}(\theta^2)$, independently. Give an explicit expression for the MLE $\hat{\theta}_n$ as a function of the data.

If $\sum_{i=1}^n X_i = \sum_{i=1}^n Y_i = 2n$, find the MLE for θ (give your answer as a fraction, or a decimal with at least 3 significant digits).

(d) Find the asymptotic distribution of $\hat{\theta}_n$ as $n \to \infty$. (Don't worry about checking any regularity conditions for this part).

(e) A simpler estimator for θ is

$$\tilde{\theta}_n = \frac{\overline{X}_n + \overline{Y}_n^{1/2}}{2},$$

where $\overline{X}_n = n^{-1} \sum_{i=1}^n X_i$ and $\overline{Y}_n = n^{-1} \sum_{i=1}^n Y_i$.

Find the asymptotic distribution of this estimator. Justify why it has the distribution you say and give its asymptotic relative efficiency.

(f) Now suppose we want to test our model against the alternative hypothesis that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are still i.i.d. pairs of independent Poisson random variables, but their means do not have the relationship we posited. In other words, in the expanded model

$$X_i \sim \text{Pois}(\theta), \quad Y_i \sim \text{Pois}(\lambda), \quad i = 1, \ldots, n,$$

Test $H_0 : \lambda = \theta^2$ against the alternative $H_1 : \lambda \neq \theta^2$, for large n. Suggest an asymptotic test from class or homework: give an explicit expression for the test statistic and an explicit rejection cutoff in terms of a quantile of a known distribution. (If you choose a well-known test that is appropriate for this kind of setting then you do **not** need to justify why your test has the correct null distribution in this case).

(**Hint:** there are at least three choices of asymptotic tests from class or homework; it might pay off to take a moment to consider which is easiest to carry out here).

2. A problem of limited means (20 points, 5 points / part).

Some useful facts for this problem:

- The uniform density Unif$[a, b]$ with parameters $a < b$ has density

$$\frac{1\{a \le x \le b\}}{b-a}, \quad \text{for } x \in \mathbb{R}.$$

  Its mean and variance are $(a+b)/2$ and $(b-a)^2/12$, respectively.

- The exponential distribution Exp(λ) with scale parameter λ has density

$$\frac{1}{\lambda} e^{-x/\lambda}, \quad \text{for } x > 0.$$

The Gaussian density is printed in the preamble of Problem 2.

Assume that we are in the Gaussian sequence model with

$$X_i \overset{\text{ind.}}{\sim} N(\mu_i, 1), \qquad \text{for } i = 1, \ldots, d,$$

with the additional assumption that $|\mu_i| \le \theta$ for some $\theta > 0$. Assume unless specified otherwise that θ is known.

(a) Give the MLE of $\mu_1, \ldots, \mu_d$ in this model.

(b) Give an unbiased estimator for the mean squared error of the MLE, as a function of $X_1, \ldots, X_d$ and θ.

(c) Now, suppose we introduce Bayesian assumptions: we assume additionally that $\mu_i \overset{\text{i.i.d.}}{\sim} \text{Unif}[-\theta, +\theta]$, still with θ known. Give an explicit expression for the Bayes estimator of $\mu_1, \ldots, \mu_d$ using squared error loss.

(d) Now, we relax the assumption that θ is known and introduce a hierarchical Bayesian model with an exponential hyperprior for θ:

$$\begin{aligned} \theta &\sim \text{Exp}(\lambda) \\ \mu_i \mid \theta &\overset{\text{i.i.d.}}{\sim} \text{Unif}[-\theta, +\theta], \quad i = 1, \ldots, d \\ X_i \mid \theta, \mu &\overset{\text{ind.}}{\sim} N(\mu_i, 1), \quad i = 1, \ldots, d. \end{aligned}$$

Suggest a Gibbs sampler algorithm to sample from the posterior distribution of $(\theta, \mu_1, \ldots, \mu_d)$. Give the update rules explicitly.

3. Gamma palooza (25 points, 5 points / part).

Some useful facts for this problem:

- For shape parameter $k > 0$ (not necessarily an integer) and scale parameter $\sigma > 0$, the Gamma(k, σ) distribution has density

$$\frac{1}{\sigma^k \Gamma(k)} x^{k-1} e^{-x/\sigma}, \quad \text{for } x > 0.$$

  Its mean and variance are $k\sigma$ and $k\sigma^2$, respectively.

- The χ^2_d distribution is Gamma$(d/2, 2)$. It is usually defined when d is an integer, but the density is still a proper density for any $d > 0$. The same is true for distributions derived from the χ^2 like t or F whose "degrees of freedom" argument(s) can take on any positive real value.

Assume that we observe independent random variables X_{ij} with

$$X_{ij} \overset{\text{ind.}}{\sim} \text{Gamma}(k_i, \sigma_j), \quad \text{for } i = 1, \ldots, n \geq 2, \text{ and } j = 1, 2.$$

Unless otherwise specified, assume all k_i and σ_j are unknown and strictly positive (different parts of the problem will consider simpler submodels). Let $S_j = \sum_{i=1}^n X_{ij}$ and $M_i = X_{i1} X_{i2}$.

(a) Show that $T(X) = (S_1, S_2, M_1, \ldots, M_n)$ is a complete sufficient statistic for this model.

(b) Assume (for this part **only**) that $k_1, \ldots, k_n$ are known. Give an explicit formula for an exact equal-tailed confidence interval for σ_2/σ_1, in terms of the sufficient statistics described above and quantiles for one or more known distributions from class.

(c) Assume instead (for this part **only**) that σ_1 and σ_2 are known, and also it is known that $k_1 = k_2 = \cdots = k_n = k$, but the common value k is unknown. Suggest a UMP test of the hypothesis $H_0 : \ k = k_0$ against the alternative $H_1 : \ k > k_0$, where k_0 is generic. Give the test statistic and explain how to calculate the rejection cutoff (give an explicit recipe that anyone can follow).

(d) Suppose (for this part **only**) that $n = 2$ with all of $k_1, k_2, \sigma_1, \sigma_2$ unknown. Suggest an exact UMPU test of $H_0 : \ k_1 = k_2$ against $H_1 : \ k_1 > k_2$. Say what test statistic you would use and give a precise mathematical description of the rejection cutoff, but you do **not** need to give an explicit expression or recipe for how to calculate it.

(e) (*) Drop all assumptions from previous parts, so n is arbitrary and no parameters of the model are known.

Suppose that we begin doubting the validity of our Gamma model, and we want to generalize it to replace the Gamma family with a generic scale family:

$$X_{ij} \overset{\text{i.i.d.}}{\sim} G_i(x/\sigma_j),$$

for a generic, unknown, continuous distribution function G_i that puts all its mass on positive values of x (i.e., $G_i(0) = 0$). We want to guarantee Type I error control no matter what $G_1, \ldots, G_n$ are.

Explain how to calculate an exact 95% confidence interval for σ_2/σ_1. Your interval must be nontrivial; we will not award any points for answers like "flip a coin and cover the entire parameter space with probability 95%."

(**Hint:** This problem is closely related to testing $H_0 : \sigma_1 = \sigma_2$ against $H_1 : \sigma_1 > \sigma_2$. The testing problem might be easier to think about at first, and partial credit will be awarded for making progress on it.)

4. Apocalypse τ (25 points, 5 points / part).

Some useful facts for this problem:

- For $\sigma^2 > 0$ and $\mu \in \mathbb{R}$, the Gaussian density for $X \sim N(\mu, \sigma^2)$ is

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad \text{for } x \in \mathbb{R}.$$

  Its mean and variance are μ and σ^2.

Assume we observe i.i.d. pairs (X_i, Y_i) for $i = 1, \ldots, n$, where $X_1, \ldots, X_n \in \mathbb{R}^k$ are sampled from a known density $q(x)$ and Y_i are real numbers with

$$Y_i = f_\tau(X_i) + \varepsilon_i, \quad \text{where } \varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2).$$

Assume the errors $\varepsilon_1, \ldots, \varepsilon_n$ are independent of $X_1, \ldots, X_n$.

The parameters $\tau \in [-1, 1]$ and $\sigma^2 > 0$ are fixed and unknown, but the real-valued function $f_\tau(x)$ is known up to its parameter τ.

Assume that

- $f_\tau(x)$ is infinitely differentiable *with respect to* τ, with first and second derivatives

$$g_\tau(x) = \frac{\partial f}{\partial \tau}(x), \quad \text{and } h_\tau(x) = \frac{\partial^2 f}{\partial \tau^2}(x).$$

- $g_\tau(x) > 0$ for all τ and x.
- $|g_\tau(x)|, |h_\tau(x)| \leq 1$ for all τ and x.

(a) Assume (for this part **only**) that X_i are fixed instead of random, while the errors still have the same distribution, $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$.

Consider testing $H_0 : \tau = 0$ against the alternative $H_1 : \tau \neq 0$ using the test statistic

$$T = \frac{\sum_{i=1}^n g_0(X_i)(Y_i - f_0(X_i))}{\hat{\sigma}\left(\sum_{i=1}^n g_0(X_i)^2\right)^{1/2}},$$

where

$$\hat{\sigma}^2 = \frac{1}{d}\left[\sum_{i=1}^n (Y_i - f_0(X_i))^2 - \frac{\left[\sum_{i=1}^n g_0(X_i)(Y_i - f_0(X_i))\right]^2}{\sum_{i=1}^n g_0(X_i)^2}\right]$$

What number should we plug in for d? Give the distribution of T under the null, and justify your answer.

(b) Now go back to assuming that X_i are random, sampled i.i.d. from an unknown distribution. Show that the test from part (a) still works; i.e. its distribution under the null is independent of $X_1, \ldots, X_n$.

(c) Assume (for this part **only**) that σ^2 is known. Show that the MLE $\hat{\tau}_n$ is consistent for τ as $n \to \infty$. (For full credit, please check appropriate conditions).

(d) Continue to assume (for this part **only**) that σ^2 is known. Assuming the MLE is consistent, and $\tau \in (-1, 1)$ (i.e. not at the boundary of the parameter space), find its asymptotic distribution as $n \to \infty$. (You do not need to check conditions for this).

(e) (*) Suppose we add an intercept to the model, so

$$Y_i \mid X_1, \ldots, X_n \overset{\text{ind.}}{\sim} N(\alpha + f_\tau(X_i), \sigma^2).$$

Can we still estimate τ consistently as $n \to \infty$ using maximum likelihood? Prove or give a counterexample.