University of Washington  
Biostatistics First-Year Theory Exam  
Statistics MS Theory Examination  
15-16 June 2021

Instructions

• This is a take-home, open-note exam. You have a total of 4 hours to complete and upload your exam: 3 hours to complete the exam and 1 hour to upload the exam. You may choose whatever 4-hour window works best for you within the following 24-hour period. The clock starts as soon as you access the exam, via the Canvas Quiz feature. The exam ends for all on June 16 (Wednesday) at 9AM PT.

• You should have received an exam ID number to be used for identifying yourself on your written exam solutions. Write your exam ID number at the top of every page of your exam solutions. Do not list your name or UW student number on your exam.

• During the chosen 4-hour exam period, you may:
  – Consult your class notes and materials from STAT 512/513, including notes from the instructor, homework exercises and keys, and notes you have written yourself.
  – Consult the textbook by Casella and Berger (Statistical Inference), either in printed or electronic form.

• During the chosen 4-hour exam period, you may not:
  – Consult any other references materials, except those listed above.
  – Communicate in any way, with anyone other than the Graduate Program or MS Theory Exam committee, about the questions in the exam or topics in STAT 512/513.
  – Search online or use electronic tools (e.g. Mathematica) to answer any of the questions. The committee will also make sure that no questions in the exam can be easily solved via a simple online search or use of electronic tools.

• There are six questions, each with multiple sub-questions. Answer any five of them, and indicate which five you have attempted in the solutions document that you will upload to Canvas. No extra credit is available for answering all six, under any circumstances. If you answer all six questions and do not specify which of the six questions should be graded, the exam committee will grade the first five questions only.

• Each question is worth 20 points, although the questions vary in difficulty.

• Your answers may be hand-written or word-processed: neither is “better”, and you should do what best fits your situation:
For hand-written solutions: Use a new page for each question, and label every page at the top with your exam ID number, along with the question and sub-question numbers, e.g. “Q1 (a)”. Write clearly, and use a sufficiently dark pen or pencil that your solutions are still legible after being scanned/photographed. Submit each question as a separate document, i.e., 5 documents for 5 questions.

For word-processed solutions: Use at least 11pt font sizes. Submit each question as a separate document, i.e., 5 documents for 5 questions. At the top of each page give your Exam ID number and the question number: give sub-question numbers as you go through your answers.

- Your exam solutions should be converted to PDF files (one PDF file per question, i.e., 5 documents for 5 questions) and uploaded to the Canvas website by the end of the chosen 4 hour window, and no later than 9AM PT on June 16 (Wednesday).

- Problems or questions during the exam:

  - If you have difficulties downloading the exam or uploading your solutions to Canvas, please send an email to the Graduate Program at the email address mstheory21-logistics@stat.washington.edu. Please document what happened by making notes, saving screens or messages from the event.

  - You can email the theory exam committee at the email address fanghan@uw.edu for questions of clarification. These are unusual, but they can happen – if some ambiguity slipped through the (extensive) vetting of the questions, for example. Note that this is different from a student not understanding the question, for which help will not be provided. Any content in the email that may violate the graders’ blindness (e.g., the way you try to solve the question) is strictly forbidden. Your answers should also not mention that you received clarification.

  - The committee will reply to clarification questions during the 24-hour exam period, except for 11PM-7AM PT. The reply will be sent only to the student who asked the question, and will come from fanghan@uw.edu. Only questions that the committee deems appropriate will be answered by the committee, i.e., questions for which the response does not give an unfair advantage to the student.

Good luck!
1. Suppose that $X$ is binomial based on $m$ independent trials and success probability $\theta_X$. Assume $Y$ is binomial based on $n$ independent trials with success probability $\theta_Y$ and is independent of $X$.

(a) What are the UMVUEs of $\theta_Y$ and $\theta_X$? Call them $\hat{\theta}_Y$ and $\hat{\theta}_X$.

(b) Suppose it is known that $\theta_X \leq \theta_Y$, but the parameters are otherwise unknown. Is $\hat{\theta}_Y - \hat{\theta}_X$ admissible for $\theta_Y - \theta_X$ under squared error loss? Explain your answer.
2. Let $X_1, \ldots, X_n$ be independent and identically distributed following $N(\mu, \sigma^2)$, with $\mu$ unknown and $\sigma$ known. Consider $t \neq 0$ to be a known real constant.

(a) Find the UMVUE of $e^{t\mu}$.

(b) Derive the variance of the UMVUE.

(c) Derive the Cramer-Rao lower bound of $e^{t\mu}$, and show that it is less than the variance of the UMVUE.

(d) Show that the ratio of the variance of the UMVUE to the Cramer-Rao lower bound approaches 1 as $n \to \infty$. 

3. Let $f(x)$ be a function that we can evaluate at any point $x \in \mathbb{R}$. Let $Z$ be a random variable with a PDF $q(z)$. Suppose that we are interested in estimating

$$\theta = E[f(Z)] = \int_{-\infty}^{\infty} f(z) q(z) dz.$$ 

(a) If we can sample from $q(z)$, a simple estimate of $\theta$ is as follows. We generate $Z_1, \cdots, Z_N$ that are IID from $q$ and estimate $\theta$ via

$$\hat{\theta} = \frac{1}{N} \sum_{\ell=1}^{N} f(Z_\ell).$$

Show that $\hat{\theta}$ is an unbiased estimator of $\theta$.

(b) When it is difficult to sample from $Z$’s distribution, we may instead sample real-valued $W_1, \ldots, W_N$ independently from a distribution with density $p(w)$, where $p(w) > 0$ for all $w$. We then use the estimate

$$\tilde{\theta} = \frac{1}{N} \sum_{\ell=1}^{N} \frac{f(W_\ell) q(W_\ell)}{p(W_\ell)}.$$

Show that $E[\tilde{\theta}]$ is unbiased for $\theta$, and that

$$\text{Var}[\tilde{\theta}] = \frac{1}{N} \int_{-\infty}^{\infty} \frac{[f(w)q(w) - \theta p(w)]^2}{p(w)} dw.$$ 

(c) To illustrate some properties of this approach, consider estimating the mean (i.e., choosing $f(x) = x$) of a Gamma distribution (i.e., $Z \sim \Gamma(\alpha, \beta)$ with $\alpha > 0$ and $\beta > 0$ being shape and rate parameters, respectively), where $W$ is another Gamma distribution (i.e., $W \sim \Gamma(\alpha_0, \beta_0)$ with $\alpha_0 > 0$ and $\beta_0 > 0$). Please state the variance of $\tilde{\theta}$ in terms of $N, \alpha, \alpha_0, \beta, \beta_0$ and show that it is finite when

$$2 + 2\alpha > \alpha_0 \text{ and } 2\beta > \beta_0.$$ 

Note: a Gamma distribution with shape $\alpha > 0$ and rate $\beta > 0$ has density $\beta^\alpha z^{\alpha-1} e^{-\beta z}/\Gamma(\alpha)$, mean $\alpha/\beta$, and variance $\alpha/\beta^2$. 
4. In meta-analysis, summary results from several similar studies are used together, to give an overall estimate and corresponding inference.

Studies $i = 1, 2, \ldots, k$ provide point estimates $\hat{\beta}_i$, for which they also provide an estimated variance $\hat{\sigma}_i^2$. We assume that $\hat{\beta}_1, \ldots, \hat{\beta}_k, \hat{\sigma}_1^2, \ldots, \hat{\sigma}_k^2$ are mutually independent.

(a) Denoting $E[\hat{\beta}_i] = \beta_i$ and $\text{Var}[\hat{\beta}_i] = \sigma_i^2$, show that

$$
E \left[ \sum_{i=1}^{k} \left( \frac{\sigma_i^{-2}}{\sum_{j=1}^{k} \sigma_j^{-2}} \hat{\beta}_i \right) \right] = \sum_{i=1}^{k} \left( \frac{\sigma_i^{-2}}{\sum_{j=1}^{k} \sigma_j^{-2}} \beta_i \right),
$$

$$
\text{Var} \left[ \sum_{i=1}^{k} \left( \frac{\sigma_i^{-2}}{\sum_{j=1}^{k} \sigma_j^{-2}} \hat{\beta}_i \right) \right] = \frac{1}{\sum_{i=1}^{k} \sigma_i^{-2}}.
$$

(b) Show that

$$
\text{Var} \left[ \sum_{i=1}^{k} \left( \frac{\hat{\sigma}_i^{-2}}{\sum_{j=1}^{k} \hat{\sigma}_j^{-2}} \hat{\beta}_i \right) \right] = E \left[ \sum_{i=1}^{k} \left( \frac{\sigma_i^2}{\hat{\sigma}_i^4} \right) \right] + \text{Var} \left[ \sum_{i=1}^{k} \left( \frac{\hat{\sigma}_i^{-2}}{\sum_{j=1}^{k} \hat{\sigma}_j^{-2}} \hat{\beta}_i \right) \right].
$$

(c) There are two terms on the right hand side of the equation in (b), an expectation (which approaches $1/\sum_{i=1}^{k} \sigma_i^{-2}$ in large samples) and a variance. Suppose the $\beta_i$'s are all equal; what is the value of the variance term?
5. This question considers the joint distribution of two binary-valued variables $Y_1$ and $Y_2$. We define

$$p_{00} = P[Y_1 = Y_2 = 0], \quad p_{01} = P[Y_1 = 0, Y_2 = 1],$$
$$p_{10} = P[Y_1 = 1, Y_2 = 0], \quad p_{11} = P[Y_1 = Y_2 = 1],$$

where $p_{00} + p_{01} + p_{10} + p_{11} = 1$.

(a) The joint distribution has exchangeability if

$$P[Y_1 = y_1, Y_2 = y_2] = P[Y_1 = y_2, Y_2 = y_1]$$

for any specific values $y_1 \in \{0, 1\}$ and $y_2 \in \{0, 1\}$. Show how, for an exchangeable distribution, $\{p_{00}, p_{01}, p_{10}, p_{11}\}$ can be written only in terms of

$$q_0 = P[Y_1 + Y_2 = 0], \quad q_1 = P[Y_1 + Y_2 = 1], \quad q_2 = P[Y_1 + Y_2 = 2],$$

and show $q_0 + q_1 + q_2 = 1$.

(b) Consider the distribution of $Y_1, Y_2$ when a third latent variable is present and denoted as $Z$. Specifically,

$$Y_i \mid Z = z \overset{i.i.d.}{\sim} \text{Bernoulli}(z), \quad i = 1, 2$$
$$Z \sim F$$

for some specified distribution function $F$ of either a PMF or a PDF with support in $[0, 1]$. Show that the resulting joint distribution of $Y_1, Y_2$ is exchangeable. (Here we say $X \sim \text{Bernoulli}(z)$ if $P(X = 1) = 1 - P(X = 0) = z$.)

(c) Consider the situation in (b) where

(i) $F$ is Unif(0, 1);
(ii) $F$ is a discrete distribution taking value $z = 0, 1/2, 1$ with probabilities 1/6, 2/3, 1/6 respectively.

Show that the resulting exchangeable joint distributions for $Y_1, Y_2$ are identical.

(d) By using the Cauchy-Schwarz inequality or otherwise, show that there are exchangeable distributions for $Y_1, Y_2$ that cannot be represented in the form given in (b), for any choice of $F$ of either a PMF or a PDF with support in $[0, 1]$. 
6. Consider \( \{\epsilon_t, t = 1, \ldots, n\} \) to be independent and identically distributed following \( N(0, \sigma^2) \). Let \( X_0 \sim N(0, \sigma^2/(1-\phi^2)) \) be independent of \( \{\epsilon_t, t = 1, \ldots, n\} \) and

\[
X_t = \phi X_{t-1} + \epsilon_t, \quad t = 1, \ldots, n,
\]

with \( |\phi| < 1 \) and \( \sigma^2 \in (0, \infty) \).

(a) Calculate the mean and variance of each \( X_t, t = 1, \ldots, n \).

(b) Derive the distribution of each \( X_t, t = 1, \ldots, n \).

(c) Derive the log likelihood function of \( X_0, X_1, \ldots, X_n \) and accordingly derive a 3-dimensional sufficient statistic for \( (\phi, \sigma^2) \). (Hint: you may use the fact that \( f(x_1, \ldots, x_n) = f(x_n|x_1, \ldots, x_{n-1})f(x_{n-1}|x_1, \ldots, x_{n-2}) \cdots f(x_2|x_1)f(x_1) \) for any density function \( f \).

(d) Is the 3-dimensional sufficient statistic you derived in (c) minimal sufficient? Justify your answer.

(e) Derive the estimator of \( \phi \) that maximizes the conditional joint distribution of \( X_1, \ldots, X_n \) given \( X_0 \). (Note that the maximum may lie on the boundary of values \( \phi = -1 \) or \( \phi = 1 \).)