Math 369 / 690 Fall 2021
Midterm Examination One

Professor Adam Kapelner

Thursday, October 7, 2021

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.

By taking this exam, you acknowledge and agree to uphold this Code of Academic Integrity.

Instructions

This exam is 70 minutes (variable time per question) and closed-book. You are allowed **one** page (front and back) of a "cheat sheet", blank scrap paper and a graphing calculator. Please read the questions carefully. I recommend answering all questions that are easy first and then circling back to work on the harder ones. No food is allowed, only drinks.

Problem 1 [6min] (and 6min will have elapsed) These are conceptual questions about statistical inference and sampling.

- [14 pt / 14 pts] Record the letter(s) of all the following that are **true** in general. At least one will be true.

(a) Beginning with a sample of data, it may be difficult to clearly define the population.

(b) Beginning with a well-defined population, it may be difficult to sample uniformly from this population.

(c) In a simple random sample, the probability of each item in the population being in the sample is the same.

(d) Samples from a real-world population are never truly independent.

(e) If you had access to the entire population, population parameter values would be known exactly.

(f) Samples can sometimes be thought of realizations from a process defined by random variables, the DGP.

(g) The data $x_1, \ldots, x_n$ can be sampled only after assuming a DGP explicitly.

(h) The goal of point estimation is to find an approximation of the true value of θ.

(i) The goal of theory testing is to find an approximation of the true value of θ.

(j) The goal of statistical inference (in general) is to learn the true values of $x_1, \ldots, x_n$.

(k) The goal of statistical inference (in general) is to learn the true values of $x_1, \ldots, x_N$.

(l) Statistical estimates can only be computed after the data $x_1, \ldots, x_n$ is sampled.

(m) The sampling distribution can only be known after the data's values $x_1, \ldots, x_n$ are known.

(n) The estimator $\bar{X}$ is unbiased for all DGP's (as long as the DGP's expectation exists).

Your answer will consist of a lowercase string (e.g. `aebgd`) where the order of the letters does not matter.

Problem 2 [15min] (and 21min will have elapsed) Consider the DGP $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$ and the estimators $\hat{\theta}_B = \bar{X} + 0.1$ and $\hat{\theta}_A = \bar{X} + 0.1/n$.

- [19 pt / 33 pts] Record the letter(s) of all the following that are **true** in general. At least one will be true.

(a) The estimator $\hat{\theta}_B$ is unbiased for all $\theta \in \Theta$.

(b) The estimator $\hat{\theta}_B$ is unbiased for at least one $\theta \in \Theta$.

(c) The estimator $\hat{\theta}_B$ is asymptotically unbiased for all $\theta \in \Theta$.

(d) Every possible estimate $\hat{\hat{\theta}}_B$ is in Θ.

(e) At least one estimate $\hat{\hat{\theta}}_B$ is in Θ.

(f) The L1 loss for $\hat{\theta}_B$ is always ≥ 0.1.

(g) The L2 loss for $\hat{\theta}_B$ is always ≥ 0.1.

(h) The L1 loss for $\hat{\theta}_B$ is always ≥ 0.

(i) The L2 loss for $\hat{\theta}_B$ is always ≥ 0.

(j) The risk under L1 loss for $\hat{\theta}_B$ is 0.1.

(k) The risk under L2 loss for $\hat{\theta}_B$ is $\mathbb{E}\left[(\bar{X} - \theta)^2 + 0.2(\bar{X} - \theta)\right] + 0.01$.

(l) The risk under L2 loss for $\hat{\theta}_B$ is $\theta(1-\theta)/n + 0.01$.

(m) $\text{MSE}[\hat{\theta}_B] = \theta(1-\theta)/n + 0.01$.

(n) The sup risk under L2 loss for $\hat{\theta}_B$ is $(4n)^{-1}$.

(o) The estimator $\hat{\theta}_A$ is unbiased for all $\theta \in \Theta$.

(p) The estimator $\hat{\theta}_A$ is asymptotically unbiased for all $\theta \in \Theta$.

(q) $\text{MSE}[\hat{\theta}_A] = \theta(1-\theta)/n + 0.01/n^2$.

(r) The sup risk under L2 loss for $\hat{\theta}_A$ is $(4n)^{-1}$.

(s) The loss function $\ell(\hat{\hat{\theta}}, \theta) = |\hat{\hat{\theta}} - \theta|^{0.5}$ is a valid loss function for these both $\hat{\theta}_B$ and $\hat{\theta}_A$.

Your answer will consist of a lowercase string (e.g. `aebgd`) where the order of the letters does not matter.

Problem 3 [10min] (and 31min will have elapsed) The following are conceptual question about hypothesis testing.

- [17 pt / 50 pts] Record the letter(s) of all the following that are **true** in general. At least one will be true.

(a) If you wish to prove a theory to the wide world, it is best to let detractors prove your theory wrong and if they cannot, your theory will be widely accepted.

(b) If the value of θ were to be known, there would be no reason to do statistical hypothesis testing.

(c) Using a sample of data, it is possible to absolutely prove (or disprove) a theory.

(d) If you are trying to prove that your investment idea is correct, the null hypothesis should be that your investment idea is incorrect.

(e) Type I errors are only possible if H_0 is true.

(f) When doing a power calculation, we assume H_0 is true.

(g) The type I error rate is unaffected by the sample size.

(h) $\mathbb{P}\left(\hat{\theta} \in \text{RET}\right) = \alpha$.

(i) If you retain H_0 it is impossible to know if you made a Type I error.

(j) If you retain H_0 it is impossible to know if you made a Type II error.

(k) The lower Fisher's p-value is, the more "statistically significant" the result of the test becomes.

(l) If $\alpha = 0$, power is always 1.

(m) In the one-sample z test, if α decreases, power decreases.

(n) In the one-sample z test, if n decreases, power decreases.

(o) In the one-sample z test, if σ^2 decreases, power decreases.

(p) In the one-sample z test, if $H_0 : \theta \leq 3.6$, $H_a : \theta > 3.6$, this test would be "one-sided" and "right-tailed".

(q) In the one-sample z test, if $H_0 : \theta \leq 3.6$, $H_a : \theta > 3.6$, $\alpha = 5\%$, $n = 500$ and $\sigma^2 = 1.3$, power can be computed.

Your answer will consist of a lowercase string (e.g. `aebgd`) where the order of the letters does not matter.

Problem 4 [12min] (and 43min will have elapsed) We flip a fair coin 7 times and record the proportion of heads $\hat{\hat{\theta}}$. Let θ denote the probability of flipping heads. Below is a table of the PMF of the Binomial $(7,\, 0.5)$ rounded to three digits.

x	0	1	2	3	4	5	6	7
$p(x)$	0.008	0.055	0.164	0.273	0.273	0.164	0.055	0.008

- [13 pt / 63 pts] Record the letter(s) of all the following that are **true** in general. At least one will be true.

(a) The sample size is $n = 7$.

Assume for the following questions that we are trying to prove that the coin is unfair.

(b) $H_0 : \theta = 0.5$

(c) The size 5% is not attainable for our test scenario.

(d) If the level was 1%, the rejection region for $\hat{\hat{\theta}}$ would be $\{0, 1\}$.

(e) If the level was 7.5%, the rejection region for $\hat{\hat{\theta}}$ would be $\left\{0, \frac{1}{7}, \frac{6}{7}, 1\right\}$.

(f) If the level was 5%, and $\hat{\hat{\theta}} = 4/7$, the null hypothesis would be retained.

(g) The test in (f) is called a "Binomial Exact Test".

Assume for the following questions that we are trying to prove that the coin is biased towards heads.

(h) $H_0 : \theta = 0.5$

(i) If the level was 5%, the rejection region for $\hat{\hat{\theta}}$ would be $\left\{0, \frac{1}{7}, \frac{6}{7}, 1\right\}$.

(j) If the level was 5%, and $\hat{\hat{\theta}} = 4/7$, the null hypothesis would be retained.

(k) If the level was 5%, and $\hat{\hat{\theta}} = 4/7$, Fisher's p-val would be 22.7%.

(l) If the level was 1%, and $\hat{\hat{\theta}} = 5/7$, Fisher's p-val would be larger than Fisher's p-val when $\hat{\hat{\theta}} = 4/7$.

(m) If the level was 1%, and $\hat{\hat{\theta}} = 5/7$, you have enough information to compute power of this test.

Your answer will consist of a lowercase string (e.g. `aebgd`) where the order of the letters does not matter.

Problem 5 [13min] (and 56min will have elapsed) We suspect that pollution in a certain river is affecting the nutrition for a Atlantic Salmon and thereby making their adult size less than what is historically expected for salmon. Salmon are known to have mean length 29in with standard deviation 0.95in and for the distribution to be normal. However, we are not sure what the standard deviation for malnourished salmon will be. We collect the following lengths of salmon from the river: $\boldsymbol{x} = \langle 25.7, 27.4, 25.3, 30.2, 27.7, 25.4, 28.0 \rangle$ where $\bar{x} = 27.1$ and $\hat{\sigma}^2 = 2.71$. We would like to prove that θ, the length of fish from the polluted river, are smaller than regular salmon via a 1-sample t-test. Below is a table of CDF values for the T distribution for many different degrees of freedom (df):

df	0.1%	0.5%	1%	5%	10%	15%	20%
6	-5.21	-3.71	-3.14	-1.94	-1.44	-1.13	-0.91
7	-4.79	-3.50	-3.00	-1.89	-1.41	-1.12	-0.90
8	-4.50	-3.36	-2.90	-1.86	-1.40	-1.11	-0.89

- [11 pt / 74 pts] Record the letter(s) of all the following that are **true** in general. At least one will be true.
  (a) The value $\bar{x} = 27.1$ is a estimate realized from an unbiased estimator.
  (b) The value $\hat{\sigma}^2 = 2.71$ is a estimate realized from an unbiased estimator.
  (c) The reason why we use a 1-sample t-test is because we know $\sigma = 0.95$.
  (d) The value of $\bar{x}$ standardized is -1.855 rounded to the nearest 3 decimals.
  (e) The value of $\bar{x}$ standardized is -2.827 rounded to the nearest 3 decimals.
  (f) The value of $\bar{x}$ standardized is -3.298 rounded to the nearest 3 decimals.
  (g) The value of $\bar{x}$ standardized is -3.054 rounded to the nearest 3 decimals.
  (h) The null hypothesis is rejected at $\alpha = 1\%$.
  (i) The null hypothesis is rejected at $\alpha = 5\%$.
  (j) Fisher's p-val for this test is likely closest to 1% among all the options in the table of CDF values for the T distribution.
  (k) Fisher's p-val for this test is likely closest to 5% among all the options in the table of CDF values for the T distribution.

Your answer will consist of a lowercase string (e.g. `aebgd`) where the order of the letters does not matter.

Problem 6 [14min] (and 70min will have elapsed) Chevalier de Mere was a French nobleman writer / philosopher who lived from 1607-1684. But he is far better known for his contribution to probability theory likely due to his compulsive gambling habit. Before notions of probabilitistic independence, complement rule, random variable theory (all the familiar 241 topics) were discovered, he wrote a letter to Blaise Pascal where he accurately conjectured that the probability of winning a certain dice game. In this dice game, you "win" if you get at least one double 6 in a roll of two dice among 24 rolls of two dice. He conjectured that the probability of winning was less than 50% and the actual probability is $\theta = 49.14\%$. Assuming each of these entire 24 rolls defines one game and he plays many games, then the DGP for winning of each game truly is $\overset{iid}{\sim}$ Bernoulli (0.4914). Assuming he kept a flawless mental record of wins and losses, how many games would he need to observe to reject the null hypothesis that $\theta \geq 50\%$ at $\alpha = 5\%$ with power of 80%? The two CDF values for the standard normal distribution that are needed are: $\Phi(-1.645) = 5\%$ and $\Phi(0.84) = 80\%$.

- [12 pt / 86 pts] Select the answer choice below whose number is *closest* to your answer.

(a) 1,000
(b) 5,000
(c) 7,500
(d) 10,000
(e) 15,000
(f) 20,000
(g) 30,000
(h) 40,000
(i) 50,000
(j) 75,000
(k) 100,000
(l) 200,000

Your answer will consist of a lowercase string (e.g. `aebgd`) where the order of the letters does not matter.