MATH 369/650 Fall 2020 Homework #1

Professor Adam Kapelner

Due by email noon Friday, September 11, 2019

(this document last updated Thursday 10th September, 2020 at 7:58pm)

Instructions and Philosophy

The path to success in this class is to do many problems. Unlike other courses, exclusively doing reading(s) will not help. Coming to lecture is akin to watching workout videos; thinking about and solving problems on your own is the actual "working out." Feel free to "work out" with others; **I want you to work on this in groups.**

Reading is still *required*. For this homework set, review Math 241 concerning random variables, support, parameter space, PMF's, CDF's, bernoulli, binomial, geometric. Then read on your own about the negative binomial, convolutions and the multinomial distribution.

The problems below are color coded: green problems are considered *easy* and marked "[easy]"; yellow problems are considered *intermediate* and marked "[harder]", red problems are considered *difficult* and marked "[difficult]" and purple problems are extra credit. The *easy* problems are intended to be "giveaways" if you went to class. Do as much as you can of the others; I expect you to at least attempt the *difficult* problems. "[MA]" are for those registered for 621 and extra credit otherwise.

This homework is worth 100 points but the point distribution will not be determined until after the due date. See syllabus for the policy on late homework.

Up to 7 points are given as a bonus if the homework is typed using LaTeX. Links to instaling LaTeX and program for compiling LaTeX is found on the syllabus. You are encouraged to use `overleaf.com`. If you are handing in homework this way, read the comments in the code; there are two lines to comment out and you should replace my name with yours and write your section. The easiest way to use overleaf is to copy the raw text from hwxx.tex and preamble.tex into two new overleaf tex files with the same name. If you are asked to make drawings, you can take a picture of your handwritten drawing and insert them as figures or leave space using the "\vspace" command and draw them in after printing or attach them stapled.

The document is available with spaces for you to write your answers. If not using LaTeX, print this document and write in your answers. I do not accept homeworks which are *not* on this printout. Keep this first page printed for your records.

NAME: _____

Problem 1

We will explore sampling in some detail herein. Some of the answers can be found by reading the Wikipedia article.

(a) [easy] Define *population* and *sample* and give the sizes of each using the notation we used in class.

(b) [easy] After selecting a sample, we take a *survey*. In general, what is a survey? (It is also called *measurement*). How does it relate to random variable models and realizations from random variables?

(c) [easy] In the survey we took in class, what were the possible values of a datum for each individual?

(d) [easy] For the survey we did in class, give three answers for the "representative population".

(e) [easy] Define *simple random sample* (SRS).

(f) [harder] Pick one of your answers from (d) and call that the population. Do you believe our class survey was an SRS? Why or why not?

(g) [harder] We did not define *sampling frame* in class. What is it and why is it important in practice?

(h) [harder] What are some problems with SRS's?

(i) [easy] Explain why assuming the population size is infinite gives you an $\overset{iid}{\sim}$ DGP when you use an SRS to sample individuals.

(j) [harder] If we do not allow the population to be infinitely sized, is the DGP identically distributed? Would it be independent? Explain.

(k) [easy] Consider a Bernoulli survey. Split a population into subgroups A, B, C, D with subpopulation sizes N_A, N_B, N_C, N_D (which all may be different quantities) and unknown number of positives of χ_A, χ_B, χ_C, χ_D (which all may be different quantities). How do you calculate θ using these expressions?

(l) [harder] Under what condition(s) would sampling from subpopulation B exclusively be representative of the whole population?

Problem 2

Here we will review estimates and estimators.

(a) [easy] Define statistic / estimate / statistical estimate.

(b) [easy] What is the difference between $\hat{\hat{\theta}}$ and $\hat{\theta}$?

(c) [easy] Is the w the same function to compute both $\hat{\hat{\theta}}$ and $\hat{\theta}$? Yes / No

(d) [difficult] For any DGP $X_1, \ldots, X_n$ which is independent and identically distributed where $\theta = \mathbb{E}[X]$, we introduced the estimator $\hat{\theta} = \bar{X}$ that can be used to infer θ. Think of another estimator that can be used to infer θ.

(e) [easy] For any DGP $X_1, \ldots, X_n$ which is identically distributed where $\theta = \mathbb{E}[X]$, prove that $\hat{\theta} = \bar{X}$ is always *unbiased*.

(f) [harder] skip

(g) [easy] If the DGP is $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\theta, \sigma^2)$, show that the the risk under squared error loss for $\hat{\theta} = \bar{X}$ is the same regardless of θ. Marked easy because it uses the previous question heavily.

(h) [difficult] [MA] Consider the scenario in 1(k). Sample subpopulation A with size n_A, sample subpopulation B with size n_B, sample subpopulation C with size n_C and sample subpopulation D with size n_D, four different quantities such that $n = n_A + n_B + n_C + n_D$. Create an unbiased estimator for θ using these four samples (and prove that it is unbiased).

(i) [harder] If the DGP is $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$ and we choose the estimator $\hat{\theta}_{\text{bad}} = 1/2$ (that is, degenerately $1/2$ with probability 1). Why is this estimator a bad choice?

(j) [harder] Graph the bias of $\hat{\theta}_{\text{bad}}$ over all θ. Label your axes.

(k) [harder] Graph the risk of $\hat{\theta}_{\text{bad}}$ under absolute loss. Label your axes.

(l) [difficult] Compare $\hat{\theta}_{\text{bad}}$ to $\bar{X}$ using the sup risk with squared error loss. How much better is $\bar{X}$? With n getting larger does it get better?

(m) [difficult] [MA] Invent a loss function that would make the risk for $\hat{\theta}_{\text{bad}}$ better than $\bar{X}$.

Problem 3

Here we will review theory testing from a conceptual point of view.

(a) [easy] What is *theory testing* and how is it different from *point estimation*?

(b) [easy] Give an example where theory testing is important.

(c) [easy] What are the two ways to go about convincing someone of your theory? Write them in the order we did in class since we will be referencing the "first mode of convincing" and the "second mode of convincing" in problems further on.

(d) [easy] Which way will win you more adherents and why?

(e) [easy] Regardless of how you do your convincing, why is it fundamentally impossible to prove your theory (in an absolute sense) in our context of statistical inference?

(f) [easy] The flavor of theory testing we traditionally do in statistical inference is called *hypothesis testing*. What is a *hypothesis*?

(g) [easy] In the second mode of convincing someone of your theory, which is your theory you want to convince others of, the null hypothesis or the alternative hypothesis?

(h) [easy] In the first mode of convincing someone of your theory, which is your theory you want to convince others of: the null hypothesis or the alternative hypothesis?

(i) [easy] In order to perform a hypothesis test (in either of the two modes), do you need to assume a DGP? Yes/No.

(j) [easy] What are the three steps we mentioned to generate a test? Right them in the order we mentioned in class.

(k) [easy] For the third step, how do you know how far is a far enough departure? You need an additional parameter α that you (the statistician) is responsible for supplying to the theory testing. What is its role in the test?

(l) [easy] Although this parameter is fundamentally up to you, what is the scientific community's standard(s)?

(m) [easy] What is the *retainment region* (RET)? What is the *rejection region*? Why do you think the retainment region is sometimes called the *non-critical region* and the rejection region is sometimes called the *critical region*?

(n) [easy] What is the two possible outcomes of a hypothesis test? And which set does $\hat{\hat{\theta}}$ belong to in each of these two possible outcomes?

(o) [easy] For each of these two possible results, which error could be made? Give the name of the error and describe it conceptually.

(p) [harder] Circle the black dots next to the statement(s) which are *always* true:

- $\mathbb{P}\left(\hat{\theta} \in RET\right) = \alpha$
- $\mathbb{P}\left(\hat{\theta} \notin RET\right) = \alpha$
- $\mathbb{P}\left(\text{Type I error}\right) = \alpha$
- $\mathbb{P}\left(\text{Type II error}\right) = \alpha$
- $\mathbb{P}\left(\text{Type I error}\right) + \mathbb{P}\left(\text{Type II error}\right) = \alpha$
- If H_0 is true, you cannot make a type I error.
- If H_0 is true, you cannot make a type II error.
- If H_a is true, you cannot make a type I error.
- If H_a is true, you cannot make a type II error.
- If $\alpha = 0$, $\mathbb{P}\left(\text{Type II error}\right) = 0$
- If $\alpha = 0$, $\mathbb{P}\left(\text{Type II error}\right) = 1$

- If $\alpha = 0$ and H_a is true, $\mathbb{P}(\text{Type II error}) = 0$
- If $\alpha = 0$ and H_a is true, $\mathbb{P}(\text{Type II error}) = 1$
- If $\alpha = 1$, $\mathbb{P}(\text{Type II error}) = 0$
- If $\alpha = 1$, $\mathbb{P}(\text{Type II error}) = 1$
- If $\alpha = 1$ and H_a is true, $\mathbb{P}(\text{Type II error}) = 0$
- If $\alpha = 1$ and H_a is true, $\mathbb{P}(\text{Type II error}) = 1$
- If α increases, the $\mathbb{P}(\text{Type II error})$ increases.
- If α decreases, the $\mathbb{P}(\text{Type II error})$ increases.
- After the test is completed, there is a mathematical way to determine if we made a Type I or Type II error.

Problem 4

Here we will do a binomial exact test. We want to demonstrate that the iphone users in our class is greater than the national average (which is 52.4%). Recall that our data was as follows: for $n = 20$, the $\hat{\hat{\theta}} = 0.60$ where the estimator we chose was the sample proportion.

(a) [easy] Write down H_a then H_0.

(b) [easy] Declare your α level desired for this test. You do not need to justify it. It is what you are comfortable with.

(c) [harder] Because we want to show something is greater than a point value, it is called a right-tailed test. In any test, we need to find the distribution (or approximate the distribution of) the estimator under the null hypothesis. Because we will reject on the right, why is the most conservative value of θ to choose when deriving the sampling distribution to be largest value in the null hypothesis region (in this case $\theta = \theta_0 = 0.524$)?

This question unfortunately isn't answerable without knowing the concept of the "power function" which we didn't get to until lecture 5. So you can skip it.

(d) [easy] Regardless of if you understood the previous question or not, what is the exact sampling distribution given the null hypothesis? Marked easy because you can copy from class.

(e) [easy] Draw the PMF of the sampling distribution. Label all axes carefully and provide sufficient tick marks. Marked easy because you can copy from class.

(f) [easy] Indicate the RET and the rejection region in the above illustration. Use your α. Everyone's answer may be different!

(g) [harder] Were you able to create a rejection region at your exact level of α? Yes / no and why?

(h) [easy] Run the test. Write your conclusion in English.