MATH 369/650 Fall 2020 Final Homework (#7)

Professor Adam Kapelner

Due by email 11:59PM Sunday, December 13, 2020

(this document last updated Sunday 13th December, 2020 at 5:59pm)

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, read about the class topics: the 1-sample and 2-sample Kolmogorov-Smirnov test, the two-sample permutation test, the nonparametric bootstrap (its use in hypothesis testing and confidence interval construction), causal estimates, bias in observational study estimates, counfounding variables, two-arm randomized experiments for testing treatment vs. control, and their unbiasedness in the two recommended textbooks and online.

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 the 600-level class 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

Consider the height data from class. We sampled $n_1 = 10$ men and measured heights in inches: 67, 68, 69, 70, 70, 71, 72, 72, 73 and 73 and $n_2 = 6$ females and measured heights in inches: 59, 60, 63, 64, 64 and 64.

(a) [easy] Over the regions of Europe, North America, Australia, and East Asia, female height is found to be normally distributed with mean 64.8in and standard deviation 2.8in (see here). We wish to test if our data deviates from this distribution. State the null and alternative hypothesis.

(b) [easy] Using these mean and standard deviation values, standard the data for the n_2 female height measurements and provide the values of $z_1, ..., z_6$ below.

(c) [easy] In the following space create an illustration that plots the empirical CDF of the standardized female heights. Also on this plot, graph $F_Z(z)$, the CDF of $\mathcal{N}(0, 1)$.You will have to look up the quantiles for the standard normal from Math 241. Try to make your illustration to scale as much as possible but zoom in on the y axis more than the x axis. Make the y axis as high as the space below and have y range from 0 to 1.

(d) [easy] From the plot in (c), try to estimate D_6, the "supremum norm difference" which is the largest absolute difference between the empirical CDF and $F_Z(z)$.

(e) [easy] In the statement $\sqrt{n} D_n \overset{d}{\to} K$, what is the name of the distribution K? And who proved this result? Plot a rough sketch below of the PDF of K. Label the axes.

(f) [easy] Run the one-sample Kolmogorov-Smirnov (K-S) test for the H_a in (a) using the test statistic in (d) at $\alpha = 5\%$. Note that $F_K(1.359) = 95\%$. What is your decision?

(g) [easy] We now wish to test if the DGP's for male and female height are different. State the null and alternative hypothesis.

(h) [harder] In the following space create an illustration that plots the empirical CDF of the raw female heights (not standardized). Also on this illustration, plot the empirical CDF of the raw male heights (not standardized). Try to make your illustration to scale as much as possible but zoom in on the y axis more than the x axis. Make the y axis as high as the space below and have y range from 0 to 1.

(i) [easy] From the plot in (g), try to estimate $D_{6,10}$, the two-sample "supremum norm difference" which is the largest absolute difference between the two empirical CDFs.

(j) [easy] Run the two-sample K-S test for the H_a in (g) using the test statistic in (i) at $\alpha = 5\%$. Note that $F_K(1.359) = 95\%$. What is your decision?

(k) [harder] Is the quantile in the cases in (f) and (j) accurate given our sample size? To run these two K-S tests more accurately, what can you do?

(l) [easy] We now wish to run Fisher's permutation test to test H_a in (g). Explain the steps of the permutation test carefully. Provide the four examples from the lecture of test statistics that can be employed.

(m) [difficult] Provide an examples of a test statistic that can be employed within the permutation test that we did not discuss in class. Explain why it would work.

(n) [easy] In this dataset, what is B_{all}, i.e. how many possible unique resamplings are there? Can these number of resamplings be done comfortably in a modern computer?

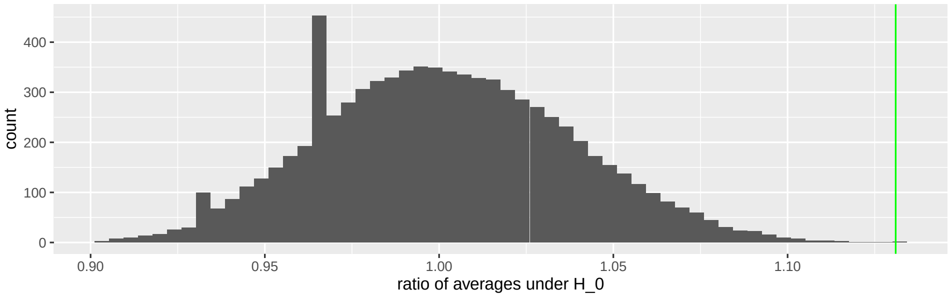
(o) [easy] Let's employ the ratio-of-averages test statistic (even though we used difference-of-averages in class). Under H_0 in (g), what do you expect the value of the ratio-of-averages to be?

(p) [difficult] Prove that the true mean over all resamplings is indeed the answer from (o).

(q) [easy] Calculate the ratio-of-averages test statistic for the original data.

(r) [easy] Do one resampling of the data manually by hand (make sure you split all n into partitions of sizes n_1 and n_2). Write the resampled data below. Calculate the ratio-of-averages test statistic.

(s) [easy] Below is an histogram of all B resamplings. The value of the ratio of averages in the original sample data is displayed as a green vertical line.



Estimate the RET at $\alpha = 5\%$. Run the test and report your decision.

(t) [harder] Estimate Fisher's p-value for this test.

(u) [difficult] When we did this in class, we got a very different result. Why? Any thoughts on advice to someone doing this again? skip this problem... it is a result of my mistake

Problem 2

In midterm I, problem 10 we looked at lifespans data rats. Mesured in years, here were the lifespans of each rat, sorted: 0.87, 1.04, 1.09, 1.33, 1.39, 1.47, 1.54, 1.69, 1.94, 2.23, 2.28, 2.44, 2.48, 2.49, 2.52, 2.57, 2.64, 2.71, 2.8, 3.08, 3.08, 3.18, 3.22, 3.28, 3.4, 3.65, 3.86, 3.96, 4.18, 4.80.

(a) [easy] For the $n = 30$ rats, $\bar{x} = 2.57$ and $s = 1.00$. Create an approximate 95% CI for the mean lifespan of these rats by inverting the Wald test.

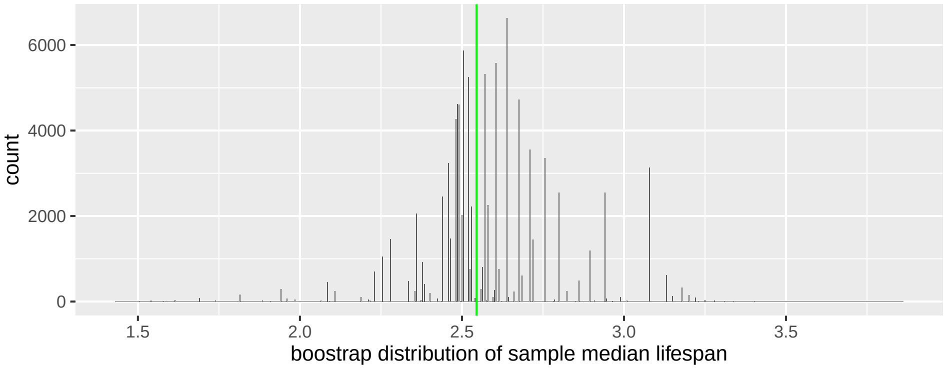
(b) [easy] For the $n = 30$ rats, calculate the sample median.

(c) [easy] The CI in (a) relied upon using the CLT and an application of the continuous mapping theorem and Slutsky's theorem. If we interested in a 95% CI for the *median*, could we rely on the same theory? Yes / no.

(d) [easy] Consider a nonparametric bootstrap of this data. Could this yield approximate inference (i.e. testing hypotheses and constructing confidence intervals) for the sample median? Yes / no.

(e) [difficult] How large is B_{all}, the total number of unique resamplings of this dataset during the nonparametric bootstrap procedure?

(f) [harder] Below is a histogram of the boostrap distribution of the sample median for $B = 100,000$. The value of the sample median is displayed as a green vertical line.



Why does the bootstrap distribution appear discrete? Is the actual estimator's distribution discrete?

(g) [easy] Provide an approximate 95% CI for the median lifespan of these rats by estimating it from the picture above.

(h) [easy] Do an approximate test of the median rat lifespan being greater than two years. State the hypotheses first and justify your answer.

Problem 3

We will now interpret observational estimates and causal estimates.

(a) [easy] For scenario 1 and scenario 2 discussed in class, write the observational and causal statements for both scenario 1 and 2. Label them 1a, 1b, 2a and 2b.

(b) [easy] What is a confounding variable? Here is some good reading material.

(c) [easy] For scenario 1, why is the causal statement not justified? Discuss.

(d) [harder] For scenario 2, why is the causal statement justified? Discuss.

Problem 4

We will now investigate the differences between observational and experimental studies. For the purposes of this class, we will focus only two arm studies; this means the treatment variable will always have only two levels which we will call treatment and control and denote T and C respectively. (Multiarm studies and continuously-valued treatment variables are beyond the scope of this course, but that doesn't mean they're not important)!

(a) [easy] What is an outcome / response / endpoint? Assume n subjects in the study. How are the n response measurements denoted in our class?

(b) [easy] What is a treatment allocation / assignment? How are the n allocations denoted in our class?

(c) [easy] Explain what an observational study is. Mention how the responses are measured from the subjects. Are the assignments truly *assigned*?

(d) [easy] Explain what an experimental study is. Mention how the responses are measured from the subjects. Are the assignments truly *assigned*?

Problem 5

We will now talk about the Rubin Causal Model.

(a) [easy] What are potential outcomes? What is its notation we discussed in the lecture?

(b) [easy] What is a subject-level (unit-level) causal effect? What is its formula using the notation discussed in (a)?

Problem 6

We will now talk about assumptions in the causal model.

(a) [easy] If the treatment effect θ is constant for all subjects (the same) and additive, express the potential outcome for subject i in the treatment arm as a function of the potential outcome for subject i in the control arm and θ

(b) [easy] Explain why $\bar{y}_T - \bar{y}_C$, the naive estimate for $\theta = \mathbb{E}[Y_T] - \mathbb{E}[Y_C]$, is not guaranteed to work. In your answer, define *selection bias.*

(c) [easy] If the confounder x is linear in the response with scaling factor β and so is an error term e, write formulas for both of the potential outcomes using the model in (a).

(d) [easy] Under the model in (c), write the naive estimate $\bar{y}_T - \bar{y}_C$ as a function of $\bar{x}_T$, $\bar{x}_C$, $\bar{e}_T$, $\bar{e}_C$, β and θ.

(e) [harder] Explain why $\bar{e}_T - \bar{e}_C \approx 0$. State all assumptions.

(f) [harder] Explain why $\bar{Y}_T - \bar{Y}_C$ is a biased estimator for θ. What is the bias term?

(g) [E.C.] Show that $\bar{Y}_T - \bar{Y}_C$ is unbiased (over all designs and errors) under the completely randomized design.

(h) [easy] What is the essential reason we use randomized experiments for causal inference?

(i) [harder] Explain why $\bar{x}_T - \bar{x}_C \approx 0$ in a completely randomized design.

(j) [harder] Why is it important to reduce the MSE of estimators for θ?