MATH 341/641 Fall 2023 Homework #4

Professor Adam Kapelner

Due by email 11:59PM Nov 2, 2023

(this document last updated Friday 20th October, 2023 at 5:08pm)

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 340 concepts: the CLT, the CMT, Slutsky's theorems.

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 5 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: _____

Consider the following data from two populations:

$$n_1 = 6, \ \bar{x}_1 = 62.3, \ s_1^2 = 2.25^2, \ n_2 = 10, \ \bar{x}_2 = 70.5, \ s_2^2 = 2.07^2$$

We will provide inference for this data under different assumptions.

(a) [easy] If you assume both populations are iid normal DGP's where the variances are known: $\sigma_1^2 = 3.5^2$ and $\sigma_2^2 = 4^2$. Run a hypothesis test which attempts to prove unequal means. Is this test an exact or approximate test?

(b) [harder] In the previous question, did you rely on any limit theorems? If so, which ones?

(c) [easy] If you assume both populations come from an iid DGP but the distribution is unknown, run a hypothesis test which attempts to prove unequal means. Is this test an exact or approximate test?

(d) [easy] Create a 95% CI for the difference in means, $\theta_1 - \theta_2$ under the assumptions from the previous question. Is this test an exact or approximate interval?

(e) [difficult] If you assume both populations come from an iid DGP but the distribution is unknown, and that the variances are unknown *but equal*, run a hypothesis test which attempts to prove unequal means. Is this test an exact or approximate test?

(f) [harder] In the previous three questions, did you rely on any limit theorems? If so, which ones?

Problem 2

We will run inference on real data here.

(a) [harder] For the obesity study found at this hyperlink, consider the outcome metric "binge eating remission" which is binary where $1 =$ the subject no longer binge eats and $0 =$ the subject still binge eats. Identify who the two population groups. Run a test attempting to prove remission rates are unequal in the two groups.

(b) [harder] In the previous question, did you rely on any limit theorems? If so, which ones?

(c) [harder] Compute a 95% CI for the difference in remission rates among the two groups.

(d) [harder] In the previous question, did you rely on any limit theorems? If so, which ones?

(e) [easy] For the obesity study, now consider the outcome metric "binge episodes per week" which is a count metric (e.g. 1 day, 2 days, 3 days, ..., all 7 days!). Since it's real data, no moments are known! Explain why the normality assumption does not hold for the DGP in both samples.

(f) [harder] Test the theory that mean "binge episodes per week" differs in both poputions. The raw data is in Table 3 and the format is $\bar{x} \pm s$. Don't forget that standard error is not standard deviation!

(g) [harder] [MA] For two independent samples of unknown DGP's with means θ_1 and θ_2 and unknown by finite variances, prove that the Wald test is an asymptotically valid 2-sample z-test. This is the test you'll use in the next question.

Problem 3

This problem will be about the multiple testing / multiple comparisons problem in general.

(a) [harder] Let's say we define a family of m tests. Draw the 2×2 table from class that accounts for the taillies of the four possibilities (decision $\times$ truth). Indicate which quantities you observe. Indicate which quantities you do not observe. Denote random quantities with an uppercase letter. Denote constants with a lowercase later. Make up letters if we did not have letters for each of the four boxes.

(b) [easy] In the case where all m H_0's are true, redo (a).

(c) [easy] Define FWER, FDP and FDR using notation and in your own words.

(d) [harder] Describe a scenario where you would want FWER $\leq 1\%$.

(e) [harder] Describe a scenario where you would want FDR $\leq 1\%$.

(f) [easy] Prove that FWER $=$ FDR when all m H_0's are true.

(g) [easy] Describe the Simes procedure in detail. What is the threshold it rejects at?

(h) [easy] Describe what the Benjamini-Hochberg procedure accomplishes in detail (not the procedure itself, as the procedure itself is the Simes procedure).

(i) [E.C.] Prove that Simes controls FWER when all m H_0's are true.

(j) [easy] Recall the IPMC data from research into mouse sexual dimorphism in genetic knockouts. There are $m = 172,328$ tests and we investigated the naive, Bonferroni, Sidak and Simes for weak FWER control and the Benjamini-Hochberg procedure for FDR control. We wanted FWER and FDR control of 5% in this demo.

We looked at the illustration below during lecture. Identify the red line, the yellow line (which is actually two different things), the green line and the black line by writing atop the illustration. Then, indicate and give a numerical estimate to the number of rejections for the naive procedure of setting $\alpha = 5\%$ for all m tests. Then indicate and give a numerical estimate to the number of rejections for the Bonferroni procedure. Then indicate and give a numerical estimate to the number of rejections for the Simes / Benjamini-Hochberg procedure.
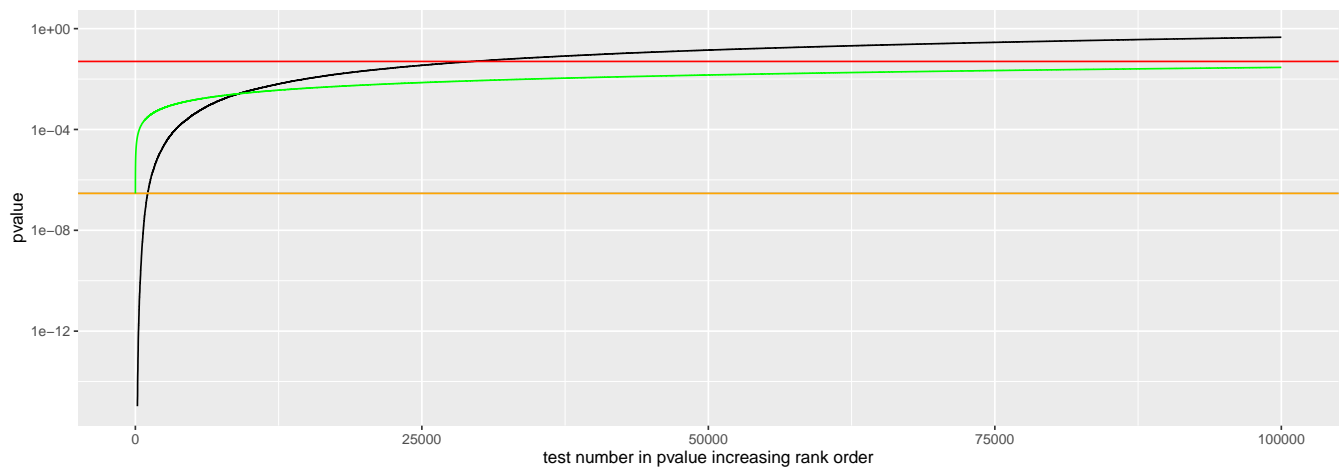
(k) [easy] Compute the Bonferroni threshold. Ensure that the Bonferroni is smaller than the Sidak.

(l) [easy] Compute the Sidak α threshold. Ensure that the Bonferroni threshold is smaller than the Sidak threshold.

(m) [easy] The Simes α threshold is 0.00262. Would that yield more rejections than Bonferroni? Yes / No.

(n) [easy] Employing the Benjamini-Hochberg procedure, what does your number of rejections mean? Explain and be specific.

(o) [harder] Why do you think the Benjamini-Hochberg procedure to control FDR has had such a huge impact on science?

(p) [easy] We looked at the illustration below during lecture, the histogram of the pvals.
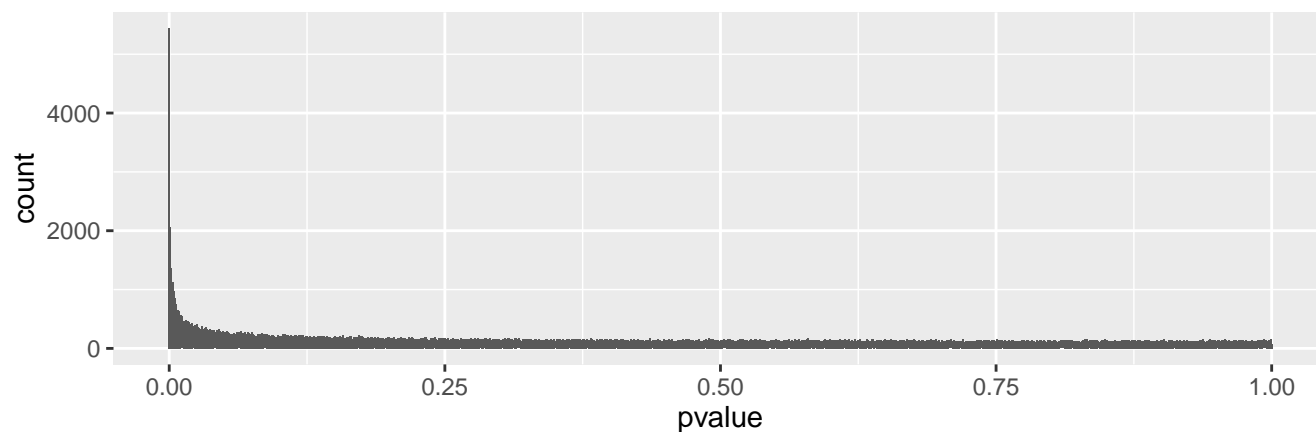
Do you believe that all H_0's are true? Yes / No.

(q) [difficult] Do you think that Bonferroni / Sidak / Simes are more conservative now that you've seen the plot? Explain

Problem 4

This is about the core thm.

(a) [easy] State the theorem's results for MLE's and MM estimators

(b) [easy] Assuming $\hat{\theta}_{\mathrm{MLE}} \xrightarrow{p} \theta$ and the "technical conditions" we glossed over (e.g. the MLE is not on a boundary of Θ), prove that the MLE is asymptotically normal with mean θ and variance equal to the CRLB on variance.

Consider height data from 2020's MATH 369 class. We sample $n_1 = 10$ 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, standardize 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 (the estimate, $\hat{F}$) 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 your 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 $\sqrt{n}D_n \xrightarrow{\mathcal{D}} K$, 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?