MATH 341/641 Fall 2025 Homework #3

Professor Adam Kapelner

Due by email 11:59PM at the date on the homepage

(this document last updated Friday 19th September, 2025 at 4:04pm)

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: random variables, PMF's, PDF's and the normal 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 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: ____

Problem 1

Remember the male student height data from a previous year: $n = 13$ and $\bar{x} = 68.85''$. We want to test at $\alpha = 1\%$ if the population that this sample was drawn from has a *different* mean than the American male height mean of 69.92" for ages 20-29 (this measurement is from more exact studides gathered from this article). You can assume that the true variance of the population if 4in^2 but *do not assume* the DGP is normal!

(a) [easy] Write the alternative and null hypotheses.

(b) [harder] Do we know the null distribution exactly? Why or why not?

(c) [easy] Write the approximate null sampling distribution on the original scale (i.e. in inches).

(d) [easy] Write the RET region as a set on the original scale.

(e) [easy] Is it possible to provide the exact $\mathbb{P}$ (Type I error)? Yes / no

(f) [easy] What is the approximate $\mathbb{P}$ (Type I error) in this test?

(g) [easy] Is it possible to provide the exact p-val? Yes / no

(h) [easy] Calculate the approximate p-val for this test.

(i) [easy] Is the dataset's estimate "statistically significant"? Yes / no

(j) [easy] Was this an exact test? Yes / no

(k) [easy] Write a conclusion of this test in English.

(l) [harder] Regardless of what the test's decision came out to be, assume H_0 is rejected. Is the dataset's estimate "practically significant" (or "clinically significant")? Discuss.

(m) [harder] With a very large sample size, would H_0 always be rejected? Discuss.

(n) [easy] Imagine you could assume the sample was drawn $\overset{iid}{\sim}$ from a normal distribution. Create a 95% 2-sided CI for the mean height for all 300-level STEM courses at CUNY.

(o) [harder] Without assuming the sample was drawn $\overset{iid}{\sim}$ from a normal distribution, what would be the problem with building a CI?

(p) [easy] Does the CI include θ_0? Yes / no / maybe

(q) [harder] Does the CI include θ? Yes / no / maybe

Problem 2

Here we will investigate MLE's and UMVUEs.

(a) [in the notes] Assume the DGP: $X_1, \ldots, X_n \stackrel{iid}{\sim} \text{Bernoulli}(\theta)$. Find the MLE for θ.

(b) [in the notes] Prove the MLE from from the previous problem is a UMVUE.

(c) [harder] Assume the DGP: $X_1, \ldots, X_n \stackrel{iid}{\sim} \text{Exp}(\theta) := \theta e^{-\theta x} \mathbb{1}_{x>0}$. Find the MLE for θ.

(d) [harder] Now assume a different parameterization of the DGP, $X_1, \ldots, X_n \stackrel{iid}{\sim} \text{Exp}\left(1/\theta\right)$. Find the MLE for θ.

(e) [difficult] Prove the MLE from from the previous problem is a UMVUE.

Problem 3

Here we will get more practice with MM estimators

(a) [difficult] Consider the DGP $X_1, \ldots, X_n \stackrel{iid}{\sim} \text{Gamma}\left(\theta_1,\, \theta_2\right)$. Below are some facts about this distribution that I took from wikipedia:

$$\begin{aligned}
\text{Gamma}\left(\theta_1,\,\theta_2\right) &:= \frac{\theta_2^{\theta_1}}{\Gamma\left(\theta_1\right)} x^{\theta_1-1} e^{-\theta_2 x} \mathbb{1}_{x>0}, \quad \text{Supp}\left[X\right] = (0,\infty), \quad \theta_1, \theta_2 \in (0,\infty), \\
\mathbb{E}\left[X\right] &= \int_0^\infty x f^{old}(x)dx = \frac{\theta_1}{\theta_2}, \\
\mathbb{V}\text{ar}\left[X\right] &= \int_0^\infty (x - \mathbb{E}\left[X\right])^2 f^{old}(x)dx = \frac{\theta_1}{\theta_2^2}
\end{aligned}$$

Find MM estimators for both parameters. Hint: leave expressions in terms of $\hat{\sigma}^2$.

(b) [easy] Provide point estimates $\hat{\hat{\theta}}_1$ and $\hat{\hat{\theta}}_2$ for the unknown parameters θ_1 and θ_2 given the dataset $\boldsymbol{x} =< 10.8, 8.5, 13.2, 9.1, 13.5, 11.2, 7.1 >$ for the $\overset{iid}{\sim}$ Gamma $(\theta_1,\, \theta_2)$ DGP. No need to show work.

(c) [difficult] [MA] In Math 241 you learned about expectation and variance where expectation was a measure of central tendency of a distribution and variance is a measure of dispersion around that central tendency. The next most important metric for rv's is probably its *skewness* defined as $\gamma := \mathbb{S}\text{kew}\left[X\right] := \mathbb{E}\left[\left(\frac{X-\mathbb{E}[X]}{\mathbb{SD}[X]}\right)^3\right]$ where SD refers to standard deviation. Skewness is technically the third standardized moment since $\frac{X-\mathbb{E}[X]}{\mathbb{SD}[X]}$ is the distribution standardized and then the third power is taken. Skewness is a metric of which tail of the distribution is longer and by how much as seen in the first figure here. Since third powers are both positive and negative, skewness can be both positive and negative (and zero if the distribution is symmetric with right and left tails the same). In class, we derived nonparametric MM estimators $\bar{X}$ and $\hat{\sigma}^2$ for the expectation and variance (nonparametric meaning that the derivation for them was for *all* iid DGP's). Show that the nonparametric MM estimator for skewness is:

$$\hat{\gamma} = \sqrt{n} \frac{\sum_{i=1}^{n} (X_i - \bar{X})^3}{\left(\sum_{i=1}^{n} (X_i - \bar{X})^2\right)^{3/2}}$$

Hint: assume a iid DGP with density / mass function $f(\theta_1, \theta_2, \theta_3)$ where θ_1 is the expectation, θ_2 is the variance and θ_3 is the skewness.

Problem 4

We will prove some of the main theorems of this class (and some other relevant facts) here.

(a) [in the notes] Prove the CRLB from scratch. Justify each step. List assumptions.

(b) [in the notes] Prove that Fisher Information which is defined as $I(\theta) := \mathbb{E}\left[\ell'(\theta; X)^2\right]$, the expected score squared, is equal to $\mathbb{E}\left[-\ell''(\theta; X)\right]$. If you make any assumptions proving this, indicate it so.

Problem 5

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.

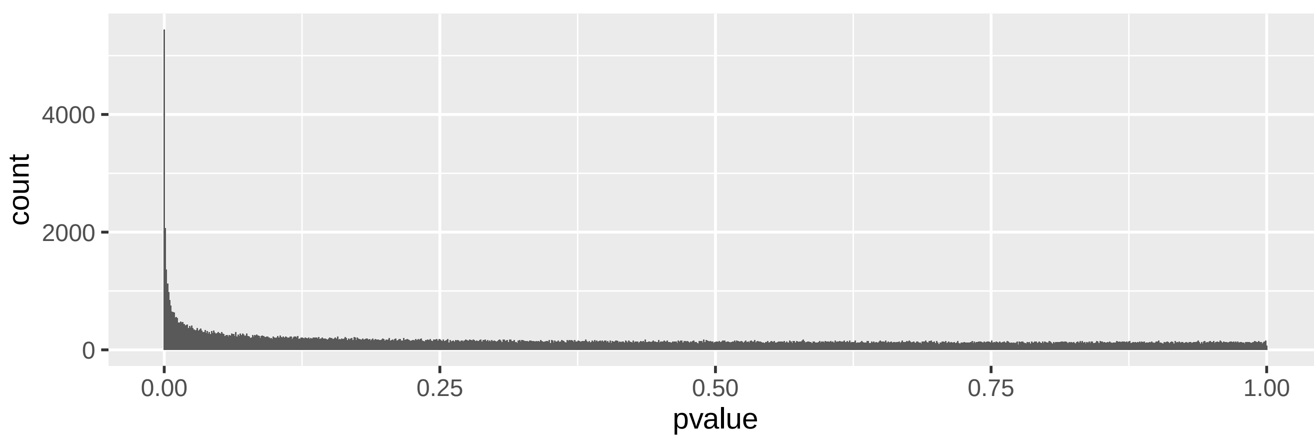
(l) [easy] Compute the Sidak α threshold. Ensure that the Bonferroni threshold is smaller than the Sidak threshold.

(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