

MATH 341/641 Fall 2024 Homework #4

Professor Adam Kapelner

Due by email 11:59PM Oct 28, 2024

(this document last updated Saturday 26th October, 2024 at 9:15pm)

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 L^AT_EX. Links to installing L^AT_EX and program for compiling L^AT_EX 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 L^AT_EX, 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 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.
- (b) [easy] Was the test above an exact or approximate test?
- (c) [harder] In the test above, did you rely on any limit theorems? If so, which ones?
- (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) [easy] Now, assume both populations come from an iid DGP but the distribution is unknown, run a hypothesis test which attempts to prove unequal means.

(f) [easy] Was the test above an exact or approximate test?

(g) [harder] In the test above, did you rely on any limit theorems? If so, which ones?

(h) [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?

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

(j) [easy] Was the test above an exact or approximate test?

(k) [harder] In the test above, did you rely on any limit theorems? If so, which ones?

(l) [harder] 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?

(m) [difficult] If you assume both populations come from an iid DGP but the distribution is unknown, and that the variances are unknown *but now assumed equal*, run a hypothesis test which attempts to prove unequal means.

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 test above, 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] 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 an assumption that the DGP is normal in both samples does not hold. This question requires 340 thinking.
- (e) [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, s/\sqrt{n} , is not standard deviation, s !

- (f) [harder] [MA] For two independent samples of unknown DGP's with means θ_1 and θ_2 and unknown but finite variances, show that $(\bar{X}_1 - \bar{X}_2)/\sqrt{S_1^2/n + S_2^2/n}$ can be used as an asymptotic Z test when $H_a : \theta_1 - \theta_2 \neq 0$ via invoking Slutsky's B, Slutsky's A, CMT and CLT.

Problem 3

This is about the core / monster theorem for MM's and MLE's.

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

- (b) [in the notes] Assuming $\hat{\theta}_{\text{MLE}} \xrightarrow{p} \theta$ 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.

Problem 4

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, \dots, z_6 below.

- (c) [harder] 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 $\sqrt{n}D_n \xrightarrow{d} K$, 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) [in the notes] Did we have enough n for this test to be trusted? Yes/no
- (h) [easy] We now wish to test if the DGP’s for male and female height are different. State the null and alternative hypothesis.
- (i) [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.

- (j) [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.
- (k) [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?

Problem 5

Consider the following DGP: $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Exp}(\theta) := \theta e^{-\theta x} \mathbb{1}_{x>0}$.

- (a) [harder] Derive $\hat{\theta}_{\text{MLE}}$.

- (b) [harder] Derive $I(\theta)$.

- (c) [easy] Since $\hat{\theta}_{\text{MLE}} \neq \bar{X}$, we cannot rely on the CLT to provide asymptotic Z-testing. What theorem can we rely to provide asymptotic Z-testing?

- (d) [harder] Derive \hat{Z} , the asymptotically normal test statistic to test $H_a : \theta \neq \theta_0$. Your answer must be a function of $X_1, \dots, X_n, n, \theta_0$ only.
- (e) [easy] Consider the dataset $\mathbf{x} = \langle 0.156, 0.257, 0.454, 0.559, 0.305, 0.431, 0.599, 0.665, 0.185, 0.15 \rangle$. Compute $\hat{\theta}_{\text{MLE}}$ to the nearest 3 decimals.
- (f) [harder] For the dataset above test if $H_a : \theta > 1$ as $\alpha = 5\%$.
- (g) [harder] For the dataset above, create a 95% CI for θ .

Problem 6

This problem will cover the Score Test and the Likelihood Ratio Test when testing the parameter in the iid Bernoulli DGP. Consider the MLE, $\hat{\theta}_{\text{MLE}} = \bar{X}$ and the null hypothesis $H_0 : \theta = 0.5$.

- (a) [in the notes] Provide the asymptotically normal estimator that is the basis of the score test for one parameter for any iid DGP $f(x; \theta)$.

(b) [in the notes] Show that the score test is equivalent to the Wald test in the case where the DGP is iid Bernoulli (θ). This means the estimator is the same. Find in the lectures where we derived $I(\theta)$ for the iid Bernoulli (θ) DGP. Then it's algebraic simplification from there.

(c) [easy] Run the score test for the iid Bernoulli (θ) DGP where $n = 100$ and $\sum x_i = 61$.

(d) [in the notes] Prove the Likelihood Ratio (LR) test for one parameter for any iid DGP $f(x; \theta)$, i.e., declare the estimator and provide its exact or approximate distribution.

- (e) [in the notes] Show that the LR test is *not* equivalent to the Wald test / Score test in the case where the DGP is iid Bernoulli(θ). This means the estimator is *not* the same.
- (f) [easy] Run the LR test for the iid Bernoulli(θ) DGP where $n = 100$ and $\sum x_i = 61$. Ensure it's different but similar to the result of the score test.
- (g) [in the notes] Plot a log-likelihood function vs $\hat{\theta}$. Mark $\hat{\theta}_{\text{MLE}}$ and θ_0 , a value you're testing against in H_a . Also illustrate the distance wa that corresponds to the numerator of the statistic used in the Wald test for the MLE, the distance sc that corresponds to the numerator of the statistic used in the score test and lr which corresponds to half the likelihood ratio statistic.
- (h) [E.C.] Prove the asymptotic equivalence of the Wald, Score and LR tests.