MATH 341/641 Fall 2023 Homework #5

Professor Adam Kapelner

Due by email 11:59PM Nov 16, 2023

(this document last updated Thursday 2nd November, 2023 at 10:55pm)

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: ____

Problem 1

In lecture, we did two-sided two-sample z and t tests. We will repeat these tests now but do them one sided. For extra practice, I will make them left-sided. To do this, I will switch the indexing of the two populations. The female population is now considered population #1 and the male population is now considered population #2. We assume the DGP for female height measurements is $\overset{iid}{\sim} \mathcal{N}(\theta_1, \sigma_1^2)$ independent of the DGP for male height measurements assumed to be $\overset{iid}{\sim} \mathcal{N}(\theta_2, \sigma_2^2)$.

The sample sizes, point estimates for the mean and point estimates for the variance computed from an in-class student survey are:

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

We will now assume that the variances are equal i.e. $\sigma_1^2 = \sigma_2^2 = \sigma^2$ but its value is *unknown*.

(a) [easy] Write the exact or approximate distribution of the standardized estimator under the null hypothesis which we denote $(\hat{\theta}_1 - \hat{\theta}_2)/SE \mid H_0$. Write the SE as a mathematical expression.

(b) [harder] Will this test be an *exact test* or instead an *approximate test*? Explain.

(c) [harder] Compute the retainment region. Note that $\mathbb{P}(T_{14} \leq -1.76) = 5\%$ where T_{14} denotes a standard Student's t rv with 14 degrees of freedom.

(d) [easy] Run the test and write your conclusion using an English sentence.

(e) [harder] Find the p-value of our estimate by writing a statement like $\mathbb{P}(T_{df} < t)$ or $\mathbb{P}(T_{df} > t)$. You need to solve for df, t.

(f) [easy] Without computing the p-value explicitly, would it be above or below $\alpha = 5\%$? Is the estimate *statistically significant*?

We will now assume that the variances are unequal i.e. $\sigma_1^2 \neq \sigma_2^2$ and both values are *unknown*. This is known as the Behrens-Fisher problem.

(g) [easy] Write the exact or approximate distribution of the standardized estimator under the null hypothesis which we denote $(\hat{\theta}_1 - \hat{\theta}_2)/SE \mid H_0$. Write the SE as a mathematical expression.

(h) [easy] Assume you know the exact distribution (which was solved in 2018 and can be found in this paper). Will this be an exact test or an approximate test?

(i) [easy] Assume you instead use the Welch-Satterthwaite test. Will this be an exact test or an approximate test?

(j) [harder] Compute the retainment region. Note that $\mathbb{P}(T_{9.94} \leq -1.81) = 5\%$.

(k) [easy] Run the test and write your conclusion using an English sentence.

Problem 2

In this question, we will use the *univariate delta method.*

(a) [easy] State the univariate delta method. List assumptions.

(b) [easy] Prove the univariate delta method. Justify each step.

(c) [difficult] Assume the $\overset{iid}{\sim}$ Bernoulli DGP with mean θ. Sometimes researchers are interested in the following parameterization: the log odds against the event ocurring, i.e. $\phi := \ln\left(\frac{1-\theta}{\theta}\right)$, a metric that can be any number in $\mathbb{R}$. Derive an asymptotically normal estimator for ϕ.

(d) [harder] Given the previous answer, write a formula for $\hat{\hat{CI}}_{\phi,1-\alpha}$ where ϕ is the log odds against the event ocurring.

(e) [easy] Recall the PUFA-Atrial Fibrilation after open heart surgery study from many of our lectures (click here to find the study online). In class we derived a confidence interval for the odds against getting Atrial Fibrilation in the control (non-PUFA) group. Find a point estimate for ϕ.

(f) [harder] Test at $\alpha = 5\%$ that log odds against is nonzero.

(g) [harder] Compute a $\hat{\hat{CI}}_{\phi,95\%}$ where ϕ is the log odds against getting Atrial Fibrilation in the control group and round to 3 digits.

Problem 3

We will review (a) the equivalence of the two-sided z test and the χ^2 test and (b) the equivalence of the two-sided t test and the F test.

(a) [easy] Fill in the blank:

$$\frac{\hat{\theta} - \theta}{\mathbb{SE}[\hat{\theta}]} \overset{\mathcal{D}}{\to} \mathcal{N}(0, 1) \quad \Rightarrow \quad \frac{(\hat{\theta} - \theta)^2}{\mathbb{V}\text{ar}[\hat{\theta}]} \overset{\mathcal{D}}{\to}$$

(b) [easy] Fill in the blank:

$$\sqrt{n}\frac{\hat{\theta} - \theta}{S} \sim T_{n-1} \quad \Rightarrow \quad n\frac{(\hat{\theta} - \theta)^2}{S^2} \sim$$

(c) [easy] For the PUFA-Atrial Fibrilation after open heart surgery study, test the hypothesis that AF incidence is unequal between the PUFA and non-PUFA groups using an approximate χ^2 test at $\alpha = 5\%$. Note that $F_{\chi^2_1}(3.84) = 95\%$.

Problem 4

This example is a famous one and you can find it on p161 of AoS. Gregor Mendel was a scientist and abbott in what's now modern-day Czech Republic. In 1866 he published his work on a theory of genetic inheritance. He conjectured that if phenotypes, i.e. what you can see in an organism, were binary (e.g. ear lobe attached to your face or separated from the face) it was controlled by a pair of "genes". He proposed that the constituents of the pairs were either "recessive" or "dominant". If one or both were dominant, the dominant phenotype would be expressed. If both were recessive, the recessive phenotype would be expressed. See this illustration.

In his famous pea experiment, he looked at two binary phenotypes of peas: shape (round vs. wrinkled) and color (yellow vs. green) which he assumed independent. He conjectured that the round was the dominant shape and yellow was the dominant color. He also conjectured that the initial expression of the genes were 50-50 dominant recessive. Thus, you would get 3/4 of the peas be round (dominant-dominant, dominant-recessive, recessive-dominant), 1/4 of the peas be wrinkled (recessive-recessive only), 3/4 of the peas be yellow (dominant-dominant, dominant-recessive, recessive-dominant) and 1/4 of the peas be green (recessive-recessive only).

Putting it all together, 9/16 of all peas should be yellow and round, 3/16 should be yellow and wrinkled, 3/16 should be green and round and only 1/16 should be green and wrinkled. Between 1856 and 1863 he sampled $n = 556$ peas growing in his garden.

(a) [harder] Assume that the DGP is $\boldsymbol{X} \sim \text{Multinom}(n, \boldsymbol{\theta})$ where $n = 556$. Formulate Mendel's conjecture as a null and alternative hypothesis.

(b) [easy] Assuming the null hypothesis, what are the expected counts in each of the four groups for the $n = 556$ peas?

(c) [easy] Of the $n = 556$ peas, he found 315 were yellow and round, 101 were yellow and wrinkled, 108 were green and round and 32 were green and wrinkled. Calculate the value of the χ^2 goodness-of-fit test statistic to two digits which gauges the data's departure from H_0.

(d) [easy] Run "Pearson's χ^2 goodness of fit test" at $\alpha = 5\%$ and state whether there is sufficient evidence to reject Mendel's theory of genetic inheritance. Note that $F_{\chi^2_3}(7.81) = 95\%$.

(e) [harder] We could've also run a χ^2 test of independence. Define the values of θ here and formulate a null and alternative hypothesis.

(f) [harder] Run the test of independence and state whether there is sufficient evidence to reject Mendel's theory of genetic inheritance. Note that $F_{\chi^2_1}(3.84) = 95\%$. Write a concluding sentence.

Problem 5

In class we spoke about the relationship between hair color and eye color for men. Here is an analogous dataset for women:

	Brown	Blue	Hazel	Green
Black	36	9	5	2
Brown	66	34	29	14
Red	16	7	7	7
Blond	4	64	5	8

(a) [easy] Write the null hypothesis for hair and eye color being independent. Use the notation $\theta_{i\cdot}$ and $\theta_{\cdot j}$ from class.

(b) [harder] Under the null hypothesis, *estimate* the expected frequencies in all 16 groups.

(c) [harder] Calculate the χ^2 test statistic which gauges the data's departure from H_0. No need to show work.

(d) [harder] Run a chi-squared test of hair and eye color being independent at $\alpha = 5\%$. Note that $F_{\chi^2_9}(16.92) = 95\%$. Write a concluding sentence.

(e) [harder] [MA] Here's frequency data on men and women's hair color:

	Brown	Blue	Hazel	Green
Male	98	101	47	33
Female	122	114	46	31

We wish to run a χ^2 test of homogeneity. Write the hypotheses below and run the test at $\alpha = 5\%$ and provide a concluding sentence. Note that $F_{\chi^2_3}(7.81) = 95\%$.

Problem 6

Herein we will practice the model selection theory and techniques we learned in class. Consider the following dataset with $n = 10$: -0.67, -0.58, 0.57, -0.34, -0.22, 0.60, -0.42, -0.01, 0.76, 0.80. Consider the following $M = 4$ candidate iid DGPs / models similar to the lecture:

MOD 1: $\mathcal{N}(\theta_1, \theta_2)$

MOD 2: $\text{Cauchy}(\theta_1, \theta_2)$

MOD 3: $\text{Logistic}(\theta_1, \theta_2)$

MOD 4: $\text{Laplace}(\theta_1, \theta_2)$

After using maximum likelihood, we find the following estimates and AIC metrics for each DGP / model:

MOD 1: $\mathcal{N}(0.050, 0.303)$. AIC $= 20.427$

MOD 2: $\text{Cauchy}(-0.182, 0.391)$. AIC $= 26.899$

MOD 3: $\text{Logistic}(0.028, 0.345)$. AIC $= 21.689$

MOD 4: $\text{Laplace}(-0.176, 0.496)$. AIC $= 23.843$

(a) [harder] Compute $\ell\left(\hat{\hat{\theta}}_1^{\text{MLE}}, \hat{\hat{\theta}}_2^{\text{MLE}}; x_1, \ldots, x_{10}\right)$ for MOD 1 without using the AIC value. This is nothing but some computation. Remember θ_2 in the $\mathcal{N}(\theta_1, \theta_2)$ notation is the variance not the standard deviation!

(b) [harder] [MA] Compute $\ell\left(\hat{\hat{\theta}}_1^{\text{MLE}}, \hat{\hat{\theta}}_2^{\text{MLE}}; x_1, \ldots, x_{10}\right)$ for MOD 3 without using the AIC value.

(c) [easy] Compute the AIC for MOD1 given your answer in (a). Is it the same that I computed using software?

(d) [easy] According to the AIC metric, which model fits this dataset the best?

(e) [easy] Calculate the $M = 4$ Akaike weights. If the true model was among these four candidate models, what is the probablity the true model is normally distributed?

(f) [easy] Compute all AICc metrics. According to the AICc metric, which model fits this dataset the best?

(g) [easy] Why should AICc be employed in this case instead of AIC?

Problem 7

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 $H_0 : \theta = \theta_0$.

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

(b) [harder] 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 simplication from there.

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

(d) [easy] 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.

(e) [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.

(f) [easy] Plot $\hat{\hat{\theta}}$ vs ℓ and illustrate how the Wald, Score and LR tests can be visualized.