MATH 369/650 Fall 2020 Homework #5

Professor Adam Kapelner

Due by email 11:59PM Monday, November 9, 2020

(this document last updated Sunday 8th November, 2020 at 8:50am)

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 (e.g. two proportion z test, Fisher Information, asymptotically normal estimators, delta method, confidence intervals, chi-squared tests of goodness of fit, independence and homogeneity, model section with AIC / AICc) 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

We will review the two-proportion z-test here.

(a) [easy] For two independent samples of Bernoulli DGP's with parameters θ_1 and θ_2, prove that the 2-sample z-test using the sample proportion estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ we developed in class is asymptotically valid.

(b) [harder] For the obesity study found at `https://www.biologicalpsychiatryjournal.com/article/S0006-3223(06)01009-2/fulltext`, 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 of significance to test if remission rates are unequal in the two groups.

Problem 2

We will review some Wald tests here.

(a) [harder] 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.

To do this, you will need the following theorem. For constants c, d and r.v. sequences A_n and B_n where $A_n \xrightarrow{p} a$ and $B_n \xrightarrow{p} b$ then $cA_n + dB_n \xrightarrow{p} ca + db$. Then you use the continuous mapping theorem (what we've been calling Thm 5.5.4) and Slutsky's (i.e. follow HW4 1e exactly). Remember $S_1 \xrightarrow{p} \sigma_1$ and $S_2 \xrightarrow{p} \sigma_2$.

(b) [easy] For the same obesity study as in 1(b), 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.

(c) [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!

(d) [harder] Assume a DGP of $\overset{iid}{\sim}$ Gumbel$(\theta, 1)$. In HW 4 you proved that $\hat{\theta}^{\text{MLE}} = \ln\left(n/\sum_{i=1}^{n} e^{-X_i}\right)$ and $I(\theta) = 1$ (the $I(\theta) = e^{-2\theta}$ in lecture 12 was my mistake so please ignore that). Provide the asymptotic distribution of $\hat{\theta}^{\text{MLE}}$ and run the test from lecture 12 again given the data there as a 2-sided test.

Problem 3

We will discuss theory of confidence intervals here.

(a) [easy] Provide the definition of an interval estimator and interval estimate.

(b) [easy] What is the $\mathbb{P}\left(\theta \in CI_{\theta,1-\alpha}\right)$ if CI refers the interval estimator?

(c) [difficult] What is the $\mathbb{P}\left(\theta \in CI_{\theta,1-\alpha}\right)$ if CI refers the interval estimate?

(d) [harder] Let's say you run $L = 37$ experiments where you compute an exact $CI_{\theta,1-\alpha}$ in each one. What is the probability none of the confidence intervals contain θ?

(e) [harder] Find an approximate $CI_{\theta,1-\alpha}$ by inverting the Wald test of 2(d). This means you must prove the approximate coverage is valid.

(f) [harder] For the same obesity study as in 1(b), consider the outcome metric "binge episodes per week", define θ_1 and θ_2 as the unknown populations' mean parameter and find an exact $CI_{\theta_1-\theta_2,1-\alpha}$ assuming both population DGP's are normal and σ^2 is the same in both populations.

(g) [harder] For the same obesity study as in 1(b), consider the outcome metric "binge episodes per week", define θ_1 and θ_2 as the unknown populations' mean parameter and find an approximate $CI_{\theta_1-\theta_2,1-\alpha}$ assuming both population DGP's are normal and σ^2 is *not* the same in both populations.

(h) [harder] For the same obesity study as in 1(b), consider the outcome metric "binge episodes per week", define θ_1 and θ_2 as the unknown populations' mean parameter and find an approximate $CI_{\theta_1-\theta_2,1-\alpha}$ *without* assuming both population DGP's are normal.

Problem 4

We will talk a little statistical philosophy here.

(a) [easy] For a single point estimate, what can a skeptic of the field of Statistics argue?

(b) [easy] For a single hypothesis test result, what can a skeptic of the field of Statistics argue?

(c) [difficult] For the previous question, consider the case where H_0 is rejected, if the p-value is 1 in 1,000,000,000, does the skeptic have any argument? Thanks to Robin for making me think about this.

(d) [easy] For a single confidence interval estimate, what can a skeptic of the field of Statistics argue?

(e) [easy] How do you define "wrong" in point estimation?

(f) [easy] How do you define "wrong" in confidence interval construction?

(g) [easy] How do you define "wrong" in hypothesis testing?

Problem 5

In this question, we will use the *univariate delta method* and *multivariate 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 a different parameterization, the log odds against the event ocurring, i.e. $\ln\left(\frac{1-\theta}{\theta}\right)$, a metric that can be any number in $\mathbb{R}$. Derive an asymptotically normal estimator for the log odds.

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

(e) [harder] Recall the PUFA-Atrial Fibrilation after open heart surgery study from many of our lectures (especially lecture 14). In class we derived a confidence interval for the odds against getting Atrial Fibrilation in the control group. Compute a $CI_{\phi,95\%}$ where ϕ is the log odds against getting Atrial Fibrilation in the control group and round to 3 digits.

(f) [easy] State the multivariate delta method. List assumptions.

(g) [E.C.] Prove the multivariate delta method. Justify each step.

(h) [difficult] Assume two independent populations both with $\overset{iid}{\sim}$ DGP's with means θ_1, θ_2 respectively and variances σ_1^2 and σ_2^2 respectively. Using the multivariate delta method, derive an asymptotically normal estimator for the ratio of the means $\frac{\theta_1}{\theta_2}$.

(i) [harder] Given the previous answer, write a formula for $CI_{\phi,1-\alpha}$ where ϕ is the ratio of the means $\frac{\theta_1}{\theta_2}$.

(j) [harder] For the same obesity study as in 1(b), consider the outcome metric "binge episodes per week", define θ_1 and θ_2 as the unknown populations' mean parameter and find an approximate $CI_{\phi,1-\alpha}$ where ϕ is the ratio of the means $\frac{\theta_1}{\theta_2}$.

(k) [difficult] List all the reasons why the CI you constructed in the previous problem is approximate. How much should you trust it?

Problem 6

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:

$$\frac{\hat{\theta}-\theta}{\mathbb{SE}[\hat{\theta}]} \xrightarrow{d} \mathcal{N}(0,\,1) \quad \Rightarrow \quad \frac{(\hat{\theta}-\theta)^2}{\mathbb{Var}[\hat{\theta}]} \xrightarrow{d}$$

(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 7

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] Formulate Mendel's conjecture as a null and alternative hypothesis. Construct your own notation.

(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 $F_{\chi^2_3}(7.81) = 95\%$.

Problem 8

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.

(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.

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

(e) [difficult] [MA] There is also a "chi-squared test of homogeneity". Here we are testing the equality of proportion among many groups at once. If there was only one group, then you would use a two-proportion z-test. It turns out the test statistic is the same and it is asymptotically χ^2 with degrees of freedom being the number of tests minus 1 (just like in the setting of the goodness of fit test). 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

Write H_0 homogeneity between male and female across hair colors and H_a: heterogeneity between male and female across hair colors using this class's canonical notation. Then run the test.

Problem 9

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: Cauchy(θ_1, θ_2)

MOD 3: Logistic(θ_1, θ_2)

MOD 4: Laplace(θ_1, θ_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: Cauchy$(-0.182, 0.391)$. AIC $= 26.899$

MOD 3: Logistic$(0.028, 0.345)$. AIC $= 21.689$

MOD 4: Laplace$(-0.176, 0.496)$. AIC $= 23.843$

(a) [harder] Compute $\ell\left(\hat{\theta}_1^{\text{MLE}}, \hat{\theta}_2^{\text{MLE}}; x_1, \ldots, x_{10}\right)$ 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 $\ell\left(\hat{\theta}_1^{\text{MLE}}, \hat{\theta}_2^{\text{MLE}}; x_1, \ldots, x_{10}\right)$ 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?

(h) [E.C.] Prove the bias term from the lecture is K_m. State all assumptions