MATH 369/650 Fall 2020 Homework #6

Professor Adam Kapelner

Due by email 11:59PM Tuesday, Decber 1, 2020

(this document last updated Wednesday 25th November, 2020 at 4:48pm)

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. practical significance, multiple comparison problem, familywise error rate control, false discovery rate control, ...) 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

Herein will examine data from an interesting experiment on ESP and psychic control. For those of you, who took / will take the 341 class, we examined / will examine it there too. The example comes from page 19 of a 2010 paper by Jose Bernardo, a world-famous statistician:

> "... the results reported by Jahn et al. (1987) using a random event generator based in a radioactive source, and arranged so that one gets a random sequence of 0's and 1's with theoretically equal probability for each outcome. A subject then attempted to mentally 'influence' the results so that, if successful, data would show a proportion of 1's significantly different from 0.5. There were $n =$ 104,490,000 trials resulting in ... 52,263,471 successes."

We wish to test if the subject in Jahn et al. (1987) was able to "psychically influence" the radioactivity and provide $H_a : \theta \neq 0.5$. In this problem we will employ a 2-sided 1-proportion z test of the binomial proportion (it's good review for the final exam!)

(a) [easy] Find $\hat{\hat{\theta}}$ and show the standardized estimate is $\hat{\hat{\theta}}_{std} = 3.61$.

(b) [easy] A single hypothesis test with a standardized z-score estimate of 3.61 rejects at any reasonable α. In fact, the p value ≈ 0.0003. Is this result "statistically significant"? Yes / no

(c) [easy] Calculate the experimental *effect size* which in a 2-tailed test is defined by the absolute difference of $\hat{\hat{\theta}}$ minus the value of θ_0, the null hypothesis value.

(d) [difficult] Assume the experiment was perfectly executed with no other source of bias or cheating by the investigators whatsoever. Is the *effect size* they found "practically significant"? In other words, is the subject in the study truly a "psychic"? Discuss. There is no mathematics here. And there is no "right" answer but you must defend your opinion clearly using the concepts we discussed in the lecture.

Problem 2

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

(a) [easy] 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 × 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.

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

(c) [easy] In the case where all m H_0's are true and the m tests are independent, prove that the m p-values are realizations from $\mathcal{P}_1, \ldots, \mathcal{P}_m \overset{iid}{\sim} \mathrm{U}\left(0,\, 1\right)$ (a).

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

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

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

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

(h) [easy] Prove an upper bound on FWER when all m H_0's are true but the tests are dependent. Using this bound, give an expression for α, the p-value rejection threshold for an individual test. What is this expression called?

(i) [easy] Prove an upper bound on FWER when all m H_0's are true but the tests are *in*dependent. Using this bound, give an expression for α, the p-value rejection threshold for an individual test. What is this expression called?

(j) [easy] Describe the Simes procedure in detail.

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

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

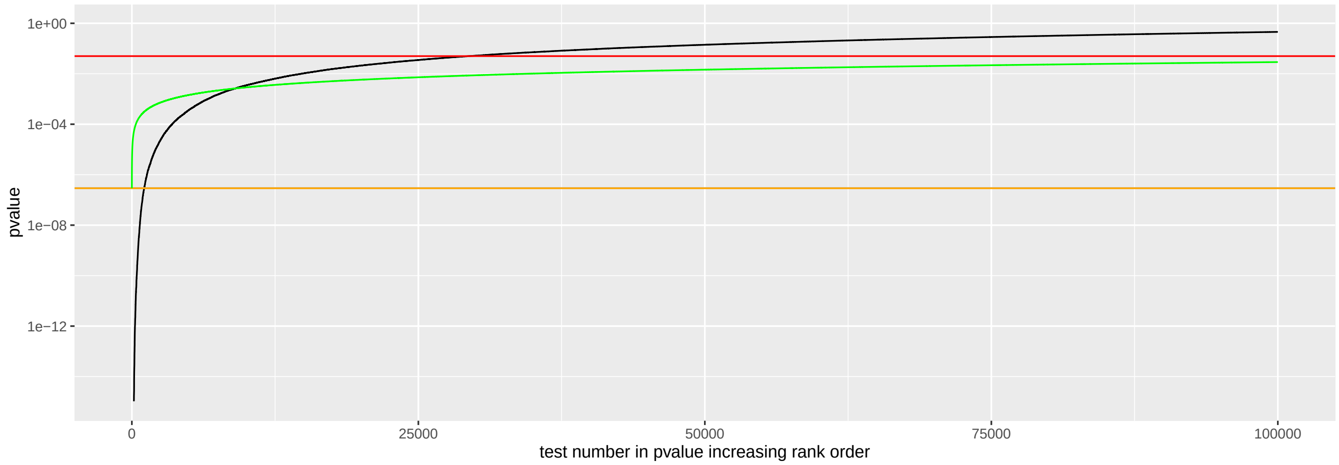
(m) [E.C.] Prove that Benjamini-Hochberg controls FDR regardless of how many m H_0's are true.

Problem 3

This problem will be about the multiple testing / multiple comparisons problem in the context of the IPMC data on investigating 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.

(a) [easy] 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.



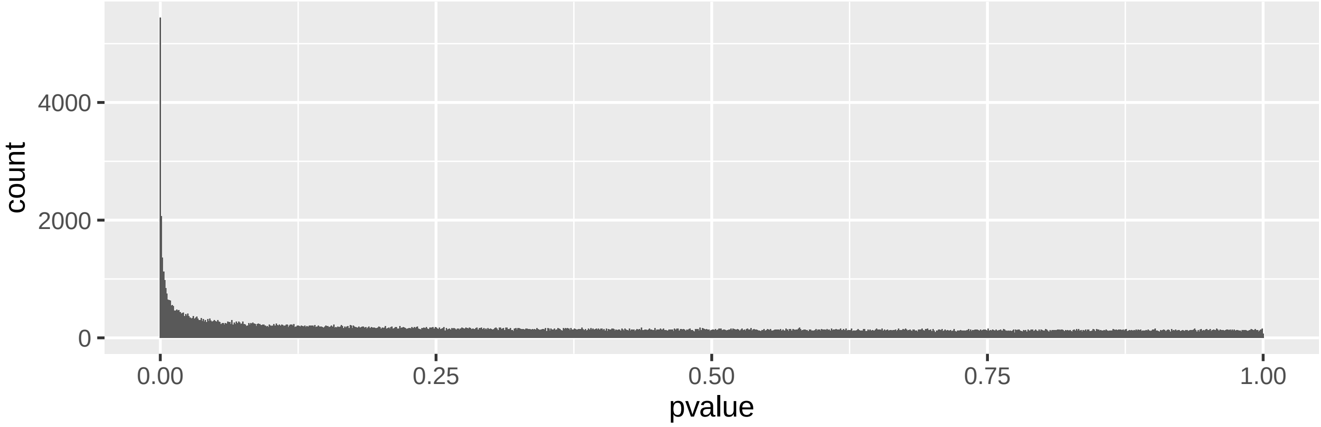
(b) [easy] Compute the Bonferroni and Sidak α thresholds. Ensure that the Bonferroni is smaller than the Sidak.

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

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

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

(f) [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.

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

Problem 4

This problem will cover the Wald Test for the MLE, 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 $H_0 : \theta = \theta_0$. We know by the central limit theorem (and also by the asymptotic normality of the MLE theorem) that under H_0 the standardized sampling distribution denoted Z is asymptotically normal:

$$Z = \frac{\bar{X} - \theta_0}{\sqrt{\frac{\theta_0(1-\theta_0)}{n}}} \overset{d}{\to} \mathcal{N}(0,\, 1)$$

Since the Wald test is defined as a z-test based on an asymptotically normal estimator, the Z above is the estimator for the Wald test. And thus the 1-proportion z-test is the Wald test in this setting.

Below are critical values for the chi-squared distribution that will be of use throughout the rest of the homework:

df	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
$F_{\chi^2_{df}}(\cdot) = 95\%$	3.84	5.99	7.81	9.49	11.07	12.59	14.07	15.51	16.92	18.31	19.68	21.03	22.36	23.68	25.00	26.30	27.59

(a) [easy] The data is $n = 100$ and $\bar{x} = 61$. We are testing against $H_0 : \theta = \frac{1}{2}$. Show that the Wald test statistic (the estimate) is $z = 2.2$. Would you reject the null hypothesis at $\alpha = 5\%$ using the Wald test?

(b) [easy] If you square the Wald test statistic, you get an equivalent test

$$Z^2 = \frac{(\bar{X} - \theta_0)^2}{\frac{\theta_0(1-\theta_0)}{n}} \overset{d}{\to} \chi^2_1$$

Some textbooks define this as Wald test. Compute the test statistic (the estimate) for the data in (a) and show you reach the same decision in your hypothesis test.

(c) [easy] Prove the Score test for one parameter for any iid DGP $f(x;\theta)$, i.e., declare the estimator and provide its exact or approximate distribution (lecture 19).

(d) [harder] Show that the score test is equivalent to the Wald test in the case where the DGP is iid $\text{Bern}\,(\theta)$. This means the estimator is the same. For all the formulas you need, see middle of lecture 10 where we derived $I(\theta)$ for the iid $\text{Bern}\,(\theta)$ DGP. Then it's algebraic simplication from there.

(e) [easy] 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 (lec 20).

(f) [easy] Show that the LR test is *not* equivalent to the Wald test / Score test in the case where the DGP is iid $\text{Bern}\,(\theta)$. This means the estimator is *not* the same. This is marked easy because it appears at the end of lecture 19.

(g) [easy] Compute the test statistic (the LR test statistic which we denote $\hat{\hat{\Lambda}}$ in lecture) for the data in (a). Since the likelihood ratio test is not the same as the Wald test, your answer should be different than the answer in (b). But since the Wald, Score and Likelihood Ratio tests are all asymptotically equivalent, this means the answer here should be *approximately* the same numeric value as the answer you got in (b). Does the difference in value change your decision (do you still retain/reject H_0 as you did previously)?

(h) [easy] Plot a log-likelihood function vs $\hat{\hat{\theta}}$. Mark $\hat{\hat{\theta}}^{\text{MLE}}$ and θ_0, a value you're testing aginst 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.

(i) [E.C.] Prove the asymptotic equivalence of the Wald, Score and LR tests.

Problem 5

We will talk about the generalized Likelihood Ratio test here for testing the difference between a *full model* and a *reduced model* where the DGP is shared between the full model and the reduced model and the reduced model is said to be *nested in* the full model.

Consider the generalized logistic DGP:

$$f(x;\theta_1,\theta_2,\theta_3) = \frac{\theta_3 e^{-\frac{x-\theta_1}{\theta_2}}}{\theta_2\left(1+e^{-\frac{x-\theta_1}{\theta_2}}\right)^{\theta_3+1}}$$

The reason why it's called the "generalized" logistic model is because it adds another parameter to the logistic model allowing for more flexibility.

(a) [easy] If we were testing against $H_0 : \theta_1 = \theta_{1_0}$ and $\theta_2 = \theta_{2_0}$ and $\theta_3 = \theta_{3_0}$ via the LR test, what would be the asymptotic distribution of $\hat{\Lambda}$? What is the critical threshold value to reject at $\alpha = 5\%$?

(b) [easy] If we were testing against $H_0 : \theta_1 = \theta_{1_0}$ and $\theta_2 = \theta_{2_0}$ via the LR test, what would be the asymptotic distribution of $\hat{\Lambda}$? What is the critical threshold value to reject at $\alpha = 5\%$?

(c) [easy] If we were testing against $H_0 : \theta_1 = \theta_{1_0}$ via the LR test, what would be the asymptotic distribution of $\hat{\Lambda}$? What is the critical threshold value to reject at $\alpha = 5\%$?

(d) [easy] If we set $\theta_3 = 1$, the generalized logistic gives us back the vanilla logistic model:

$$f(x;\theta_1,\theta_2,1) = \frac{e^{-\frac{x-\theta_1}{\theta_2}}}{\theta_2\left(1+e^{-\frac{x-\theta_1}{\theta_2}}\right)^{2}}$$

We wish to test the full model (the generalized logistic) vs the reduced model (the logistic which is the full model restricted to $\theta_3 = 1$). What would be the null hypothesis of this test?

(e) [difficult] Derive the LR test statistic $\hat{\hat{\Lambda}}$ for the test in (d) given a sample size n of iid data. Simplify as much as you can. I suggest you use the notation a, b, c, d, e, etc for the different maximum likelihood estimates. Remember the MLE's are different in the numerator and the denominator even though they are estimating the same parameter!

(f) [harder] For the full model let $\hat{\theta}_1^{\text{MLE}} = 12.22$, $\hat{\theta}_2^{\text{MLE}} = 4.03$ and $\hat{\theta}_3^{\text{MLE}} = 3.49$. For the reduced model let $\hat{\theta}_1^{\text{MLE}} = 18.65$ and $\hat{\theta}_2^{\text{MLE}} = 3.09$. Let $x_1 = 21.86$ $x_2 = 20.71$ and $x_3 = 16.11$. Although $n = 3$ is definitely not a large enough sample size for the asymptotic distribution to kick in, nevertheless compute $\hat{\hat{\Lambda}}$. This will be some boring calculation.

(g) [harder] For the JFK windspeed data on midterm 2, question 8, the log-likelihood for the generalized logistic model (the full model) is -1129.654 and the log-likelihood for the logistic model (the reduced model) is -1138.298. Calculate the LR test's statistic $\hat{\hat{\Lambda}}$. Do you reject or retain H_0? Can you explain what your rejection or retainment means in a few sentences?

Problem 6

This problem is further about the LR test. Below I describe a problem setting that is motivated by Abhinav's question on slack. You do not need to understand what's below and you can skip it if you wish but this is an example of how the LR test is used by real practicing statisticians.

Imagine a clinical trial which is a randomized experiment testing a treatments for depression (T_1: therapy vs C: no treatment). The usual goal is to measure the *treatment effect*i.e. the difference between these two treatments (which we call $\theta_T - \theta_C$) and then ascertain if the treatment does better than the control (i.e. reject $H_0 : \theta_T - \theta_C = 0$). We will be talking about this classic setting during the last two lectures of the course. It is very important outside of clinical trials by the way: Amazon is running 1000's of experiments all the time!

To run the hypothesis test, the standard methodology is linear regression which is taught in ECON 382 and MATH 342. As a secondary goal, we also wish to measure the effect of the subjects' characteristics. In this study we measure ten of them e.g. is this person married? does this person have prior drug usage? how bad was their depression symptoms when the study began? etc. So the total number of parameters in the model is ten plus a intercept to allow for an overall average plus a nuisance parameter for the variance (like we've seen in our testing as well) for a total of 13 parameters.

However, Abhinav was interested in seeing whether the treatment effect differs based on the subjects. To do this test, the standard methodology is to interact the treatment with the ten characteristics creating ten more parameters for a total of 23 parameters (the full model). We then ask the question: is it truly a better explanatory model to add this complexity? Can we get away with having only the original model (which is now termed the reduced model).

(a) [easy] As described above, the full model has 23 parameters and the reduced model as 13 parameters. We're testing if we need the full model to explain our data. Hence the null hypothesis is that the reduced model is true. What is the approximate distribution of the the LR test statistic $\hat{\Lambda}$?

(b) [easy] We fit the models using maximum likelihood and then compute the log likelihoods numericaly. The full model has log likelihood -473.3 and the reduced model has log likelihood -489.2. Run the test and provide your conclusion and write a couple sentences to explain it.