MATH 341 / 650 Spring 2017 Homework #1

Professor Adam Kapelner

Due *in class*, Thursday, February 9, 2016

(this document last updated Tuesday 31^{st} January, 2017 at 11:28pm)

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 241 concerning random variables, support, parameter space, PMF's, PDF's, CDF's, Bayes Rule, read about parametric families and maximum likelihood estimators on the Internet, read the preface and ch 1 and 4 of Bolstad and read the preface and Ch1 of McGrayne.

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.

Problems marked "[MA]" are for the masters students only (those enrolled in the 650 course). For those in 390, doing these questions will count as extra credit.

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 10 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

These exercises will review the Bernoulli model.

(a) [easy] If $X \sim \text{Bernoulli}(\theta)$, find $\mathbb{E}[X]$, $\mathbb{V}\text{ar}[X]$, $\text{Supp}[X]$ and Θ. No need to derive from first principles, just find the formulas.

(b) [harder] If $X \sim \text{Bernoulli}(\theta)$, find $\text{median}[X]$.

(c) [harder] If $X \sim \text{Bernoulli}(\theta)$, write the "parametric statistical model" below using the notation we used in class only.

(d) [harder] Explain what the semicolon notation in the previous answer indicates.

(e) [easy] If $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$, find the likelihood, $\mathcal{L}$, of θ.

(f) [difficult] Given the likelihood above, what would $\mathcal{L}$ be if the data was $< 0, 1, 0, 1, 3.7 >$? Why should this answer have to be?

(g) [easy] If $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}\,(\theta)$, find the log-likelihood of θ.

(h) [difficult] [MA] If $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$, explain why the log-likelihood of θ is normally distributed if n gets large.

(i) [easy] If $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}\,(\theta)$, find the score function (i.e the derivative of the log-likelihood) of θ.

(j) [harder] If $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}\,(\theta)$, find the maximum likelihood estimator for θ.

(k) [easy] If $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}\,(\theta)$, find the maximum likelihood *estimate* for θ.

(l) [easy] Given the previous two questions, describe the difference between a random variable and a datum.

(m) [easy] If your data is $< 0, 1, 1, 0, 1, 1, 0, 1, 1, 1 >$, find the maximum likelihood estimate for θ.

(n) [easy] Given this data, find a 99% confidence interval for θ.

(o) [easy] Given this data, test $H_0 : \theta = 0.5$ versus $H_a : \theta \neq 0.5$.

Problem 2

We will review the frequentist perspective here.

(a) [difficult] Why do frequentists have an insistence on θ being a fixed, immutable quantity?

(b) [easy] What are the three goals of inference?

(c) [easy] What are the three reasons why *frequentists* (adherents to the frequentist perspective) use MLEs i.e. list three properties of MLEs that make them powerful.

(d) [difficult] [MA] Give the conditions for asymptotic normality of the MLE,

$$\frac{\hat{\theta}_{\text{MLE}} - \theta}{\mathbb{SE}\left[\hat{\theta}_{\text{MLE}}\right]} \overset{\mathcal{D}}{\to} \mathcal{N}\left(0,\, 1\right).$$

You can find them online.

(e) [difficult] [MA] In class we said that $\mathbb{S}\text{E}\left[\hat{\theta}_{\text{MLE}}\right]$ cannot be found without θ so we substituted $\hat{\theta}_{\text{MLE}}$ into $\mathbb{S}\text{E}\left[\hat{\theta}_{\text{MLE}}\right]$ and called it $\hat{\mathbb{S}\text{E}}\left[\hat{\theta}_{\text{MLE}}\right]$ (note the hat over the SE). Show that this too is asymptotically normal, *i.e.*

$$\frac{\hat{\theta}_{\text{MLE}} - \theta}{\hat{\mathbb{S}\text{E}}\left[\hat{\theta}_{\text{MLE}}\right]} \overset{\mathcal{D}}{\to} \mathcal{N}\left(0,\, 1\right)$$

You need the continuous mapping theorem and Slutsky's theorem.

(f) [easy] [MA] Explain why the previous question allows us to build asymptotically valid confidence intervals using $\left[\hat{\theta}_{\text{MLE}} \pm z_{\alpha/2}\hat{\mathbb{S}\text{E}}\left[\hat{\theta}_{\text{MLE}}\right]\right]$.

(g) [harder] Why does all of frequentist inference break down if n isn't large?

(h) [easy] Write the most popular two frequentist interpretations of a confidence interval .

(i) [harder] Why are each of these unsatisfactory?

(j) [easy] What are the two possible outcomes of a hypothesis test?

(k) [harder] What is the weakness of the interpretation in both outcomes?

(l) [difficult] [MA] What is the weakness of the interpretation of the p-val?

Problem 3

We review and build upon conditional probability here.

(a) [easy] Explain why $\mathbb{P}(B \mid A) \propto \mathbb{P}(A, B)$.

(b) [easy] if B represents the hypothesis or the putative cause and A represents evidence or data, explain what Bayesian Conditionalism is, going from which probability statement to which probability statement.

(c) [harder] In class we presented the posterior odds form of Bayes Theorem. Prove it below.

(d) [harder] Show that the Bayes Factor is the ratio of posterior odds of the hypothesis to prior odds of the hypothesis.

(e) [easy] On the wikipedia page about Bayes Factors, Harrold Jeffreys (who we will be returning to later in the semester) gave interpretations of Bayes Factors (which is denoted K there and B in Bolstad's book on page 70). Give the ranges of K here (not in terms of powers of 10, but as a pure number) for his interpretations i.e. "negative," "strong," etc.

(f) [difficult] [MA] Conceptually why should the likelihood being greater than $\mathbb{P}(A)$ imply that the hypothesis is more likely after observing the data than before?

Problem 4

We examine here paternity testing (i.e. answering the question "is this guy the father of my child?") via the simplistic test using blood types. These days, more advanced genetic methods exist so these calculations aren't made in practice, but they are a nice exercise.

First a crash course on basic genetics. In general, everyone has two alleles (your genotype) with one coming from your mother and one coming from your father. The mother passes on each of the alleles with 50% probability and the father passes on each allele with 50% probability. One allele gets expressed (your phenotype). So one of the genes shone through (the dominant one) and one was masked (the recessive one). Dominant blood types are A

Genotype	Frequency
OO	0.52
AA	0.0196
AO	0.2016
BB	0.0196
BO	0.2016
AB	0.0392

and B and the recessive type is o (lowercase letter). The only way to express phenotype o is to have genotype oo i.e. both genes are o. There is an exception; A and B are codominant meaning that blood type AB tests positive for both A and B.

In this case consider a child of blood type B and the mother of blood type A. Using this hereditary guide, we know that the mother's type must be Ao so she passed on an o to the child thus the child got the B from the father. Thus the father had type AB, BB or Bo. I got the following data from this paper (so let's assume this case is in Nigeria in 1998).

(a) [easy] Bob is the alleged father and he has blood type B but his genotype is unknown. What is the probability he passes on a B to the child?

(b) [easy] What is the probability a stranger passes on a B to the child?

(c) [easy] Assume our prior is 50-50 Bob is the father, the customary compromise between a possibly bitter mother and father. What is the prior odds of Bob being the father? Don't think too hard about this one; it is marked easy for a reason.

(d) [difficult] We are interested in the posterior question. What is the probability Bob is the father given the child with blood type B?

(e) [difficult] What is the Bayes Factor here? See (a) and (b).

(f) [easy] What is the probability Bob is not the father given the child with blood type B? Should be easy once you have (c) and (e).