MATH 341 / 650.3 Spring 2019 Homework #1

Frank Palma Gomez

Tuesday 11th February, 2020

Problem 1

These are questions about McGrayne's book, preface, chapter 1, 2 and 3.

(a) [easy] Explain Hume's problem of induction with the sun rising every day.

Hume believed that we can't be absolutely certain about anything that involves cause and effect, traditional beliefs, and habitual relationships. For that reason, Hume disregards the notion of the sun rising every day because we are not absolutely certain what the next day will bring.

(b) [easy] Explain the "inverse probability problem."

The inverse probability problem states that instead of arriving to an effect through a cause, you could instead have the effect lead to the cause. Bayes did so by inventing a number called it a guess. As he gathered information he refined the guess in order for it to be as close as the actual value.

(c) [easy] What is Bayes' billiard table problem?

Baye's billiard table problem involves a square table that if a ball a thrown on it, it would have the same chance of landing on one spot as on any other.

(d) [difficult] [MA] How did Price use Bayes' idea to prove the existence of the deity?

(e) [easy] Why should Bayes Rule really be called "Laplace's Rule?"

Laplace had made probability-based statistics commonplace. Laplace put Bayes rule into modern terms far beyond the theoretical gambling problems that Bayes was putting the formula to use.

(f) [difficult] Prove the version of Bayes Rule found on page 20. State your assumption(s) explicitly. Reference class notes as well.

(g) [easy] Give two scientific contexts where Laplace used inverse probability theory to solve major problems.

Laplace used the inverse probability theory to calculate the gravitational attraction on the motion of the moon and to calculate the motions of Jupiter and Saturn.

(h) [difficult] [MA] Why did Laplace turn into a frequentist later in life?

(i) [easy] State Laplace's version of Bayes Rule (p31).

The probability of a hypothesis (given information), equals to the initial estimate of its probability times the probability of each new piece of information under the hypothesis, divided by the sum of the probabilities of the data all possible hypothesis.

(j) [easy] Why was Bayes Rule "damned" (pp36-37)?

Bayes rule was "dammed" due to the religious views that the critics themselves had. They claimed that it was filled with subjective and objective views.

(k) [easy] According to Edward Molina, what is the prior (p41)?

Edwards Molina definition of the prior is the collateral information. Since statisticians were forced to make decisions based on insufficient data, in such cases they relied on prior knowledge, called collateral information.

(l) [easy] What is the source of the "credibility" metric that insurance companies used in the 1920's?

The source of "credibility" was the industry wide experience and the local businesses for new data.

(m) [easy] Can the principle of inverse probability work without priors? Yes/no.

No

(n) [difficult] In class we discussed / will discuss the "principle of indifference" which is a term I borrowed from Donald Gillies' Philosophical Theories of Probability. On Wikipedia, it says that Jacob Bernoulli called it the "principle of insufficient reason". McGrayne in her research of original sources comes up with many names throughout history this principle was named. List all of them you can find here.

(o) [easy] Jeffreys seems to be the founding father of modern Bayesian Statistics. But why did the world turn frequentist in the 1920's? (p57)

Due to Fisher's argumentative nature, the book describes Jeffrey's as a mild character. At the same, Quantum Mechanics became popular and scientists used frequencies in order to know the location of an electron.

Problem 2

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.

- $\mathbb{E}[X] = \theta$
- $\mathbb{V}\text{ar}[X] = \theta(1-\theta)$
- $\text{Supp}[X] = \{k \in [0,1]\}$
- $\Theta = [0,1]$

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

$$\begin{cases} 0 & \theta < 1/2 \\ [0,1] & \theta = 1/2 \\ 1 & \theta > 1 \end{cases}$$

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

If $X \sim \text{Bernoulli}(\theta)$, then $\mathcal{F} = \{\ P(x;\theta) : \theta \in \Theta\ \}$

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

Represents the conditional distribution of X given θ

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

$$\mathcal{L}(\theta; x) = \prod_{i=1}^{n} p(x_i;\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}$$

(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?

$$\mathcal{L}(\theta, x) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{5.7}(1-\theta)^{-1.7}$$

Since $3.7 \notin \text{Supp}[X]$, $\mathcal{L}(\theta, x) = 0$

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

$$\ell(\theta) = \ln(\theta)(n\bar{X}) + \ln(1-\theta)(n - n\bar{X})$$

(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.

By the law of large numbers and the central limit theorem the log-likelihood converges to θ as $n \to \infty$

$$\sqrt{n}(\bar{X} - \theta) \to \mathcal{N}\left(0, \mathbb{V}\text{ar}\left[X_i\right]\right) = \mathcal{N}\left(0, \theta\right)$$

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

$$\ell'(\theta) = n(\frac{\bar{X}}{\theta} - \frac{1-\bar{X}}{1-\theta})$$

(j) [harder] If $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}\left(\theta\right)$, find the maximum likelihood estimator for θ. An "estimator" is a random variable. Thus, it will be an uppercase letter.

$$\hat{\theta} = argmax\{\mathcal{L}(\theta)\}$$

(k) [easy] If $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}\left(\theta\right)$, find the maximum likelihood *estimate* for θ. An "estimate" is a number. Thus, it will be a lowercase letter.

$$\hat{\theta}_{MLE} = \frac{1}{n}\sum_{i=1}^{k} X_i$$

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

$$\hat{\theta}_{MLE} = \bar{X} = \frac{1}{n}\sum_{i=1}^{k} X_i = \frac{0+1+1+0+1+1+0+1+1+1}{10} = \frac{7}{10} = 0.7$$

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

Recall Confidence Interval:

$$CI_{\theta,1-\alpha} = [\hat{\theta}_{MLE} \pm z_{\frac{\alpha}{2}} SE[\hat{\theta}_{MLE}]]$$

We know the value of $\hat{\theta}_{MLE} = 0.7$
We know the value of $\alpha = 1$ - $0.99 = 0.01$
Thus, $Z = 1 - \frac{\alpha}{2} = 1 - \frac{0.01}{2} = 0.995$

Therefore

$$[0.7 - (0.995)(0.144), 0.7 + (0.995)(0.144)] = [0.55672, 0.84328]$$

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

(o) [easy] Write the PDF of $X \sim \mathcal{N}(\theta, 1^2)$.

$$\mathcal{N}(\theta, 1^2) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(x-\theta)}{2}\right)$$

(p) [difficult] Find the MLE for θ if $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\theta, 1^2)$.

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(X_i-\theta)}{2}\right)$$

$$\ell(\theta) = \ln(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(X_i-\theta)}{2}\right)) = \sum_{i=1}^{n} \ln(\frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(X_i-\theta)}{2}\right))$$

$$= \sum_{i=1}^{n} [-\ln(\sqrt{2\pi}) - \frac{1}{2}(X_i - \theta)^2]$$

$$\ell'(\theta) = -n\ln(\sqrt{2\pi}) - \sum_{i=1}^{n} \frac{1}{2}(X_i - \theta)^2 = \sum_{i=1}^{n} X_i - n\theta$$

$$\ell'(\theta) = 0$$

$$\hat{\theta}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

(q) [difficult] [MA] Find the MLE for θ if $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. Solve the system of equations $\frac{\partial}{\partial \mu}[\ell(\theta)] = 0$ and $\frac{\partial}{\partial \sigma^2}[\ell(\theta)] = 0$ where $\ell(\theta)$ denotes the log likelihood. You can easily find this online. But try to do it yourself.

Problem 3

We will review the frequentist perspective here.

(a) [difficult] Why do frequentists have an insistence on θ being a fixed, immutable quantity? We didn't cover this in class explicitly but it is lurking behind the scenes. Use your reference resources.

(b) [easy] What are the three goals of inference? Give short explanations.

- Provides an estimate for θ. Point estimation
- Provides a confidence set which allows us to see a range of possible values for θ
- Testing allows us to test hypothesis. We define a hypothesis and allow the data to tell us if we are correct or not

(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.

- Provides *consistency* as $\hat{\theta}_{MLE} \approx \theta$
- MLE's have an *assymptotic normality*. $\hat{\theta}_{MLE} \approx \mathcal{N}\left(\theta, SE(\hat{\theta}_{MLE})\right)$
- Among all consistent estimators $\hat{\theta}_{MLE}$ has the lowest variance therefore it is efficient

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

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

You can find them online.

(e) [difficult] [MA] The standard error of the estimator, $\mathbb{SE}\left[\hat{\theta}_{\text{MLE}}\right]$ cannot be found without the true value of θ. If we had the true value of θ we wouldn't be doing inference! So we substituted $\hat{\theta}_{\text{MLE}}$ (the point estimate) into $\mathbb{SE}\left[\hat{\theta}_{\text{MLE}}\right]$ and called it $\hat{\mathbb{SE}}\left[\hat{\theta}_{\text{MLE}}\right]$ (note the hat over the SE). Show that this too is asymptotically normal, *i.e.*

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

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 $\left[\hat{\theta}_{\text{MLE}} \pm z_{\alpha/2}\hat{\mathbb{SE}}\left[\hat{\theta}_{\text{MLE}}\right]\right]$.

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

When n is small $\hat{\theta}_{MLE}$ does not approximate to $\mathcal{N}\left(\theta, SE(\hat{\theta}_{MLE})\right)$

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

- If we repeat the experiment multiple times with a 95% the confidence interval will include θ
- A confidence interval provides a set of θ values that can be tested.

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

Because both of these methods require a significant amount of data in order to produce true positive values.

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

Null Hypothesis

$$H_0 : \theta = \theta_0$$

Alternative Hypothesis

$$H_a : \theta \neq \theta_0$$

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

- If the sample size is large enough, it is possible to have a significant p-value.
- If the sample size is small there can be statistical insignificance dispite having a large maginitude of association.

- There might be a statistical significance even if the data demostrate a bias which affects the accuracy of the data.

Problem 4

We review and build upon conditional probability here.

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

If various events are equally likely, and an event is observed, then the probability for the alternative events are proportional to the probabilities of that the observed event would have occurred under those other alternate events.

(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.