Bayesian Data Science: Assignment 1

Instructions:

- Please show all relevant working and use R for all statistical programming and analysis. This assignment will cover Part 1 of the course: Chapters 1-5 "Doing Bayesian Data Analysis", second edition by JK. This assignment will be graded on 100 points.
- Use "R Mark down" to construct your assignment answers. Use appropriate Latex formulae (https://en.wikibooks.org/wiki/LaTeX/Mathematics) and r code chunks.
    - Please submit your Rmd document, html and pdf knit documents.
    - Make sure all of the files pertaining to the assignment are in the same directory.
    - I will run your Rmd document and check that your code runs
    - The functions you make must have arguments suitable to solve the problem and produce the desired output. **It is up to YOU to determine what arguments are needed**.
- This assignment assumes that you have already installed the following software (all of which are free)
    - R
    - R Studio
    - Latex (best to have a full distribution)

Questions:

1. A coin is tossed 10 times with 4 heads and 6 tails. Suppose we know that the trials are all independent and the coin is unaltered from trial to trial. Suppose also that the event "Head" is a success. Using this information answer the following questions:

a. Use classical methods to obtain a point estimate for p, the probability of a success.
b. Use classical methods and the normal approximation to the binomial to find a 95% interval estimate for p.
c. Make an R function called **`mycoin()`** that will produce:
    i. A Bayes' box table for the above binomial experiment. The Bayes' box must have the following column headings: p, prior, likelihood, h, posterior. Where "h" is the prior*likelihood.
    ii. The Bayes' box should include appropriate totals to check the probabilities conform to distributional properties.
    iii. A graph with all three curves: The prior, likelihood and posterior. It should be color coded and have a legend. You may use base R graphics or ggplot.
    iv. The posterior mean (the point estimate for p)
    v. A (1-alpha)100% BCI (Bayesian credible interval) for "p"
    vi. A classical estimate for "p"
    vii. A classical (1-alpha)100% ci for "p"
    viii. These outputs (Table, all estimates) must be placed inside a list that is released to the command line when the function is called.
    ix. Also the Bayes' box table must be saved to a csv file in the working directory
    x. The plot should also be saved as a .jpg file to the working directory.
    xi. The function arguments will include a vector of p values (use seq(0,1,length), x, n, prior vector, alpha.
    xii. Give the output when the following is called:

1. `mycoin(p = seq(0,1,length=20), prior =rep(1/20, 20), n=10, x=4, alpha = 0.05)`
2. `mycoin(p = seq(0,1,length=40), prior =rep(1/40, 40), n=10, x=4, alpha = 0.05)`
3. `mycoin(p = seq(0,1,length=20), prior =rep(1/20, 20), n=10, x=4, alpha = 0.1)`
4. `mycoin(p = seq(0,1,length=40), prior = pr, n =10, x=4, alpha=0.05)` and

2. If $X \sim Bin(n.p)$ show using the definition $V(X) = E(X-\mu)^2$ that the variance of X is npq where $q = 1-p$.

3. The moment generating function (mgf) is defined as $M_X(t) = E(e^{Xt}) = \sum_X e^{Xt}p(x)$

a. Show that if $X \sim Bin(n,p)$ then $M_X(t) = (q+pe^t)^n$ where $q = 1-p$.

b. One property of the mgf is the following: $\frac{d^k M_X(t)}{dt^k}$ evaluated at t=0 yields $E(X^k)$.

i. Using this property find $E(X^2)$

ii. Show that $V(X) = E(X^2) - \mu^2$ using the definition of $V(X) = E[(X-\mu)^2]$

iii. Using these results show $V(X) = npq$

4. One very important continuous distribution is the Normal. Say that $Y \sim N(\mu, \sigma^2)$. Make an R function called **`mynorm()`** that will do the following:

a. Make a plot of the normal density. (You may use Base R)

i. The plot should show an area above the interval whose x co-ordinates are `c(a,b)`.

ii. Has x, y and main titles.

iii. Has a legend explaining the shaded area.

iv. The area should have a 4 decimal place approximation to the area. You may wish to use **`pnorm()inside your function.`**

b. Command line output should include a list of the following objects:

i. The shaded area to 4 dec places.

ii. The $\frac{\alpha}{2}, 1-\frac{\alpha}{2}$ quantiles where $\alpha \in (0,1)$

iii. Hint: You may wish to use **`qnorm()`**

c. Invoke your function with $\mu = 10, \sigma = 8,\ a = 8, b = 11, \alpha = 0.10$ – this will test that your function works.

5. One Classical method for point estimation is the method of maximum likelihood (mle) developed originally by R.A. Fisher. In this method data are assumed the result of an independent and identically distributed process. The method forms a connecting thread between the classical approach and the Bayesian paradigm and for this reason it will be well for us to review it and use it. See https://en.wikipedia.org/wiki/Maximum_likelihood_estimation . The maximum likelihood algorithm is essentially carried out in the following way:

a. Express the joint distribution of the data as a product (independence assumption)

b. Look at this as a function of the parameters to be estimated. Call it the likelihood $L(\theta)$

c. Take the log (base e) of this, call it $l(\theta) = \log(L(\theta))$

d. Find the value of θ that maximizes this function i.e -- $l'(\hat{\theta}) = 0$

e. Check that $\hat{\theta}$ corresponds to a maximum.

Now you will use this algorithm to find the maximum likelihood estimate of λ in the following problem.

Let $X_1, \ldots, X_n$ be a random sample from a $Pois(\lambda)$ distribution where λ is an unknown positive parameter. Find the maximum likelihood estimator $\hat{\lambda}$.

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}, x = 0,1,2, \ldots$$

1. Find $\hat{\lambda}$ as a formula.
2. Find the second derivative of $l(\lambda)$ as a formula.
3. Show $\hat{\lambda}$ is a maximum. (how will you do this?)
4. In the algorithm we form the log likelihood – show using calculus that the max of $L(\theta)$ is the same as the maximum of $l(\theta)$.
5. Make a function called **`myml(x, ...)`** that will draw the graph of $L(\lambda)$ and $l(\lambda)$ with the x = vector of data. Make sure the function produces:
    a. Plot of l and L
    b. $\hat{\lambda}$ plotted on the graph with 4 decimal places of accuracy – you can use text(), round()
    c. Command line output that has a list containing the estimate.
    d. Give the output of your function when x = {3,4,3,5,6}.
6. Explain how the maximum likelihood method differs from the Bayesian approach. Hint: Write down Bayes' formula – identify the parts.

6. An estimator $\hat{\theta}$ is said to be unbiased if $E(\hat{\theta}) = \theta$ using your result above for $\hat{\lambda}$ decide whether it is unbiased or not -- show working.

7. Our work on Bayesian theory will require us to look closely at the Binomial distribution (see question 1,2 above). If $X \sim Bin(n, \theta)$ where there is one Binomial experiment of n Bernoulli trials and X successes – find $\hat{\theta}$ the mle of θ as a formula.

8. Many problems work fine and analytical results will be productive. Sometimes this is not the case and numerical methods may in fact be the only way we can obtain estimates. Problems often reduce to finding the roots of functions. One such numerical method is the Newton Raphson algorithm.

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

Where f is the function we want to find the root of and n is the nth iteration.
Make an R function that will find the roots of a function f call it **`mynr(f,fdash,x0, ...)`** where f is the function , fdash its derivative and x0 the initial guess for the zero.
The function should do the following:

a. Produce a plot of the function (well labelled etc)
b. Show where the function cuts the x axis
c. Give the value of the root to whatever number of decimal places required.
d. Command line output that has the root in it.
e. Give the output of your function when $f(x) = x^2 - 5x + 6$ and $x_0 = 5$ with 4 dec places of accuracy.

9. Now alter your function **`mynr()and call it mynrml()`** to produce maximum likelihood estimates for the problem in question 5 with x = {3,4,3,5,6}. Show the results when the function is called.

10. Now make a NR function that does not require fdash. Alter **`mynr()`** and call it **`mynrf()`** so that it does not require fdash. Hint you can use an approximation to the derivative

$$f'(x) \approx \frac{f(x+h) - f(x)}{h}$$

Call **`mynrf()`** with $f(x) = x^2 - 5x + 6, x_0 = -5$