Problem Set #1

Due Wednesday January 29th

ANALYTICAL EXERCISES

1. Consider an urn containing N balls each labeled with a unique number $1,2,...,N$. M of these balls are colored red and the remaining $N-M$ are white, $0 \leq M \leq N$. Understand that these facts represents our state of knowledge, or information I. Let $R_i \equiv$ red ball on the i^{th} draw and $W_i \equiv$ white ball on the i^{th} draw. According to our knowledge of the composition of the urn, only red or white can be drawn, thus, for the i^{th} draw it must be that $p[R_i|I] + p[W_i|I] = 1$.

   a) Find the probability of drawing and red or white ball on the first draw, i.e. $p[R_1|I]$? and $P[W_1|I]$

   b) Find the probability of drawing a red ball on the first two draws, assuming you are sampling without replacement, i.e. find $p[R_1R_2|I]$

   c) Following this logic find the probability for drawing a red ball on the first $m \leq M$ consecutive draws. Show your logic.

   d) Find the probability for drawing exactly $m \leq M$ red balls in $n \leq N$ draws, regardless of order.

2. A *sample* is a set of n numbers $x = x_1, x_2, ..., x_n$. The *sample mean* is the average of the sample, $m[x] = \frac{x_1+...+x_n}{n}$, the *sample variance* is the average squared deviation of the sample values from the sample mean $s[x]^2 = \frac{(x_1-m[x])^2+...+(x_n-m[x])^2}{n}$, and the *sample standard deviation* is the square root of the sample variance.

   a) Suppose we model the sample as a constant, μ. In general the likelihood that all the numbers in the sample will be to equal μ is zero. In order to allow for deviations, suppose that we assume that likelihood of the sample is proportional to

$exp[-\frac{(x_1-\mu)^2+...+(x_n-\mu)^2}{n\sigma}]$, where σ is a second model parameter. Derive the expression for the posterior probability $p[\mu,\sigma|x]$, given a prior $p[\mu,\sigma]$.

b) *Jeffreys' prior* is $p[\mu,\sigma] = d\mu\frac{d\sigma}{\sigma}$. Letting $d\mu = d\sigma = 1$ Use Jeffreys' prior to find the maximum posterior probabilities, $(\hat{\mu},\hat{\sigma})$. (Hint: Use the first order conditions. You should get familiar looking results)

3. I'm thinking of a number between 1 and 100. How many bits (yes/no questions) of information are necessary in order to identify the exact number? Explain your reasoning.

4. Find the maximum entropy distribution for a random variable $X \in \mathbb{R}$ constrained to have a given mean $\int_x f[x]x dx = \mu$ and variance $\int_x f[x]x^2 dx = \sigma^2$.

5. Let $p[y,x]$ be given by

$X\downarrow$/Y$\rightarrow$	0	1
0	1/3	1/3
1	0	1/3

Find:

a) $H[X]$, $H[Y]$

b) $H[X|Y]$, $H[Y|X]$

c) $H[X,Y]$

d) $I[X;Y]$

6. Let the random variable X have three possible outcomes $\{a,b,c\}$ Consider two distributions on this random variable:

Symbol	$p[x]$	$q[x]$
a	1/2	1/3
b	1/4	1/3
c	1/4	1/3

Find:

a) $H[p]$, $H[q]$

b) $D[p||q]$, $D[q||p]$

c) Verify that $D[p||q] \neq D[q||p]$

R Exercises

Please email me your final cleaned R script annotated with comments.

1. A bank has made 100 mortgages of a new type (say it's 2005 and they are subprime mortgages), and all have been outstanding 5 years. Of these 100, 5 of them have defaulted. The bank would like to estimate the probability θ of default in the first five years for this type of mortgage, and get some idea of how much uncertainty there is about the probability, given the observed data. These being a new type of mortgage, the bank assigns a uniform prior over θ.

   a) Plot the likelihood in R and indicate on the plot (e.g. use the abline() function) the location of the maximum likelihood value of θ as well as the expected value of θ.

   b) Using the quantile function $qbeta()$ calculate and indicate a symmetric 95% confidence interval (cut off 2.5% of the left and right tail). Does this look like a reasonable confidence interval?

   c) Find the shortest interval with 95% probability (Hint: Feel free to use a package such as "TeachingDemos" that calculates the the highest posterior density region).

   d) Compare graphically the 95% confidence interval and the shortest 95% interval. Which do you favor and why?