Lab: Giving a good estimate for $p = p(H)$

This lab, worth 15 points, is due on Friday, November 15 in class.

Tabulated in the table to the right are some data based on flipping an unfair coin where $p = p(H)$ is the probability that *Heads* was flipped.

Sample size	Number H
10	1
100	16
1000	207
n	n_1

THE **Frequentist** APPROACH

1. Give the best approximation to p you can for each of the four sample sizes. Then briefly explain why this is the 'frequentist' approach and which estimate of p you believe to be the most accurate.

   A frequentist would likely suggest $\hat{p} = \frac{n_1}{n}$ for a good point estimator since the data shows this to be the proportion of H in the trials. For the sample sizes above, we find $\hat{p}_1 = .1$, $\hat{p}_{10} = .16$, $\hat{p}_{1000} = .207$. The estimate corresponding to $\hat{p}_{1000} = .207$ is the 'best' since a frequentist believes (and common sense dictates) that the more independent and identical trials you have, the better your estimate for the model parameter p is.

THE **Bayesian** APPROACH: Now you will give a Bayesian posterior estimate for p. This will be a density function for the random variable p. The next questions will step you through this.

2. *Preliminary:* Maximizing a function of the form $y = x^m(1-x)^n$ on the interval $[0,1]$ when $m, n > 1$.

   The quick way to maximize this function is to take its log and since $\log(y)$ is an increasing function, the value of x that maximizes $\log(y)$ is the same value that maximizes y. (Indeed, you *know* that y has a global maximum on $[0,1]$ by our study of Beta-distributed random variables so just finding critical points should be enough.) For starters, the value(s) of x that maximizes $y = x^m(1-x)^n$ is the same as the x that maximizes $\log(y) = \log(x^m(1-x)^n) = m\log(x) + n\log(1-x)$. Find the maximizer x of function y above in terms of m and n. (This is a Calculus I problem, find the critical points, etc.) Store this information away in your brain for use multiple times in the next problem.

   If $g = \log\left(x^m(1-x)^n\right)$, then $g' = \frac{m}{x} + \frac{n}{1-x}$ and setting this equal to zero we find that the critical point is $x = \frac{m}{m+n}$. Since the equation for y, up to a normalizing constant, is the density for a Beta-distributed random variable, this value of x is a maximizer, and there is no need to check the end points 0 and 1. Note that the mean of the Beta distribution is also the mode.

   The maximizer is $x = \underline{\frac{m}{m+n}}$.

3. If the sample size is n, then Y: *number of H in n coin tosses* is modeled by a binomial random variable $Y \sim \text{Binom}(n, p)$. Letting n_1 denote the number of H in n tosses and n_2 the number of T (so that $n_1 + n_2 = n$), then the likelihood function $L(p \mid \text{data}) = L(p \mid n_1, n_2) = P(\text{data} \mid p)$ is

$$L(p \mid n_1, n_2) = \binom{n}{n_1} p^{n_1}(1-p)^{n_2},$$

   and the Likelihood function is proportional to $L(p \mid \text{data}) \propto p^{n_1}(1-p)^{n_2}$.

   Stating Bayes' Rule in this context we have

$$P(p \mid \text{data}) = \frac{L(p \mid \text{data})\,\boldsymbol{\pi}(p)}{P(\text{data})}$$

   where the constituent parts are indicated in the following color-coded equation.

   Likelihood function ↓ Prior ↙

$$\text{Posterior probability of } p \to \; P(p \mid \text{data}) = \frac{L(p \mid \text{data})\,\boldsymbol{\pi}(p)}{P(\text{data})}$$

   The first simplification we will make is that the *posterior probability* of p, informally 'the posterior', is a **density** function. In particular, if we can find a formula $\varphi(p)$ for the posterior $P(p \mid \text{data})$ that is correct up to

a normalizing constant, then we can find the posterior exactly by integrating $\varphi(p)$ over the support $p \in [0,1]$ to find the posterior. That is, simply divide (or multiply) $\varphi(p)$ by an appropriate normalizing constant to get the posterior distribution for p.

To this end, the posterior of p is proportional to $\varphi(p)$, the product of the likelihood $L(p \mid \text{data})$ and the prior $\boldsymbol{\pi}(p)$ via

$$P(p \mid \text{data}) \;\propto\; \varphi(p) = p^{n_1}(1-p)^{n_2}\,\boldsymbol{\pi}(p).$$

Your goal is to find 1) the formula for the posterior distribution $P(p \mid \text{data})$; 2) plot the posterior density; and 3) find the mode of the posterior density for each of the three sample sizes $n = 1, 100, 1000$ and for two choices of prior $\boldsymbol{\pi}(p) \sim \text{Unif}(0,1)$ and $\boldsymbol{\pi}(p) \sim \text{Beta}(2,8)$. Then you will comment on your findings. To be clear, there are **SIX** posterior probabilities you are finding, one for each choice of sample size and prior.

SOLUTIONS

Below are the formulas and plots of the posterior probability densities, labelled by the choice of sample size and prior.

1. Sample size $n = 10$ and $\boldsymbol{\pi}(p) \sim \text{Unif}(0,1)$ so that $\boldsymbol{\pi}(p) = 1$ on $[0,1]$.

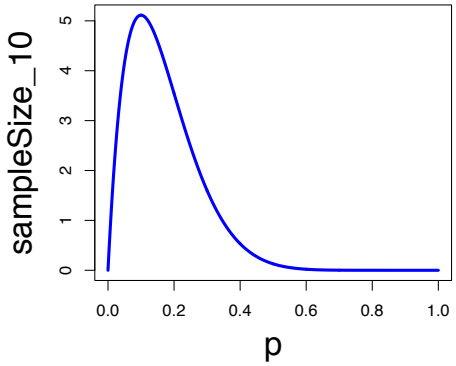
   A quick calculation shows that the posterior density is proportional to

   $$P(p \mid n_1 = 1) \;\propto\; \varphi(p) = p^1\,(1-p)^9 \cdot 1 = p\,(1-p)^9 \text{ for } 0 \le p \le 1.$$

   Thus, φ is, up to a constant, $\varphi(p) \sim \text{Beta}(2,10)$ and we compute that multiplying by $C = 1/\,\text{Beta}(2,10) = \dfrac{\Gamma(12)}{\Gamma(2)\Gamma(10)} = 11 \cdot 10 = 110$ will give us total probability 1. (The 110 is the reciprocal of the normalizing constant since we multiply by it, instead of dividing.) Putting this together, we find that the posterior is

   $$P(p \mid n_1 = 1) = 110\,p\,(1-p)^9 \text{ for } 0 \le p \le 1,$$

   whose graph is shown below.



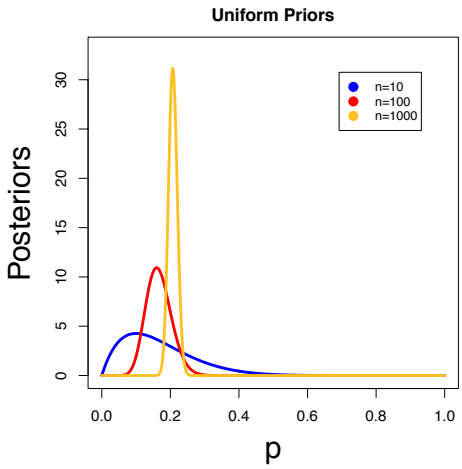
The global maximum of the posterior occurs at $p = .1$ and values of p around that value are also highly likely to be close to the true value of p.

The solutions to questions 2 and 3, and 4-6 have been combined. Notation used was set out above. Since the uniform distribution has density 1, the function $\varphi(p)$ is directly proportional to the likelihood which is of the form $\text{Beta}(n_1+1, n_2+1)$. Thus, it is easy to determine the normalizing constant: we simply multiply by $C = \frac{\Gamma(n+2)}{\Gamma(n_1+1)\Gamma(n_2+1)}$. (This was found by looking at the inside back cover of the textbook.) The support of the density is always $(0,1)$ and will not be mentioned explicitly.

2. Sample size $n = 100$ and $\boldsymbol{\pi}(p) \sim \text{Unif}(0,1)$ so that $\boldsymbol{\pi}(p) = 1$ on $[0,1]$.

3. Sample size $n = 1000$ and $\boldsymbol{\pi}(p) \sim \text{Unif}(0,1)$ so that $\boldsymbol{\pi}(p) = 1$ on $[0,1]$.

Sample size	n_1	$\varphi(p)$	C	Posterior $P(p \mid n_1)$
100	16	$p^{16}(1-p)^{84}$	$\frac{\Gamma(102)}{\Gamma(17)\Gamma(85)} \approx 1.36 \cdot 10^{20}$	C $p^{16}(1-p)^{84} \sim \text{Beta}(17,85)$
1000	207	$p^{207}(1-p)^{793}$	$\frac{\Gamma(1002)}{\Gamma(208)\Gamma(794)} \approx 9.21 \cdot 10^{222}$	C $p^{207}(1-p)^{793} \sim \text{Beta}(208,794)$



The plots of the three posterior probability densities for the uniform prior are shown to the left. Note that as the sample size increases, the plots get narrower, I.e., the variance decreases. All three plots suggest a probability $p \approx \frac{n_1}{n}$. That is to say, the *mode* of the density is the frequentist's estimate $\hat{p} = \frac{m}{m+n}$.

4. Sample size $n = 10$ and $\boldsymbol{\pi}(p) \sim \text{Beta}(2, 8)$ so that $\boldsymbol{\pi}(p) =$ _____ on $[0, 1]$.
5. Sample size $n = 100$ and $\boldsymbol{\pi}(p) \sim \text{Beta}(2, 8)$ so that $\boldsymbol{\pi}(p) =$ _____ on $[0, 1]$.
6. Sample size $n = 1000$ and $\boldsymbol{\pi}(p) \sim \text{Beta}(2, 8)$ so that $\boldsymbol{\pi}(p) =$ _____ on $[0, 1]$.

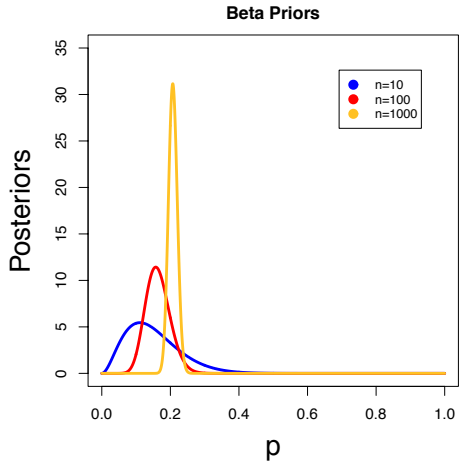
Solutions 4-6:

The density for the prior is proportional to $p(1-p)^7$ so $\varphi(p)$ is

$$\varphi(p) = p^{n_1}(1-p)^{n_2} \cdot p(1-p)^7 = p^{n_1+1}(1-p)^{n_2+7}$$
$$\sim \text{Beta}(n_1 + 2, n_2 + 8)$$

Again, it is easy to determine the normalizing constant: we simply multiply by $C = \frac{\Gamma(n+10)}{\Gamma(n_1+2)\Gamma(n_2+8)}$.

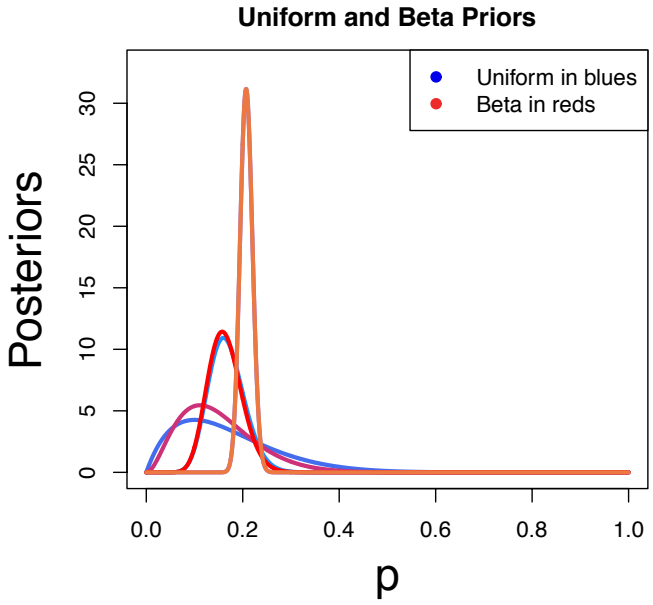
Sample size	n_1	$\varphi(p)$	C	Posterior $P(p \mid n_1)$	mode $= \frac{n_1+1}{n+8}$
10	1	$p^2(1-p)^{16}$	$\frac{\Gamma(20)}{\Gamma(3)\Gamma(17)} = 2907$	$\text{Beta}(3, 17)$	.11
100	16	$p^{17}(1-p)^{91}$	$\frac{\Gamma(110)}{\Gamma(18)\Gamma(92)} \approx 3.00 \cdot 10^{21}$	$\text{Beta}(18, 92)$	.16
1000	207	$p^{208}(1-p)^{800}$	$\frac{\Gamma(1010)}{\Gamma(209)\Gamma(801)} \approx 2.26 \cdot 10^{224}$	$\text{Beta}(209, 801)$	.21



The plots of the three posterior probability densities for the beta prior are shown to the left, and again as the sample size increases, the plots get narrower, I.e., the variance decreases. All three plots suggest a probability close to $p \approx .1, .16, .207$ respectively, but the modes are slight different. They are computed using $m/(m+n)$ and have value $2/18 \approx 0.11, 17/(17+91) = 17/108 \approx .16, 208/1008 \approx .21$. The effect of the Beta prior which has mean .2 is to move them slightly from $\hat{p} = n_1/n$.

7. Commentary: What do you notice about the posteriors as the sample size $n \to \infty$? Do you prefer a uniform prior or a beta prior? Why? What do you think about the Frequentist approach versus the Bayesian one? What have you learned? ....

So many possibilities here .... One remarkable thing to notice is that the choice of prior (uniform vs. beta) did not have all that much influence on the posterior density. Other good comments include noticing the narrowing of the posterior as $n \to \infty$, the relationship between the mode and $\hat{p}$, whether you like the idea of a *density* for estimating p or a point estimate, etc.


Uniform and Beta Priors
Uniform in blues
Beta in reds
Posteriors
0
5
10
15
20
25
30
0.0
0.2
0.4
0.6
0.8
1.0
p