2022 STAT/BIOSTAT 570: Final

To be handed in on Monday 12th December, 2022.

This is an exam, so no collaboration.

Consider the cycle to pregnancy data in Table 1. These data are from a study described in Baird and Wilcox (1985) in which women with planned pregnancies were interviewed and asked how many cycles it took them to get pregnant. Women were classified as current smokers if they reported smoking at least an average of 1 cigarette a day during the first cycle thy were trying to get pregnant.

Time (Cycle)	1	2	3	4	5	6	7	8	9	10	11	12	>12
Cycle to Pregnancy (Smokers)	29	16	17	4	3	9	4	5	1	1	1	3	7
Cycle to Pregnancy (Non-Smokers)	198	107	55	38	18	22	7	9	5	3	6	6	12

Table 1: Number of cycles until pregnancy for two groups of women, smokers and non-smokers.

We will begin by describing a model for a single group only.

1. We describe a simple model for these data. Let p $(0 < p < 1)$ denote the probability of conception during a particular cycle, and T the random variable describing the cycle at which pregnancy was achieved. Then T may be modeled as a geometric random variable:

$$\text{Pr}(T = t \mid p) = \begin{cases} p(1-p)^{t-1}, & t = 1, 2, 3, \ldots \\ 0 & \text{otherwise.} \end{cases}$$

Let Y_t represent the number of women that conceive in cycle t, $t = 1, \ldots, N$, and Y_{N+1} the number of women that have not conceived by cycle N.

(a) **5 marks** Show that the likelihood function is,

$$L(p) = \left\{ \prod_{t=1}^{N} \left[p(1-p)^{t-1} \right]^{Y_t} \right\} \times \left[(1-p)^N \right]^{Y_{N+1}}. \tag{1}$$

(b) **3 marks** Find an expression for the MLE, $\widehat{p}$.

(c) **3 marks** Find the form of the *observed* information and hence the asymptotic variance of the MLE.

(d) **4 marks** For the data in Table 1, calculate the MLEs $\widehat{p}_1$ (Smokers) and $\widehat{p}_2$ (Non-Smokers). the variance of $\widehat{p}_j$, and asymptotic 95% confidence intervals for p_j, $j = 1, 2$.

(e) **6 marks** We now consider a Bayesian analysis for a single group. The conjugate prior for p is a beta distribution, Be(a, b). State the form of the posterior with this choice. Give the form of the posterior mean and write as a weighted combination of the MLE and the prior mean.

(f) **3 marks** Suppose we wish to fix the parameters of the prior, a and b, so that the mean is μ and the prior standard deviation is σ. Obtain expressions for a and b in terms of μ and σ^2.

(g) **6 marks** For the data in Table 1, assume we wish to have a beta prior with $\mu = 0.2$ and $\sigma = 0.08$ for each of the two groups of women. State the posterior for the prior corresponding to this choice and give the posterior means and 95% credible intervals for each group. Provide representations of the posterior distributions.

2. (a) **5 marks** A more complex likelihood for these data would assume that the i-th component had their own probability p_i, with the p_i's arising from a distribution $\pi(p)$. Show that

$$\mathsf{Pr}(T = t) = \mathsf{E}[(1-p)^{t-1}] - \mathsf{E}[(1-p)^t],$$

and

$$\mathsf{Pr}(T > t) = \mathsf{E}[(1-p)^t].$$

(b) **5 marks** Obtain expressions for $\mathsf{Pr}(T = t \mid \alpha, \beta)$ and $\mathsf{Pr}(T > t \mid \alpha, \beta)$ with $\pi(\cdot)$ taken as the beta distribution, $\mathsf{Be}(\alpha, \beta)$.

(c) **5 marks** Using the previous part, write down the likelihood function $L(\alpha, \beta)$ based on data $\{Y_t, t = 1, \ldots, N+1\}$.

(d) **8 marks** Find the MLEs, $\widehat{\alpha}_j, \widehat{\beta}_j$, for the data of Table 1 for $j = 1$ (Smokers) and $j = 2$ (Non-Smokers).

[Hint: you might find the `R` function `lgamma` useful.]

3. (a) **10 marks** Show that the likelihood (1) can be written as a product of binomial distributions, the form of which you should precisely give.

(b) **5 marks** Fit the binomial model, and show that the estimates of the probabilities p_j, $j = 1, 2$, are identical to that under the previous MLE analysis. Obtain 95% asymptotic confidence intervals for p_1 and p_2.

(c) **8 marks** Carry out a Bayesian analysis with independent Beta(1,1) priors on p_1 and p_2 and produce a histogram representation of the posterior distribution of $p_1 - p_2$.

(d) **2 marks** What is the posterior probability that $p_2 > p_1$?

(e) **6 marks** Obtain predictive distributions of the cycle until conception for women who smoke and for women who do not smoke.