Bayesian Analysis I, FALL 2023, TAKE-HOME Assignment

INSTRUCTIONS: Be sure to include your **codes, output and brief comments** for each problem in your write-up. When submitting, convert your write-up into a single file (preferably in PDF format) that includes your *name* and *student number.* You may use any books, references, notes, but are **not allowed** to discuss these problems with any person other than the instructor until the due date. No credit will be given if any collaboration is detected.

1. In Olympic games there may be potential benefits that a host country may experience in terms of performance and support. We want to explore such an advantage by examining the rate of medals per participant in the host year compared to their previous Olympics.

		Medals won		Participants	
Host country	Year	Previous	Host	Previous	Host
Finland	1952	24	22	129	258
Australia	1956	11	35	81	294
Italy	1960	25	36	135	280
Japan	1964	18	29	162	328
Mexico	1968	1	9	94	275
West Germany	1972	26	40	275	423
Canada	1976	5	11	208	385
Soviet Union	1980	125	195	410	489
United States	1984	94	174	396	522
South Korea	1988	19	33	175	401
Spain	1992	4	22	229	422
United States	1996	108	101	545	647
Australia	2000	41	58	417	617
Greece	2004	13	16	140	426
China	2008	63	100	384	599
Great Britain	2012	47	65	304	530
Brazil	2016	17	19	236	462
Japan	2021	41	51	395	621

Let X_{i1} be the number of medals won by the host country during the Olympics i and X_{i0} be the number of medals won by the county in the previous Olympics. Similarly, let N_{i0} and N_{i1} be the number of participants from the country in the corresponding Olympics.

For example, Finland hosted the Olympics in 1952. They had $N_{11} = 258$ participants in 1952 and $N_{10} = 129$ participants in the 1948; they won $X_{11} = 22$ medals in 1952 and $X_{10} = 24$ medals in the 1948.

a) To assess broadly for a host country advantage, first combine data across all years for the host county in the host year, $X_1 = \sum_{i=1}^{18} X_{i1}$ and $N_1 = \sum_{i=1}^{18} N_{i1}$.

Conduct a Bayesian analysis of λ_1, the expected number of medals per participant in their home country, assuming a Poisson sampling model, Poisson$(N_1\lambda_1)$ and the prior distribution Gamma(0.1, rate=0.1). (Note : Gamma(0.1, rate=0.1) = Gamma(0.1, scale=10) : parametrization used in our course.)

Repeat this analysis using the data from the previous year, $X_0 = \sum_{i=1}^{18} X_{i0}$ and $N_0 = \sum_{i=1}^{18} N_{i0}$ to do analysis of λ_0. Compare these two posterior distributions in a figure.

b) Conduct a Bayesian test of the hypothesis that there is a home-county advantage, $\lambda_1 > \lambda_0$, i.e. compute the the posterior probability that $\lambda_1 > \lambda_0$. (Direct sampling through Monte Carlo simulation with e.g. 100,000 sampling offers one easy method to approximate such a probability.)

Are your results sensitive to the prior? Compute the posterior probability over (3-4) different sets of hyper-prior parameters.

c) Choose 5 different countries and conduct an analysis separately by country by comparing the posterior distribution of the ratio $r = \lambda_1/\lambda_0$ for each county. Is there evidence that the home-country advantage differs by country?

(Combine the data across the two Olympics for Australia, Japan and USA. Then there are a total of 15 counties.)

(Extra-credit) d) The next Olympics will be held in France in 2024. Predict the number medals France will win in the 2024 Olympics and quantify your uncertainty about this prediction by providing an interval estimate.

France had 398 participants and won 33 medals in 2021, but the number of participants in 2024 is not known yet. (You can select a reasonable number for this.)

2. A new process is proposed for preparing chicken breasts. One measure of its success will be the level of wastage; the standard process produces an average 4.7 % wastage. The process is tested by two operators on 6 chicken breasts each. The data were as follows (in % wastage):

Operator 1	4.3, 4.3, 2.7, 3.6, 3.5, 4.5
Operator 2	3.9, 4.0, 4.5, 2.9, 5.2, 4.8

It is thought that the mean μ should be the same for both operators, but that they may have different variances. If y_{ij} is the wastage for observation j from operator i, the model is that

$$y_{i,j}|\mu, \tau_1, \tau_2 \sim N(\mu, \tau_i)$$

independent. Assume that the parameters μ, τ_1, and τ_2 have independent prior distributions.

$$\mu \sim N(4.7, \nu), \quad \frac{1}{\tau_i} \sim \text{ Gamma}(4, \text{rate} = 5)$$

a) The elicited value of ν is 0.2. Analyze these data for the case $\nu = 0.2$, with particular reference to inference about whether $\mu < 4.7$ and about the posterior distribution of $\phi = \frac{\tau_2}{\tau_1}$. [Your solution should include a listing of your WinBUGS/RJAGS/Stan program and output.]

b) Explore sensitivity of the analysis to ν, considering the range $0.15 \leq \nu < \infty$. i.e. select several values of ν in the given range and check how the results change according to them.

c) Write a short report on your findings for the process developer who is assumed not to be a statistician. i.e. Explain your conclusions in simple language for the developer.

(Extra-credit)
3. The lengths (in millimeters) of ancient skulls found at two sites in Tibet are

y : 190.5, 172.5, 167.0, 169.5, 175.0, 177.5, 179.5, 179.5, 173.5, 162.5, 178.5, 171.5, 180.5, 183.0, 169.5, 172.0, 170.0

and z : 182.5, 179.5, 191.0, 184.5, 181.0, 173.5, 188.5, 175.0, 196.0, 200.0, 185.0, 174.5, 195.5, 197.0, 182.5

The assumption is that the variance is expected to be the same for both sites, although there could be differences in their means.

$$y_i|\mu_1, \tau, \sim N(\mu_1, \tau) \quad \text{independent of} \quad z_j|\mu_2, \tau, \sim N(\mu_2, \tau)$$

The noninformative (independent) priors for μ_1, μ_2 and τ can be used : $\mu_1 \sim \text{Normal}(100, 10000)$, $\mu_2 \sim \text{Normal}(100, 10000)$, and $1/\tau \sim$ Gamma(0.1, rate= 0.1).

Conduct a Bayesian analysis for $\mu_1 - \mu 2$ and give some comments on your results. Compare the above model with the model where two variances are assumed to be different. Briefly describe your findings.