Biostats VHM 801 Course
Winter 2018, Atlantic Veterinary College, PEI
Henrik Stryhn

Supplementary exercises

The purpose of these exercises is to allow a student seeking challenges beyond the standard curriculum of VHM 801 to explore some questions of both theoretical and practical relevance.

Problem 1. Proportion of "suspected outliers" in different distributions.

How many observations the rule for "suspected outliers" based on the interquartile range will flag depends on the distributional shape. It is useful to have an intuitive understanding of this relationship. The task is therefore to compute the proportion beyond the cut-off of the rule for a number of theoretical distributions of different shapes.

1) Start by a standard normal distribution $N(0, 1)$; the answer can be checked with the value given in Lecture 1.
2) To represent symmetrical distributions with less tails than the normal distribution, compute for the uniform distribution and the triangular distribution; both of these distributions have smaller kurtosis than $N(0, 1)$.
3) To represent symmetrical distribution with heavier tails than the normal distribution, compute for t-distributions with varying degrees of freedom; these distributions have larger kurtosis than $N(0, 1)$.
4) To represent right-skewed distributions, compute for log-normal distributions with varying standard deviation σ; the (positive) skewness increases with values of σ.

Problem 2. Comparison of classical (frequentist) and Bayesian inference for two independent samples.

In this exercise, we compare estimates, confidence intervals and tests from classical statistical analysis for two independent samples with their counterparts in Bayesian analysis. Data from Supplementary Exercise 7.80 for IPS7 will be used.

1) Complete Exercise 7.80 using classical statistical methods. Make sure to explicitly state the statistical model. Include results for both versions of the two-sample inference based on the assumptions made for variance parameters, and compare the results.
2) Carry out MCMC Bayesian estimation with vague (non-informative) priors for all parameters, possibly (but not necessarily) using the default priors of your software. For mean parameters (and the mean difference), use the mean, standard deviation and 95% range from the MCMC samples for comparison with the values obtained by classical methods. Reflect on which of the two versions of classical procedures the results should be compared with.
3) Inspect the diagnostics from the MCMC estimation of the previous question, and comment on your findings. In addition, explore how sensitive your results from 2) are to from assumptions regarding the prior distributions; concentrate on the prior distributions for the two mean parameters.

4) Although Bayesian statistical methods have no direct counterpart of classical significance testing, one simple approach to determine "significance" against a two-sided alternative is to determine whether the 95% range of the posterior distribution contains the target value. The approach can be repeated for other proportions of the posterior distribution. Determine in this way an approximate P-value for the hypothesis test of interest, against a two-sided alternative, and compare with the corresponding classical P-value.

5) Use the same procedure to determine approximate P-values against the one-sided alternative from part 1); here, one-sided 95% ranges of the posterior should be employed. Also here, compare with the P-value from classical statistics.

Problem 3. Bayesian inference for three binomial samples.

In this exercise, we compare estimates and confidence intervals from classical statistical analysis with the corresponding quantities from Bayesian analysis, with and without informative priors. A small experiment was performed measuring the risk of a certain kind of tumor (endometrial stromal polyps) in groups of rats given different doses of a certain drug. The goal of the study was to estimate relation between the dose and the probability of tumor development, and in particular the rate at which the tumor risk increases (or decreases) as a function of dose. The data are given in the table below.

dose level	number of rats	number of rats with tumors
0	14	4
1	34	4
2	34	2

1) Set up a suitable statistical model, and carry out a standard (frequentist) analysis of these data, including model validation. Give estimates and confidence intervals for the parameters, as well as intervals for the tumor probabilities at each of the three doses. As this part does not involve Bayesian methods, you may report it briefly while concentrating on the main points.

2) Carry out a similar analysis as in 1) using the Bayesian approach and "non-informative" prior distributions, using software defaults. Compute intervals for the same quantities as above, and compare the intervals of the two analyses. Explain the difference in their interpretation. Make sure to include proper assessment of the validity of the MCMC estimation.

3) Explore the implications of the "non-informative priors" used in 2) for the prior proportions of rats with tumors in each for the three groups. Do these seem as meaningful prior distributions in the absence of knowledge about the occurrence of tumors? If not, try to develop alternative prior distributions that you find more intuitive, and explore the impact of using these priors on the results by comparing with those you obtained in 1) and 2).

4) Data from previous experiments with rats of the same type (strain) existed, however only for rats that were not exposed to the drug. The table below gives data for 70 previous experiments under similar laboratory conditions with non-exposed rats. Note that e.g. ($\times 7$) means that 7 experiments gave this outcome.

number of rats with tumor(s) out of the total number of rats
0/20 (×7), 0/19 (×4), 0/18 (×2), 0/17, 1/20 (×4), 1/19 (×2), 1/18 (×2), 2/25, 2/24, 2/23, 2/20 (×6), 1/10, 5/49, 2/19, 5/46, 3/27, 2/17, 7/49, 7/47, 3/20 (×2), 2/13, 9/48, 10/50, 4/20 (×7), 10/48, 4/19 (×3), 5/22, 11/46, 12/49, 5/20 (×2), 6/23, 5/19, 6/22, 6/20 (×3), 16/52, 15/47, 15/46, 9/24

Use the historical data (available in the dataset `rattumor`) to form an informative prior distribution for at least one of the parameters in the previously used model, and carry out a Bayesian analysis along the same lines as above with this prior distribution. Answer to the same points as in 2); however, you don't need to give the full details of the MCMC assessment. Describe in detail how you constructed the prior distribution(s). Compare the results with the previous analyses, and discuss your findings.