Bayesian Statistics short course – week 1

Jonathan J Forster and Jakub Bijak, University of Southampton

Homework

Task 1

(*Example inspired by the Understanding Uncertainty blog of Prof. David Spiegelhalter, University of Cambridge, which is worth visiting in its own right: https://understandinguncertainty.org*)

In the overall population, on average 10 people in 1000 have a Disease. For every 1000 people with Disease, the screening test gives a positive result in 850 cases and negative in 150 cases. For every 1000 people without Disease, the test gives a negative result in 900 cases, and positive in 100 cases.

Use Bayes's theorem to answer the following question: If for a particular person the screening test gives a positive result, what is the probability that this person actually has Disease?

Task 2

(*A practical task, requiring JASP*)

The enclosed csv file (`NZ_crime.csv`) contains information on crimes in the regions of New Zealand in 2015. [Source: Statistics New Zealand, via: http://www.stats.govt.nz/tools_and_services/releases_csv_files/csv-files-other.aspx]

Using the Bayesian features of JASP, verify the hypothesis that the crime rate per 10,000 inhabitants (variable 'Rate_per_10000_population') in New Zealand varies between the urban and rural areas of the country (variable 'Urban_area_type'). For the size of the effect (difference between the rural and urban areas), evaluate the key characteristics of the posterior distribution and compare them with the prior. Assess the sensitivity of the results to the choice of the prior distribution.

Task 3*

(*A more advanced, optional task, requiring some knowledge of integrals*)

In a statistical model, let the likelihood function $p(x|m)$ for a single observation x be described by a uniform distribution, $U(0, m)$, where m is the unknown parameter denoting the upper bound of the possible values of x. In other words, let $p(x|m) = 1/m$ for $0 \le x \le m$, and $p(x|m) = 0$ otherwise.

Let n be the number of independent data points in the sample, the largest of them equal to 1. For the parameter m, assume *a priori* the following Pareto distribution: $p(m) = 2/m^3$ for $m \ge 1$, and $p(m) = 0$ otherwise.

Derive the posterior distribution for m and comment on its properties.

Solutions:

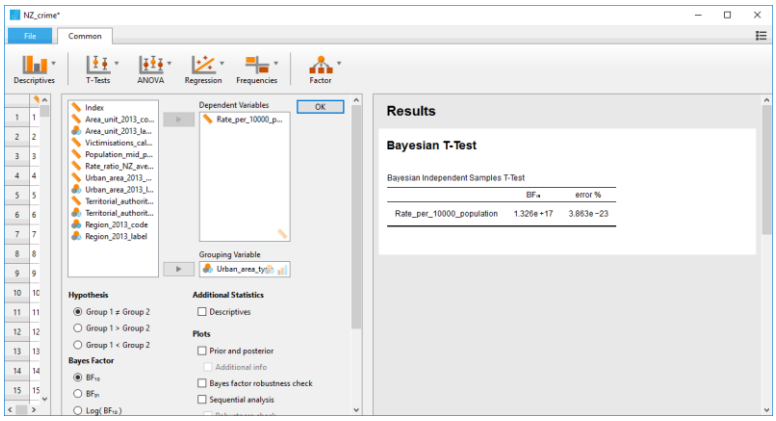
Task 1 (*corrected answer – well spotted*)

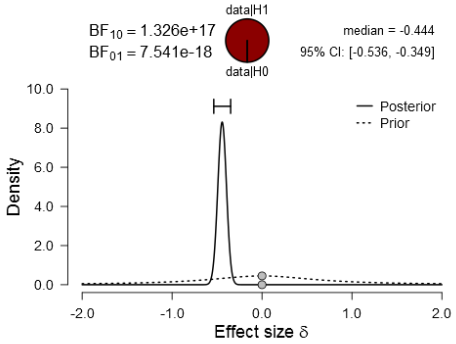
Baseline (prior) risk: P(Disease) = 0.01, hence P(No disease) = 0.99
P(Positive|Disease) = 0.85, P(Negative|Disease) = 0.85
P(Negative|No disease) = 0.9, P(Positive|No disease) = 0.1

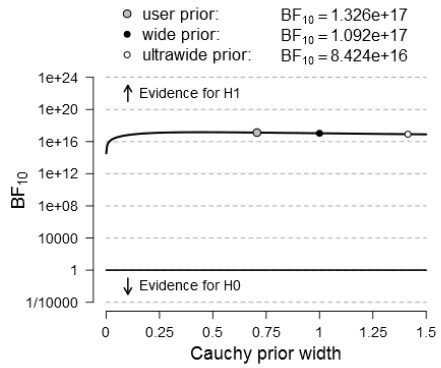
Hence, from Bayes's theorem: P(Disease|Positive) = 0.85 x 0.01 / (0.85 x 0.01 + 0.1 x 0.99) = **0.07907**

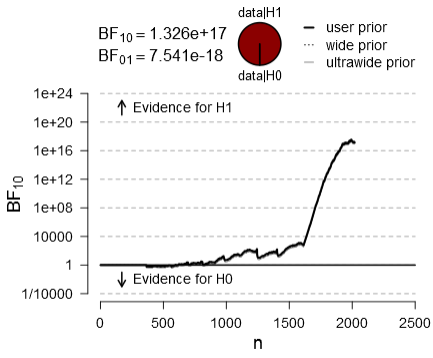
Task 2

Bayesian independent samples t-test gives very strong evidence of differences between the urban and rural areas. The results are largely insensitive to the choice of priors (see screenshot below).









Task 3*

Given: $p(m) = 2 / m^3$, $p(x|m) = 1 / m^n$. A general form of the Pareto distribution is: $p(m) = K\, b^K / m^{K+1}$.

From Bayes's theorem: $p(m|x) = p(x|m)\, p(m) / C$, where $C = \int_{[1,\infty)} p(x|m)\, p(m)\, dm$, since all $x \leq 1$.

Hence: $C = 2 \int_{[1,\infty)} m^{-(n+3)}\, dm = -2 / (n+2) \left[m^{-(n+2)} \right]_{[1,\infty)} = (n+2) / 2$, and thus $p(m|x) = (n+2) / m^{(n+3)}$

The posterior is also a Pareto distribution (conjugate), with parameters b = 1 and K = n+2.