5 Other models

5.1 Poisson model

Poisson-distribution is one of the most commonly used models in e.g. reliability research and epidemiology. It is used for describing number of 'rare events'. Poisson distribution can be derived as a limiting case of binomial distribution $\text{Bin}(N_k, r_k)$ when $N_k \to \infty$ and $r_k \to 0$ so that the product $N_k r_k \to \lambda$, when $k \to \infty$. Then, the (Poisson) distribution of a single observation $X \in \{0, 1, 2, 3, \ldots\}$ is

$$P(X \mid \lambda) = \frac{\lambda^X}{X!} e^{-\lambda}.$$

The Poisson distribution also emerges from Poisson process (a special case of a stochastic process) with constant intensity λ. If, e.g. accidents occur with constant intensity λ per time unit, then the expected number of accidents in a time unit is λ and the number of them (per time unit) follows Poisson distribution with parameter λ, which is both the mean and the variance of Poisson distribution. Due to additivity of Poisson variables, if $X \sim \text{Poisson}(\lambda_1)$ and $Y \sim \text{Poisson}(\lambda_2)$, then $X + Y \sim \text{Poisson}(\lambda_1 + \lambda_2)$. Likewise, the number of events during time T has Poisson distribution $\text{Poisson}(\lambda T)$. In a Poisson process with constant intensity λ, the waiting time until next event is exponentially distributed with mean $1/\lambda$, regardless of the past history, (if λ given).

As a conjugate distribution, the prior of λ is $\text{Gamma}(\alpha, \beta)$-density

$$\pi(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda},$$

which leads to the posterior:

$$\pi(\lambda \mid X) \propto \frac{\lambda^X}{X!} e^{-\lambda} \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda},$$

which is, up to a normalizing constant, the same as

$$\lambda^{x+\alpha-1} e^{-(1+\beta)\lambda}.$$

In other words: recognized to be $\text{Gamma}(X + \alpha, 1 + \beta)$-density. The posterior mean is thus

$$E(\lambda \mid X, \alpha, \beta) = \frac{X + \alpha}{1 + \beta} = \frac{1}{1+\beta} X + \frac{\beta}{1+\beta} \frac{\alpha}{\beta}$$

which is a **weighted average of prior mean** α/β **and** X. If we have a series of observations $X_1, \ldots, X_n$, an analogous result can be derived.

Informative prior would need to be elicited from some useful knowledge, by e.g. specifying the most probable value of λ and some upper limit (e.g. 95% percentile), and solving parameters of gamma-prior from this. An **uninformative prior** would obviously have 'small' values α, β. In the limit, these could be set to zero, so that the posterior then only depends on data. However, the prior is then not proper density and it would not be possible to get a prior predictive distribution. Also, e.g. with the single observation X, it could happen that $X = 0$, in which case the posterior would be $\text{Gamma}(0, 1)$

- not proper. (With improper priors, always check if posterior distribution is proper).

Posterior predictive distribution can be analytically solved when the prior is conjugate (Gamma) distribution. The idea is the same as before (as always: follow the probability calculus):

$$P(X^{\text{next}} \mid X) = \int_0^\infty \text{Poisson}(X^{\text{next}} \mid \lambda)\text{Gamma}(\lambda \mid X)\mathbf{d}\lambda$$

where $\text{Gamma}(\lambda \mid X)$ is the posterior distribution of λ, based on earlier data X. By doing this integration we get Negative Binomial distribution. The mean and variance of NegBin distribution can be found as before with BetaBinom distribution, by using the law of total expectation and total variance:

$$E(X) = E(\underbrace{E(X \mid \lambda)}_{\text{from Poisson}}) = \underbrace{E(\lambda)}_{\text{from Gamma}} = \alpha/\beta$$

and

$$V(X) = E(\underbrace{V(X \mid \lambda)}_{\text{from Poisson}}) + V(\underbrace{E(X \mid \lambda)}_{\text{from Poisson}}) = \underbrace{E(\lambda) + V(\lambda)}_{\text{from Gamma}} = \frac{\alpha}{\beta} + \frac{\alpha}{\beta^2}$$

In the posterior predictive distribution, the parameters α, β would correspond to the parameters of the posterior distribution for λ, shown above for a single observation X. Again, if we think of the Gamma distribution with just some parameters α, β, this can be used for **modeling overdispersed data** when the Poisson model as such is not adequate. (Poisson has mean $= \lambda =$ variance, which can be restrictive. NegBin distribution would depend on the two Gamma parameters α, β which allows more flexibility). In general, similar effects can be obtained by modeling the intensity λ of a Poisson process as a random parameter or some function over time, so that we have a *nonhomogeneous Poisson process* with intensity $\lambda(t)$. This is further specified depending on the 'second order' model structure ('=model of parameters') so that the intensity at one time point could depend on other intensities at other time points and other parameters - but then we have a more complex model, and the posterior distribution might not have a closed form anymore.

5.1.1 Example: asthma mortality

Epidemiological Example from Gelman [5]: Poisson model parameterized in terms of rate and exposure:

$$X_i \sim \text{Poisson}(E_i\theta)$$

where X_i is the number of e.g. disease cases in a group with exposure E_i and θ is the unknown (common) parameter of interest, the 'underlying rate'. The 'exposure' E_i could be the person years in the ith group 'under risk' and X_i is then the observed disease cases as an outcome of that. The rate would then be interpreted as the underlying disease incidence per person per year.

(For rare diseases or 'once-in-a-lifetime' events, we do not expect more than one occurrence per person. Especially, if we model incidence of death. More exactly, such incidence for a person should be modeled as θ before the event, and zero after the event has occurred. Otherwise, the Poisson process with intensity θ assumes that any number of events is possible. In a large population with small number of cases X, it does not matter much if the diseased are subtracted after their onset of disease from the total exposure group or not. See more: *survival models* below).

In epidemiological applications, it is this underlying incidence θ we aim to estimate. A simple point estimate would be $\hat{\theta} = X_i/E_i$. For example, if the group i is the population of a town with 50000 inhabitants, and if the population does not vary significantly during a year, then $E_i \approx 50000$. If some disease count is $X_i = 30$ for some year, we get point estimate $30/50000$, or '60 per 100000'. Assuming all individuals are 'exchangeable', then only the person years at risk matters. Similarly, if we estimate the failure rate of light bulbs: only the total time in use would matter, not the number of light bulbs. Two light bulbs in use for one year, then one burned, would give the same estimate (1 per 2×1 years $= 0.5$ per year) as one light bulb in use for two years, then burned.

The conditional probability of the data from N different groups (cities, etc.) $X = (X_1, \ldots, X_N)$ each with different exposures E_i is

$$\pi(X \mid \theta) \propto \theta^{\sum_{i=1}^{N} X_i} \exp((-\sum_{i=1}^{N} E_i)\theta)$$

With the conjugate prior Gamma(α, β), the posterior is

$$\pi(\theta \mid X) = \text{Gamma}(\alpha + \sum_{i=1}^{N} X_i, \beta + \sum_{i=1}^{N} E_i)$$

Assume there were $X = 3$ deaths due to asthma in a city during a year, out of a population of 200000. Hence the crude estimate per 100000 per year would be 1.5 cases. The model for the observed count could be

$$X \sim \text{Poisson}(2\theta) = \text{Poisson}(E\theta)$$

where θ represents the 'underlying mortality rate' per 100000 per year, and E 'exposure'. To compute the posterior $\pi(\theta \mid X)$, we choose a conjugate prior $\pi(\theta) = \text{Gamma}(\alpha, \beta)$ by choosing (α, β) so that the prior represents reasonable background information. According to literature, the typical asthma mortality rate in Western countries would be around 0.6 per 100000. It is also known that values above 1.5 are rare. Hence, Gamma$(3, 5)$ prior has mean 0.6, standard deviation 0.35, and this prior also has $P(\theta < 1.44) = 97.5\%$. All this seems to fulfill both prior specifications. (The prior parameters can be chosen by trial and error). The posterior distribution is then Gamma$(6, 7)$, which has mean 0.86. That is substantial shrinkage towards prior distribution.

Estimating several θ_i instead of one common θ:
The same idea is exploited further e.g. in spatial epidemiology, where we wish to estimate disease incidences θ_i in different geographical areas, instead of assuming there is a common incidence θ everywhere. The exposure E_i can be very small in some geographical areas due to low population. The risk estimates based on local data would be very unstable because - by chance - there could be ± 1 case, which already could cause a high/low point estimate. Therefore, results from small population groups are analyzed with respect to expected results based on the larger population. The former gives the likelihood part, the latter the prior $\rightarrow$ the result is a shrinkage towards the prior.

Examples are sometimes seen in daily news, reporting 'exceptional rise of crime rate in a small area'. With many small areas with small populations, most of them would show observed incidence of zero, but by chance some would have one or two cases which would give observed incidence much higher

than the national incidence. This should not be interpreted as if the local risk is *really* that much bigger. Likewise, a small university department might be unfairly reported to score badly if for one year there happens to be no doctoral dissertations. Similarly, salmonella subtypes detected annually in some species could show so low counts that most types have zero frequency whereas few types would have only 1 or 2 counts. Does it mean that the other types are absolutely nonexistent - just because they have not been detected this year? Or, in the cracking of the Enigma code: having not yet observed some bits of code does not mean that their probability is absolutely zero. A sensible prior distribution would put some positive probability for events not yet seen (if they logically can exist). Otherwise, the model would exclude their possibility completely and could not update those probabilities.

In spatial models in health geography, the disease rates (estimated Poisson intensities) are therefore **smoothed** over the map, so that any local estimate is taken as a compromise between the local data and its neighborhood (=prior). If the local data are heavy, the final estimate will be determined by that, but if the local data are scarce, the estimate is more influenced by neighborhood. This is also called borrowing strength. The same is often done with time series data, to get smoothed temporal rates. In other applications, we might have a stratification of a population so that the population counts in some strata are too low, hence the estimates would be smoothed towards other data. Smoothing can be either towards neighborhood mean or towards global mean.

Different alternatives would be built in the posterior of the local disease rates:

$$\pi(\theta_1, \ldots, \theta_N \mid X_i, \ldots, X_N) \propto \prod_{i=1}^{N} \pi(X_i \mid \theta_i)\pi(\theta_i \mid \text{something})$$

by choosing the definition of prior $\pi(\theta_i \mid \text{something})$ as needed:

either 1: the prior of each disease rate θ_i (of $i = 1, \ldots, N$ subpopulations) is based on background information about larger population. (With this prior, we expect disease rates in different geographic areas to be as the national rate is - or as the rate in surrounding area - or whatever reference group thought to be relevantly informative). Here, the prior is given as a fixed distribution.

or 2: the prior can be further generalized to allow prior parameters to be unknown, so that they are estimated too as part of a posterior distribution. This is accomplished by letting the prior to be e.g. $\pi(\theta_i \mid \mu)$ with a further prior $\pi(\mu)$ which could represent global mean, around which local means θ_i are distributed (θ_i are conditionally independent, given μ). In turn, local observed disease counts X_i would be conditionally independent, given θ_i. This makes a hierarchical model where the smoothing depends on the unknown parameters in $\pi(\theta_i \mid \mu)$. A 'tight' prior makes shrinkage towards the global μ (which itself is uncertain) whereas a 'loose' prior lets each θ_i to be close to the observed incidence X_i/E_i.

or 3: same as before, but the prior of θ_i could depend on other θ_j, so the prior could e.g. specify θ_i to be distributed around a local mean of other θ_j, $j \neq i$ in the neighborhood of the ith geographical area. Or in temporal models: conditionally on the θ_{i-1} of the previous time interval.

In any case, the structure of smoothing becomes specified by the choice of prior.
(•) In the first case, a prior for each parameter is set independently of other parameters and it is a fixed distribution: $\pi(\theta_i)$. This model cannot borrow strength from other data because the prior is a

fixed distribution and cannot change. Every θ_i is estimated separately from the others.
(•) In the latter cases, prior of θ_i depends on a parameter(s) that is also estimated, and this common parameter will get estimated from information about all other θ_j too, which finally are informed by their specific data X_i. Every θ_i is estimated with influence from other θ_j as well, via the common parameter(s) which need 'hyper prior': a prior for the parameters of a prior.

Large models could have several layers of priors, describing the structural assumptions of conditional dependency we make.

5.2 Exponential distribution

Assume a single observation $X \in \mathbb{R}^+$ (typical example: waiting times, time of next event) for which the conditional distribution is exponential:

$$\pi(X \mid \theta) = \theta \exp(-X\theta).$$

As a conjugate prior of θ, we choose Gamma(α, β), so that the posterior $\pi(\theta \mid X)$ becomes Gamma$(\alpha + 1, \beta + X)$. The posterior mean is

$$E(\theta \mid X, \alpha, \beta) = \frac{\alpha + 1}{\beta + X}$$

With a set of observations $X_1, \ldots, X_n$ (mean $\bar{X} = \sum_{i=1}^n X_i/n$) we get

$$\pi(X \mid \theta) = \theta^n \exp(-n\bar{X}\theta)$$

which leads to the posterior Gamma$(\alpha + n, \beta + n\bar{X})$, so that the Gamma$(\alpha, \beta)$ prior can be thought as equivalent of α prior observations $X_1^0, \ldots, X_\alpha^0$ for which the sum $\sum X_i^0$ equals to β. In the posterior distribution, these are updated by size of data n (number of observations) and the data sum $\sum X_i$, respectively.

5.2.1 Survival models

Consider a non-homogeneous Poisson process with intensity (hazard) function $\lambda(t)$. The definition of hazard is the limit of a conditional probability

$$\lambda(t) = \lim_{\delta t \to 0} \frac{P(t < X \leq t + \delta t \mid t < X)}{\delta t}$$

In other words:

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{-S'(t)}{S(t)}$$

where $f(t)$ is the probability density function of the event time $X \in [0, \infty)$ and $S(t) = 1 - F(t)$ is the *survival function*, (F is the cumulative probability function).

Solving the differential equation $\lambda(t) = -S'(t)/S(t) = -\,\mathbf{d}\log(S(t))/\mathbf{d}t$ gives

$$S(t) = e^{-\int_0^t \lambda(\tau)\mathbf{d}\tau}.$$

If the intensity is just constant $\lambda(t) = \lambda$, this results to the cumulative probability function of exponential distribution

$$S(t) = e^{-\lambda t}$$

so that the event time X is exponentially distributed, conditionally on λ. An observed data of several event times (life times, failure times, etc.) gives the likelihood function

$$\prod_{i=1}^{N} e^{-\lambda X_i} \lambda^N = \lambda^N e^{-N\bar{X}\lambda}$$

which can be seen as the product of probabilities to (1) 'survive' up to time X_i and (2) to 'die' at time X_i, for all data points $X_1, \ldots, X_N$ representing the event times for individuals. *Each individual contributes to the likelihood only for the length of time he/she survived.* If all these event times are observed, and with Gamma-prior for λ, bayesian inference for the intensity is straightforward, leading to a Gamma-posterior. Typically, some of the event times are not observed. This is called censoring.

5.2.2 Censored data

In survival analysis and reliability applications, it is common that the 'failure times' (times of death, infections, illness, etc.) are exactly known for only some individuals. For others, the time can be censored, which means that we only know that the event has not happened before some known time point. (This is also information!). Often, the censoring time can be the ending time of the follow-up period, or ending time of the study, T. The probability for such event is written via the *survival probability*: $P(X_i > T \mid \lambda) = 1 - P(X_i < T \mid \lambda) = 1 - F(T \mid \lambda) = \exp(-\lambda T) = S(T \mid \lambda)$. The conditional probability of the whole data is then of the form

$$P(X \mid \lambda) = \prod_{i=1}^{k} \lambda \exp(-\lambda X_i) \times S(T \mid \lambda)^{n-k} = \lambda^k \exp(-\lambda[\sum_{i=1}^{k} X_i + (n-k)T]).$$

Applying the Gamma(α, β)-prior, the posterior is then Gamma$(\alpha + k, \beta + \sum_{i=1}^{k} X_i + (n-k)T)$.

More generally, we may know that for some individuals the event occurred before some given time, or between two given times. In each case, this information should be included by writing the corresponding conditional probability. (This is sometimes called as the 'full likelihood'). For example, if some events are only known to have been before time T_1 and some are known to be after time T_2, and for the rest we know the exact time, then **the full likelihood** would be of this form

$$P(X \mid \lambda) = \prod_{i \in E_1} \underbrace{F(T_1 \mid \lambda)}_{P(X_i < T_1 \mid \lambda)} \times \prod_{i \in E_2} \underbrace{S(T_2 \mid \lambda)}_{P(X_i > T_2 \mid \lambda)} \times \prod_{i \in E_3} \underbrace{\lambda \exp(-\lambda X_i)}_{P(X_i \mid \lambda)}.$$

and this could still be expanded by interval censored data, by incorporating probabilities of the type $P(L_1 < X_i < L_2 \mid \lambda)$. All these censored observations give likelihood contributions involving integrations of the density function of X_i.

Whatever the expression of the full likelihood, the principle is generally the same: to compute posterior distribution, conditionally to 'full data' (=censored and exact event times):

$$\pi(\lambda \mid \text{full data}) \propto \pi(\lambda) P(\text{full data} \mid \lambda)$$

which might take a form that does not reduce to a known density function in closed form!

Note: by using the cumulative probability function F, probability expressions for all different situations of censoring might be written.

Note: when the event time is known, the conditional probability of this observation is $P(X_i \mid \lambda) = \lambda \exp(-\lambda X_i)$, but when the censoring time is known, the observation can be interpreted as a Bernoulli variable (indicator variable!) that was one:

$$Y_i = \begin{cases} 0 & \text{if } X_i < T \\ 1 & \text{if } X_i > T \end{cases}$$

so that $P(Y_i = 1 \mid \lambda) = S(T \mid \lambda)$.

Censoring can happen also with other variables than event times. In microbiology, bacterial concentrations are measured and some (or many) measurements can be censored. As an example of what a heavily censored data could look like, the following represents observed bacterial concentrations in samples. There is only one exactly observed measurement. All others are somehow censored:

Number of samples	Concentration (CFU/g (Colony Forming Units per gram))
54	<0.04
2	<100
26	0.04-10
1	15
8	0.04 -100
2	>100
1	<1
1	>1
7	0.04 -1
1	1-100

Reference: P Busschaert, AH Geeraerd, M Uyttendaele, JF Van Impe. Estimating distributions out of qualitatite and (semi)quantitative microbial contamination data for use in risk assessment. International Journal of Food Microbiology. 138 (2010), 260-269.

5.3 Approximating posterior density

Posterior distributions could be approximated by finding out the posterior mean and variance, and then using normal distribution

$$\mathrm{N}(E(\theta \mid X), V(\theta \mid X))$$

in place of the exact posterior density. Moreover, Posterior density can be approximated focusing on the mode as:

$$\pi(\theta \mid X) \approx \mathrm{N}(\hat{\theta}, [I(\hat{\theta})]^{-1}),$$

where $\hat{\theta}$ is the posterior mode and $I(\theta)$ is so called *observed information*

$$I(\theta) = -\frac{\mathbf{d}^2}{\mathbf{d}\theta^2} \log \pi(\theta \mid X).$$

The approximation is based on Taylor series expansion of $\log \pi(\theta \mid X)$ centered at the posterior mode, $\hat{\theta}$. For a scalar valued θ this is

$$\log \pi(\theta \mid X) = \log \pi(\hat{\theta} \mid X) + \underbrace{[\frac{\mathbf{d}}{\mathbf{d}\theta} \log \pi(\theta \mid X)]_{\theta=\hat{\theta}}}_{=0} \frac{(\theta - \hat{\theta})}{1!} + [\frac{\mathbf{d}^2}{\mathbf{d}\theta^2} \log \pi(\theta \mid X)]_{\theta=\hat{\theta}} \frac{(\theta - \hat{\theta})^2}{2!} + \cdots,$$

where the first derivative at posterior mode $\hat{\theta}$ is zero. When θ is near the mode, the higher order terms are small compared to the first terms. As a function of θ, the first term in the expression is constant whereas the 2nd order term is proportional to the logarithm of a normal density, which provides the approximation. For a vector valued θ, the Taylor series would be

$$\log \pi(\theta \mid X) = \log \pi(\hat{\theta} \mid X) + \frac{1}{2}(\theta - \hat{\theta})^T \Big[\frac{\mathbf{d}^2}{\mathbf{d}\theta^2} \log \pi(\theta \mid X)\Big]_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \cdots.$$

This normal approximation (modal approximation) can be a useful benchmark and it gives a quick approximation of the posterior density. For final results, more accurate computations are usually needed. Even so, the first rough estimates can be obtained from the approximation, if only as realistic starting values for more complicated calculations.

6 Multiparameter models

In nearly all inference problems there is more than one unknown quantity. Often, only one of them is of interest and the others are *nuisance parameters*. Assume there are two unknown parameters θ_1, θ_2 (both can be vectors) and some set of data X. The posterior density is

$$\pi(\theta_1, \theta_2 \mid X) \propto \pi(X \mid \theta_1, \theta_2)\pi(\theta_1, \theta_2),$$

and the marginal density of θ_1 is

$$\pi(\theta_1 \mid X) = \int \pi(\theta_1, \theta_2 \mid X)\mathbf{d}\theta_2,$$

which can also be calculated as

$$\pi(\theta_1 \mid X) = \int \pi(\theta_1 \mid \theta_2, X)\pi(\theta_2 \mid X)\mathbf{d}\theta_2.$$

This integral is usually not computed directly, but it shows an important structure that is used when hierarchical models are constructed, and also when MCMC algorithms are implemented.

Note: the unknown parameters θ can be 'unknown model parameters', or missing data variables, or variables to be predicted, or unobservable latent (hidden) variables. They are all simply unknown, and in bayesian inference they are all treated as unknown quantities, so that we aim to compute the posterior:

$$P(\text{'all unknowns'} \mid \text{'all known things'})$$

Note: it is difficult to visualize a posterior density for three or more unknown quantities. Therefore, we often plot one-dimensional marginal distributions, or two-dimensional marginal distributions for selected quantities of interest. This is always based on the full posterior density that can be multidimensional.

6.1 Multinomial model, unknown $r_1, \ldots, r_k$

Binomial model can be generalized to multinomial model by considering outcomes of several types instead of two types. For example, in a large bag there are balls of k different colours. The proportions of these are $r = r_1, \ldots, r_k$. A sample of N balls is drawn, and we observe the number of balls of each colour $X_1, \ldots, X_k$. The goal is now to solve the posterior density:

$$\pi(r_1, \ldots, r_k \mid X_1, \ldots, X_k).$$

Note that the unknown proportions have to sum to one: $\sum r_i = 1$. The conditional distribution of the data is now

$$P(X_1, \ldots, X_k \mid r_1, \ldots, r_k, N) = \binom{N}{X_1, \ldots, X_k} r_1^{X_1} \times \cdots \times r_k^{X_k}.$$

The conjugate prior density is $\text{Dir}(\alpha) = \text{Dirichlet}(\alpha_1, \ldots, \alpha_k)$:

$$\pi(r_1, \ldots, r_k) = \frac{\Gamma(\alpha_1, \ldots, \alpha_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} r_1^{\alpha_1 - 1} \times \cdots \times r_k^{\alpha_k - 1},$$

so that the posterior density will also be Dirichlet, with parameters $(\alpha_1 + X_1, \ldots, \alpha_k + X_k)$:

$$\propto r_1^{\alpha_1 + X_1 - 1} \times \cdots \times r_k^{\alpha_k + X_k - 1}.$$

Again, prior parameters $\alpha_1, \ldots, \alpha_k$ can be interpreted to represent 'prior data' so that the 'prior sample size' is $\sum \alpha_i$. A usual uninformative prior choice is $\text{Dir}(1, \ldots, 1)$, which is the generalization of $\text{Beta}(1, 1)$. The posterior means can be written as weighted mean of prior and data proportions

$$E(r_i \mid X, \alpha) = \frac{\alpha_i + X_i}{\sum(\alpha_i + X_i)} = \frac{\sum \alpha_i}{\sum(X_i + \alpha_i)} \frac{\alpha_i}{\sum \alpha_i} + \frac{\sum X_i}{\sum(X_i + \alpha_i)} \frac{X_i}{\sum X_i}$$

Note also that if $r \sim \text{Dir}(\alpha)$, then the marginal distribution of each r_j is $\text{Beta}(\alpha_j, \sum_i \alpha_i - \alpha_j)$, with variance $\alpha_j(\sum_i \alpha_i - \alpha_j)/((\sum_i \alpha_i)^2(\sum \alpha_i + 1))$. To simplify notations, write $A = \sum_i \alpha_i$. Then the marginal variance may be written as $\frac{\alpha_j}{A}(1 - \frac{\alpha_j}{A})/(A + 1)$.

The marginal posterior distribution (here solved as Beta) allows to make probability statements of any single parameter, while accounting for the uncertainty in all parameters.

If dirichlet distribution is not found in a software, the following result can be useful:

$$Z_i \sim \text{Gamma}(\alpha_i, 1) \quad \Rightarrow \quad (\frac{Z_1}{\sum Z_i}, \ldots, \frac{Z_k}{\sum Z_i}) \sim \text{Dir}(\alpha_1, \ldots, \alpha_k).$$

References

[1] McGrayne S B: The theorem that would not die. Yale University Press. 2011.

[2] Berger J: The Case for Objective Bayesian Analysis. Bayesian Analysis, 2006, Vol 1, 3, 385-402.

[3] Goldstein M: Subjective Bayesian Analysis: Principles and Practice. Bayesian Analysis, 2006, Vol 1, 3, 403-420.

[4] Christensen R, Johnson W, Branscum A, Hanson E: Bayesian Ideas and Data Analysis. CRC Press. 2011.

[5] Gelman A, Carlin J B, Stern H S, Rubin D B: Bayesian data analysis, 2nd edition. Chapman & Hall/CRC. 2004.

[6] Jaynes E T: Probability theory: the logic of science. Cambridge university press. 2003.

[7] Sivia D S: Data Analysis, a Bayesian tutorial, 2nd edition. Oxford university press. 2006.

[8] Robert C P, Casella G: Monte Carlo Statistical Methods. Springer 1999.

[9] Congdon P: Bayesian Statistical Modelling. John Wiley & Sons, Ltd. 2001.

[10] Congdon P: Applied Bayesian Modelling. John Wiley & Sons, Ltd. 2003.

[11] Bernardo J M, Smith A F M: Bayesian Theory. John Wiley & Sons, Ltd. 2000.

[12] Gelman A: Prior distributions for variance parameters in hierarchical models. Bayesian Analysis, 1, No 3, pp. 515-533. 2006.