# APPLIED BAYESIAN STATISTICS SUMMARY

ANDREW TULLOCH

## 1. Probability and Bayes theorem for discrete observables

**Definition.** *Suppose we want to predict a random quantity $X$, and we do so by providing a probability distribution $P$. Suppose we observed a specific value $x$, then a scoring rule $S$ provides a reward $S(P, x)$. If the true distribution of $X$ is $Q$, then the expected score is denoted $S(P, Q)$, where $S(P, Q) = intSP(, x)Q(x)dx$. A proper scoring rule has $S(Q, Q) \geq S(P, Q)$ for all $P$, and is **strictly proper** if $S(Q, Q) = S(P, Q)$ if and only if $P = Q$.*

**Theorem.** For a null hypothesis $H_0, H_1$ as "not $H_0$",

$$\frac{p(H_0|y)}{p(H_1|y)} = \frac{p(y|H_0)}{p(y|H_1)} \times \frac{p(H_0)}{p(H_1)}, \tag{1.1}$$

posterior odds equals the likelihood ratio times prior odds.

**Definition.** *We have observed quantities $y$ (the data), have an unknown quantity taking on a set of discrete values $\theta_i, i \in 1, \dots, n$. We specify a sampling model $p(y|\theta)$, a probability distribution $p(\theta_i)$, and together define $p(y, \theta_i) = p(y|\theta_i)p(\theta_i)$ - a "full probability model".*

*Then, use Bayes theorem to botain the conditional probability distribution for unobserved qunaitites given the data,*

$$p(\theta_i|y) = \frac{p(y|\theta_i)p(\theta_i)}{\sum_k p(y|\theta_k)p(\theta_k)} \propto p(y|\theta_i)p(\theta_i) \tag{1.2}$$

*or equivalently, the posterior is proportional to the likelihood times the prior.*

**Definition.** *$\theta \sim$ Beta$(a, b)$ represents a Beta distribution with properties*

$$p(\theta|a, b) = \frac{\Gamma(a, b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}), \theta \in (0, 1) \tag{1.3}$$

$$\mathbb{E}(\theta\ a, b) = \frac{a}{a+b} \tag{1.4}$$

$$\mathbb{V}(\theta|a, b) = \frac{ab}{(a+b)^2(a+b+1)} \tag{1.5}$$

$$mode = \frac{a-1}{a+b-2}(a, b > 0) \tag{1.6}$$

$$\tag{1.7}$$

*where $\Gamma(a) = (a-1)!$ is $a$ is an integer.*

**Theorem.** Our parametric sampling distribution $p(y|\theta)$ with uncertainty about $\theta$ given by a distribution $p(\theta)$ gives a predictive distribution $p(y) = \int p(y|\theta)p(\theta)d\theta$. The mean and variance of a predictive distribution can be obtained using

$$\mathbb{E}(Y) = \mathbb{E}_\theta(\mathbb{E}(Y|\theta)) \tag{1.8}$$

$$\mathbb{V}(Y) = \mathbb{E}_\theta(\mathbb{V}(Y|\theta)) + \mathbb{V}_\theta(\mathbb{E}(Y|\theta)) \tag{1.9}$$

**Theorem.** For two random variables with joint density $p(x, y)$, then $\mathbb{E}(Y) = \mathbb{E}_X(\mathbb{E}(Y|x))$ and $\mathbb{V}(Y) = \mathbb{E}_X(\mathbb{V}(Y|x)) + \mathbb{V}_X(\mathbb{E}(Y|x))$.

**Definition.** *Suppose $\theta \sim$ Beta$(a, b)$, $Y \sim$ Binomial$(\theta, n)$. The exact predictive distribution for $Y$ is known as the BetaBinomial with*

$$p(y) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\binom{n}{y}\frac{\Gamma(a+y)\Gamma(b+n-y)}{\Gamma(a+b+n)}, y = 0, 1, 2, \dots, n \tag{1.10}$$

*If $a = b = 1$ (the prior is uniform on $0, 1$), then $p(y)$ is uniform on $0, 1, \dots, n$.*

*The mean and variance of the BetaBinomial is given as $\mathbb{E}(Y) = \frac{na}{a+b}$ and $\mathbb{V}(Y) = n\frac{ab}{(a+b)^2}\frac{(n+a+b)}{(1+a+b)}$*

**Definition.** *The Gamma distribution is a flexible distribution for positive quantities. If $Y \sim$ Gamma$(a, b)$, then*

$$p(y|a, b) = \frac{b^a}{\Gamma(a)}y^{a-1}e^{-by}, y \in (0, \infty) \tag{1.11}$$

$$\mathbb{E}(Y|a, b) = \frac{a, b}{den} \tag{1.12}$$

$$\mathrm{V}(Y|a, b) = \frac{a}{b^2} \tag{1.13}$$

*The* Gamma$(1, b)$ *distribution is exponential with mean $\frac{1}{b}$. The* Gamma$(\frac{\nu}{2}, \frac{1}{2})$ *is a chi-squared $\chi^2_\nu$ with $\nu$ degrees of freedom.*

**Theorem.** Suppose $\theta \sim$ Gamma$(a, b)$, $Y \sim$ Poisson$(\theta)$, then the exact predictive distribution of $Y$ is **negative-binomial** with

$$p(y) = \frac{\Gamma(a+y)}{\Gamma(a)\Gamma(y+1)}\frac{b^a}{(b+1)^{a+y}}, y = 0, 1, 2, \dots \tag{1.14}$$

$$\mathbb{E}(Y) = \frac{a}{b} \tag{1.15}$$

$$\mathbb{V}(Y) = \frac{a}{b} + \frac{a}{b^2} \tag{1.16}$$

**Theorem.** Consider the general one-parameter exponential family

$$p(y|\theta) = \exp(a(y) + b(\theta) + u(\theta)t(y)) \tag{1.17}$$

where $u(\theta)$ is a **natural** or canonical parameter, and $t(y)$ is the **natural sufficient statistic**.

Suppose we have a conjugate prior distribution of the form $p(\theta) = \frac{1}{c(n_0, t_0)}\exp(n_0 b(\theta) + t_0\mu(0))$ where $c(n_0, t_0) = \int \exp(n_0 b(\theta) + t_0 u(\theta))d\theta$. Then the predictive distribution is

$$p(y) = e^{a(y)}\frac{c(n_0+1, t_0+t(y))}{c(n_0, t_0)}. \tag{1.18}$$

## 2. Conjugate Analysis

**Theorem.** Suppose we have a independent sample of data $y_i \sim$ Normal$(\mu, \sigma^2)$, $i = 1, \dots, n$, with $\sigma^2$ known and $\mu$ unknown. The conjugate prior for the normal mean is also normal, $\mu$, $\mu \sim$ Normal$(\gamma, \tau^2)$, where $\gamma$ and $\tau^2 = \frac{\sigma^2}{n_0}$ are specified. The posterior distribution is

$$p(\mu\ y) \propto p(\mu)\prod_{i=1}^n p(y_i|\mu) = \text{Normal}(y_n, \tau_n^2) \tag{2.1}$$

where $\gamma_n = \frac{n_0\gamma + n\overline{y}}{n_0+n}$ and $\tau_n^2 = \frac{\sigma^2}{n_0+n}$.

The posterior predictive distribution is thus Normal$(\gamma_n, \sigma^2 + \tau_n^2)$.

**Theorem.** Suppose again $y_i \sim N(\mu, \sigma^2)$, but $\mu$ is known $\sigma^2$ is unknown.

If we use precision $\omega = \frac{1}{\sigma^2}$, we have the conjugate prior for $\omega$ as $\omega \sim$ Gamma$(\alpha, \beta)$, so $p(\omega) \propto w^{\alpha-1}\exp(-\beta\omega)$. $\sigma^2$ has an **inverse-gamma** distribution.

The posterior distribution has the form $p(\omega|\mu, y) =$ Gamma$(\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum_{i=1}^n(y_i - \mu)^2)$.

**Theorem.** If we have $I$ possible prior distributions $p_i(\theta)$ with weights $q_i$, then the mixture prior is $p(\theta) = \sum_i q_i p_i(\theta)$. If we now observe data $y$, the posterior for $\theta$is $p(\theta|y) = q_i' p(\theta|y, H_i)$, where $p(\theta|y, H_i) \propto p(y|\theta)p(\theta|H_i)$, where $q_i' = p(H_i|y) = \frac{q_i p(y|H_i)}{\sum_i q_i p(y|H_i)}$ where $p(y|H_i) = \int p(y|\theta)p(\theta|H_i)d\theta$ is the predictive probability of the data $y$ assuming $H_i$.

**Theorem.** In a general one-parameter exponential family, we have $p(y|\theta) = \exp(\sum_i a(y_i) + nb(\theta) + u(\theta)\sum_i t(y_i))$ and prior $p(\theta) \propto \exp(n_0 b(\theta) = t_0 u(\theta))$ so the posterior distribution is

$$p(\theta|y) \propto \exp((n+n_0)b(\theta) + u(\theta)(\sum_i t(y_i) + t_0)) \tag{2.2}$$

which is in the same family as the prior distribution. $t_0$ can be thought of as the sum of $n_0$ imaginary distributions.

## 3. Prior Distributions

**Theorem.** If $\frac{1}{\sigma^2}|y \sim$ Gamma$(\alpha, \beta)$, then $\frac{2\beta}{\sigma^2} \sim \chi^2_{2\alpha}$.

If $Z \sim$ Normal$(0, 1)$, $X \sim \frac{\chi^2_\nu}{\nu} \sim t_\nu$.

**Definition.** *A Jeffreys prior is compatible with a Jeffrey's prior for any $1-1$ transformation $\phi = f(\Theta)$.*

$p(\theta) \propto I(\theta)^{\frac{1}{2}}$ *where* $I(\theta)$ *is the Fisher information for* $\theta$,

$$I(\theta) = -\mathbb{E}_{Y|\theta}(\frac{\partial^2 \log p(Y|\theta)}{\partial\theta^2}) = E_{Y|\theta}((\frac{\partial \log p(Y|\theta)}{\partial\theta})^2) \quad (3.1)$$

*This is invariant to re-parameterization as*

$$\mathbb{E}_{Y|\phi}(\frac{\partial \log p(Y|\phi)}{\partial\phi})^2 = \mathbb{E}_{Y|\theta}(\frac{\partial \log p(Y|\theta)}{\partial\theta})^2|\frac{\partial\theta}{\partial\phi}|^2 = I(\theta)|\frac{\partial\theta}{\partial\phi}|^2 \quad (3.2)$$

**Definition.** *For location parameters,* $p(y|\theta)$ *is a function of* $y-\theta$*, and the distribution of* $y-\theta$ *is independent of* $\theta$*, hence* $p_J(\theta) \propto C$ *constant. Can us* `dflat()` *in winbugs or a proper distribution such as* `dunif(-100, 100)`

**Definition.** *For count/rate parameters, the Fisher information for* POISSON *data is* $I(\theta) = \frac{1}{\theta}$*, and so the Jeffreys prior is* $p_J(\theta) \propto \frac{1}{\sqrt{\theta}}$*, which can be approximated by a* `dgamma(0.5, 0.000001)` *distribution in BUGS.*

*This same prior is appropriate if* $\theta$ *is a rate parameter per unit time — so* $Y \sim Poisson(\theta t)$.

**Definition.** $\sigma$ *is a scale parameter if* $p(y|\sigma) = \frac{1}{\sigma} f(\frac{y}{\sigma})$ *for some function* $f$*, so that the distribution of* $\frac{Y}{\sigma}$ *does not depend on* $\sigma$*. The Jeffreys prior is* $p_J(\sigma) \propto \sigma^{-1}$*. This implies that* $p_j(\sigma^k) \propto \sigma^{-k}$*, for any choice of* $k$*, and thus for the precision of the normal distribution, we should have* $p_J(\omega) \propto \omega^{-1}$*, which can be approximated by* `dgamma(0.0001, 0.0001)` *in BUGS (an inverse-gamma distribution on the variance* $\sigma^2$*).*

## 4. Multivariate Distributions

**Definition.** *Array of counts* $(n_1, \ldots, n_k)$ *in* $K$ *categories — the multinomial density is* $p(n|q) = \frac{(\sum n_k)!}{\prod n_k!}\prod_{k=1}^K q_k^{n_k}$*, with likelihood propertional to* $\prod_{k=1}^K q_k^{n_k}$*. The conjugate prior is a* DIRICHLET$(\alpha_1, \ldots, \alpha_k)$ *distribution with*

$$p(q) = \frac{\Gamma(\sum \alpha_k)}{\prod \Gamma(a_k!)} q_k^{a_k - 1} \quad (4.1)$$

*with* $\sum_k q_k = 1$*. The posterior is* $p(q|n = $ DIRICHLET$(\alpha_1 + n_1, \ldots, \alpha_k))$*. The Jeffreys prior is* $p(q)\alpha \prod_k q_k^{-\frac{1}{2}}$*.*

**Definition.** *The multivariate normal for a p-dimensional vector* $y \sim Normal_p(\mu, \Sigma)$*, or using* $\Omega = \Sigma^{-1}$*, so* $p(y|\mu, \Omega) \propto \exp(-\frac{1}{2}(y-\mu)^T\Omega(y-\mu))$*, and conjugate prior for* $\mu$ *is also a multivariate normal,*

$$p(\mu|\psi_0, \Omega_0) \propto \exp(-\frac{1}{2}(\mu - \gamma_0)^T\Omega_0(\mu - \gamma_0)). \quad (4.2)$$

*We then have* $\mu \sim Normal_p(\gamma_n, \Omega_n^{-1})$ *where* $\Omega_n = \Omega_0 + n\Omega$ *and* $\gamma_n = (\Omega_0 + n\Omega)^{-1}(\Omega_0\gamma_0 + n\Omega\overline{y})$*.*

**Definition.** *The conjugate prior on the precision matrix of a multivariate normal is the Wishart distribution (analogous to Gamma/*$\chi^2$*).*

*The Wishart distribution* $W_p(k, R)$ *for a symmetric positive definite* $p \times p$ *matrix* $\Omega$ *is* $p(\Omega) \propto |R|^{\frac{k}{2}}|\Omega|^{\frac{k-p-1}{2}}\exp(-\frac{1}{2}\operatorname{tr}(R\Omega))$*.*

*The sampling density of a MVN with known mean and unknown matrix is* $p(y_1, \ldots, y_n|\mu, \Omega) \propto |\Omega|^{\frac{n}{2}}\exp(-\frac{1}{2}\operatorname{tr}(S\Omega))$ *where* $S = \sum_i (y_i - \mu)(y_i - \mu)^T$*, and therefore*

$$p(\Omega|y) \propto |\Omega|^{\frac{n+k-p-1}{2}\exp(-\frac{1}{2}\operatorname{tr}((S+R)\Omega))} \quad (4.3)$$

*which is a* $W_p(k+n), R+S$ *distribution.*

*The Jeffreys prior is* $p(\Sigma) \propto |\Sigma|^{-\frac{p+1}{2}}$*, equivalently* $k \to 0$*.*

## 5. Regression Models

Assume for a set of covariates $x_{i1}, \ldots, x_{ip}$, $\mathbb{E}(Y_i) = x_i'\beta$, and $Y_i \sim N(\sum \beta_i x_i, \sigma^2)$. Assuming $Y_i$are conditionally independent given $\beta, \sigma^2$, we can write $Y \sim N_n(X\beta, \sigma^2 I_n)$. The least squares estimate and MLE is

$$\hat{\beta} = (X^TX)^{-1}X^Ty \quad (5.1)$$

$$\hat{\beta} \sim N_p(\beta, \sigma^2(X^TX)^{-1}) \quad (5.2)$$

With known variances, assume $\beta \sim N_p(\gamma_0, \sigma^2 V)$. Then $p(\beta|y) \propto \exp(-\frac{1}{2\sigma^2}((\beta - \gamma_n)^T D^{-1}(\beta - \gamma_n))$ where $D^{-1} = X^TX + V^{-1}$, $\gamma_n = D(X^Ty + V^{-1}\gamma_0) = D(X^TX\hat{\beta} + V^{-1}\gamma_0)$, so $\beta|y \sim N_p(\gamma_n, \sigma^2 D)$. As $V^{-1} \to 0$, we have $B|y \sim N_p(\hat{\beta}, (X^TX)^{-1}\sigma^2)$.

With $p(\beta) \propto C$ and $p(\sigma^2) \propto \sigma^{-2}$, then conditional on $\sigma^2$, from the preceding general model $\beta|y, \sigma^2 \sim N_n(\hat{\beta}, (X^TX)^{-1}\sigma^2)$ where $\hat{\beta} = (X^TX)^{-1}X^Ty$

Since $\beta|y, \sigma^2 \sim N_p(\hat{\beta}, (X^TX)^{-1}\sigma^2)$, a single regression coefficient $\beta_i$ has posterior $\beta_i|y, \sigma^2 \sim N(\hat{\beta}_i, s_i^2\sigma^2)$, where $s_i^2 = (X^TX)^{-1}_{ii}$.

## 6. Categorical Data, Prediction, and Ranking

Suppose $N$ individuals are classified according to two binary variables, into a $2 \times 2$ table. We have three situations — one margin fixed, both margins fixed, and the overall total fixed.

If one margin is fixed, then $n_i$, and $n_2$ are fixed. Then $y_{i1} \sim$ BINOMIAL$(n_i, p_i)$.

If no margins are fixed, we only fix the total $N = \sum y_{ij}$. With a full multinomial model $Y \sim$ MULTINOMIAL$(q, N)$. Note if we just take a single row, we have standard BETABINOMIAL updating, as $Y_{11}|n_1 \sim$ BINOMIAL$(n_1, \frac{q_{11}}{q_{1.}})$ from the properties of the multinomial, and $\frac{q_{11}}{q_{1.}}$ from the properties of the Dirichlet.

**Definition.** *Recall if* $Y_k \sim$ POISSON$(\mu_k)$*, and* $\sum_k Y_k = N$*, then* $Y|N \sim$ MULTINOMIAL$(q, N)$*. Letting* $Y_k \sim Poission(\mu_k)$ *and using log-link function* $\log u_k = \lambda + \alpha_k$*, give a uniform prior to* $\lambda$*. This is equivalent to assuming a multinomial distribution for* $Y$ *with parameters* $q_k = \frac{e^{\alpha_k}}{sum_k}e^{\alpha_k}$*,* $N = \sum_k Y_k$*.*

For a $2 \times 2$ table, we can assume $Y_{ij} \sim$ POISSON$(\mu_{ij})$ and assume $\log \mu_{ij} = \phi + \alpha_i + \beta_j + \gamma_{ij}$ with the corner constraints$\alpha_1 = \beta_1 = \gamma_{12} + \gamma_{11} = \gamma_{21} = 0$.

Assuming we have multinomial observations $Y_i \sim$ MULTINOMIAL$(q_i, N_i)$ with covariates $x_i = x_{i1}, \ldots, x_{iP}$. Then we can express log odds of a category $k$ relative to a baseline category as $\phi_{k1} = \log\frac{q_{ik}}{q_{i1}} = \sum_{p=1}^P \beta_{kp}x_{ip}$, with category probabilities $q_{ik} = \frac{\exp(\sum_p \beta_{kp}x_{ip})}{\sum_k \exp(\sum_p \beta_{kp}x_{ip})}$.

**Definition.** *For ranking, assume* $O_i \sim$ POISSON$(\lambda_i E_i)$*, with* $\lambda_i$ *a standardized mortality rate, with Jeffreys prior* $\propto \frac{1}{\sqrt{\lambda_i}}$*.*

## 7. Sampling Properties in Relation to Other Methods

**Definition.** *Formally, an exchangeable sequence of random variables is a finite or infinite sequence* $X_1, X_2, \ldots$ *of random variables such that for any finite permutation* $\sigma$ *of the indices* $1, 2, 3, \ldots$*, (the permutation acts on only finitely many indices, with the rest fixed), the joint probability distribution of the permuted sequence*

$X_{\sigma(1)}, X_{\sigma(2)}, X_{\sigma(3)}, \ldots$ *is the same as the joint probability distribution of the original sequence.*

**Theorem.** If an infinite sequence of binary variables is exchangeable, then it implies that any finite set $p(y_1, \ldots, y_n) = \int \prod_{i=1}^n p(y_i|\theta)p(\theta)d\theta$ for some density $p(\theta)$ (with regularity conditions)

**Definition.** *The likelihood principle: all information about* $\theta$ *provided by data* $y$ *is contained in the likelihood* $\propto p(y|\theta)$*.*

**Theorem.** The statistic $t(Y)$ is sufficient for $\theta$ if and only if we can express the density $(y|\theta)$ in the form $p(y|\theta) = h(y)g(t(y)|\theta)$.

Trivially, the Bayesian posterior distribution only depends on the sufficient statistic.

## 8. Criticism and Comparison

**Definition.** *The Bayes factor comparison of models* $M_0$ *and* $M_1$ *are given as*

$$\frac{p(M_0|y)}{p(M_1|y)} = \frac{p(M_0)}{p(M_1)}\frac{p(y|M_0)}{p(y|M_1)} \quad (8.1)$$

*or in words — posterior odds of* $M_0$ *equals the Bayes factor* $(B_{01})$ *times the prior odds of* $M_0$*. This quantifies the weight of evidence in favor of the hypothesis* $H_0 : M_0$ *is true.*

*If both models are equally likely a priori, the Bayes factor is the posterior odds in favor of* $M_0$*.*

**Definition.** *The Bayesian Information Criterion (BIC) is*

$$BIC = -2\log p(y|\hat{\theta}) + k\log n \quad (8.2)$$

*where* $\hat{\theta}$ *is the MLE.* $BIC_0 - BIC_1$ *is intended to approximate* $-2\log B_{01}$

**Definition.** *The deviance of a sampling distribution is defnied as* $D(\theta) = 2\log p(y|\theta)$*.*

**Definition.** *The AIC is given as* $-2\log p(y|\hat{\theta}) + 2k$ *where kis the dimensionality of* $\theta$*.*

*Asmyptitocally, AIC is equivalent to leave-on-out cross-validation.*

**Definition.** *Model dimensionality can be measured as* $p_D = \mathbb{E}_{\theta|y}(-2\log p(y|\theta)) + 2\log p(y|\tilde{\theta}(y))$*. If we take* $\tilde{\theta} = \mathbb{E}(\theta|y)$*, then* $P_D$ *is equal to the posterior mean deviance minus the deviance of the posterior means.*

*We can approximate* $P_D \approx \text{tr}(-L''_\theta C)$*, where* $C = \mathbb{E}\Big((\theta-\overline{\theta})(\theta-\overline{\theta})^T\Big)$ *is the posterior covariance matrix of* $\theta$*.*

*Thus* $p_D$ *can be thought of the ratio of infomration in the likelihood about the parameters as a fraction of the total information in the posterior. We an also think of* $p_D$ *as the franction of total information in the posteriro that is identified for the prior.*

*For general normal regression models, we have this is exact, and* $p_D = \text{tr}((X^TX)(X^TX+V^{-1})^{-1})$*.*

*If there is* ***vague*** *prior infomration,* $\hat{\theta} \approx \overline{\theta}$ *(the MLE), and so* $D(\theta) \approx D(\overline{\theta}) - (\theta-\overline{\theta})^T L''(\hat{\theta})(\theta-\overline{\theta}) = D(\overline{\theta}) + \chi^2_p$*, and so* $p_D = \mathbb{E}(\chi^2_p) = p$*, the true number of parameters.*

**Definition.** *The* ***DIC*** *is defined as* $DIC = D(\overline{\theta}) + 2p_D = \overline{D} + p_D$*.*

## 9. Heirachcial Models

**Definition.** *Suppose* $y_{ij}$ *is outcome for individual* $j$*, unit* $i$*, with unit-specific parameter* $\theta_i$*. The assumption of partial exchangability of individuals within units can be represented by the following model —* $y_{ij} \sim p(y_{ij}|\theta_i, x_{ij})$*,* $\theta_i \sim p(\theta_i)$*.*

*Assumption of exchangability of units can be represented by the model* $\theta_i \sim p(\theta_i|\phi)$*,* $\phi \sim p(\phi)$ *- a common prior for all units (but a prior with unknown parameters.)*

*Excchangability is a* ***judgement*** *based on our knowledge of the context.*

*Assuming* $\theta_1, \ldots, \theta_I$ *are drawn from some common prior distribution whose parameters are unknown is known as a* ***hierarchical*** *model.*

**Definition.** *The normal-normal model is givens* $y_{ij} \sim N(\theta_i, \sigma^2)$*,* $j = 1, \ldots, n_i$*,* $i = 1, \ldots, I$*,* $\theta_i \sim N(\mu, \tau^2)$*,* $i = 1, \ldots, I$*,* $\mu \sim Uniform$*. Assume* $\sigma, \tau$ *known for the moment and express* $\tau^2$ *as* $\tau^2 = \frac{\sigma^2}{n_0}$*. From standard results,*

$$p(\theta_i|y,\mu,\tau,\sigma) = \textsc{Normal}\left(\frac{n_0\mu + n_i\overline{y}_i}{n_0+n_i}, \frac{\sigma^2}{n_0+n_i}\right) \tag{9.1}$$

*Now the marginal distribution of* $\overline{Y}_{i.}$ *is* $\overline{Y}_{i.} \sim N(\mu, \sigma^2(n_i^{-1} + n_0^{-1}))$*. Writing* $[\sigma^2(n_i^{-1}+n_0^{-1})]^{-1}$ *as* $\pi_i$*, the precision, we have* $\mu|y,\tau \sim N(\hat{\mu}, V_\mu)$ *where* $\hat{\mu} = \frac{\sum_i \pi_i \overline{y}_{i.}}{\sum_i \pi_i}$*,* $V_\mu = \frac{1}{\sum_i \pi_i}$*.*

*We can then show (reasonably easily) that* $\mathbb{E}(\pi_i|y,\tau,\sigma) = \frac{n_0\hat{\mu}+n_i\overline{y}_i}{n_0+n_i}$ *— an appropriate weighted average of the observed individual group mean and estimated population mean.*

**Definition.** *For normal hierarchical models the Jeffreys prior can be inconvenient. Assume* $y_i \sim N(\theta_i, \sigma_i^2)$*,* $\sigma_i^2$ *known, and* $\theta_i \sim N(\mu, \tau^2)$*,* $i = 1, \ldots, I$*. Then, integrating out the* $\theta_i$*, we get* $y_i|\mu, \tau^2 \sim N(\mu, \sigma_i^2 + \tau^2)$ *which are conditionally independent given* $\mu, \tau^2$*.*

*The posterior is* $p(\tau^2|y) \propto p(y|\mu,\tau^2)p(\tau^2)$ *where* $p(y|\mu,\tau^2) \propto \prod_i(\sigma_i^2 + \tau^2)^{-\frac{1}{2}}\exp(-\frac{1}{2}$
$frac(y_i-\mu)^2\sigma_i^2+\tau^2)$*.*

*Letting* $\tau^2 \to 0$*,* $p(y|\mu,\tau^2)$ *tends to a non-zero constant* $c$*, so* $p(\tau^2 < \epsilon|y) \propto cP(\tau^2 < \epsilon)$*.*

*Using an improper Jeffreys prior* $p(\tau^2 \propto \tau^{-2})$*,* $p(\tau^2 < \epsilon)$ *is unbounded, and so* $p(\tau^2 < \epsilon|y)$ *is unbounded, hence the posterior is improper.*

*Note that* $\frac{1}{\tau^2} \sim \textsc{Gamma}(\epsilon, \epsilon)$ *is proper, but inference can be sensitive to the choice of* $\epsilon$*.*

**Definition.** *Empirical Bayes methods proceed as before* $y_{ij} \sim p(y_{ij}|\theta_i)$*,* $\theta_i \sim p(\theta_i|\phi)$*, but do not put a prior on* $\phi$*. Estimate* $\phi$ *by, for example, maximum marginal likelihood — the value* $\hat{\phi}$ *that maximizes the marginal likelihood*

$$p(y|\phi) = \prod_i \int \prod_j p(y_{ij}|\theta_i)p(\theta_i|\phi)d\theta_i, \tag{9.2}$$

*known as the* ***Type II Maximum Likelihood****. Then use* $\hat{\phi}$ *as a "plug-in" estimate, as if the prior distribution was known.*

*Can think of it as* ***estimating*** *prior from the data — understates uncertainty since it ignores uncertainty in* $\hat{\phi}$ *— for large number of units and observations, have similar results to the "full Bayes" approach.*

## 10. Robustness and Outlier Detection

**Definition.** *If we assume, say* $Y \sim t_k(\theta, \tau)$*, then estimates will be less influenced by outliers. If we want to simultaneously find outliers, we can fit a t-distribution as a mixture of normals. Recall if* $Y \sim Norm(\theta, \sigma^2)$*, and* $\sigma^2 = \frac{\tau^2 k}{X^2}$*, where* $X^2 \sim \chi^2_k$*, then* $Y \sim t_k(\theta, \tau)$*. So an equivalent model to* $Y \sim t_k(\theta, \tau)$ *is to assume* $Y \sim \textsc{Normal}(\theta, \sigma_i^2)$*,* $\sigma_i^2 = \frac{\tau^2 k}{X_i^2}$*,* $X_i^2 \sim \chi^2_k$*, and monitor* $s_i = \frac{k}{X_i^2}$ *— values of* $s_i$ *much great than 1 indicate outliers.*

## 11. Miscellaneous

## References

Table 1. Conjugate Prior Distributions

| L | P | ConjP | Posterior | Predictive | Interpretation |
|---|---|---|---|---|---|
| Binomial | $\theta$ | Beta$(a, b)$ | $a + y,\, b + n - y$ | BetaBinomial$(y)$ | $\alpha - 1$ successes, $\beta - 1$ failures |
| Poisson | $\theta$ | Gamma$(a, b)$ | $a + y,\, b + n$ | NegativeBinomial$(y)$ | $\alpha$ total occurences in $\beta$ intervals |
| Normal | $\mu$ | Normal$(\gamma, \frac{\sigma^2}{n_0})$ | $\frac{n_0\gamma + n\overline{y}}{n_0+n}, \sigma_n^2 = \frac{\sigma^2}{n_0+n}$ | Normal$(\gamma_n, \sigma^2 + \sigma_n^2)$ | $n_0$ observations with sample mean $\gamma$ |
| Normal | $\mu$ | Normal$(\gamma, \tau_0)$ (precision) | $\frac{\tau_0\gamma + n\overline{y}}{\tau_0+n\tau}, \tau_n = \tau_0 + n\tau$ | Normal$(\gamma_n, \frac{1}{\tau_n} + \frac{1}{\tau})$ | |
| Normal | $\sigma^2 = \frac{1}{\omega}$ | $\omega \sim$ Gamma$(\frac{n_0}{2}, \frac{n_0\sigma_0^2}{2})$ | $\frac{n_0+n}{2}, \frac{n_0\sigma_0^2}{2} + \frac{1}{2}\sum(y_i - \mu)^2$ | | |
| Multinomial | $p_1, \ldots, p_k$ | Dirichlet$(\alpha_1, \ldots, \alpha_k)$ | $\alpha_1 + n_1, \ldots, \alpha_k + n_k$ | | $\alpha_i - 1$ occurences of category $i$ |

Table 2. Distributions

| Distribution | Density | Mean | Variance |
|---|---|---|---|
| Normal$(\mu, \sigma^2)$ | $\frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(x-\mu)^2}{2\sigma^2})$ | $\mu$ | $\sigma^2$ |
| Poisson$(\lambda)$ | $\frac{e^{-\lambda}\lambda^k}{k!}$ | $\lambda$ | $\lambda$ |
| Gamma$(a, b)$ | $\frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$ | $\frac{a}{b}$ | $\frac{a}{b^2}$ |
| Beta$(a, b)$ | $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$ | $\frac{a}{a+b}$ | $\frac{ab}{(a+b)^2(a+b+1)}$ |
| Dirichlet$(\alpha_1, \ldots, \alpha_K)$ | $\propto \prod_{i=1}^{K} x_i^{\alpha_i - 1}$ | $\frac{\alpha_i}{\sum_k \alpha_k}$ | |