1 Bayes

1.1 Bayes Theorem

$$P(A \cap B \cap I) = P(A|B \cap I)P(B \cap I) = P(A|B \cap I)P(B|I)P(I)$$

and:

$$P(A \cap B \cap I) = P(B|A \cap I)P(A \cap I) = P(B|A \cap I)P(A|I)P(I)$$

thus:

$$P(A|B \cap I) = \frac{P(B|A \cap I)P(A|I)\cancel{P(I)}}{P(B|I)\cancel{P(I)}}$$

Application:

$$P(D^-|T^+) = \frac{P(T^+|D^-)P(D^-)}{P(T^+)}$$

Law of total probabilities:

$$= \frac{P(T^+|D^-)P(D^-)}{P(T^+|D^-)P(D^-) + P(T^+|D^+)P(D^+)}$$

$$= \frac{(1 - P(T^-|D^-))(1 - P(D^+))}{(1 - P(T^-|D^-))(1 - P(D^+)) + P(T^+|D^+)P(D^+)}$$

with Sensitivity $P(T^+|D^+)$, Specificity $P(T^-|D^-)$, Prevalence $P(D^+)$

1.2 Bayes Factor

$$\boxed{BF_{01}(y) = \frac{f(y|H_0)}{f(y|H_1)}} \quad (1)$$

Is a direct quantitative measure of how data y have increased or decreased the odds of H_0 and is referred as the strength of evidence for or against H_0.

$$P(H_i|y) = \frac{f(y|H_i)P(H_i)}{P(y)} \rightarrow \underbrace{\frac{P(H_0|y)}{P(H_1|y)}}_{\text{Posterior Odds}} = \underbrace{\frac{f(y|H_0)}{f(y|H_1)}}_{BF_{01}(y)} \underbrace{\frac{P(H_0)}{P(H_1)}}_{\text{Prior Odds}}$$

$$\frac{P(H_0|y)}{1 - P(H_0|y)} = BF_{01}(y)\frac{P(H_0)}{1 - P(H_0)} \rightarrow P(H_0|y) = \frac{BF_{01}\frac{P(H_0)}{1-P(H_0)}}{1 + BF_{01}\frac{P(H_0)}{1-P(H_0)}}$$

Levels of Evidence against H_0: weak(1 to 1/3), moderate (1/3 to 1/10), substantial (1/10 to 1/30), strong (1/30 to 1/100), very strong (1/100 to 1/300) and decisive (<1/300).The minimum Bayes factor is the smallest Bayes factor within a certain class of alternative hypotheses. Minimum Bayes factors are very interesting because they quantify the maximal evidence of a p-value against a point H_0 within a certain class of alternative hypotheses.

Multiple hypothesis comparison:

$$BF_{01}(y)BF_{12}(y) = \frac{f(y|H_0)}{f(y|H_1)}\frac{f(y|H_1)}{f(y|H_2)} = \frac{f(y|H_0)}{f(y|H_2)} = BF_{02}(y)$$

Marginal Likelihood under H_0:

$$\boxed{f(y|H_1) = \int f(y|\theta)f(\theta|H_1)d\theta} \quad (2)$$

Application:

$Y|\mu \sim \mathcal{N}(\mu, \kappa^{-1})$ with known κ^{-1}

- $H_0 : \mu = \mu_0 \rightarrow N(\mu, V = \kappa^{-1})$
- H_1 : suppose that μ is known with prior distribution $\mu \sim \mathcal{N}(\nu, \lambda^{-1})$

$$f(y|H_1) = \int \sqrt{\frac{\kappa}{2\pi}} \exp\left(-\frac{\kappa}{2}(y-\mu)^2\right) \cdot \sqrt{\frac{\lambda}{2\pi}} \exp\left(-\frac{\lambda}{2}(\mu-\nu)^2\right) d\mu$$

$$= \int \sqrt{\frac{\kappa\lambda}{4\pi^2}} \exp\left(-\frac{1}{2}\left(\kappa y^2 - 2\kappa y\mu + \kappa\mu^2 + \lambda\mu^2 - 2\lambda\mu\nu + \lambda\nu^2\right)\right) d\mu$$

$$= \int \sqrt{\frac{\kappa\lambda}{4\pi^2}} \exp(-\frac{1}{2}($$

$$(\kappa+\nu)(\mu^2 - 2\mu\frac{\kappa y + \lambda\nu}{\kappa+\nu} + \left(\frac{\kappa y + \lambda\nu}{\kappa+\nu}\right)^2 - \left(\frac{\kappa y + \lambda\nu}{\kappa+\nu}\right)^2) + \kappa y^2 + \lambda\nu^2))d\mu$$

$$= \int \sqrt{\frac{\kappa\lambda}{2\pi}}\sqrt{\frac{\kappa+\lambda}{2\pi}}\frac{1}{\sqrt{\kappa+\lambda}} \exp\left(-\frac{1}{2}(\kappa+\lambda)\left(\mu - \frac{\kappa y + \lambda\nu}{\kappa+\nu}\right)^2\right) \cdot$$

$$\exp\left(\frac{(\kappa y + \lambda\nu)^2 - (\kappa y^2 + \lambda\nu^2)(\kappa+\lambda)}{2(\kappa+\lambda)}\right) d\mu$$

$$= \sqrt{\frac{\kappa\lambda}{2\pi(\kappa+\lambda)}} \exp\left(\frac{-\kappa\lambda(y-\nu)^2}{2(\kappa+\lambda)}\right) \cdot$$

$$\underbrace{\int \sqrt{\frac{\kappa+\lambda}{2\pi}} \exp\left(-\frac{1}{2}(\kappa+\lambda)\left(\mu - \frac{\kappa y - \lambda\nu}{\kappa+\lambda}\right)\right) d\mu}_{=1}$$

$$\rightarrow \boxed{Y|H_1 \sim \mathcal{N}\left(\nu, \frac{\kappa+\lambda}{\kappa\lambda}\right)}$$

1.3 Calibration of p-value

```
(dd<-matrix(c(9,5,14,1),ncol=2,dimnames=list(c("A","B"), c("Ca","Con"))))
```

```
##   Ca Con
## A  9  14
## B  5   1
```

```
library(pCalibrate);(result <- twoby2Calibrate(dd))
```

```
## $minBF
## [1] 0.3684956
##
## $p.value
##       p.pb       p.ce       p.bl      p.mid      p.lie
## 0.08007663 0.13908046 0.08007663 0.07586207 0.06580151
```

```
c(formatBF(result$minBF),round(BF2pp(result$minBF,0.5),3))
```

```
## [1] "1/2.7" "0.269"
```

The Bayes factor of 1/2.7 means weak evidence against H_0. Given $P(H_0) + P(H_1) = 1$, and assuming $P(H_0) = 0.5$ the BF_{01} yields a posterior $P(H_0|y)$=0.269.

2 Posterior Distribution

2.1 Conjugate Bayes

Likelihood	Conjugate Prior
Bin(n,π)	$\pi \sim Be(\alpha, \beta)$
$f(y \mid \pi) = \binom{n}{y}\pi^y(1-\pi)^{n-y}$	$f(\pi) = \frac{\pi^{\alpha-1}(1-\pi)^{\beta-1}}{B(\alpha,\beta)}$
Geom(π)	$\pi \sim Be(\alpha, \beta)$
$f(y \mid \pi) = \pi(1-\pi)^{y-1}$	$f(\pi) = \frac{\pi^{\alpha-1}(1-\pi)^{\beta-1}}{B(\alpha,\beta)}$
Po(λ)	$\lambda \sim G(\alpha, \beta)$
$f(y \mid \lambda) = \frac{\lambda^y \exp(-\lambda)}{y!}$	$f(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}\exp(-\beta\lambda)$
Exp(λ)	$\lambda \sim G(\alpha, \beta)$
$f(y \mid \lambda) = \lambda\exp(-\lambda y)$	$f(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}\exp(-\beta\lambda)$
N(μ, σ^2_{known})	$\mu \sim N(\nu, \tau^2)$
$f(y \mid \mu) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$	$f(\mu) = \frac{1}{\sqrt{2\pi\tau^2}}\exp\left(-\frac{(\mu-\nu)^2}{2\tau^2}\right)$
N(μ_{known}, σ^2)	$\sigma^2 \sim IG(\alpha, \beta)$
$f(y \mid \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$	$f(\sigma^2) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{-\alpha-1}\exp\left(-\frac{\beta}{\sigma^2}\right)$

Assume: $y_{1:n} \sim N(m, \kappa^{-1})$ and prior $m \sim N(\mu, \lambda^{-1})$:

$$\underbrace{f(m|y_1,...,y_n)}_{\text{posterior}} \propto \underbrace{f(y_1,...,y_n|m)}_{\text{likelihood}}\underbrace{f(m)}_{\text{prior}} / f(y_1,...,y_n)$$

$$\boxed{f(m|y_1,...,y_n) = \frac{f(y_1,...,y_n|m)f(m)}{\underbrace{\int_{-\infty}^{\infty} f(y_1,...,y_n|m)f(m)dm}_{f(y_1,..y_n)}}}$$

Likelihood:

$$f(y_1,...,y_n|m) = \prod_{i=1}^{n}\left(\frac{\kappa}{2\pi}\right)^{1/2}\exp\left(-\frac{\kappa}{2}(y_i-m)^2\right)$$

$$= \left(\frac{\kappa}{2\pi}\right)^{n/2}\exp\left(-\frac{\kappa}{2}\sum_{i=1}^{n}(y_i-m)^2\right)$$

Prior:

$$f(m) = \left(\frac{\lambda}{2\pi}\right)^{1/2}\exp\left(-\frac{\lambda}{2}(m-\mu)^2\right)$$

Posterior (only has to depend on m):

$$f(y_1,...,y_n|m) = \left(\frac{\kappa}{2\pi}\right)^{n/2}\left(\frac{\lambda}{2\pi}\right)^{1/2}\exp\left(-\frac{\kappa}{2}\sum_{i=1}^{n}(y_i-m)^2-\frac{\lambda}{2}(m-\mu)^2\right)$$

$$= \left(\frac{\kappa}{2\pi}\right)^{n/2}\left(\frac{\lambda}{2\pi}\right)^{1/2}\exp\left(-\frac{\kappa}{2}\sum_{i=1}^{n}(y_i^2-2y_im+m^2)-\frac{\lambda}{2}(m^2-2m\mu+\mu^2)\right)$$

$$= \left(\frac{\kappa}{2\pi}\right)^{n/2}\left(\frac{\lambda}{2\pi}\right)^{1/2}$$

$$\cdot\exp\left(-\frac{1}{2}\left[(\lambda+\kappa n)\left(m^2-2m\frac{\mu\lambda+\kappa n\bar{y}}{\lambda+n\kappa}\right)+\kappa\sum_{i=1}^{n}y_i^2+\lambda\mu^2\right]\right)$$

$$= \left(\frac{\kappa}{2\pi}\right)^{n/2}\left(\frac{\lambda}{2\pi}\right)^{1/2}$$

$$\cdot\exp\left(-\frac{1}{2}\left[(\lambda+\kappa n)\left(m-\frac{\mu\lambda+\kappa n\bar{y}}{\lambda+n\kappa}\right)^2\right]\right)$$

$$\cdot\exp\left(-\frac{1}{2}\left[-(\lambda+\kappa n)\left(\frac{\mu\lambda+\kappa n\bar{y}}{\lambda+n\kappa}\right)^2+\kappa\sum_{i=1}^{n}y_i^2+\lambda\mu^2\right]\right)$$

$$\propto \exp\left(-\frac{\lambda+\kappa n}{2}\left(m-\frac{\mu\lambda+\kappa n\bar{y}}{\lambda+n\kappa}\right)^2\right)$$

$$\rightarrow \boxed{m|y_1,...,y_n \sim N\left(\frac{\mu\lambda+\kappa n\bar{y}}{\lambda+n\kappa},(\lambda+\kappa n)^{-1}\right)}$$

More examples!!!

2.2 Bayesian Learning:

We will end at the same stage if we analyse a trial sequentially. Updating prior belief by accumulated data is the same as treating all of them as component of the likelihood.

$$\begin{aligned}P(\theta|y_1,y_2,y_3,I) &\propto P(y_3|\theta,y_1,y_2,I)P(\theta|,y_1,y_2,I)\\ &\propto P(y_3|\theta,y_1,y_2,I)P(y_2|\theta,y_1,I)\underbrace{P(y_1|\theta,I)P(\theta|I)}_{\propto P(\theta|y_1,I)}\\ &\propto \underbrace{\prod_{i=1}^{3}P(y_i|\theta,I)}_{\text{pooled likelihood}}\underbrace{P(\theta|I)}_{prior}\end{aligned}$$

2.3 Prior and Posterior Effective Sample size

Beta distribution:

$$f(p) = \frac{1}{B(\alpha,\beta)}p^{\alpha-1}(1-p)^{\beta-1}$$

- $B(x,y) = \int_0^1 u^{x-1}(1-u)^{y-1}du = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$

$$X\sim \text{Beta}(\alpha,\beta) \qquad \begin{cases}\mathbb{E}[X] = \dfrac{\alpha}{\alpha+\beta}\\ \text{Var}[X] = \dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \dfrac{\mathbb{E}[X]\cdot(1-\mathbb{E}[X])}{\alpha+\beta+1}\end{cases}$$

Prior effective sample size:

$$\text{PriESS} \approx \frac{1}{\text{Var}[X]} \approx \frac{\mathbb{E}[X]\cdot(1-\mathbb{E}[X])-\text{Var}[X]}{\text{Var}[X]} = \alpha+\beta$$

Posterior effective sample size:

$$\text{PostESS} \approx \frac{1}{\text{Var}[p\mid y_1,\cdots,y_n]} \approx \alpha+n\bar{y}+\beta+n-n\bar{y} = \alpha+\beta+n$$

This informs us about the weight of the prior (skeptical, neutral, enthusiastic)

$$\text{PriESS} \approx \underbrace{\frac{1}{Var(p)}}_{\text{prior precision}}$$

where $Var(p)$ comes from the prior distribution.

$$\text{PostESS} \approx \underbrace{\frac{1}{Var(p|y_1,...y_n)}}_{\text{posterior precision}}$$

2.4 Credible Intervals:

Confidence interval:
For repeated random samples from a distribution with unknown parameter θ, a $(1-\alpha)100\%$ confidence interval will cover θ in $(1-\alpha)100\%$ of all cases. (θ is fixed but unknown)

Credible interval:
The posterior probability of p (has now distribution and is not fixed) lies between x and y with 95% , when a specific prior is assumed.

Equi-tailed credible interval:
An equi-tailed (1-α) credible interval has $\frac{\alpha}{2}, 1-\frac{\alpha}{2}$ quantiles of $f(\theta|\boldsymbol{y})$ at its endpoints. Discards equal amounts of the posterior probability on either side of the interval.

- intuitively straightforward
- Easy to compute from MC and MCMC samples
- Nice invariance properties

Highest posterior density (HPD) interval:
Provides the shortest possible (1-α) credible interval

- For symmetric distributions coincides with equi-tailed credible interval
- Harder to compute
- Invariance property does not hold

3 Lecture 3

3.1 Predictive distribution:

Predictive distributions for binary data:

$$\begin{aligned}&\text{Prior predictive distribution} = f(y_1,\cdots,y_n)\\ &= \int_0^1 f(y_1,\cdots,y_k,p)dp\\ &= \int_0^1 \underbrace{f(y_1,\cdots,y_k\mid p)}_{\text{Binomial likelihood}}\underbrace{f(p)}_{\text{Beta prior}}\,dp\\ &= \binom{k}{k\bar{y}^{(k)}}\frac{\text{B}(\alpha+k\bar{y}^{(k)},\beta+k-k\bar{y}^{(k)})}{\text{B}(\alpha,\beta)}\\ &= \binom{1}{y}\frac{\text{B}(\alpha+y,\beta+1-y)}{\text{B}(\alpha,\beta)} \qquad \text{for } k=1\\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\cdot\frac{\Gamma(\alpha+y)\Gamma(\beta+1-y)}{\Gamma(\alpha+\beta+1)}\\ &= \frac{1}{\alpha+\beta}\cdot\frac{\Gamma(\alpha+y)}{\Gamma(\alpha)}\cdot\frac{\Gamma(\beta+1-y)}{\Gamma(\beta)}\\ &= \begin{cases}\frac{\alpha}{\alpha+\beta}, & \text{if } y=1\\ \frac{\beta}{\alpha+\beta}, & \text{if } y=0\end{cases}\end{aligned}$$

$$\begin{aligned}&\text{Posterior predictive distribution}\\ &= f(\underbrace{y_{n+1},\cdots,y_{n+k}}_{\text{future observations}}\mid \underbrace{y_1,\cdots,y_n}_{\text{known observations}})\\ &= \int_0^1 f(y_{n+1},\cdots,y_{n+k},p\mid y_1,\cdots,y_n)dp\\ &\overset{i.i.d.}{=} \int_0^1 f(y_{n+1},\cdots,y_{n+k}\mid p)f(p\mid y_1,\cdots,y_n)dp\\ &\because\quad p\mid y_1,\cdots,y_n \sim \text{Beta}(\alpha+n\bar{y}^{(n)},\beta+n-n\bar{y}^{(n)})\\ &= \binom{k}{k\bar{y}^{(k)}}\frac{\text{Beta}(\alpha+n\bar{y}^{(n)}+k\bar{y}^{(k)},\beta+n-n\bar{y}^{(n)}+k-k\bar{y}^{(k)})}{\text{Beta}(\alpha+n\bar{y}^{(n)},\beta+n-n\bar{y}^{(n)})}\\ &\text{for } k=1\end{aligned}$$

$$\text{Be}\left(\frac{\alpha+n\bar{y}^{(n)}}{\alpha+\beta+n}\right) = \begin{cases}\frac{\alpha+n\bar{y}^{(n)}}{\alpha+\beta+n}, & \text{if } y_{n+1}=1\\ \frac{\beta+n-n\bar{y}^{(n)}}{\alpha+\beta+n}, & \text{if } y_{n+1}=0\end{cases}$$

Predictive distributions for normal data:

$$\begin{aligned}\text{Likelihood:}\quad & y_1,\cdots,y_n\mid m \overset{i.i.d.}{\sim} \text{N}(m,\kappa^{-1})\\ \text{Prior:}\quad & m\sim\text{N}(\mu,\lambda^{-1})\\ \text{Posterior:}\quad & m\mid y_1,\cdots,y_n \sim \text{N}\left(\frac{\kappa n\bar{y}+\lambda\mu}{n\kappa+\lambda},(n\kappa+\lambda)^{-1}\right)\end{aligned}$$

Prior predictive distribution: $y \sim \mathrm{N}(\mu, \lambda^{-1} + \kappa^{-1})$

Posterior predictive distribution: $y_{n+1} \mid y_1, \cdots, y_n \sim \mathrm{N}(\mu_{\text{post}}, \lambda_{\text{post}}^{-1} + \kappa^{-1})$

An alternative proof of the prior predictive distribution:

$$\begin{aligned} Y \mid m &\sim \mathrm{N}(m, \sigma^2) \\ m &\sim \mathrm{N}(\mu, \tau^2) \\ Y &\sim \mathrm{N}(\mu, \tau^2 + \sigma^2) \end{aligned}$$

3.2 Monte Carlo simulation:

We can sample independent $u \sim U(0,1)$ and transform it with $F^{-1}(.)$ to get samples of x.

$$\boxed{P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x)}$$

$$\mathbb{E}(Y) = \mathbb{E}_m[\mathbb{E}_Y(Y \mid m)] = \mathbb{E}_m(m) = \mu$$

$$\begin{aligned} \mathrm{Var}(Y) &= \mathrm{Var}_m[\mathbb{E}_Y(Y \mid m)] + \mathbb{E}_m[\mathrm{Var}_Y(Y \mid m)] \\ &= \mathrm{Var}_m[m] + \mathbb{E}_m[\sigma^2] \\ &= \tau^2 + \sigma^2 = \lambda^{-1} + \kappa^{-1} \end{aligned}$$

Poisson-Gamma

$$\underbrace{f(\lambda \mid y_{1:n})}_{\text{Posterior}} \propto \underbrace{f(y_{1:n} \mid \lambda)}_{\text{Likelihood}} \cdot \underbrace{f(\lambda)}_{\text{Prior}}$$

Likelihood:

$$f(y_{1:n} \mid \lambda) = \prod_{i=1}^{n} \frac{\lambda^{y_i} \exp(-\lambda)}{y_i!} \propto \lambda^{\sum_{i=1}^n y_i} \exp(-n\lambda)$$

Prior:

$$f(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} \exp(-\beta\lambda) \propto \lambda^{\alpha-1} \exp(-\beta\lambda)$$

Posterior:

$$\begin{aligned} f(\lambda \mid y_{1:n}) &\propto f(y_{1:n} \mid \lambda) \cdot f(\lambda) \\ &\propto \lambda^{\sum_{i=1}^n y_i} \exp(-n\lambda) \cdot \lambda^{\alpha-1} \exp(-\beta\lambda) \\ &= \lambda^{\sum_{i=1}^n y_i + \alpha - 1} \exp(-(n+\beta)\lambda) \end{aligned}$$

$$f(\lambda \mid y_{1:n}) \propto \lambda^{(\alpha + \sum_{i=1}^n y_i) - 1} \exp(-(\beta + n)\lambda)$$

Hence:

$$\lambda \mid y_{1:n} \sim \mathrm{G}\left(\alpha + \sum_{i=1}^{n} y_i, \beta + n\right)$$

Likelihood:

$$y_i \mid \lambda \sim \mathrm{Po}(\lambda)$$

Prior:

$$\lambda \sim \mathrm{G}(\alpha, \beta)$$

Prior predictive distribution:

$$\begin{aligned} f(y_i) &= \int_0^\infty f(y_i \mid \lambda) \cdot f(\lambda) \mathrm{d}\lambda \\ &= \int_0^\infty \frac{\lambda^{y_i} \exp(-\lambda)}{y_i!} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} \exp(-\beta\lambda) \mathrm{d}\lambda \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \frac{1}{y_i!} \int_0^\infty \lambda^{y_i + \alpha - 1} \exp(-(1+\beta)\lambda) \mathrm{d}\lambda \\ &= \frac{\beta^\alpha}{(\beta+1)^{\alpha+y_i}} \cdot \frac{\Gamma(\alpha+y_i)}{\Gamma(\alpha)} \cdot \frac{1}{y_i!} \underbrace{\int_0^\infty \frac{(\beta+1)^{\alpha+y_i}}{\Gamma(\alpha+y_i)} \lambda^{(\alpha+y_i)-1} \exp(-(\beta+1)\lambda) \mathrm{d}\lambda}_{\text{integrates to 1}} \\ &= \frac{\beta^\alpha}{(\beta+1)^{\alpha+y_i}} \cdot \frac{\Gamma(\alpha+y_i)}{\Gamma(\alpha)} \cdot \frac{1}{y_i!} \end{aligned}$$

Exponential-Gamma Likelihood:

$$y \mid \lambda \sim \mathrm{Exp}(\lambda)$$

$$f(y \mid \lambda) = \lambda \exp(-\lambda y)$$

Prior:

$$\lambda \sim \mathrm{G}(\alpha, \beta)$$

$$f(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} \exp(-\beta\lambda)$$

Posterior:

$$f(\lambda \mid y) \propto f(y \mid \lambda) \cdot f(\lambda)$$

$$\begin{aligned} &= \lambda \exp(-\lambda y) \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} \exp(-\beta\lambda) \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha} \exp(-(y+\beta)\lambda) \end{aligned}$$

$$\lambda \mid y \sim \mathrm{G}(\alpha + 1, \beta + y)$$

Geometric-Beta Likelihood:

$$y \mid \pi \sim \mathrm{Geom}(\pi)$$

$$f(y \mid \pi) = \pi(1-\pi)^{y-1}$$

Prior:

$$\pi \sim \mathrm{Be}(\alpha, \beta)$$

$$f(\pi) = \frac{1}{\mathrm{B}(\alpha, \beta)} \pi^{\alpha-1} (1-\pi)^{\beta-1}$$

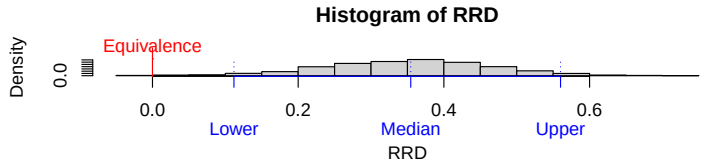
Posterior:

$$f(\pi \mid y) \propto f(y \mid \pi) \cdot f(\pi)$$

$$\begin{aligned} &= \pi(1-\pi)^{y-1} \cdot \frac{1}{\mathrm{B}(\alpha, \beta)} \pi^{\alpha-1} (1-\pi)^{\beta-1} \\ &= \frac{1}{\mathrm{B}(\alpha, \beta)} \pi^{\alpha} (1-\pi)^{\beta+y-2} \end{aligned}$$

$$\pi \mid y \sim \mathrm{G}(\alpha + 1, \beta + y - 1)$$

```
set.seed(13);n <- 1000
#1. MC sample of posterior response rates
secukinumab_posterior <- rbeta(n=n,shape1=14.5, shape2=10)
placebo_posterior <- rbeta(n=n,shape1=12 , shape2=37)
# 2. Response rate differences (RRD)
RRD <- secukinumab_posterior-placebo_posterior
RRD_estimate <- mean(RRD)
RRD_CrI <- quantile(RRD,probs = c(0.025, 0.975))
# 3. Median, 95% CrI
measures <- quantile(RRD,probs = c(0.025, 0.5, 0.975))
names(measures) <-c("Lower","Median","Upper")
# 4 Posterior Probability of superiority (RRD > 0)
PPS <- sum(RRD>0)/length(RRD) #or mean(RRD>0)
# 5 Standard error of Posterior Probability of superiority (RRD > 0)
se_PPS <- sqrt(var(RRD>0)/length(RRD))
measures

##     Lower    Median     Upper
## 0.1118431 0.3545441 0.5601402
```



Interpretation: *We randomly draw samples from both, the secukinumab and placebo response rate posterior distribution and look at the difference in response rate (POC requires that the response rate is of secukinumab is larger then that of placebo POC > 90%). One therefore creates a sample of the response differences and this sample takes on a specific distribution. A Null-hypothesis would says that there is no difference between the two treatments and would be reflected in the distribution taking a mean at 0. From the data we can compute the 2.5%, 50% (median) and 97.5% quantiles that represent the 95% Credible Interval and thus the most promising values for the actual difference in response and we see that the reference value for the Null-Hypothesis $H_0 : diff = 0$ is not included in the interval. The posterior probability of success which is 0.997 with a standard error of 0.00173 suggests also high evidence against H_0 because it is significantly larger than 95% (not contained in the wald confidence interval [0.9936, 1]).*

4 Markov Chain Monte Carlo

The construction of a Markov Chain $f(\theta^{(t+1)}|\theta^{(t)})$ (should be easy to generate) is an iterative process, where the generated value $\theta^{(t+1)}$ on one step depends one the value of the previous step $\theta^{(t)}$. This eventually converges to the target distribution (stationary or equilibrium with $t \to \infty$ assures that the samples are identically distributed) which is the posterior distribution $f(\theta|\boldsymbol{y})$. MCMC samples are not anymore independent but we can still approximate areas. A Markov Chain is:

- Irreducible: No matter where you start, the chain is able to get reach any other point in a finite number of iterations with positive probability (no absorbing state)
- Positive-recurrent: The expected return time to any state is finite
- Aperiodic: No flip flop through states!

If all 3 criteria hold, the chain has converged and thus possesses a unique stationary distribution. If only aperiodic & positive recurrent this means it is ergodic!

4.1 Metropolis Hasting algorithm & Gibbs:

The aim of the Metropolis-Hastings algorithm is to generate a collection of states (here parameters α, β) that describe our target distribution, which is the logistic regression $f(\alpha, \beta|\boldsymbol{y}, \boldsymbol{n}, \boldsymbol{x})$. Via a Markov process, the state (the parameters) should converge to a stationary state which is the desired target distribution.

The algorithm starts with the so called condition of detailed balance and means simply that each transition, for example $\alpha \to \alpha'$, is reversible and happens with the same probability. More specifically this means the probability to be in state α and transitioning to state α' is equal to the probability to be in state α' and transitioning to state α.

$$\underbrace{P(\alpha'|\alpha)}_{\text{Transitioning to state } \alpha' \text{ given one is in } \alpha} \cdot \underbrace{f(\alpha, \beta|\boldsymbol{y}, \boldsymbol{n}, \boldsymbol{x})}_{\text{P to be in state } \alpha} = P(\alpha|\alpha') f(\alpha', \beta|\boldsymbol{y}, \boldsymbol{n}, \boldsymbol{x}$$

$$\frac{P(\alpha'|\alpha)}{P(\alpha|\alpha')} = \frac{f(\alpha', \beta|\boldsymbol{y}, \boldsymbol{n}, \boldsymbol{x})}{f(\alpha, \beta|\boldsymbol{y}, \boldsymbol{n}, \boldsymbol{x})}$$

For simplicity we focus her on one transition, namely $P(\alpha'|\alpha)$ but the same procedures also hold for the back-transition $P(\alpha|\alpha')$. Now, the transition $P(\alpha'|\alpha)$ has to be split into two steps. The first one is to "propose" the next value (α') and the second one is to accept this proposal or not. The proposal distribution $q(\alpha'|\alpha)$ is in our case here a normal distribution, so $\alpha' \sim \mathcal{N}(\alpha, \sigma_\alpha^2)$ (random walk proposal) and this means that the proposed value α' is depending on the current value α but the spread is defined by a tuning parameter σ_α^2 which kind of defines the update step. The acceptance distribution $A(\alpha', \alpha)$ does not have to be defined here as we will see. We can now rewrite:

$$\frac{A(\alpha', \alpha) \cdot q(\alpha'|\alpha)}{A(\alpha, \alpha') \cdot q(\alpha|\alpha')} = \frac{f(\alpha', \beta|\boldsymbol{y}, \boldsymbol{n}, \boldsymbol{x})}{f(\alpha, \beta|\boldsymbol{y}, \boldsymbol{n}, \boldsymbol{x})}$$

$$\frac{A(\alpha', \alpha)}{A(\alpha, \alpha')} = \frac{f(\alpha', \beta|\boldsymbol{y}, \boldsymbol{n}, \boldsymbol{x}) \cdot q(\alpha|\alpha')}{f(\alpha, \beta|\boldsymbol{y}, \boldsymbol{n}, \boldsymbol{x}) \cdot q(\alpha'|\alpha)}$$

In case of the Metropolis-Hastings algorithm the acceptance rate $\frac{A(\alpha', \alpha)}{A(\alpha, \alpha')}$ can be written as A^α and it is forced to be lower or equal to 1:

$$A^\alpha = \min\left(1, \frac{f(\alpha', \beta|\boldsymbol{y}, \boldsymbol{n}, \boldsymbol{x}) \cdot q(\alpha|\alpha')}{f(\alpha, \beta|\boldsymbol{y}, \boldsymbol{n}, \boldsymbol{x}) \cdot q(\alpha'|\alpha)}\right)$$

Bayes theorem:

$$= \min\left(1, \frac{f(\boldsymbol{y}, \boldsymbol{n}, \boldsymbol{x}|\alpha', \beta) \cdot f(\alpha', \beta) \cancel{f(\boldsymbol{y}, \boldsymbol{n}, \boldsymbol{x})^{-1}} \cdot q(\alpha|\alpha')}{f(\boldsymbol{y}, \boldsymbol{n}, \boldsymbol{x}|\alpha, \beta) \cdot f(\alpha, \beta) \cancel{f(\boldsymbol{y}, \boldsymbol{n}, \boldsymbol{x})^{-1}} \cdot q(\alpha'|\alpha)}\right)$$

Priors are independent:

$$= \min\left(1, \frac{f(\boldsymbol{y}, \boldsymbol{n}, \boldsymbol{x}|\alpha', \beta) \cdot f(\alpha') \cancel{f(\beta)} \cdot q(\alpha|\alpha')}{f(\boldsymbol{y}, \boldsymbol{n}, \boldsymbol{x}|\alpha, \beta) \cdot f(\alpha) \cancel{f(\beta)} \cdot q(\alpha'|\alpha)}\right)$$

We can also make a bivariate proposal $\binom{\alpha'}{\beta'}$<-rmvnorm($\binom{\alpha}{\beta}, \Sigma$):

$$A^{\alpha,\beta} = \min\left(1, \frac{f(\boldsymbol{y}, \boldsymbol{n}, \boldsymbol{x}|\alpha', \beta') \cdot f(\alpha') f(\beta') \cdot q(\alpha, \beta|\alpha', \beta')}{f(\boldsymbol{y}, \boldsymbol{n}, \boldsymbol{x}|\alpha, \beta) \cdot f(\alpha) f(\beta) \cdot q(\alpha', \beta'|\alpha, \beta)}\right)$$

Now, if we get a proposed value α' that is more probable then the current value α this means that the acceptance rate A^α is higher than 1 and we should always accept this value. On the other hand, if we get a proposed value that is less likely we should sometimes (to not get stuck in a plateau) accept it. This is achieved by adding a stochastic component in form of a random sample from the uniform distribution. If the acceptance rate is higher then the random sample we still accept this point. If it is lower we reject and stick with the current value α.

The Gibbs sampler is a special case where the proposal distribution is the current posteriori distribution. This means that A is always 1 and thus each step is always accepted (there is no tuning parameter):

$$q(\alpha|\alpha') = f(\alpha, \beta|\boldsymbol{y}, \boldsymbol{n}, \boldsymbol{x})$$

$$A^\alpha = \min\left(1, \frac{f(\alpha', \beta|\boldsymbol{y}, \boldsymbol{n}, \boldsymbol{x}) \cdot f(\alpha, \beta|\boldsymbol{y}, \boldsymbol{n}, \boldsymbol{x})}{f(\alpha, \beta|\boldsymbol{y}, \boldsymbol{n}, \boldsymbol{x}) \cdot f(\alpha', \beta|\boldsymbol{y}, \boldsymbol{n}, \boldsymbol{x})}\right) = 1$$

```r
#Define the parameters of the prior distributions
mu0 <- -3
sigma2_0 <- 4
a0 <- 1.6
b0 <- 0.4
# initialization
set.seed(44566)
n.iter <- 10000
n.burnin <- 4000
n.thin <- 1
#n.thin <- floor((n.iter-n.burnin)/500)
n.chains <- 1
parameters <- c("mu", "sigma2", "inv_sigma2")
n.parameters <- length(parameters)
n.tot <- n.burnin + n.iter*n.thin

gibbs_samples <- matrix(NA, nrow = n.iter, ncol = n.parameters)
mu.sim <- rep(NA, length = n.tot)
sigma2.sim <- rep(NA, length = n.tot)
inv.sigma2.sim <- rep(NA, length = n.tot)
colnames(gibbs_samples) <- parameters

#Set the initial value
sigma2.sim[1] <- 1/runif(n.chains)
k <- 0
#Run the for loop (only one chain)
for(i in 2:(n.burnin+n.iter*n.thin)){
  mu.sim[i] <- rnorm(1,
    mean=(sum(y)/sigma2.sim[i-1]+mu0/sigma2_0)/(n/sigma2.sim[i-1]+1/sigma2_0),
    sd=sqrt(1/(n/sigma2.sim[i-1]+1/sigma2_0))
        )
  sigma2.sim[i] <- 1/rgamma(1, shape = n/2 + a0,
                            scale = 1 / (sum((y-mu.sim[i])^2)/2 + b0)
                            )
  inv.sigma2.sim[i] <- 1/sigma2.sim[i]

  # after the burnin save every n.thin'th sample
  if((i > n.burnin) && (i%%n.thin == 0)){
    gibbs_samples[k,] <- c(mu.sim[i], sigma2.sim[i], inv.sigma2.sim[i])
    k <- k + 1
  }
}
```

4.2 MH code:

The parameters of interest in this task are the intercept (α) and the slope (β) of the logistical regression:

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta x_i$$

- p_i is the estimated relative frequency of deaths, which can be obtained from the data ($p_i = y_i/n_i$)
- Prior: $f(\alpha_0) = \mathcal{N}(0, 10^4)$, $f(\beta_0) = \mathcal{N}(0, 10^4)$
- Proposal distribution: $q(\alpha'|\alpha) = \mathcal{N}(\alpha, \sigma_\alpha^2)$, $q(\beta'|\beta) = \mathcal{N}(\beta, \sigma_\beta^2)$

$\to$ since both proposal distributions are symmetric ($q(i'|i) = q(i'|i)$) they cancel out when assessing the rejection!

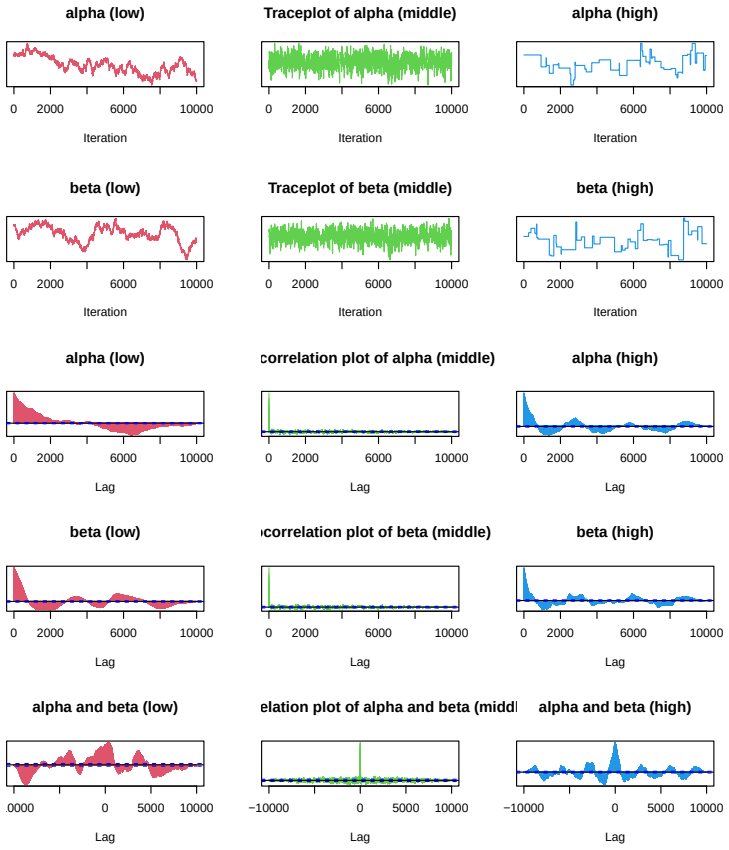
$$A(\alpha) = \min\left(1, \frac{f(\mathbf{y}, \mathbf{n}, \mathbf{x} \mid \alpha') \cdot f(\alpha')}{f(\mathbf{y}, \mathbf{n}, \mathbf{x} \mid \alpha) \cdot f(\alpha)}\right) = \min\left(1, \frac{\text{Bin}(n, \frac{y}{n} \mid \alpha') \cdot \text{N}(0, 10^4)}{\text{Bin}(n, \frac{y}{n} \mid \alpha) \cdot \text{N}(0, 10^4)}\right)$$

Tunning parameters $\sigma_\alpha, \sigma_\beta$ to get the optimal acceptance rate of 0.2-0.4:

- **Small:** The algorithm takes only very small steps and thus the acceptance rate is very high! Because they are too close, it is heavily cross- and auto-correlated and it fails to explore the whole parameter space!
- **High:** The algorithm takes only large jumps and thus the acceptance rate is very low and therefore the values stay mostly the same! It is heavily cross- and auto-correlated and it fails to explore the whole parameter space!

```r
# inverse logit: logit^(-1)(alpha + beta*x)
mypi <- function(alpha, beta, x){
  tmp <- exp(alpha + beta*x)
  pi <- tmp/(1+tmp)
  return(pi)
}
MH.sampler <- function(x,               # covariate values
                       y,               # number of mice deaths
                       n,               # total number of mice
                       sigma2 = 10^(4), # variance of normal priors
                       n.iter = 10000,  # number of MCMC iterations
                       n.burnin = 4000, # burnin length
                       n.thin = 1,      # thinning parameter
                       alpha = 0,       # starting point
                       beta = 0,        # starting point
                       s_alpha = 1,     # SD for normal proposal
                       s_beta = 60      # SD for normal proposal
                       ) {

  alpha_samples <- c();beta_samples <- c()
  # number of accepted proposals
  alpha_yes <- 0;beta_yes <- 0
  # counter
  count <- 0
  alpha_yes_history <- rep(0, n.burnin+n.iter*n.thin+1)
  beta_yes_history <- rep(0, n.burnin+n.iter*n.thin+1)
  # start the MCMC algorithm (the first iteration after the burn-in is 1)
  for(i in -n.burnin:(n.iter*n.thin)){
    count <- count + 1

    ## update alpha: generate a new proposal for alpha
    alpha_star <- rnorm(1, alpha, sd=s_alpha)

    # NOTE: it is more stable to calculate everything on the log scale
    enum<-sum(dbinom(y,size=n,prob=mypi(alpha_star,beta,x),log=TRUE)) +
      dnorm(alpha_star, mean=0, sd=sqrt(sigma2), log=TRUE)
    denom<-sum(dbinom(y,size=n,prob=mypi(alpha,beta,x),log=TRUE))+
      dnorm(alpha, mean=0, sd=sqrt(sigma2), log=TRUE)

    # log acceptance rate (since we use a random walk proposal there is no
    # proposal ratio in the acceptance probability)
    logacc <- enum - denom
    if(log(runif(1)) <= logacc){
      # accept the proposed value
      alpha <- alpha_star
      alpha_yes <- alpha_yes + 1; alpha_yes_history[count] <- 1
    }
```

```
    ## update beta: generate a new proposal for beta
    beta_star <- rnorm(1, beta, sd=s_beta)

    enum<-sum(dbinom(y,size=n,prob=mypi(alpha,beta_star,x),log=TRUE))+
      dnorm(beta_star, mean=0, sd=sqrt(sigma2), log=TRUE)
    denom<-sum(dbinom(y,size=n,prob=mypi(alpha,beta,x),log=TRUE))+
      dnorm(beta, mean=0, sd=sqrt(sigma2), log=TRUE)
    # log acceptance rate
    logacc <- enum - denom

    if(log(runif(1)) <= logacc){
      # accept the proposed value
      beta <- beta_star
      beta_yes <- beta_yes + 1;beta_yes_history[count] <- 1
    }

    # after the burnin save every kth sample (depending on thinning parameter)
    if((i > 0) && (i%%n.thin == 0)){
      alpha_samples <- c(alpha_samples, alpha)
      beta_samples <- c(beta_samples, beta)
    }
  }
  # output:
  output <- list("alpha_samples"=alpha_samples,
                 "beta_samples"=beta_samples,
                 "alpha_yes"=alpha_yes, "beta_yes"=beta_yes,
                 "alpha_yes_history"= alpha_yes_history,
                 "beta_yes_history"= beta_yes_history)
return(output)
}
```



5 JAGS and CODA

5.1 JAGS (Just Another Gibbs Sampler)

R/Stan: $\mathcal{N}(\mu, sd = \sigma)$ — BUGS, JAGS, INLA: $\mathcal{N}\left(\mu, \tau = \frac{1}{\sigma^2}\right)$ wants the precision!

What are graphical models and why are they useful?

- Breaks down complex models into simple components
- Communication of the essential structure of the problem
- Squares are data (observed), nodes are random variables and edges represent the relations

```
library(rjags); data_jags <- list(y=y, x=x, n=n)
pl1_modelString <- "model{
  # likelihood
  for (i in 1:length(y)){
  # 1. Calculating likelihood for each observation
  #    given the proposed parameter
y[i] ~ dbin(p[i],n[i]);
p[i] <- ilogit(alpha + beta * x[i])

  }
  # 2. Prior for parameter alpha
  alpha ~ dnorm(0, 1.0E-04)
  beta  ~ dnorm(0, 1.0E-04)
}"; writeLines(pl1_modelString, con = "pl1_modelString.txt")
```

```
# Initialize starting points (let JAGS initialize) and set seed
inits.jags <- list(
  list(.RNG.name="base::Wichmann-Hill", .RNG.seed=314159),
  list(.RNG.name="base::Marsaglia-Multicarry", .RNG.seed=159314),
  list(.RNG.name="base::Super-Duper", .RNG.seed=413159),
  list(.RNG.name="base::Mersenne-Twister", .RNG.seed=143915))

# Compile JAGS model
N_iter <- 10000
N_thin <- 1
N_burnin <- 4000
model.jags <- jags.model(file = "pl1_modelString.txt",quiet=T,
                         data = data_jags,
                         inits = inits.jags,
                         n.chain = 4,
                         n.adapt = 4000)#somehow not needed!
# burn-in
update(model.jags, n.iter = N_burnin)
# posterior sampling
fit.model.jags <- coda.samples(model = model.jags,
                               variable.names = c("alpha","beta"),
                               n.iter = N_iter,
                               thin = N_thin)
```

```
s <- summary(fit.model.jags)
s$statistics


##               Mean         SD    Naive SE Time-series SE
## alpha   -0.9609658  0.2299865 0.001149933      0.00228776
## beta  -142.2333579 24.7146657 0.123573329      0.24368319
```

5.2 CODA (Convergence Diagnostic)

Without convergence the properties do not hold! We must verify that the MCMC is ergodic! A MCMC will eventually converge but we dont know how quickly!

We have one or several unknown target distributions (posteror) and we want to learn about:

- Marginal posterior densities (only one parameter)
- $E(\theta_i|\boldsymbol{y})$
- $\sqrt{Var(\theta_i|\boldsymbol{y})}$
- Percentiles

Two main types of convergence are relevant for MCMC sampling:

- Convergence to stationary distribution: Burn in problem (when to cut?)
- Convergence of the sample statistics to the truth: Moment based convergence means that the mean, variance, quantiles (sample statistics) are close enough to the truth. Requires the central limit theoem to be valid for a Markov Chain.

5.2.1 (Markov chain) Monte carlo Data set:

b burn-in		stochastic nodes	
		θ_i^*	θ_j^*
$t=1$	$b+1$	θ_{1i}	θ_{1j}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$t=M$	$b+M$	θ_{Mi}	θ_{Mj}

The MCMC sample size M does not depend on the thinning parameter but the time to get it will be affected. If we say that after the burn-in of b samples stationarity has been achieved then:

- Sample mean: $\boxed{\bar{\theta}_{\cdot i}^*|\boldsymbol{y} = \frac{1}{M}\sum_{t=1}^{M}\theta_{ti}^*}$

  which is an approximation of $E(\theta_i^*|\boldsymbol{y}) = \int_{\infty}^{\infty} u f_{\theta_i^*|\boldsymbol{y}}(u)du$

- Sample standard deviation: $\boxed{\text{SD}(\theta_i^*|\boldsymbol{y}) = \sqrt{\frac{1}{M-1}\sum_{t=1}^{M}(\theta_{ti}^* - \bar{\theta}_{\cdot i}^*)^2}}$

- Sample percentiles: $\hat{F}_{\theta_i^*|\boldsymbol{y}}(q) = \frac{1}{M}\sum_{t=1}^{M} I[\theta_{ti}^* \leq q] = 0.025$

  solve for q

These estimates are consistent, which means that they can be made arbitrary close to the truth as $M \to \infty$

5.3 CODA:

Questions we are trying to answer:

1. What should be the starting value?
2. How long should the burn-in period be? When does the chain reach stationarity?
3. How long do we need to monitor the chain to get results of sufficient MC-Accuracy? (< 0.001)

Question 1 and 2:

1. Several different starting values (recommendend to use 4)
2. Start somewhere near a centre of the posterior distribution (mean,mode) determined by using non-informative prior.
   $\rightarrow$ this is problematic if the posterior is multimodal!

Question 2: How do we know if a Markov Chain has converged (Potential danger)?

1. We dont know $\rightarrow$ can look for evidence that it has not converged (H_0 : "Has converged!")!
2. Type II errors (β): Diagnostic test is non-significant but it has not converged in reality!! Is severe!

Conv. to stationarity	Conv. to ergodic average
• traceplots • Heidelberger & Welch (run length control based on mean) • Gewecke (lack of convergence)	• NA • ESS (Effective Sample Size) • Raftery & Lewis (run length control based on quantiles)
• rank plots • NA	• NA • BRG (Brooks, Rubin, Gelman),R (Lack of convergence using multiple chains)

5.3.1 Graphical Stationarity Diagnostic (Trace & Rank):

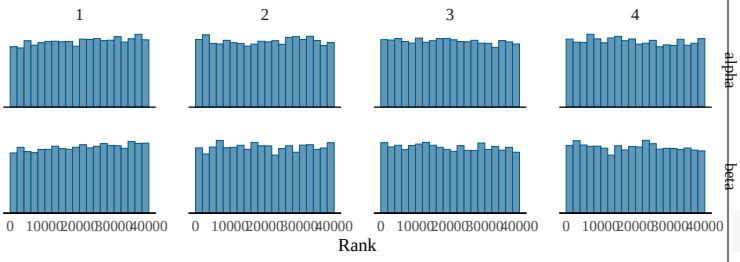
If a Markov Chain has converged to stationarity this means that it should not depend on the initial starting point. This means that different chains should be "equally" or mix well!

If we have only 1 chain we look at the traceplot which should reveal no pattern (stationary). If we have several chains we use rank plots. Here all posterior draws from all chains are mixed and then ranked and then again splitted into their original chain. Plotting of the ranks of each chaing should give a uniform plot.

Solutions:

- Tuning of acceptance rate (0.2-0.4) by adjusting proposal distribution parameters
- Thinning (keep only every kth observation)

```
library(bayesplot)
bayesplot::mcmc_rank_hist(fit.model.jags)
```



Interpretation: Uniformity for all chains indicates that they have converged (mix well).

5.3.2 Effective Sample Size and Autocorr. (ergodic average):

If the autocorrelation is large, we are learning about $P(\boldsymbol{\theta}|\boldsymbol{y})$ at a slower pace than if it was iid. The effective sample size addresses the question "Is the effective sample size large enough to get stable estimates of uncertainty?" and measures the amount of information in M samples from a Markov chain:

$$\text{ESS} = N_{eff} = \frac{M}{1 + 2\sum_{k=1}^{\infty} \text{ACF}(k)} = \frac{M}{1 + 2\sum_{k=1}^{\infty} \text{corr}(\theta_t^*, \theta_{t+k}^*)}$$

where we can simplify the lag $k = 1, ..., w$ so that $\text{ACF}(w) < 0.1 (\text{or } < 0.05)$ and $\text{ACF}(w) > \text{ACF}(w+1)$. The ESS shows how many independent draws contain the same amount of information as the dependent sample we have obtained by MCMC sampling. The higher the ESS the better and it must be large enough to get stable inferences for quantities of interest.

5.3.3 Gelman/Rubin/Brooks (ergodic average):

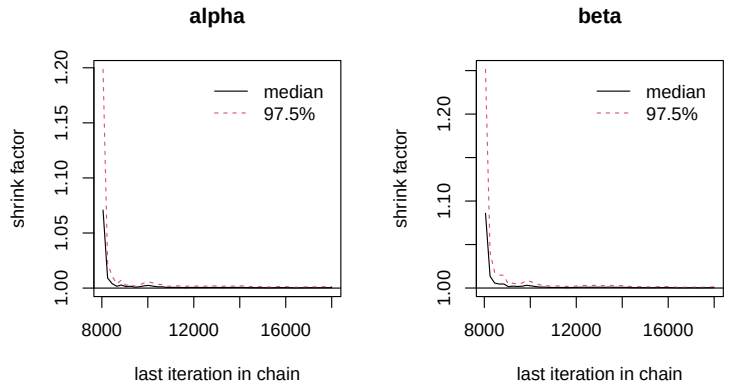
The idea is that multiple chains have not converged, then the between-chain-variability (B) is large in relation to the within-chains-variability of the pooled chain ($B + W$).

Shrink factor or potential-scale reduction factor:

$$\sqrt{\hat{R}} \approx \sqrt{\frac{W+B}{W}} \to 1, \text{ if } B \to 0$$

- Uses latter half of each chain (recommended to use 4)
- Needs widely differing starting points
- Uses normal approximation to derive $\hat{R}$
- Gives the shrink factor $\hat{R}$ and the Upper Confidence Interval
- Recommended threshold: $\hat{R} \leq 1.01$
- ESS should be high enough!

```
coda::gelman.diag(fit.model.jags)

## Potential scale reduction factors:
##
##       Point est. Upper C.I.
## alpha          1          1
## beta           1          1
##
## Multivariate psrf
##
## 1

coda::gelman.plot(fit.model.jags)
```



Interpretation: The Shrinkage factor decreases to one with increasing number of iterations. A Shrinkage factor factor of 1 ($\hat{R} < 1.01$) means that the between chain variability B is close to zero which means that the chains have converged to stationarity (mix well).

Vats and Knudson:

$$\sqrt{\hat{R}} \approx \sqrt{1 + \frac{\texttt{n.chain}}{\text{ESS}}}$$

```
library(stableGR)
stableGR::stable.GR(fit.model.jags)
```

```
## $psrf
## [1] 1.000002 1.000004
##
## $mpsrf
## [1] 1.000062
##
## $means
##        alpha         beta
##   -0.9608901 -142.2202900
##
## $n.eff
## [1] 17806.81
##
## $blather
## [1] FALSE
```

Interpretation: Extendend (Multivariate) potential scale reduction factor which should be smaller than 1.01 like in Gelman.

5.3.4 Raftery & Lewis (ergodic average):

- Each chain separately
- N_{min}: Sample size needed if independent
- M: Burn-in needed
- I: Independence factor indicates to which extent autocorrelation inflates the required sample size. $I > 5$ indicates strong autocorrelation. It is a crude estimate of the thinning interval
- N: Total sample size needed $I = \frac{M+N}{N_{min}} \rightarrow N_{min} = \frac{M+N}{I}$

```
coda::raftery.diag(fit.model.jags)[[1]]#for all chains!


##
## Quantile (q) = 0.025
## Accuracy (r) = +/- 0.005
## Probability (s) = 0.95
##
##        Burn-in  Total Lower bound  Dependence
##        (M)      (N)   (Nmin)       factor (I)
##  alpha 12       12590 3746         3.36
##  beta  7        8197  3746         2.19
```

Interpretation: The dependence factor is < 5 and thus no strong autocorrelation exists but it suggests a thinning of 4.

5.3.5 Heidelberger & Welch (Conv. to stationarity)

Test the null hypothesis that the sampled values come from a stationary distribution. The test is successively applied, firstly to the whole chain, then after discarding the first 10%, 20%, ... of the chain until either the null hypothesis is accepted, or 50% of the chain has been discarded. The latter outcome constitutes 'failure' of the stationarity test and indicates that a longer MCMC run is needed. If the stationarity test is passed, the number of iterations to keep and the number to discard are reported.

The half-width test calculates a 95% confidence interval for the mean, using the portion of the chain which passed the stationarity test. Half the width of this interval is compared with the estimate of the mean. If the ratio between the half-width and the mean is lower than eps, the halfwidth test is passed. Otherwise the length of the sample is deemed not long enough to estimate the mean with sufficient accuracy.

```
coda::heidel.diag(fit.model.jags,eps=0.1, pvalue=0.05)[[1]]


##
##       Stationarity start     p-value
##       test         iteration
## alpha passed       1         0.877
## beta  passed       1         0.926
##
##       Halfwidth Mean     Halfwidth
##       test
## alpha passed      -0.953 0.00896
## beta  passed    -141.317 0.95695
```

Interpretation: Stationarity test is passed since p-values are larger than 0.05. Also sufficient samples are drawn for enough MC-Accuracy.

5.3.6 Geweke (Conv. to stationarity)

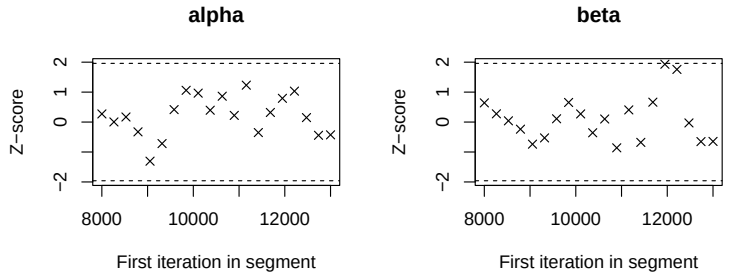
Detects cases when stationarity has not been reached. Tests equality of means between early and late sections of the chain (essentially an independent sample t-test)

- A: Early (10%) of the chain (unclear)
- B: Next (50%) of the chain (unclear)
- $|Z| > 2$?

$$Z = \frac{\bar{\theta^*}^A - \bar{\theta^*}^B}{\sqrt{V(\theta^{*A}) + V(\theta^{*A})}}$$

The first half of the Markov chain is divided into nbins - 1 segments, then Geweke's Z-score is repeatedly calculated. The first Z-score is calculated with all iterations in the chain, the second after discarding the first segment, the third after discarding the first two segments, and so on. The last Z-score is calculated using only the samples in the second half of the chain.

```
coda::geweke.diag(fit.model.jags)[[1]]#for all chains!


##
## Fraction in 1st window = 0.1
## Fraction in 2nd window = 0.5
##
##  alpha   beta
## 0.2719 0.6407


coda::geweke.plot(fit.model.jags[[1]],nbins = 20, ask =F)
```



Interpretation: Since the crosses lay within the range [-2,2] there is not enough evidence to reject the Null-hypothesis saying that the sections of a chain are different (no stationarity not rejected!)

6 Bayesian Meta-Analysis

$$f(\theta|y) \propto \int f(y|\theta)f(\theta|\psi)f(\psi)d\psi$$

$$f(\theta|y) = \int f(\theta|y,\psi)f(\psi|y)d\psi \propto \int f(\theta|y,\psi)f(y|\psi)f(\psi)d\psi$$

The Bayesian NNHM (normal-normal hierarchical model) is useful for synthesis of evidence from several studies. Similarity of those studies is given by the between study standard deviation τ:

- $\tau = 0$ (pooling) means that all studies are independent and identical realizations for the same underlying process and there is no additional heterogeneity between the studies at all
- $\tau = \infty$ independent and unrelated studies

Bayesian normal-normal hierarchical model (NNHM):

$$f(y_1, ..., y_k) = \int \prod_{i=1}^{k} f(y_i|\theta)f(\theta)d\theta$$

6.0.1 Full Bayes

The full bayesian meta-analysis provides inference on random effects $\theta_1, ..., \theta_k$ that lays between two models: one assuming they are fully indepdendent $\tau = \infty$ and that they are from the same distribution (pooled) $\tau = 0$.

- Likelihood (within study sd): $y_i \sim N(\theta_i, \sigma_i^2)$
- Random effect (between study sd): $\theta_i \sim N(\mu, \tau^2)$
- Priors: $\mu \sim N(\nu = 0, \gamma^2 = (4)^2)$ and $\tau \sim |N(0, A^2 = (0.5)^2)|$
- Within study standard deviation of the i-th study σ_i is assumed to be known!

θ_i is the trial specific mean which is measured with an uncertainty σ_i^2 (y_i and σ_i^2 are reported). The θ_i again differ from trial to trial and are distributed around a common mean μ with standard deviation τ.

Bayes theorem:

$$\underbrace{f(\mu, \tau, \boldsymbol{\theta}|(y_1, \sigma_1), ..., (y_k, \sigma_k))}_{\text{Posterior}} \propto \underbrace{f((y_1, \sigma_1), ..., (y_k, \sigma_k)|\boldsymbol{\theta})}_{\text{Likelihood}} \underbrace{f(\boldsymbol{\theta}|\mu, \tau)}_{\text{Random effect}} \underbrace{f(\mu)f(\tau)}_{Priors}$$

To get the exact posterior we need to divide it by the normalization constant which is the same but we havee to integrate out $\mu, \tau, \boldsymbol{\theta}$.

Maybe the Normal example here!

6.0.2 Pooled $\tau = 0$: (=fixed effect)

- $\theta_1, ..., \theta_k = \theta$
- Likelihood (within study sd): $y_i \sim N(\theta, \sigma_i^2)$
- Priors: $\theta \sim N(\nu = 0, \gamma^2 = (4)^2)$

This would then lead to the posterior distribution of:

$$\theta|y_1, ..., y_n \sim N\left(\frac{\sum_{i=1}^k \frac{y_i}{\sigma_i^2} + \frac{\nu}{\gamma^2}}{\sum_{i=1}^k \frac{1}{\sigma_i^2} + \frac{1}{\gamma^2}}, \left(\sum_{i=1}^k \frac{1}{\sigma_i^2} + \frac{1}{\gamma^2}\right)^{-1}\right)$$

NOTE: If the prior of μ is very flat $\gamma \to \infty$ we get the classical overall mean estimate: an average of individual estimates each weighted by its precision.

Uniform prior:

- Likelihood (within study sd): $y_i \sim N(\theta, \sigma_i^2)$
- Priors: $\theta \sim U$
- more heterogeneous estimates than full bayes

This would then lead to the posterior distribution of: $f(\theta_i|y_i) \propto f(y_i|\theta_i) \to \theta_i|y_i \sim N(y_i, \sigma_i^2)$

6.0.3 Independent $\tau = \infty$:

Means that knowledge gained from one study is irrelevant for the other one. Thus, pooling does not make sense!

6.1 Empirical Bayes

Results are closer to marginal posterior estimates of random effects provided by the full Bayesian meta-analysis. EB lies within full Bayesian and classical methods. Can be dangerous if the data generation process is not known.

Empirical Bayes:

$$\boxed{f(\theta|y) \approx f(\theta|y, \hat{\psi})}$$

- Avoids computation of the integral
- Avoids "choosing" a distribution for the hyperprior $f(\hat{\psi})$ because it is given by the data
- Uses the data y twice:
  - To select the hyperparameters of the prior $\hat{\psi}$
  - To compute the posterior according to the Bayes formula $f(\theta|y, \hat{\psi})$
- Has also good frequentist properties
- Likelihood (within study sd): $y_i \sim N(\theta_i, \sigma_i^2)$
- Random effect (between study sd): $\theta_i \sim N(\mu, \tau^2)$
- Posterior for the i-th random effect: $\boxed{\theta_i|y_i \sim N\left(\frac{\frac{y_i}{\sigma_i^2} + \frac{\mu}{\tau^2}}{\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}}, \left(\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}\right)^{-1}\right)}$

This means the Confidence Intervals are a bit narrower than assuming complete independence ($\tau = \infty$) because the precision is slightly larger. The location of each random effect $\theta_i|y_i$ is shrunk towards the prior mean μ:

$$\frac{\frac{y_i}{\sigma_i^2} + \frac{\mu}{\tau^2}}{\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}} = \frac{\tau^2}{\sigma_i^2 + \tau^2} y_i + \left(1 - \frac{\tau^2}{\sigma_i^2 + \tau^2}\right)\mu = \frac{\tau^2}{\sigma_i^2 + \tau^2} y_i + \underbrace{\frac{\sigma_i^2}{\sigma_i^2 + \tau^2}}_{\text{degree of shrinkage}} \mu$$

6.1.1 Parameter estimation $\hat{\mu}, \hat{\tau}$

$\hat{\mu}, \hat{\tau}$ are needed for the posterior and thus are estimated from the data by maximizing the prior predictive for the observed data (maximizing the marginal likelihood).

Assuming no data has been collected so far, the prior predictive distribution for one further observation is:

$$f(y_i|m) \sim N(m, \sigma_i^2) \text{ and } f(m) \sim N(\mu, \tau^2)$$

$$f(y_i) = \int_\infty^\infty f(y_i|m) f(m) dm \to y_i \sim N(\mu, \underbrace{\sigma_i^2 + \tau^2}_{\text{precision } w_i})$$

where $\frac{1}{w_i} = \sigma_i^2 + \tau^2$ and σ_i^2 is known from the individual studies. For all the k observations this leads then to a marginal log-likelihood:

$$l(\mu, \tau) = \sum_{i=1}^k \log(f(y_i)) = -\frac{1}{2}\left(\sum_{i=1}^k w_i(y_i - \mu) - \sum_{i=1}^k \log(w_i)\right)$$

$$= \sum_{i=1}^k \log(f(y_i)) = -\frac{1}{2}\left(\sum_{i=1}^k \frac{1}{\sigma_i^2 + \tau^2}(y_i - \mu) - \sum_{i=1}^k \log\left(\frac{1}{\sigma_i^2 + \tau^2}\right)\right)$$

this can then be optimized to get $\hat{\mu}, \hat{\tau}$ based on data $(y_1, \sigma_1^2), ..., (y_k, \sigma_k^2)$.

6.2 Method of moments:

Simple method where the moments (mean $E(X)$ and variance $V(X)$) of the data is matched to fit a distribution. Example: $Y|\lambda \sim P(\lambda)$ with $\lambda \sim G(\alpha, \beta)$

$$\boxed{E(Y) = E_\lambda[E_Y(Y|\lambda)]} \qquad \boxed{V(Y) = V_\lambda(E_Y(Y|\lambda)) + E_\lambda[V_Y(Y|\lambda)]}$$

$$E(Y) = E_\lambda[\lambda] = \frac{\alpha}{\beta}$$

$$V(Y) = V_\lambda(\lambda) + E_\lambda[\lambda] = \frac{\alpha}{\beta^2} + \frac{\alpha}{\beta} = \frac{\alpha(1+\beta)}{\beta^2}$$

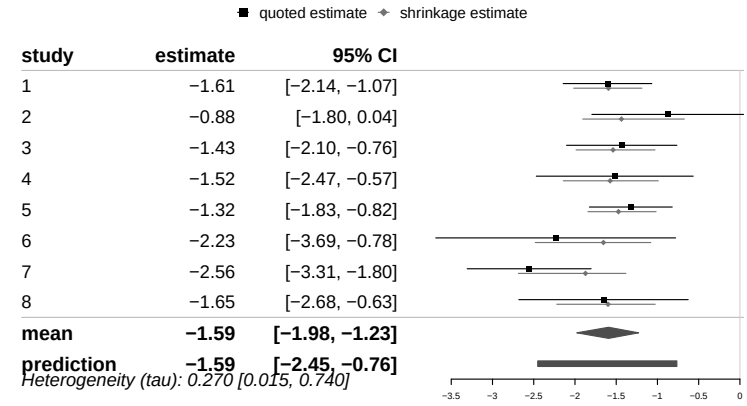
$$\beta = E(Y)/\alpha$$

$$\alpha = \frac{V(Y)\beta^2}{(1+\beta)V(Y)} = \frac{V(Y)(E(Y)/\alpha)^2}{(1+E(Y)/\alpha)V(Y)}$$

6.3 `bayesmeta`-Full Bayes

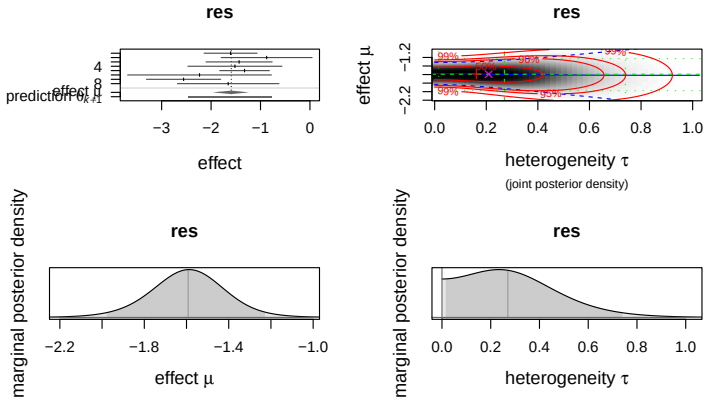
```
library(bayesmeta)
res <- bayesmeta(y = df[ , "y"],
                 sigma = df[ , "sigma"],
                 labels = df[ , "labels"],
                 mu.prior.mean = 0,
                 mu.prior.sd = 4,
                 tau.prior = function(t) {dhalfnormal(t, scale = 0.5)},
                 interval.type = "central")
```

```
forestplot(res)
```



Credible interval of the heterogeneity measure τ does not include 0 and thus assuming fixed effect (pooled) is not appropriate. Shows the original data (black) and the shrunken estimates (grey). The mean is μ and prediction is θ.

```
par(mfrow = c(2,2))
plot(res)
```



Shows joint posterior density and the marginal posterior densities.

6.4 JAGS Meta-Analysis (Mixed model)

This is a mixed model! Likelihoods:

$$y_j \sim \text{Bin}(n_j, p_j)$$
$$\text{logit}(p_j) = \mu + \beta \cdot C1_j + \eta_j$$
$$\eta_j \sim N(0, V = 1/\tau_{prec})$$

for $j = 1, ..., N$, where $\tau_{\text{prec}} = 1/\tau^2$

Priors:

$$\mu \sim U(-10, 10) \qquad \beta \sim U(-10, 10) \qquad \tau \sim U(0, 10)$$

Here, we use a $U(0, 10)$ distribution as a prior for heterogeneity, accepting that the means of included studies may not be identical. Moreover, a uniform distribution prior indicates that we have no knowledge regarding the differences between the studies. It is a weakly informative prior because it accepts all possible scenarios of heterogeneity with equal probabilities.

```
pl1.data<-list(N = 16,
y = c(23,12,19,9,39,6,9,10,120,18,107,26,82,16,126,23),
n = c(107,44,51,39,139,20,78,35,208,38,150,45,138,20,201,34),
C1 = c(0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1))
pl1_modelString <- "model{
# sampling model (likelihood)
for (j in 1:N) {
  y[j] ~ dbin(p[j],n[j])
  logit(p[j]) <- mu + beta*C1[j] + eta[j]
  eta[j] ~ dnorm(0, tau.prec)

  # prediction for posterior predictive checks
  y.pred[j] ~ dbin(p[j],n[j])
  PPC[j] <- step(y[j]-y.pred[j])-0.5*equals(y[j],y.pred[j])
}
# priors
mu ~ dunif(-10,10)
beta ~ dunif(-10,10)
tau ~ dunif(0,10)# prior for heterogeneity!
tau.prec <- 1/tau/tau

# population effect
p1 <- 1/(1+exp(-mu))
p2 <- 1/(1+exp(-mu-beta))

# predictive distribution for new study effect
eta.star ~ dnorm(0,tau.prec)
p1.star <- 1/(1+exp(-mu-eta.star))# placebo (as prior for next!)
p2.star <- 1/(1+exp(-mu-beta-eta.star))# treatment (as prior for next!)
}"; writeLines(pl1_modelString, con="TempModel.txt")
```

```
# model initiation
rjags.pl1 <- jags.model(file = "TempModel.txt",quiet=T,
data = pl1.data, n.chains = 4, n.adapt = 4000,inits = inits.jags)
# burn-in
update(rjags.pl1, n.iter = N_burnin)
# posterior sampling
fit.rjags.pl1 <- coda.samples(model = rjags.pl1,
variable.names=c("mu", "beta", "tau", "p1.star", "p2.star"),
n.iter = N_iter,
thin = 1)
summary(fit.rjags.pl1)


##
## Iterations = 8001:18000
## Thinning interval = 1
## Number of chains = 4
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##            Mean      SD  Naive SE Time-series SE
## beta     1.6152 0.21637 0.0010819      0.0054951
## mu      -1.1101 0.15966 0.0007983      0.0037626
## p1.star  0.2537 0.06762 0.0003381      0.0007762
## p2.star  0.6204 0.07973 0.0003986      0.0006798
## tau      0.2924 0.13616 0.0006808      0.0036490
##
## 2. Quantiles for each variable:
##
##             2.5%     25%     50%     75%   97.5%
## beta     1.17763  1.4812  1.6151  1.7523  2.0447
## mu      -1.42748 -1.2127 -1.1110 -1.0099 -0.7904
## p1.star  0.13545  0.2123  0.2479  0.2869  0.4088
## p2.star  0.44335  0.5784  0.6235  0.6675  0.7741
## tau      0.04152  0.2012  0.2826  0.3709  0.5899
```

7 Diverse:

7.1 Change of variable:

$$\boxed{f_Y(y) = f_X(g^{-1}(y))|\frac{dg^{-1}(y)}{dy}|}$$

$X = \frac{1}{\sigma^2} \sim G(a, b) \rightarrow f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx)$

$y = \frac{1}{x} = g(x)$, then $x = \frac{1}{y} = g^{-1}(y) \rightarrow \frac{dg^{-1}(y)}{dy} = -\frac{1}{y^2}$

$$f_Y(y) = \frac{b^a}{\Gamma(a)} \left(\frac{1}{y}\right)^{a-1} \exp\left(-b\frac{1}{y}\right)|-\frac{1}{y^2}|$$
$$= \frac{b^a}{\Gamma(a)} y^{-a-1} \exp\left(-\frac{b}{y}\right)$$

7.2 Delta rule (logit)

Apply the logit function:

$$\phi = h(\pi) = \text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$$

Invariance of the MLE gives

$$\hat{\phi}_{\text{ML}} = \text{logit}(\hat{\pi}_{ML}) = \log\left(\frac{\hat{\pi}_{ML}}{1-\hat{\pi}_{ML}}\right) = \log\left(\frac{x}{n-x}\right)$$

The standard error of $\hat{\phi}_{\text{ML}}$ can be computed with the delta method:

$$\text{se}(\hat{\phi}_{\text{ML}}) = \text{se}(\hat{\pi}_{\text{ML}}) \cdot |\frac{dh(\hat{\pi}_{\text{ML}})}{d\pi}|$$

where

$$\frac{dh(\hat{\pi}_{\text{ML}})}{d\pi} = \frac{1-\pi}{\pi} \cdot \frac{1-\pi+\pi}{(1-\pi)^2} = \frac{1}{\pi(1-\pi)}$$

$$\text{se}(\hat{\phi}_{\text{ML}}) = \sqrt{\frac{\hat{\pi}_{\text{ML}}(1-\hat{\pi}_{\text{ML}})}{n}} \cdot \frac{1}{\hat{\pi}_{\text{ML}}(1-\hat{\pi}_{\text{ML}})}$$
$$= \frac{1}{\sqrt{n \cdot \hat{\pi}_{\text{ML}}(1-\hat{\pi}_{\text{ML}})}} = \sqrt{\frac{n}{x(n-x)}}$$
$$= \sqrt{\frac{1}{x} + \frac{1}{n-x}}$$

Log odds ratio:

$$y = \log(\text{OR}) = \log\frac{x_{\text{P}}}{n_{\text{P}} - x_{\text{P}}} - \log\frac{x_{\text{T}}}{n_{\text{T}} - x_{\text{T}}}$$

$$\sigma = \text{SE}\left(\log(\text{OR})\right) = \sqrt{\frac{1}{x_{\text{P}}} + \frac{1}{n_{\text{P}} - x_{\text{P}}} + \frac{1}{x_{\text{T}}} + \frac{1}{n_{\text{T}} - x_{\text{T}}}}$$

7.3 Moment matching:

7.4 Stuff:

$$p(\theta) \propto \exp(-\frac{1}{2}(a\theta^2 - 2b\theta))$$
$$\propto \exp(-\frac{1}{2}a(\theta^2 - 2b\theta/a))$$
$$\propto \exp(-\frac{1}{2}a((\theta - 2b/a)^2 - (b/a)^2)$$
$$\propto \exp(-\frac{1}{2}a((\theta - 2b/a)^2) \cdot \exp(\frac{b^2}{2a})$$
$$\propto \exp(-\frac{1}{2}a((\theta - 2b/a)^2)$$
$$\theta \sim N(b/a, 1/a)$$

7.5 Gibbs Sampler:

Likelihood:

$$f(y_{1:n}|\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mu)^2\right)$$
$$= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right)$$

Prior (informative and independent):

$$f(\mu, \sigma^2) = f(\mu)f(\sigma^2)$$
$$= \left(\frac{1}{2\pi\sigma_0^2}\right)^{1/2} \exp\left(-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right) \cdot$$