Data exclusion based SBR:NMR ratio

ZW and LA

Two step approach:

1. Fit a model to the high quality data to estimate range of true ratios.
    - Important: we need some kind of model set-up to capture "true" variability across ratios (ie we don't want an overall mean across settings because some settings truly have lower SBRs than others)
    - Candidate set-up:
      For observed ratio r_i, assume
      $$\log(r_i) = \theta_i + \varepsilon_i,$$
      with
        - Random error $\varepsilon_i \sim N(0, v_i)$ with variance v_i calculated as per earlier approach(model details in next part)
        - Random effect θ_i, i.e. $\theta_i \sim N(\mu, \sigma^2)$, where σ^2 refers to variability across settings
2. For an observed ratio r_i (from full data base), check if an observed ratio is plausible or not to decide on exclusion.
    - Proposal: Calculate the probability of observing something more extreme then the observed ratio r_i under the fitted model for log-ratios from step 1:
        - Calculate $p_i = \int_{-\infty}^{\log(r_i)} \phi(r)dr$, where $\phi(r)$ is the predictive density for log(ratio) from model in step 1 using observation-specific error variance. Based on candidate model in step 1, the predictive distribution is given by6:
          $$N(\hat{\mu}, \hat{\delta}^2 + \hat{\sigma}^2 + v_i),$$
          where $\hat{\mu}$ is the point estimate for μ and $\hat{\delta}^2$ its posterior variance, and $\hat{\sigma}^2$ is the point estimate for the variance of the random effects.
        - Decision rule: if $p_i < x$, exclude observation i. We set $x = 0.05$ in calculation below.

Calculation of variance

For an observation i, we used a monte carlo approximation to calculate value v_i. We assumed

- stillbirths $\sim$ Bin(total births, observed sbr),
- neonatal deaths $\sim$ Bin(live births, observed nmr)

Generate S random samples form above distributions. For each sample

$$log(ratio_s) = log(\frac{SBR_s}{NMR_s})$$

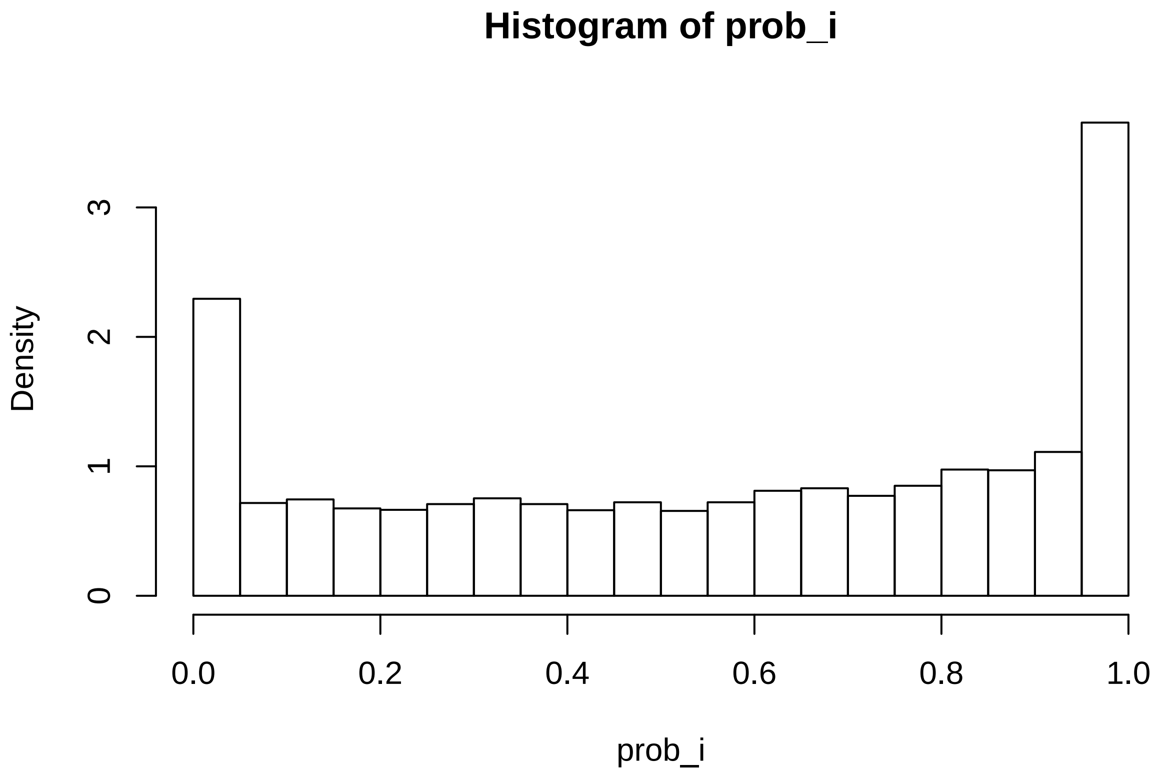
- Then get $var(log(ratio))$ of S samples $log(ratio_s), s = 1, ..., S$.

```
## Inference for Stan model: study_ratio_cutoff.
## 4 chains, each with iter=2000; warmup=1000; thin=1;
## post-warmup draws per chain=1000, total post-warmup draws=4000.
```

```
##
##         mean se_mean   sd  2.5%   25%   50%   75% 97.5% n_eff Rhat
## mu    -0.19       0 0.03 -0.25 -0.21 -0.19 -0.16 -0.12   410 1.02
## sigma  0.28       0 0.03  0.23  0.26  0.27  0.29  0.34   603 1.01
##
## Samples were drawn using NUTS(diag_e) at Tue Oct 29 17:38:14 2019.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

Histogram of step 2 p_is is displayed below. About 11.5% observation will be excluded if the cut-off x is set to 5%.

```
hist(prob_i,freq = FALSE, breaks = 20)
```



```
round(mean(prob_i<0.05,na.rm = T),digits = 3)
```

```
## [1] 0.115
```

Note that cut-off value in terms of observed SBR:NMR ratio depends on the variance of the observations v_i. The cutoff value is about 0.53 when $v_i = 0$, i.e. for observations with negligible stochastic uncertainty, and decreases as the variance increases.

```
cutoff_bound <- exp(qnorm(0.05,mu.hat,sigma))
round(cutoff_bound,digits = 2)
```

```
## [1] 0.53
```