Exact adaptive confidence intervals for small areas

Introduction

In a recent paper, Burris and Hoff (2018) detailed a procedure for constructing confidence intervals for small area (subpopulation) means with area-specific coverage. On average, these intervals will be shorter than intervals based on direct estimates alone. The purpose of this document is to highlight the importance of area-specific coverage, provide a brief introduction to the FAB procedure for constructing narrow confidence intervals with area-specific coverage, and illustrate how to use models from existing software packages to implement the procedure.

Direct Confidence Intervals and Area-Specific Coverage

Suppose we have the area-specific *sampling model*

$$y_j \sim N(\theta_j, \sigma_j^2),\ j = 1, \cdots m, \tag{1}$$

where y_j is a design-unbiased and consistent direct estimate of θ_j, the jth area mean, and σ_j^2 is the variance of the direct estimate under the sampling design. From the Central Limit Theorem, this assumption holds asymptotically as the sample size within each area increases. Even for areas with small sample sizes, this assumption is reasonable, provided that the distribution of responses within the target area is nearly normal.

For a specific area j, when the sampling σ_j^2 is assumed known, the direct $1-\alpha$ confidence interval for θ_j is

$$C_D^j(y_j) = \left\{\theta : y_j + \sigma_j z_{\alpha/2} < \theta < y_j + \sigma_j z_{1-\alpha/2}\right\}, \tag{2}$$

where z_p is the pth quantile of the standard normal distribution. This direct confidence interval procedure has the important property of *area-specific coverage* under the above sampling model, since

$$\Pr(\theta_j \in C_D^j(y_j) \mid \theta_j) = 1 - \alpha, \tag{3}$$

for all θ_j and $j \in 1, \ldots, m$. That is, regardless of the value of the target quantity θ_j, the probability that θ_j will be contained in its confidence interval is equal to $1-\alpha$ under the sampling design.

Why is area-specific coverage important?

In many applications, uncertainty intervals for group means are used to make intervention and policy decisions. For example, a government may decide to conduct an investigation if they believe it is plausible that pollution levels within a given area exceed a specified threshold. A political organization may organize a get-out-the-vote campaign in precincts where they believe increased voter turnout is likely to be beneficial to its preferred candidates. However, the end use of uncertainty intervals may be opaque to the scientist constructing them. When this is the case, it is important that uncertainty intervals have area-specific coverage, so that the study has sufficient power to detect extreme values of the target quantity, regardless of what it may be.

Consider the example of a Bayesian credible interval, which does not have $1-\alpha$ area-specific coverage. To construct a credible interval for θ_j, a prior distribution needs to be specified for θ_j (i.e, $\theta_j \sim N(\mu_j, \tau_j^2)$). It can be shown that the Bayesian credible interval has $1-\alpha$ population level coverage **on average** with respect to the prior distribution. For values of θ_j with high prior density, the credible interval procedure has greater than $1-\alpha$ coverage. For values of θ_j with low prior density, it can have far less (see the figure below). The credible interval procedure has exact $1-\alpha$ coverage for only two values of θ_j!

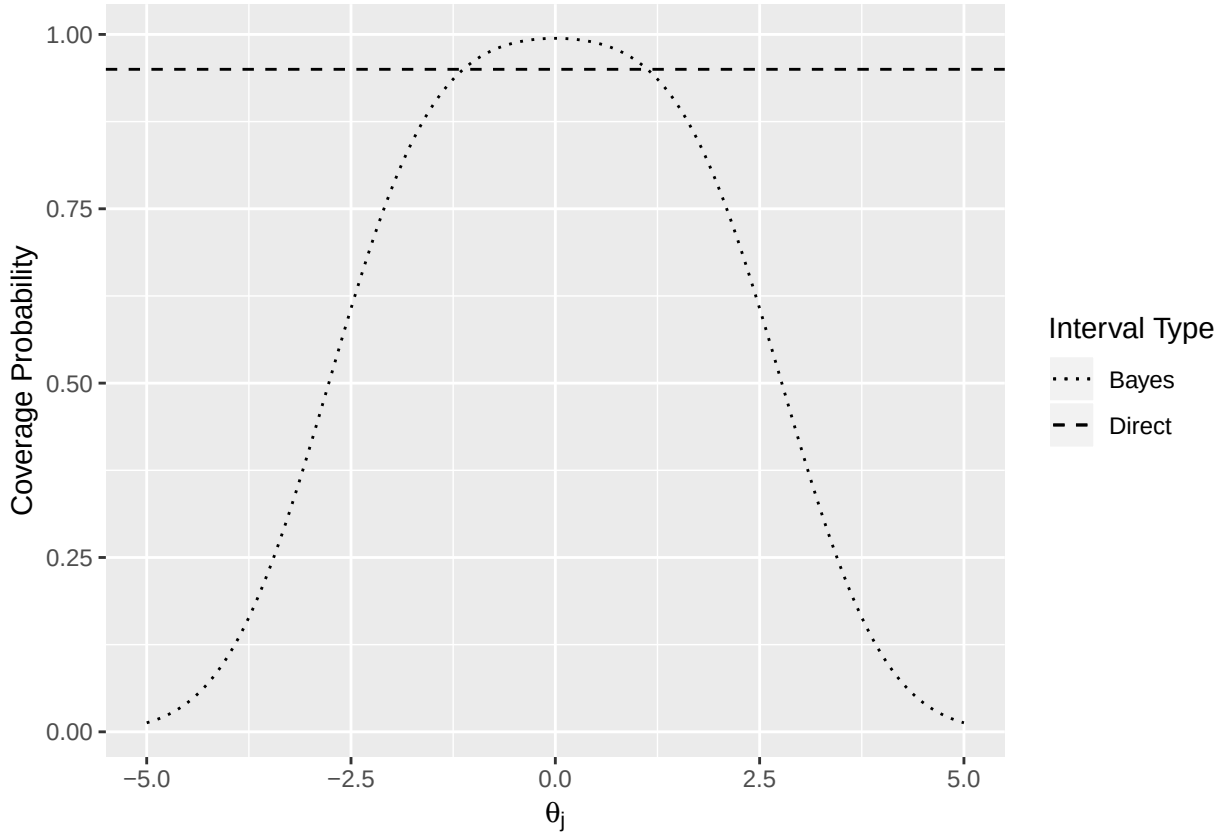


Figure 1: Coverage probability for various values of θ_j under a $N(0,1)$ prior distribution ($\sigma_j^2 = 1$).

Suppose that an intervention is required if $\theta_j \geq c$, where c is large. Then a decision procedure that intervenes when the $1-\alpha$ interval contains a value greater than c will have low sensitivity. In part for this reason, practitioners often attempt to specify a non-informative prior distribution for θ_j, since as the prior variance increases, there are more values for θ_j for which the credible interval procedure has close to $1-\alpha$ coverage.

The problem of small areas

In the context of a large-scale survey, there are often areas for which only a small number of samples are available. Although direct confidence intervals have area-specific coverage, they may be very wide. When additional precision is needed, researchers typically construct confidence intervals that borrow information from other areas and utilize auxiliary information about the area of interest. They do this by specifying a *linking model*

$$\boldsymbol{\theta} \sim F(\boldsymbol{X}, \boldsymbol{\psi}),$$

where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_m)^\top$ is a vector of area means, $\boldsymbol{X}$ is an $m \times p$ matrix representing auxiliary information about each area and $\boldsymbol{\psi}$ is a vector of model parameters. The parameters of the linking model are estimated and then a credible interval is specified for θ_j based on its posterior distribution under the sampling model and prior (linking model).

Though this approach is predominant in problems involving the analysis of small areas because it reduces interval width, it makes a substantial tradeoff: the uncertainty intervals no longer have area-specific coverage. They only have coverage on average across areas if the linking model perfectly describes the across-area heterogeneity. However, the Frequentist, Assisted by Bayes (FAB) interval procedure makes it possible to have the best of both worlds: area-specific coverage and information sharing across areas.

FAB Intervals

The direct confidence interval given above is a special case of a more general class of confidence intervals, each with $1-\alpha$ area-specific coverage. In particular, for any non-decreasing *spending function* s_j mapping $\mathbb{R}$ to the unit interval $[0, 1]$,

$$C^j_{s_j} = \left\{\theta : y_j + \sigma_j z_{\alpha(1-s_j(\theta))} < \theta < y_j + \sigma_j z_{1-\alpha s_j(\theta)}\right\} \tag{4}$$

is a valid $1-\alpha$ frequentist confidence interval. This satisfies the area-specific coverage property so long as s_j is chosen independently of y_j. The standard direct interval corresponds to $s_j(\theta) = 1/2$.

Now suppose that, based on data from other areas and a linking model, we believe θ_j is likely to be near some value μ_j. We encode this belief with a normal probability distribution $\theta_j \sim N(\mu_j, \tau_j^2)$. Given such information, we may prefer an area-specific interval procedure that, relative to the direct interval, is more precise (has shorter expected width) for values of θ_j near μ_j, at the expense of having longer expected width for values of θ_j far away from μ_j. Pratt (1963) showed how to choose s_j that minimizes the expected confidence interval width with respect to a normal prior distribution. Yu and Hoff (2016) extended this to the case where the sampling variance σ_j^2 is unknown, with prior beliefs about σ_j^2 encoded by an inverse-gamma probability distribution.

Such a confidence interval procedure has wide applications to problems involving the analysis of small areas. In particular, for a given area j, we can use data from other areas and a linking model to obtain a normal prior distribution for θ_j (and an inverse-gamma prior distribution for σ_j^2 if it is unknown). We can then find the optimal spending function and construct the corresponding confidence interval, which will have area-specific coverage and be shorter on average than the direct interval with respect to the linking model. In sum, constructing a FAB interval for the jth area mean is a five-step process:

- Specify a linking model $\boldsymbol{\theta} \sim F(\boldsymbol{X}, \boldsymbol{\psi})$. If σ_j^2 is unknown, specify a linking model $(\sigma_1^2, \cdots, \sigma_m^2) \sim G(\boldsymbol{X}, \boldsymbol{\omega})$.

- Estimate $\boldsymbol{\psi}$ and $\boldsymbol{\omega}$ using the data $\{y_1, \cdots, y_{j-1}, y_{j+1}, \cdots, y_m\}$ and $\{\widehat{\sigma}_1^2, \cdots, \widehat{\sigma}_{j-1}^2, \widehat{\sigma}_{j+1}^2, \cdots, \widehat{\sigma}_m^2\}$.
- Using estimates $\widehat{\boldsymbol{\psi}}$ and $\widehat{\boldsymbol{\omega}}$, obtain a normal prior distribution for θ_j and an inverse-gamma prior distribution for σ_j^2 if it is unknown via the method of moments.
- Obtain the optimal s_j function based on the prior distribution.
- Calculate the corresponding confidence interval for area j.

The first two steps can utilize existing software packages to estimate linking models commonly used in small area estimation. The last two steps can be done with help from the `fabCI` R package.

Examples

Dyestuff

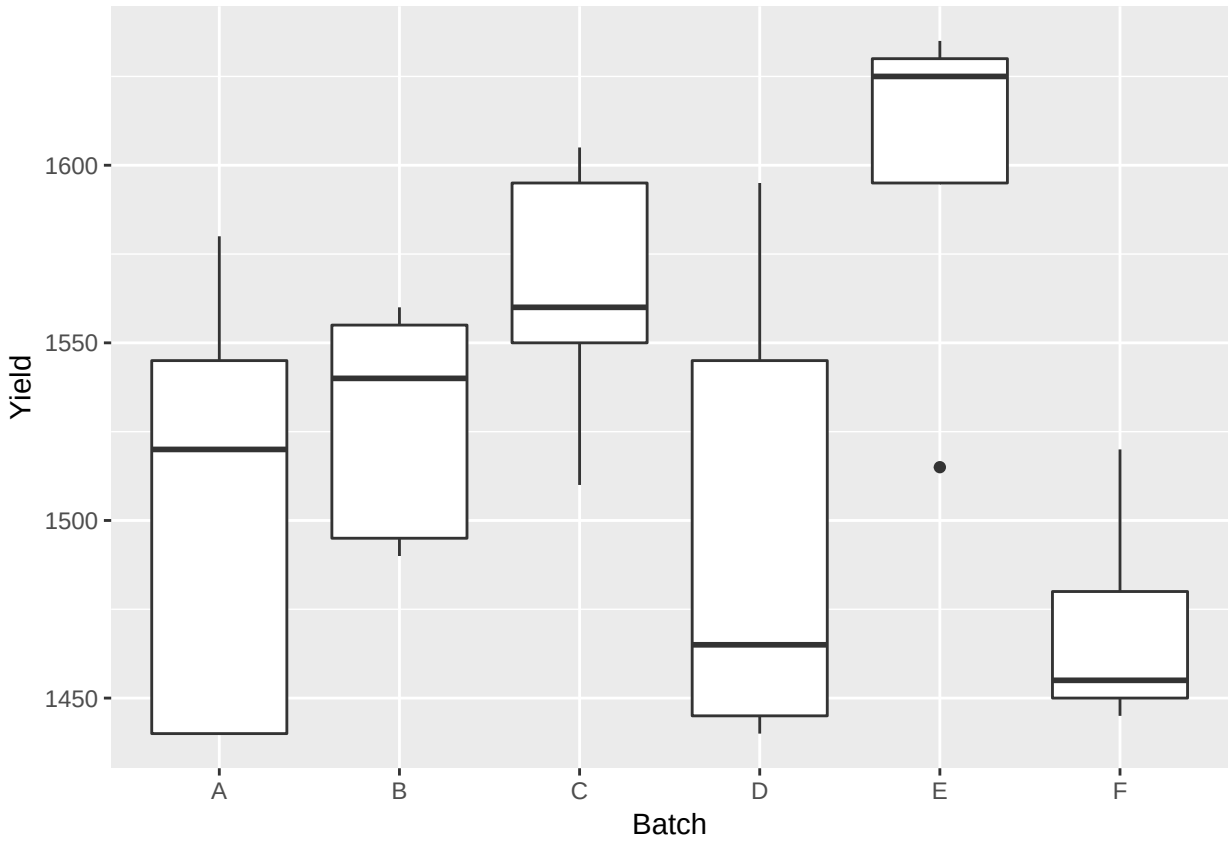
The Dyestuff data are described in Davies and Goldsmith (1972) as coming from "an investigation to find out how much the variation from batch to batch in the quality of an intermediate product (H-acid) contributes to the variation in the yield of the dyestuff (Naphthalene Black 12B) made from it. In the experiment six samples of the intermediate were obtained, and five preparations of the dyestuff were made in the laboratory from each sample. The equivalent yield of each preparation as grams of standard colour was determined by dye-trial." This dataset is made available in the `lme4` R package (Bates et al. 2015).

```
library(lme4)
library(fabCI)
library(ggplot2)

data(Dyestuff)
str(Dyestuff)
```

```
## 'data.frame':    30 obs. of  2 variables:
##  $ Batch: Factor w/ 6 levels "A","B","C","D",..: 1 1 1 1 1 2 2 2 2 2 ...
##  $ Yield: num  1545 1440 1440 1520 1580 ...
```

```
ggplot(Dyestuff, aes(x = Batch, y = Yield)) +
  geom_boxplot()
```



Suppose that we would like to construct a 95% confidence interval for the average yield of each batch. Because the within-batch variance of dyestuff yields is unknown, we could use the standard t-interval to easily obtain direct confidence intervals.

```
library(dplyr)
alpha <- 0.05

direct_intervals <- Dyestuff %>%
  group_by(Batch) %>%
  summarise(mean = mean(Yield),
            lower = mean(Yield) - qt(1- alpha/2, (n() - 1))*sd(Yield)/sqrt(n()),
            upper = mean(Yield) + qt(1- alpha/2, (n() - 1))*sd(Yield)/sqrt(n())) %>%
  mutate(length = round(upper - lower, 0))

direct_intervals
```

```
## # A tibble: 6 x 5
##   Batch  mean lower upper length
##   <fct> <dbl> <dbl> <dbl>  <dbl>
## 1 A      1505 1427. 1583.    157
## 2 B      1528 1487. 1569.     83
## 3 C      1564 1517. 1611.     94
## 4 D      1498 1413. 1583.    171
## 5 E      1600 1538. 1662.    124
## 6 F      1470 1431. 1509.     77
```

However, these confidence intervals are very wide, since they are based on data from only five samples and do not borrow information across batches. Now let's calculate a FAB interval for each batch mean θ_j,

$j = 1, \cdots, 6$. The variance of the yields for each area is unknown, so we'll need a linking model for both the means and the variances. We choose the simple model

$$\theta_j \sim N(\mu, \tau^2) \qquad 1/\sigma_j^2 \sim IG(\nu_0/2, \nu_0 s_0^2/2)$$

In general, $\{\mu, \tau^2, \nu_0, s_0^2\}$ could be jointly estimated via maximum likelihood or via the EM algorithm. However, it is not necessary to estimate the linking model precisely, since it is only being used to construct a prior distribution. Even if the parameters of the linking model are not precisely estimated, the corresponding FAB confidence interval will still have area-specific coverage.

As such, we illustrate the use of the `lme4` R package to obtain a prior distributions for θ_j and σ_j^2 for each batch $j = 1, \cdots, 6$. We then use the `fabtCI` function to calculate the corresponding FAB t-interval.

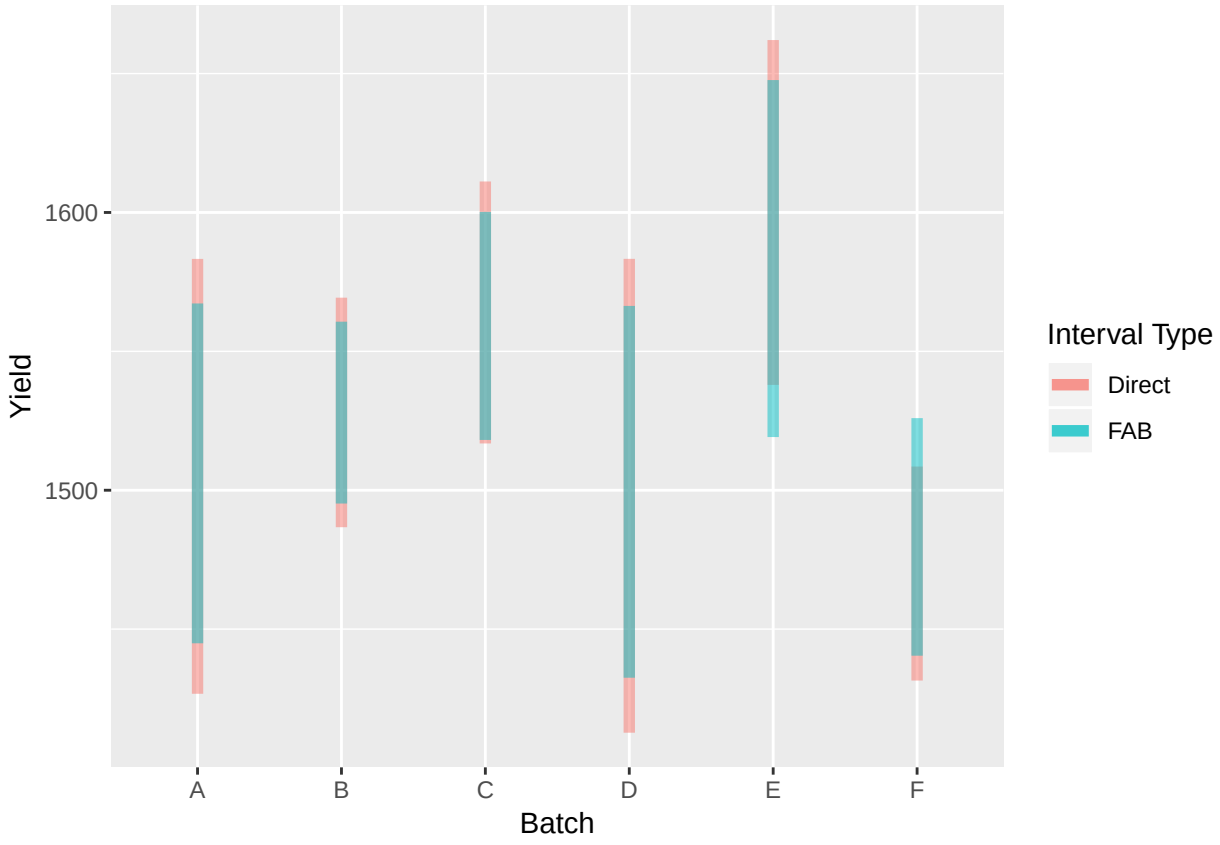
```
batches <- unique(Dyestuff$Batch)
# Construct a FAB Interval for each batch
fab_intervals <- as.data.frame(t(sapply(batches, function(batch) {
  # Only consider data not from the area of interest when estimating linking model
  Dyestuff_j <- Dyestuff %>%
    filter(Batch != batch) %>%
    droplevels()
  # Estimate linking model
  linking_model <- lmer(Yield ~ (1 | Batch), data = Dyestuff_j)
  # Construct prior distributions for area means
  mu <- fixef(linking_model)
  tau2 <- as.data.frame(VarCorr(linking_model))[1, "vcov"]
  # Construct prior distributions for area variances (method of moments)
  s20 <- sigma(linking_model) ^ 2
  nu0 <- length(batches) - 1
  # Calculate FAB Intervals
  y <- Dyestuff %>%
    filter(Batch == batch) %>%
    pull(Yield)
   fabtCI(y, psi = c(mu, tau2, s20, nu0), alpha = 0.05)
})))
colnames(fab_intervals) <- c("lower", "upper")
fab_intervals <- fab_intervals %>%
  mutate(Batch = batches, length = upper - lower) %>%
  select(Batch, lower, upper, length)
fab_intervals
```

```
##   Batch    lower    upper    length
## 1     A 1444.889 1567.245 122.35602
## 2     B 1495.206 1560.674  65.46843
## 3     C 1518.107 1600.211  82.10408
## 4     D 1432.498 1566.361 133.86246
## 5     E 1519.142 1647.670 128.52846
## 6     F 1440.421 1525.955  85.53391
```

The FAB intervals are substantially narrower for batches A-D, since the sample means for those areas are closer to the overall mean. The FAB intervals are slighly wider for batches E and F with the tail of the intervals extending towards the overall mean.

```
ggplot() +
  geom_segment(data = direct_intervals, aes(x = Batch, xend = Batch, y = lower, yend = upper,
                                            col = "Direct"), alpha = 0.5, lwd = 2) +
    geom_segment(data = fab_intervals, aes(x = Batch, xend = Batch, y = lower, yend = upper,
```

```
                                          col = "FAB"), alpha = 0.5, lwd = 2) +
  scale_color_discrete(labels = c("Direct", "FAB"), name = "Interval Type") +
  ylab("Yield")
```



Grapes

The `grapes` dataset is a synthetic dataset of grape production for 274 municipalities in Tuscany, Italy and made available in the `sae` R package (Molina and Marhuenda 2015). For each municipality, we have the following four variables.

- `grapehect`: direct estimators of the mean agrarian surface area used for production of grape (in hectares) for each Tuscany municipality.
- `area`: agrarian surface area used for production (in hectares).
- `workdays`: average number of working days in the reference year (2000).
- `var`: sampling variance of the direct estimators for each Tuscany municipality (assumed known).

We also have `grapeprox`, a proximity matrix based on the first-order neighbors of the municipalities. For this dataset, we specify a spatial Fay-Herriot model, which is discussed in more detail by Burris and Hoff (2018). Below, we first calculate 95% direct confidence intervals and then FAB confidence intervals for each area.

```
library(sae)
data(grapes)
data(grapesprox)

alpha <- 0.05

# Direct Intervals
```

```r
direct_intervals <- grapes %>%
  mutate(lower = grapehect + qnorm(alpha / 2) * sqrt(var),
         upper = grapehect + qnorm(1 - alpha / 2) * sqrt(var),
         length = upper - lower) %>%
  dplyr::select(lower, upper, length)

# FAB Intervals
fab_intervals <- as.data.frame(t(sapply(1:nrow(grapes), function(area) {
  # Only consider data not from the area of interest when estimating linking model
  grapes_j <- grapes[-area, ]
  grapes_prox_j <- grapesprox[-area, -area]
  # Estimate linking model
  linking_model <- eblupSFH(grapehect ~ area + workdays - 1, var, grapes_prox_j,
 method= "ML", data = grapes_j)
  X <- model.matrix(grapehect ~ area + workdays - 1, data = grapes)
  # Construct prior distributions for area means
  mod_tau <- linking_model$fit$refvar
  mod_rho <- linking_model$fit$spatialcorr
  pred_mean <- X %*% linking_model$fit$estcoef[,"beta"]
  Impw <- as.matrix(diag(nrow(grapes)) - mod_rho * grapesprox)
  Sigma <- mod_tau * solve(Impw %*% t(Impw))
  Sigma22 <- solve(Sigma[-area, -area])
  mu <- pred_mean[area, ] + Sigma[area, -area] %*%
    Sigma22 %*% (linking_model$eblup - pred_mean[-area, ])
  tau2 <- Sigma[area, area] - Sigma[area, -area] %*% Sigma22 %*% Sigma[-area, area]
  # Calculate FAB Intervals
  y <- grapes$grapehect[area]
  s2 <- grapes$var[area]
  fabzCI(y, mu, tau2, s2, alpha = alpha)
})))
colnames(fab_intervals) <- c("lower", "upper")
fab_intervals <- fab_intervals %>%
  mutate(length = upper - lower) %>%
  dplyr::select(lower, upper, length)
```

Comparing the means of the direct interval procedure and the FAB interval procedure, we see that the average length of the FAB interval is much narrower than the average direct interval.

```r
mean(direct_intervals$length)
```

```
## [1] 112.9369
```

```r
mean(fab_intervals$length)
```

```
## [1] 95.99183
```

FAB intervals are narrower than the corresponding direct intervals for around 92% of areas.

```r
mean(fab_intervals$length <= direct_intervals$length)
```

```
## [1] 0.919708
```

The above procedure is somewhat computationally expensive, since a separate model must be estimated for each area. An interval that achieves approximately $1 - \alpha$ area-specific coverage can be constructed by using a common model for all areas. Although this means that the spending function will not be independent of the data from the target area, the impact will be negligible as the number of areas increases.

References

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67 (1): 1–48. doi:10.18637/jss.v067.i01.

Burris, Kyle, and Peter Hoff. 2018. "Exact Adaptive Confidence Intervals for Small Areas."

Molina, Isabel, and Yolanda Marhuenda. 2015. "sae: An R Package for Small Area Estimation." *The R Journal* 7 (1): 81–98. https://journal.r-project.org/archive/2015/RJ-2015-007/RJ-2015-007.pdf.

Pratt, J. W. 1963. "Shorter Confidence Intervals for the Mean of a Normal Distribution with Known Variance." *Ann. Math. Statist.* 34 (2). The Institute of Mathematical Statistics: 574–86.

Yu, Chaoyu, and Peter D. Hoff. 2016. "Adaptive Multigroup Confidence Intervals with Constant Coverage." *Biometrika* 105 (2): 319–35.