Required sample sizes for estimating means and areal fractions of soil fertility parameters for districts in Andhra Pradesh

Dick Brus

2020-12-09

Contents

1 Introduction

The Soil Health Card (SHC) project in India involves soil sampling at a very high density every two years. For example, Andhra Pradesh state (162 975 km^2) in Cycle 2 (2017/18 - 2018/19) recorded 2 393 8875[^1] observations, a density of 14.7 km^{-1}), i.e., one per 6.8 ha. These data are used for soil fertilization recommendations at the field level. Due to the very high sampling density the current soil sampling survey is expensive: labour costs for collecting the soil samples and analyzing these soil samples in laboratories are high. The question is whether this high investment in soil survey pays. Would a reduction of the number of sampling locations also suffice? This document describes statistical methods for adapting the existing sampling campaign.

In this document we focus on estimation of means and areal fractions of soil fertility variables within an administrative unit, a district in the state Andhra Pradesh. For this aim a *design-based* sampling approach is recommended (Brus and de Gruijter, 1997). In this approach sampling locations are selected by *probability sampling*. Probability sampling involves the use of a random number generator. Probability samples can be selected in many ways. The simplest selection method, also called a *sampling design*, is *simple random sampling* (SRS). In SRS each possible sample of n sampling locations (n is referred to as the *size* of the sample) has equal probability of being selected.

As an illustration we compute the required sample size for estimating the means of Zn and areal fractions with Zn-deficiency within each district of Andhra Pradesh. These estimated population parameters are of practical importance. They can, for example, be used to prioritize districts for policy interventions. The quality criterion that is used to compute these sample sizes is the width of a confidence interval of the population parameter.

The required sample sizes are computed with three approaches: the frequentist approach, the fully Bayesian approach and the mixed Bayesian-likelihood approach (Joseph et al., 1995, Joseph and Belisle, 1997). To compute the required sample sizes for estimating the mean of Zn within districts, the Zn concentrations are log-transformed. The probability distribution of log-transformed Zn concentrations are much closer to a normal distribution, which is required for the Bayesian and mixed Bayesian-likelihood approach. As a critical

Zn-concentration we use 0.9, i.e., if the Zn-concentration at a location is less than 0.9, we consider that this location is deficient of Zn, so that the application of Zn fertilizer is recommended.

With the fully Bayesian and mixed Bayesian-likelihood approach required sample sizes are computed for three criteria: the average width crietrion, the average coverage criterion and the worst outcome criterion (Joseph et al., 1995, Joseph and Belisle, 1997)..

2 Reading the data

The cycle 1 SHC data collected in 2015-2017 are used to compute summary statistics for each district.

```
##          district     n         mean       var  fraction
## 1       Anantapur 49114 -0.725297326 0.9556266 0.7679277
## 2        Chittoor 37978 -0.060070260 0.4066594 0.4946021
## 3  East Godavari 30353  0.239702652 0.6963413 0.3251408
## 4          Guntur 63956 -0.368286248 0.8161529 0.6118269
## 5          Kadapa 21739 -0.658311838 0.5992403 0.7723446
## 6         Krishna 30481 -0.050333924 0.7972063 0.3915554
## 7         Kurnool 79775 -0.388149609 1.1241939 0.5891319
## 8         Nellore 48053 -1.217143951 1.1920288 0.8576155
## 9        Prakasam 50392 -0.638136119 1.3453333 0.6716741
## 10     Srikakulam 40823  0.008696663 0.4371163 0.4009994
## 11 Visakhapatnam  8678 -0.405452646 0.9664817 0.5746716
## 12  Vizianagaram 28321 -0.350607401 0.4598608 0.6402316
## 13 West Godavari 20211  0.368697951 0.7655733 0.2586215
```

3 Required sample sizes for estimating the mean of ln(Zn)

Given a maximum width $w_{\max}$ of a $100(1-\alpha)\%$ confidence interval of the poulation mean, in the frequentist approach the required sample size can be computed with

$$n = \left(u_{(1-\alpha/2)}\frac{\tilde{\sigma}}{w_{\max}/2}\right)^2$$

The sample variance as computed with the cycle 1 SHC data of 2015-2017 (see output above) is used for $\tilde{\sigma}^2$.

As we are uncertain about the population standard deviation σ, in the fully Bayesian and mixed Bayesian-likelihood approach a prior distribution is assigned to this parameter. It is convenient to assign a gamma distribution as a prior distribution to the reciprocal of the population variance, referred to as the precision parameter $\lambda = 1/\sigma^2$. More precisely, a prior *bivariate* normal-gamma distribution is assigned to the population mean and the precision parameter. With this prior distribution, the *posterior* distribution of the population mean is fully defined, i.e. both the type of distribution and its parameters are known. We say that the prior distribution is *conjugate* with the normal distribution.

A prior gamma distribution is assigned to the reciprocal of the population variance $\lambda = 1/\sigma^2$. This gamma distribution has two parameters a and b. The mean of a gamma distribution equals a/b, the standard deviation equals $\sqrt{a/b^2}$. The mean of the gamma distribution was set equal to the reciprocal of the legacy sample variance of ln(Zn), $a/b = 1/\sigma^2$ (Table **??**). A second equation with a and b is needed to derive parameters a and b. In this second equation the coefficient of variation of the gamma distribution, $cv(\lambda)$, is set equal to a user-specified value. Solving the two equations with two unknowns gives $a = 1/\{cv(\lambda)\}^2$ and $b = a\,\sigma^2$. Required sample sizes are computed for a coefficient of variation of 0.25 of the gamma distributions for the precision parameter.

```
library(SampleSizeMeans)
wmax <- 0.2 #maximum width (=length) of confidence interval
```

```r
conflevel <- 0.95
worstlevel <- 0.80
lambda <- 1/sigma2 #prior estimate of precision parameter
cv <- 0.25 #coefficient of variation of gamma distribution for lambda
nreq.freq <- nreq.alc.bayes <- nreq.alc.mbl <- nreq.acc.bayes <- nreq.acc.mbl <- nreq.woc.bayes <- nreq
for (i in 1:length(districts)) {
  a <- 1/cv^2
  b <- a/lambda[i]
  nreq.freq[i] <- mu.freq(len=wmax, lambda=lambda[i], level=conflevel)
  nreq.alc.bayes[i] <- mu.alc(len=wmax, alpha=a, beta=b, n0=0, level=conflevel)
  nreq.alc.mbl[i] <- mu.mblalc(len=wmax, alpha=a, beta=b, level=conflevel)
  nreq.acc.bayes[i] <- mu.acc(len=wmax, alpha=a, beta=b, n0=0, level=conflevel)
  nreq.acc.mbl[i] <- mu.mblacc(len=wmax, alpha=a, beta=b, level=conflevel)
  nreq.woc.bayes[i] <- mu.modwoc(len=wmax, alpha=a, beta=b, n0=0, level=conflevel, worst.level=worstleve
  nreq.woc.mbl[i] <- mu.mblmodwoc(len=wmax, alpha=a, beta=b, level=conflevel, worst.level=worstlevel)
}
(df <- data.frame(district=districts, lambda = round(lambda,2),
                  freq = nreq.freq,
                  alc = nreq.alc.bayes,
                  alc.mbl = nreq.alc.mbl,
                  acc = nreq.acc.bayes,
                  acc.mbl = nreq.acc.mbl,
                  woc = nreq.woc.bayes,
                  woc.mbl = nreq.woc.mbl))
```

```
##         district lambda freq alc alc.mbl acc acc.mbl woc woc.mbl
## 1      Anantapur   1.05  368 386     388 397     399 466     473
## 2       Chittoor   2.46  157 165     166 169     173 197     204
## 3  East Godavari   1.44  268 282     283 289     292 339     346
## 4         Guntur   1.23  314 330     331 339     342 398     404
## 5         Kadapa   1.67  231 243     244 249     252 292     298
## 6        Krishna   1.25  307 323     324 331     334 388     396
## 7        Kurnool   0.89  432 454     455 467     470 548     556
## 8        Nellore   0.84  458 482     483 495     497 581     588
## 9       Prakasam   0.74  517 544     545 559     561 656     662
## 10    Srikakulam   2.29  168 177     179 182     185 212     219
## 11 Visakhapatnam   1.03  372 391     392 402     405 471     479
## 12  Vizianagaram   2.17  177 187     188 191     194 223     230
## 13 West Godavari   1.31  295 310     311 318     320 373     379
```

4 Required sample sizes for estimating the areal fraction with Zn deficiency

Given a maximum width $w_{\max}$ of a $100(1-\alpha)\%$ confidence interval of the areal fraction, in the frequentist approach the required sample size can be computed with

$$n = \left(u_{(1-\alpha/2)} \frac{\sqrt{\tilde{\pi}(1-\tilde{\pi})}}{w_{\max}/2} \right)^2 + 1$$

In the fully Bayesian and mixed Bayesian-likelihood approach a prior beta distribution is assigned to the design-parameter $\tilde{\pi}$. The beta distribution has two parameters α and β which correspond to the number of "successes" (1) and "failures" (0) in the problem context. The larger these numbers, the more the prior information, and the more sharply defined the probability distribution. By setting the mode of the prior beta distribution to the prior estimate of the areal fraction $\tilde{\pi}$, the parameters of the beta distribution can be computed by

$$\alpha = n_0\tilde{\pi} + 1$$
$$\beta = n_0(1 - \tilde{\pi}) + 1 \ ,$$

with n_0 the prior sample size.

```
## Loading required package: binom
```

```
##          district     f wald alc alc.mbl acc acc.mbl woc woc.mbl
## 1       Anantapur 0.768  275 223     271 226     276 263     318
## 2        Chittoor 0.495  386 335     371 335     371 343     381
## 3  East Godavari 0.325  339 299     327 301     330 336     368
## 4          Guntur 0.612  366 294     356 295     357 311     376
## 5          Kadapa 0.772  272 249     269 256     278 310     335
## 6         Krishna 0.392  368 324     353 325     354 348     378
## 7         Kurnool 0.589  373 286     364 286     364 298     379
## 8         Nellore 0.858  189 144     192 149     201 187     247
## 9        Prakasam 0.672  340 282     331 284     333 310     364
## 10     Srikakulam 0.401  371 319     358 319     359 337     379
## 11 Visakhapatnam 0.575  377 333     340 337     345 370     378
## 12  Vizianagaram 0.640  355 315     341 317     344 347     375
## 13 West Godavari 0.259  296 271     289 277     297 328     352
```

5 References

Joseph, L., Wolfson, D.B. and Du Berger, R. 1995. Sample size calculations for binomial proportions via highest posterior density intervals. Journal of the Royal Statistical Society. Series D (The Statistician): 44, 143-154.

Joseph, L. and Belisle, P. 1997. Bayesian sample size determination for normal means and differences between normal means. Journal of the Royal Statistical Society. Series D (The Statistician): 46, 209-226.