Relaxation of the parameter independence assumption in the bootComb R package

Marc Y. R. Henrion[1,2]

[1] Malawi Liverpool Wellcome Programme, Blantyre, Malawi

[2] Liverpool School of Tropical Medicine, Liverpool, UK

Key words

Biostatistics, R, confidence intervals, bootstrap, estimation

Word count

Abstract: 142 words

Main text (excluding abstract, references): 1,738 words

https://doi.org/10.1017/exp.2022.13 Published online by Cambridge University Press

Abstract

Background The bootComb R package allows researchers to derive confidence intervals with correct target coverage for arbitrary combinations of arbitrary numbers of independently estimated parameters. Previous versions (< 1.1.0) of bootComb used independent bootstrap sampling and required that the parameters themselves are independent - an unrealistic assumption in some real-world applications.

Findings Using Gaussian copulas to define the dependence between parameters, the bootComb package has been extended to allow for dependent parameters.

Implications The updated bootComb package can now handle cases of dependent parameters, with users specifying a correlation matrix defining the dependence structure. While in practice it may be difficult to know the exact dependence structure between parameters, bootComb allows running sensitivity analyses to assess the impact of parameter dependence on the resulting confidence interval for the combined parameter.

Availability bootComb is available from the Comprehensive R Archive Network (*https://CRAN.R-project.org/package=bootComb*).

Introduction

The bootcomb R package Henrion (2021) was recently published. This package for the statistical computation environment R (R Core Team, 2021) allows researchers to derive confidence intervals (CIs) with correct coverage for combinations of independently estimated parameters. Important applications include adjusting a prevalence for estimated

https://doi.org/10.1017/exp.2022.13 Published online by Cambridge University Press

test sensitivity and specificity (e.g. Mandolo et al. (2021)) or combining conditional prevalence estimates (e.g. Stockdale et al. (2020)).

Briefly, for each of the input parameters, bootComb finds a best-fit parametric distribution based on the confidence interval for that parameter estimate. bootComb then uses the parametric bootstrap to sample many sets of parameter estimates from these best-fit distributions and computes the corresponding combined parameter estimate for each set. This builds up an empirical distribution of parameter estimates for the combined parameter. Finally, bootComb uses either the percentile or the highest density interval method to derive a confidence interval for the combined parameter estimate. Full details of the algorithm are given in Henrion (2021).

A key point of the algorithm is that the best-fit distributions for the different parameters are sampled from independently. This requires the parameters to be independent. This may not be a realistic assumption in some real-world applications.

While for most practical applications the input parameters are typically estimated from independent experiments (otherwise the combined parameter could be directly estimated), the parameters themselves may not be independent. This is for instance the case when adjusting a prevalence for the diagnostic test's sensitivity and specificity. The latter two parameters are not independent: higher sensitivity can be achieved by lowering specificity and vice versa.

If the experiments estimating these parameters are sufficiently large, then the violation of the assumption of parameter independence may only have negligible impact on the resulting confidence interval for the combined parameter. However, for the sake of general

https://doi.org/10.1017/exp.2022.13 Published online by Cambridge University Press

applicability and to allow running sensitivity analyses, the author felt it was beneficial to extend bootComb to handle dependent parameters.

Methods

Copulas are multivariate distribution functions where the marginal probability distribution of each variable is the uniform distribution on the interval [0,1]. Copulas allow to specify the intercorrelation between random variables. An important probability theory result, Sklar's Theorm (Sklar, 1959), states that any multivariate probability distribution can be expressed in terms of its univariate marginal distributions and a copula defining the dependence between the variables.

Mathematically, let $X_1, X_2 \dots, X_d$ be d random variables and define $U_i = F_i(X_i), i = 1, \dots, d$. Then the copula C of $(X_1, \dots, X_d)$ is defined as the joint cumulative distribution function of $(U_1, \dots, U_d)$:

$$C(u_1, \dots, u_d) = Pr(U_1 \leq u_1, \dots, U_d \leq u_d)$$

Assume that the marginal distributions, $F_i(x) = Pr[X_i \leq x], i = 1, \dots, d$ are continuous. Then, via the probability integral transform (Angus, 1994), the random vector $(U_1, U_2, \dots, U_d)$ has marginals that are uniformly distributed on [0,1].

bootComb makes use of the fact that the above can be reversed: given a sample $(u_1, \dots, u_d)$, a sample for $(X_1, \dots, X_d)$ can be obtained by $(x_1, \dots, x_d) = \left(F_1^{-1}(u_1), \dots, F_d^{-1}(u_d)\right)$. The inverse functions $F_i^{-1}(u)$ will be defined if the marginals $F_i(x)$ are continuous. For the use of

https://doi.org/10.1017/exp.2022.13 Published online by Cambridge University Press

bootComb, where users input confidence intervals for an estimated numeric parameter, this will always be the case.

bootComb will proceed as follows to generate samples from a multivariate distribution of d dependent variables:

- Estimate best-fit distributions $F_1, \dots, F_d$ for each of the d parameters $X_1, \dots, X_d$ given the lower and upper limits of the estimated confidence intervals for each parameter.
- Sample $(z_1, \dots, z_d)$ from a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \Sigma)$ where the variances in Σ are all 1.
- Since the marginals of this normal distribution are all $\mathcal{N}(0,1)$, compute $u_i = \Phi(z_i)$ where Φ is the cumulative distribution function of the standard normal.
- Finally, for each $i = 1, \dots, d$, compute $x_i = F_i^{-1}(u_i)$ where F_i is the best-fit marginal distribution of parameter i.

The resulting vector $(x_1, \dots, x_d)$ will be a sample from the multivariate distribution of $(X_1, \dots, X_d)$. Note that the dependence structure was completely specified through the covariance matrix Σ (since the variances are assumed to be 1, this really is a correlation matrix) and marginal distributions for each parameter were specified by $F_i, i = 1, \dots, d$.

Results

I repeat the 2 examples from Henrion (2021) here, but look at the effect of specifying a dependence between the input parameters.

https://doi.org/10.1017/exp.2022.13 Published online by Cambridge University Press

All examples below use the highest density interval (HDI) method (input argument method="hdi") to derive the final confidence interval. Whether this or the percentile method is used is a user choice. The HDI derived interval will be the narrowest interval with the desired coverage and the probability density will always be higher within that interval than outside it. To note however that the HDI may not be a single interval but a set of intervals if the density is multimodal. In this case, the single interval returned by bootComb will be too wide. For this reason, users should always inspect the histogram of the sampled combined parameter when using the HDI method.

1. HDV prevalence in the general population

With an application to hepatitis D and B viruses (HDV and HBV respectively) from Stockdale et al. (2020), Henrion (2021) showed how to use bootComb to obtain a valid confidence interval for $\hat{p}_{aHDV}$, the prevalence of HDV specific immunoglobulin G antibodies (anti-HDV) in the general population.

HBV is a pre-condition for HDV and hence to derive $\hat{p}_{aHDV}$ Stockdale et al. (2020), obtained estimates of the prevalence of surface antigen of the hepatitis B virus (HBsAg), $\hat{p}_{HBsAg} = 3.5\%$, and the conditional prevalence of anti-HDV given the presence of HBsAg, $\hat{p}_{aHDV|HBsAg} = 4.5\%$:

- $\hat{p}_{HBsAg} = 3.5\%$ with 95% CI $(2.7\%, 5.0\%)$.
- $\hat{p}_{aHDV|HBsAg} = 4.5\%$ with 95% CI $(3.6\%, 5.7\%)$.

Assuming these 2 parameters to be independent, Henrion (2021) derived a 95% confidence interval for the estimate $\hat{p}_{aHDV} = \hat{p}_{aHDV|HBsAg} \cdot \hat{p}_{HBsAg}$ using bootComb, $(0.11\%, 0.25\%)$.

https://doi.org/10.1017/exp.2022.13 Published online by Cambridge University Press

If, however, the 2 input prevalences are not independent, e.g. if anti-HDV is more common among people with presence of HBsAg the higher the population prevalence of HBsAg is, then that assumption of independence would not hold. We can investigate how strong an effect dependence of the parameters can have on the resulting confidence estimate. For example, let us run the same example using bootComb with specifying the following covariance matrix for the bivariate normal copula:

$$\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

```
library(bootComb)

combFunEx<-function(pars){pars[[1]]*pars[[2]]}
bootComb(distributions=c("beta","beta"),
         qLowVect=c(0.027,0.036),
         qUppVect=c(0.050,0.057),
         combFun=combFunEx,
         Sigma=matrix(byrow=TRUE,ncol=2,c(1,0.5,0.5,1)),
         doPlot=TRUE,
         method="hdi",
         N=1e6,
         seed=123)
```

This yields the 95% confidence interval $(0.10\%, 0.26\%)$, a slightly wider interval – which makes sense, as the positive correlation means it is more likely for pairs of bootstrapped input parameters to be both near the upper (respectively lower) end of their confidence intervals.

For this particular application, a dependence between both prevalence parameters, $\hat{p}_{HBsAg}$ and $\hat{p}_{aHDV|HBsAg}$, is unlikely and I have therefore not considered this example any further.

https://doi.org/10.1017/exp.2022.13 Published online by Cambridge University Press

2. SARS-CoV-2 seroprevalence adjusted for test sensitivity and specificity

Henrion (2021) gave an example of adjusting an estimated SARS-CoV-2 seroprevalence for the estimated sensitivity and specificity of the test assay. Specifically:

- 84 out of 500 study participants tested positive for SARS-CoV-2 antibodies, yielding a seroprevalence estimate $\hat{\pi}_{raw} = 16.8\%$ with exact binomial 95% CI $(13.6\%, 20.4\%)$.
- Estimated assay sensitivity: 238 out of 270 known positive samples tested positive $\hat{p}_{sen} = 88.1\%$, 95% CI $(83.7\%, 91.8\%)$.
- Estimated assay specificity: 82 out of 88 known negative samples tested negative $\hat{p}_{spec} = 93.2\%$, 95% CI $(85.7\%, 97.5\%)$.

Assuming the sensitivity and specificity to be independent, Henrion (2021) reported an adjusted seroprevalence estimate $\hat{\pi} = 12.3\%$ with 95% CI $(3.9\%, 19.0\%)$.

However in this case, the assumption of independence is not fully realistic: there is a trade-off between sensitivity and specificity of the test assay, and as such one would expect a negative dependence between the two parameters: sensitivity can be increased at the cost of decreased specificity and vice versa.

Assuming that the sensitivity and specificity are negatively correlated with the copula correlation parameter $\rho = -0.5$ between these two parameters, using the extension of `bootComb` we can now account for the dependence of the parameters:

```
adjPrevSensSpecCI(
    prevCI=c(0.136,0.204),
    sensCI=c(0.837,0.918),
```

https://doi.org/10.1017/exp.2022.13 Published online by Cambridge University Press

```
    specCI=c(0.857,0.975),
    Sigma=matrix(byrow=TRUE,ncol=3,c(1,0,0,0,1,-0.5,0,-0.5,1)),
    doPlot=TRUE,
    prev=84/500,
    sens=238/270,
    spec=82/88,
    seed=123)
```

The reported confidence interval is now $(3.8\%, 19.4\%)$ - marginally wider than when the dependence was ignored.

If we additionally specify returnBootVals=TRUE in the function call, we can extract and plot the sampled pairs of sensitivity and specificity values to check the dependence structure. This is shown on Figure 1: as the correlation parameter ρ in the copula between the sensitivity and specificity is decreased from 0 to -1, the dependence between both parameters becomes more and more pronounced as one would expect.

This shows that a simple correlation matrix specified for the Gaussian copula results in this case in a non-trivial dependence structure between two beta-distributed variables, respecting the specified marginal distributions.

We can also visualise the effect on the estimated confidence interval, as shown on Figure 2. We can see that in this case, with a negative correlation, the width of the CI increases at the correlation becomes stronger. However, looking at the scale of the y-axis we see that this is just a marginal effect.

A more substantial effect of parameter dependence is obtained when we also allow the measured prevalence π_{raw} to be correlated with sensitivity (p_{sens}; positive correlation) and specificity (p_{spec}; negative correlation). Specifically, we can specify the following correlation matrix for the parameters $(\pi_{raw}, p_{sens}, p_{spec})$:

https://doi.org/10.1017/exp.2022.13 Published online by Cambridge University Press

$$\Sigma = \begin{pmatrix} 1 & 0.3 & -0.3 \\ 0.3 & 1 & -0.5 \\ -0.3 & -0.5 & 1 \end{pmatrix}$$

```
adjPrevSensSpecCI(
    prevCI=c(0.136,0.204),
    sensCI=c(0.837,0.918),
    specCI=c(0.857,0.975),
    Sigma=matrix(byrow=TRUE,ncol=3,c(1,0.3,-0.3,0.3,1,-0.5,-0.3,-0.5,1)),
    doPlot=TRUE,
    prev=84/500,
    sens=238/270,
    spec=82/88,
    seed=123)
```

In this case, the reported confidence interval is $(4.7\%, 18.2\%)$. This CI is 11% narrower than when the dependence structure was ignored – a substantial effect for practical purposes.

Conclusions

The R package bootComb has been extended and, using Gaussian copulas, it can now handle the case of dependent input parameters. For many applications, the effect of dependence between the parameters will be marginal or even negligible, but this is not always the case. The package now allows users to do sensitivity analyses to assess the effects of a miss-specified dependence structure between the parameters that are being combined.

At the time of publication, the most recent version of bootComb was 1.1.2.

Figure captions

Figure 1: *Scatterplots showing the bootstrapped values of sensitivity and specificity for different strenghts of dependence (from independence to perfect correlation) between*

https://doi.org/10.1017/exp.2022.13 Published online by Cambridge University Press

sensitivity and specifity. The empirical kernel density estimate for the bivariate distribution in each case is shown as orange contour lines.

Figure 2: *Width of the estimated confidence interval as a function of inreased strength of the negative correlation between sensitivity and specificity.*

Funding Information (see funding information section for more information)

This research was funded in whole, or in part, by the Wellcome Trust [grant: 206545/Z/17/Z]. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Data Availability Statement

All data to support this work are contained within the article. The software package itself is available from *https://cran.r-project.org/package=bootComb*.

Conflicts of interest

Author Marc Y. R. Henrion declares none.

https://doi.org/10.1017/exp.2022.13 Published online by Cambridge University Press

References

Angus, J. E. (1994). The Probability Integral Transform and Related Results. *SIAM Review*, *36*(4), 652–654. *http://www.jstor.org/stable/2132726*

Henrion, M. Y. (2022). *bootComb: Combine Parameter Estimates via Parametric Bootstrap* (R package version 1.1.2) [Computer software]. *https://cran.r-project.org/package=bootComb*

Henrion, M. Y. (2021). bootComb—an R package to derive confidence intervals for combinations of independent parameter estimates. *International Journal of Epidemiology*, *50*(4), 1071–1076. *https://doi.org/10.1093/ije/dyab049*

Mandolo, J. J., Henrion, M. Y. R., Mhango, C., Chinyama, E., Wachepa, R., Kanjerwa, O., Malamba-Banda, C., Shawa, I. T., Hungerford, D., Kamng'ona, A. W., Iturriza-Gomara, M., Cunliffe, N. A., & Jere, K. C. (2021). Reduction in Severity of All-Cause Gastroenteritis Requiring Hospitalisation in Children Vaccinated against Rotavirus in Malawi. *Viruses*, *13*(12), 2491. *https://doi.org/10.3390/v13122491*

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. *https://www.R-project.org/*

Sklar, M. (1959). *Fonctions de Répartition À N Dimensions Et Leurs Marges* (Issue 8, pp. 229–231). Publications de l'Institut Statistique de l'Université de Paris.

Stockdale, A. J., Kreuels, B., Henrion, M. Y. R., Giorgi, E., Kyomuhangi, I., de Martel, C., Hutin, Y., & Geretti, A. M. (2020). The global prevalence of hepatitis D virus infection: Systematic review and meta-analysis. *Journal of Hepatology*, S0168827820302208. *https://doi.org/10.1016/j.jhep.2020.04.008*

https://doi.org/10.1017/exp.2022.13 Published online by Cambridge University Press