The Journal of Open Source Software

Overlapping: a R package for Estimating Overlapping in Empirical Distributions

Massimiliano Pastore[1] **and** [1]

1 Department of Developmental and Social Psychology, University of Padova **1**

DOI: 10.21105/joss.00844

Software

- Review
- Repository
- Archive

Submitted: 23 July 2018
Published: 24 July 2018

Summary

`overlapping` is a R package for estimating the overlapping area of two or more empirical distributions. The main idea of the package is to offer an easy way to quantify the similarity (or the difference) between two or more empirical distributions. In addition, the package allows to plot density distributions, highlighting the overlapped area by using the `ggplot2` R package (Wickham 2009).

The package is available from GitHub (https://github.com/masspastore/overlapping) and CRAN (https://cran.r-project.org/package=overlapping). A full reference manual can be found at https://cran.r-project.org/web/packages/overlapping/overlapping.pdf.

Examples

Suppose we have collected data in two groups of 100 subjects each with respect to a generic variable Y, expressed by scores ranging between 0 and 30, and to be interested in assessing whether the two groups can be considered samples from populations with the same average.

We can simulate the groups scores as follows:

```
set.seed( 1 )
n <- 100
G1 <- sample( 0:30, size = n, replace = TRUE )
G2 <- sample( 0:30, size = n, replace = TRUE, prob = dbinom( 0:30, 31, .55 ) )
```

For Group 1 (`G1`) we randomly sampled `n` = 100 values from a uniform distribution; for Group 2 (`G2`) we randomly sampled 100 values from a binomial distribution. In the first group, scores range between 0 and 30 with mean 15.55 and standard deviation 8.32. In the second group, scores range between 10 and 24 with mean 16.72 and standard deviation 2.74.

We can display the scores distribution as follows:

```
library( ggplot2 )
Data <- data.frame( y = c(G1,G2), group = rep(c("G1","G2"),each=n) )
ggplot( Data, aes( y ) ) + facet_wrap( ~group ) + geom_histogram()
```

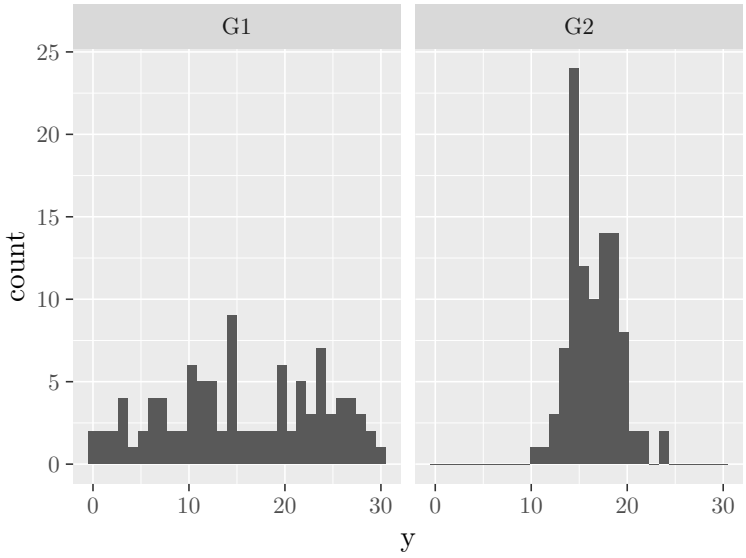

Figure 1: Score distributions of simulated groups of 100 subjects each.

obtaining the figure 1. In the left panel are depicted the scores of group 1, in the right panel the scores of group 2. From this figure it is evident the heterogeneity of the variances in the two groups. In such a case, the statistical comparison between means can be biased and not very informative; for example, with a t-test, corrected for heterogeneity, we obtain the following result: $t(120.24) = -1.34$, $p = 0.18$, from which we cannot draw any conclusion.

So, let us assume a different perspective: Rather than assessing the similarity between the two groups on the basis of averages (and standard deviations) only, we use all the information available in the data. In practice we estimate the degree of similarity by exploring the overlapping between group scores. We expect 0% to indicate the absence of overlapping (i.e., maximum distance between groups), and 100% toindicate the perfect overlap between the two distributions (i.e., groups are identically distributed). We can use the overlapping package in the following way:

```
library( overlapping )
dataList <- list( G1 = G1, G2 = G2)
overlap( dataList )$OV * 100
```

```
##    G1-G2
## 43.21998
```

With the command `library()` we load the **`overlapping`** package, next we create a `list` containing the two groups scores, and finally, by using the `overlap()` function, we compute the overlap index. The index value (43.22) is an estimate of the percentage of overlapping between density scores. We can obtain a graphical representation by adding the option `plot = TRUE` as follows:

```
overlap( dataList, plot = TRUE )
```

obtaining the figure 2. In the figure are represented the estimated densities of the two groups scores, with different colors, and the shaded part is the overlapping area of densities.

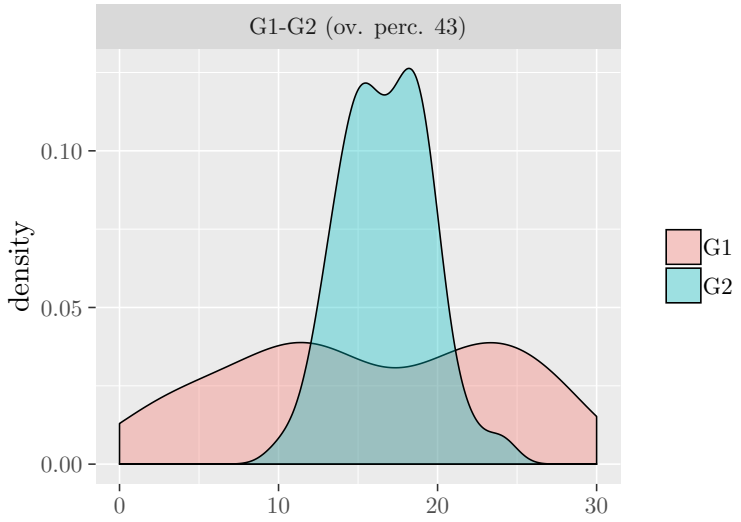

Figure 2: Comparison between densities of two groups. The overlap (43%) is represented by the shaded area.

Examples of real-world analysis

`overlapping` package was already used in different publications for many purposes, such as: 1) evaluating group invariance in questionnares, by using parameters bootstrap distributions (Lionetti, Mastrotheodoros, and Palladino 2018, Marci et al. (2018)); 2) for computing a distance index in antropological measures (Altoè, D'Amore, and Scalfari in press); 3) for identifying cut-off scores in questionnaires, estimating the intersection points of density distributions (Pluess et al. 2018, Lionetti et al. (2018)).

References

Altoè, Gianmarco, Giuseppe D'Amore, and Francesco Scalfari. in press. "Skulls and transvariation."

Lionetti, Francesca, Arthur Aron, Elaine N Aron, G Leonard Burns, Jadzia Jagiellowicz, and Michael Pluess. 2018. "Dandelions, Tulips and Orchids: Evidence for the Existence of Low-Sensitive, Medium-Sensitive and High-Sensitive Individuals." *Translational Psychiatry* 8 (1). Nature Publishing Group:24.

Lionetti, Francesca, Stefanos Mastrotheodoros, and Benedetta Emanuela Palladino. 2018. "Experiences in Close Relationships Revised Child Version (Ecr-Rc): Psychometric Evidence in Support of a Security Factor." *European Journal of Developmental Psychology* 15 (4). Taylor & Francis:452–63.

Marci, Tatiana, Francesca Lionetti, Ughetta Moscardino, Massimiliano Pastore, Vincenzo Calvo, and Gianmarco Altoé. 2018. "Measuring attachment security via the Security Scale: Latent structure, invariance across mothers and fathers and convergent validity." *European Journal of Developmental Psychology* 15 (4). Taylor & Francis:481–92.

Pluess, Michael, Elham Assary, Francesca Lionetti, Kathryn J Lester, Eva Krapohl, Elaine N Aron, and Arthur Aron. 2018. "Environmental Sensitivity in Children: Development of the Highly Sensitive Child Scale and Identification of Sensitivity Groups." *Developmental Psychology* 54 (1). American Psychological Association:51.

Wickham, Hadley. 2009. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. http://ggplot2.org.