View metadata, citation and similar papers at core.ac.uk brought to you by CORE
provided by Diposit Digital de la Universitat de Barcelona

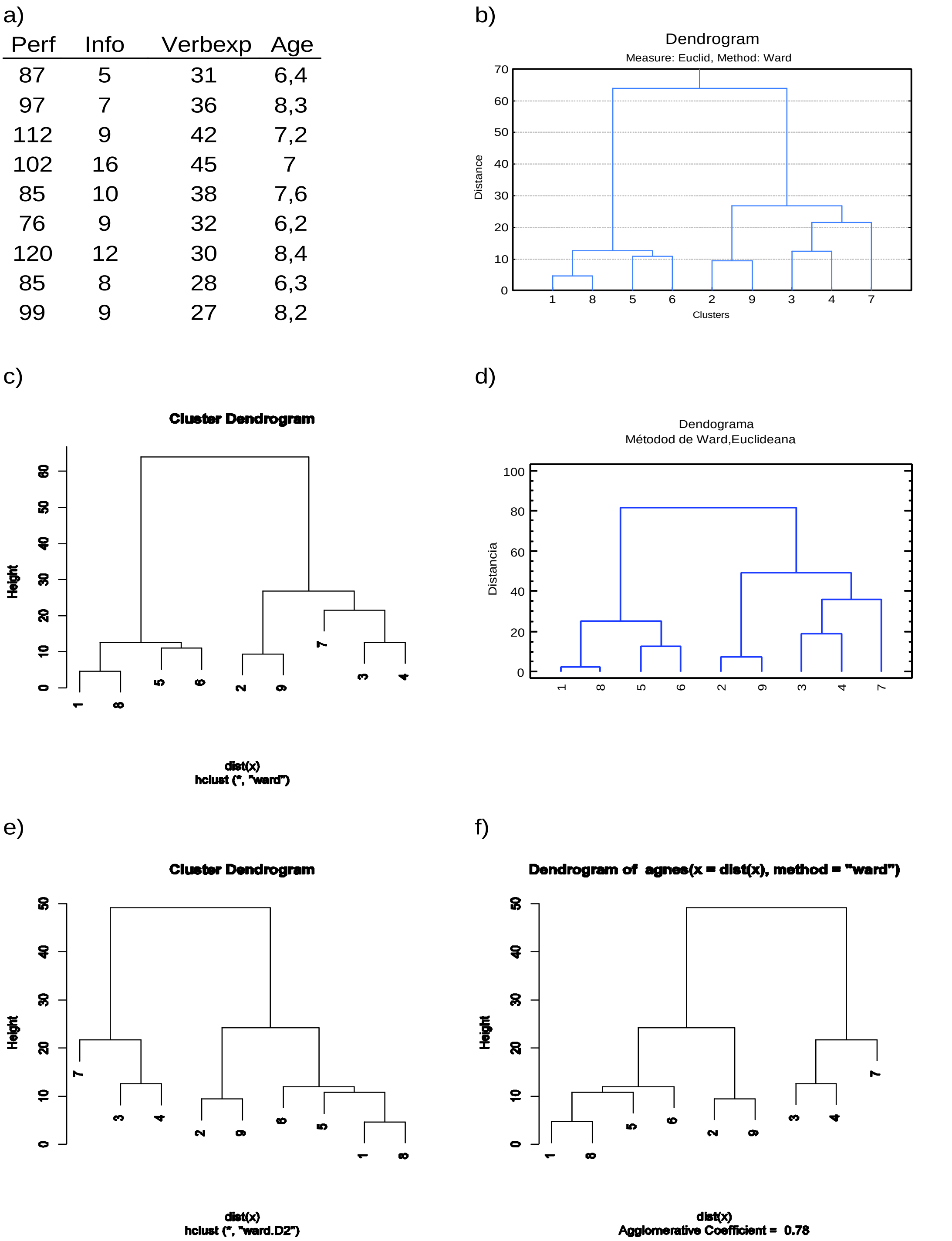
Reliable software

Concepción Arenas*
Departament d'Estadística, Facultat de Biologia
Universitat de Barcelona
carenas@ub.edu

Francesc Mestres
Departament de Genètica, Facultat de Biologia
Universitat de Barcelona
fmestres@ub.edu

Leek and Peng (2015) have presented the decisions based on P value as the final step of a data pipeline. Rightly, they describe that the final decision in a statistical study depends on a correct experimental design and data analysis. This process consists of different stages and all of them are fundamental. Thus, the authors present the following steps as essential for a proper study: experimental design, data collection, data cleaning, exploratory data analysis, exploring potential statistical models, statistical modelling, summary statistics, inference and eventually obtaining the P value. In the last years, an intense debate on the final step, the P value, has arisen (Nuzzo, 2014), but it is not our aim to discuss it in this contribution. Although the previous steps in the pipeline are usually not particularly commented, they are fundamental to complete the proposed study obtaining reliable results. On the other hand and in agreement with Leek and Peng, we think researches need a proper training in both, data analysis and software use. A poor knowledge of software could lead to wrong conclusions. For instance, if one-way ANOVA is carried out with the R function *aov*, and the variable containing the codification of the different levels is numeric, if the user is not familiar with this function and does not declare this variable as "factor" using the command *as.factor*, the program will compute a lineal regression instead of an ANOVA. These kind of mistakes are attributable to userôs lack of experience. However, we want to introduce an additional thought on this debate. In biomedicine, biodiversity and other fields of research, large databases are used. Assuming that a proper statistical procedure has been chosen, a crucial point is the selection of the right software to compute the data. The available software has to be sufficiently proven and having the guarantee that it is reliable. Currently, it is easy to obtain free software for most statistical procedures. We agree that a free software is especially useful because as a large number of researchers can take benefit of it. However, in several repositories, software has not been sufficiently proven, and could yield to erroneous results. This situation could lead to dreadful consequences, for instance, when studying cancer or complex genetic diseases. We propose that researchers should be especially accurate in their software selection, and also the control levels should be improved in order to upload new software in a public repository.

These flaws due to the program has not tested enough or not properly explained to users are also found in commercial software. One example of this situation has been recently presented by Murtagh and Legendre (2014). However, we think that their example of Wardôs hierarchical agglomerative clustering method it is still not a well-known case. An example of data from Unistat65 package is presented in Figure 1a. With this data and applying the Ward method (utilizing the Euclidean distance), different dendrograms has been obtained using different software. It can be observed that the first three trees (Figure 1b: dendrogram obtained with Unistat65; Figure 1c: dendrogram obtained using function *hclust* included in the *stats* R package selecting *method=ward*; Figure 1d: dedrogram obtained with Statgraphics and selecting as method the Ward option) are equivalent and the last two dendrograms (Figure 1e: dendrogram obtained using function *hclust* included in the *stats* R package selecting *method=ward.D2*. This option is only able from the work of Murtagh and Legendre, and is not able for example in the R version 2.15.1; Figure 1f: dendrogram obtained using function *agnes* included in the cluster R package selecting *method=ward*) are also equivalent. However, the clustering is different between both groups of trees. What is the problem? Are they different algorithms? Is there any miscalculation? Murtagh and Legendre carried out a study in depth on the operation of the distinct programs or functions in relation with this particular algorithm. This study allowed to know the programming differences and to ascertain in which cases the Ward algorithm is properly computed or not. We agree with the authors when talking about users they said that "urge to check what their favorite software is doing". However, it is evident that a statistical user will not check systematically the software that uses, and will not compare the results obtained when using different packages. Nevertheless, the verification of the proper performance of software must be carried out. For these reason, we consider a cornerstone of statistical analysis that free or commercial software had been enough tested. The verification has to include the statistical method, and also that the program is calculating what should really calculate and following the proper algorithm.



a)

Perf	Info	Verbexp	Age
87	5	31	6,4
97	7	36	8,3
112	9	42	7,2
102	16	45	7
85	10	38	7,6
76	9	32	6,2
120	12	30	8,4
85	8	28	6,3
99	9	27	8,2

Figure 1. a) Data set included as example in the Unistat 65; b) dendrogram obtained with Unistat 65; c) dendrogram obtained using function hclust included in the stats R package selecting method=ward; d) dedrogram obtained with Statgraphics and selecting as method the Ward option; e) dendrogram obtained using function hclust included in the stats R package selecting method=ward.D2; f) dendrogram obtained using function agnes included in the cluster R package selecting method=ward.

We thank Miss A. Izquierdo-Bouldsbridge (Parc Cientíc de Barcelona, Barcelona, Spain) for revising and correcting the English text.

References

[1] Leek J.T. and Peng R.D. (2015). P values are just the tip of the iceberg. Nature, 520, 612.

[2] Nuzzo R. (2014). Statistical errors. Nature, 506, 150-152.

[3] Murtagh F. and Legendre P. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? Journal of Classification, 31, 274-295.

About the authors

Concepción Arenas received a PhD in mathematics, with a specialization in statistics, from the University of Barcelona and is now a research professor in the Departament d'Estadística at the Universitat Barcelona (Spain). Her research interests include multivariate analysis as applied to bioinformatics, specifically DNA sequence analysis and microarray interpretation. She also works in biomedical statistics.

Francesc Mestres received a PhD in Genetics from the University of Barcelona and is now a research professor in the Departament de Genètica at the Universitat de Barcelona (Spain). His research interests include bioinformatics, specifically DNA sequence analysis and microarray interpretation, forensic genetics and evolution.