The extended version of Cohen's *d* index for interval-valued data

M. Asunción Lubiano, José García-García, Antonio L. García-Izquierdo, Ana M. Castaño

Abstract Nowadays, new types of data are emerging from lots of distinct real-life experiments and statistical researchers need to develop new tools to deal with them. For instance, interval-valued responses arise as an alternative to Likert-type responses in questionnaires measuring people's behavior (their attitudes, opinions, perceptions, feelings, etc.). In order to facilitate the comparison of different analysis involving several rating scales and with the aim of studying the the effect size measure for difference between two independent groups, in this paper we extend the concept of Cohen's *d* index established for real numbers to the interval-valued data context. Finally, a real-life example has been included to motivate and illustrate the problem.

1 Introduction

For many decades now, data analysts have been encouraging to enhance the presentation of research findings in the behavioral sciences by including an effect-size measure along with a statistical significance test (Cohen, 1965; Hays, 1963). Besides, the American Psychological Association (APA) Publication Manual stated "It is almost always necessary to include some measure of effect-size in the Result section" (APA, 2010, p. 34) and Wilkinson et al. (1999) highlighted the importance of including the effect size for future systematic reviews and meta-analysis.

M. Asunción Lubiano (✉) and José García-García
Department of Statistics, O.R. and D.M., University of Oviedo, C/ Federico García Lorca 18, 33007 Oviedo, Spain, e-mail: lubiano@uniovi.es,garciagarjose@uniovi.es

Antonio L. García-Izquierdo and Ana M. Castaño
Department of Psychology, Universidad de Oviedo, Plaza Feijóo, s/n, 33003 Oviedo, Spain e-mail: angarcia@uniovi.es,castanoana@uniovi.es

In Statistics, an effect-size is a quantitative measure that is independent of sample size and complements statistical hypothesis testing. This measure quantifies the magnitude of a phenomenon like the difference between populations or the relationship between explanatory and response variables and facilitates the statistical interpretation of the importance of a research result. Consequently, it allows the comparison of different results of a set of empirical studies carried out independently about a given research problem.

Nowadays, statistical data analysis methodology is constantly evolving due to the appearance of new types of real-life data that cannot be strictly classified as quantitative or qualitative ones.

In social and educational sciences and many other disciplines, *Likert-type Scales* (Likert, 1932) are the most popular rating scales considered in the literature to rate evaluations, perceptions, judgments, classifications, etc. in questionnaires. This type of data cannot be numerically measured because they concern intrinsically imprecise valued attributes and Likert scales allow the respondent to choose among a small number of predetermined 'linguistic values' (discrete scale).

To overcome the limitations of Likert-type scales, since the individual differences are almost systematically overlooked, in the last years *Interval-Valued Scales* (IVSs) are gaining strength as an alternative to Likert-type scales by allowing respondents to select a range or interval of real data and not being constrained to choose among a few pre-specified responses (see, for instance, Ellerby et al., 2021; Wagner et al., 2015; Themistocleous et al., 2019).

In this paper, we will extend the definition of one of the most known standardized mean difference effect-size measures to fuzzy approach. We will analyze the results with a real life example where interval-valued data have been gathered.

2 Preliminary concepts

Interval-valued scales make use of random intervals. In this section, we will recall the main concepts and the methodology used to analyze this type of data.

Let $\mathcal{K}_c(\mathbb{R})$ denote the class of nonempty compact intervals from $\mathbb{R}$. Each interval K in the space $\mathcal{K}_c(\mathbb{R})$ can be characterized in terms of either its infimum and supremum or its mid-point and spread or radius as follows:

$$K = [\inf K, \sup K] = [\operatorname{mid} K - \operatorname{spr} K, \operatorname{mid} K + \operatorname{spr} K].$$

When dealing with interval-valued data we use an arithmetic based on the *sum* and the *product by a scalar* operations defined as the corresponding image sets of the involved interval values (see Minkowski, 1903) which are settled for $K, K' \in \mathcal{K}_c(\mathbb{R})$ and any $\lambda \in \mathbb{R}$ as follows

$$K + K' = [\inf K + \inf K', \sup K + \sup K'],$$
$$\lambda \cdot K = \begin{cases} [\lambda \cdot \inf K, \lambda \cdot \sup K] \text{ if } \lambda \geq 0 \\ [\lambda \cdot \sup K, \lambda \cdot \inf K] \text{ if } \lambda < 0. \end{cases}$$

.

In contrast to the real-valued case, the space $\mathcal{K}_c(\mathbb{R})$ is not linear with these two operations, but only semilinear with a conical structure, because of the lack of an opposite element for the Minkowski addition. For this reason, it is not possible to treat intervals directly as two-dimensional vectors. Therefore, distances play a crucial role in statistical developments.

To measure the distance between two interval-valued data, we will make use of a metric on $\mathcal{K}_c(\mathbb{R})$ extending the Euclidean one and being easy-to-use and interpret. More precisely, we will consider the d_θ-*metric* proposed by Gil et al. (2002) which is defined for two intervals $K, K' \in \mathcal{K}_c(\mathbb{R})$ as follows,

$$d_\theta(K, K') = \sqrt{(\operatorname{mid} K - \operatorname{mid} K')^2 + \theta \cdot (\operatorname{spr} K - \operatorname{spr} K')^2},$$

where $\theta \in (0, \infty)$ weighs the relative importance assessed to deviations in imprecision in contrast to deviations in trends. It is often imposed that $\theta \in (0, 1]$, in order to weigh the deviation in location no less than the deviation in imprecision, as well as to make d_θ coincide with the metric introduced by Bertoluzza et al. (1995). Actually, the d_1 metric coincides with the 2-norm metric between intervals which has been proposed by Vitale (1985).

Compact random intervals (see Matheron, 1975) determine a well-stated and supported model for the random mechanisms generating interval-valued data within the probabilistic setting. They integrate both randomness and imprecision, so that the first one affects the generation of experimental data, whereas the second affects the nature of the experimental data which, for formal purposes, are assumed to be intrinsically interval-valued.

Following the general random set approach, given a probability space $(\Omega, \mathcal{A}, P)$, a mapping $X : \Omega \to \mathcal{K}_c(\mathbb{R})$ is said to be a **interval-valued random set** (IVRS for short) associated with it if X is measurable with respect to $\mathcal{A}$ and the Borel σ-algebra generated by the topology induced by the d_θ metric on $\mathcal{K}_c(\mathbb{R})$. Equivalently, X is a interval-valued random set if, and only if, both functions $\inf X$ and $\sup X$ (or alternatively, $\operatorname{mid} X$ and $\operatorname{spr} X$) are real-valued random variables.

As a consequence from the Borel measurability, crucial concepts in probabilistic and inferential developments, such as the (induced) distribution of a interval-valued random set or the stochastic independence of interval-valued random sets, are well-defined.

In performing inferential analysis about the distribution of interval-valued random sets, the best known involved parameters are the Aumann-type mean

value (Aumann, 1965) and the Fréchet-type variance (Korner, 1997; Lubiano et al., 2000).

We are going to recall their sample version. Given a random sample $(X_1, \ldots, X_n)$ of size n from an IVRS X and a realization $\mathbf{x} = (x_1, \ldots, x_n)$,

The *sample Aumann mean* of $\mathbf{x}$ is given by the compact interval

$$\overline{\mathbf{x}} = \frac{1}{n} \cdot (x_1 + \cdots + x_n) .$$

The *sample (d_θ Fréchet-type) variance* of $\mathbf{x}$ is given by the real number

$$s^2_{\mathbf{x}} = \frac{1}{n-1} \cdot \sum_{i=1}^{n} \left[d_\theta \left(x_i, \overline{\mathbf{x}} \right) \right]^2 .$$

The above considered (sample) mean and variance preserve all the main properties from the numerical case. All of these properties allow us to consider the mean and the variance as suitable estimates of central tendency and dispersion, respectively.

In the next section, we are going to state an effect-size measure for interval-valued data considering the extension of the most common effect-size measure for real-valued data which is defined to compare the means of two groups.

3 Standardized mean difference for interval-valued data

In research studies that involve the comparison of two groups, the standardized mean difference is one of the most frequently used effect-size measures.

Let $(\Omega, \mathcal{A}, P)$ be the probability space modeling a random experiment. Then, if X and Y are two independent IVRSs associated with $(\Omega, \mathcal{A}, P)$, we will consider the following effect size

$$\delta = \frac{d_\theta \left[E(X), E(Y) \right]}{SD},$$

where SD is the standard deviation of the population.

In the practical setting, population values are not typically known and must be estimated from sample statistics. Distinct versions of effect-sizes based on means proposed so far differ with respect to which statistics are used. The most known effect-size measure is the Cohen's d index suggested by Cohen (1969, 1988), see also Hedges (1981).

Definition 1 Let X and Y be two independent IVRSs associated with $(\Omega, \mathcal{A}, P)$ and consider a sample of independent observations from X, $\mathbf{x} = (x_1, \ldots, x_{n_1})$, and a sample of independent observations from Y, $\mathbf{y} = (y_1, \ldots, y_{n_2})$. The **extended Cohen's d index** of effect size is defined as the real number

$$d = \frac{d_\theta \left(\overline{\mathbf{x}}, \overline{\mathbf{y}} \right)}{SD_p} \quad \text{with} \quad SD_p = \sqrt{\frac{(n_1 - 1) s^2_{\mathbf{x}} + (n_2 - 1) s^2_{\mathbf{y}}}{n_1 + n_2 - 2}},$$

where SD_p is the pooled standard deviation for the two groups which is recommended if the standard deviations and sizes of the two groups differ (Cohen, 1988, p. 67).

In the next section, we are going to apply the preceding measure on a dataset obtained from a real-life situation.

4 Real-life data example

The COVID-19 pandemic has promoted a big change in the Higher Education due to the adjustment to a new scenario characterized by the need to quickly adapt from the face-to-face to the online distance modality.

An educational innovation project was carried out for the planning and improvement of teaching-learning processes of a subject of the Degree in Labor Relations and Human Resources at the University of Oviedo (Spain) for the 2020/2021 academic year. In the study a hybrid system was considered for the teaching modality (face-to-face vs. online learning).

Specifically, we are going to examine the behavior of this effect-size measure in the example with respect to the influence of respondents' sex (men *vs.* women) for both teaching modalities.

A total of 50 participants have been requested to answer a questionnaire (available by means a custom web tool, see Fig. 1) by selecting the interval that best represents their level of agreement to the statements proposed in a interval-valued scale bounded between 1 and 7.

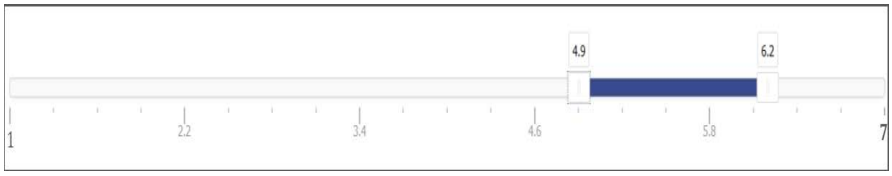

Fig. 1 Example of interval-valued based-responses to the online questionnaire

The online questionnaire was comprised of biographical information (i.e., age, gender, etc.) as well as 142 items that measured perception of lack of information and isolation (adapted from Weinert et al., 2015), perception of justice, the opportunity to carry out dishonest academic behavior, technical and contextual obstacles in monitoring of distance classes and satisfaction with the educational innovation project. We focus our attention on the seven items corresponding to the perception of lack of information and isolation displayed in Table 1.

In Table 2 we show the results of the calculation of the Cohen's d for the seven items with this teaching method when we consider the interval-valued

Table 1 Constructs and Measurement Items

	INFORMATION UNDERSUPPLY
$I.1$	I receive too little information from my classmates
$I.2$	It is difficult to receive relevant information from my classmates
$I.3$	It is difficult to receive relevant information from the teacher
$I.4$	The amount of information I receive from my classmates is very low
$I.5$	The amount of information I receive from the teacher is very low
	ISOLATION
$I.6$	I feel less integrated in my team at class
$I.7$	I feel poorly informed about the relevant issues from my team at class

scale. Between parentheses, we show the approximate p-values obtained applying the bootstrapped two-sample test about means with fuzzy rating scale-based data for independent samples (see, for instance, Lubiano et al., 2016), since interval-valued data is a particular case of trapezoidal fuzzy data.

Table 2 Analyzing the influence of respondent's sex to the perception of Information undersupply (Items 1-5) and Isolation (Items 6-7)

Cohen's d (p-value)	FACE-TO-FACE	ONLINE
$I.1$	.629 (*.027*)	.225 (*.502*)
$I.2$	.812 (*.003*)	.508 (*.082*)
$I.3$	.355 (*.196*)	.174 (*.594*)
$I.4$	.853 (*.002*)	.718 (*.012*)
$I.5$	.397 (*.134*)	.359 (*.192*)
$I.6$	.392 (*.160*)	.837 (*.002*)
$I.7$	.635 (*.021*)	.562 (*.030*)

Figure 2 show the sample means for 18 men (black) and 32 women (gray).

According to Cohen (1988), values between .2 to .49, .50 to .79, and .80 and higher are considered small, medium and large, respectively.

We can observe that respondent's sex is a non-significant factor with a small-medium effect size both face-to-face and online modality for items $I.3$ and $I.5$ related to the information undersupply received by the teacher (values of d between .174 to .397).

On the other hand, the respondent's sex is a significant factor with a medium-large effect size with both teaching modality (values of d over than .562) for items $I.4$ and $I.7$.

When we analyze the influence of sex for question $I.1$, if responses have been obtained with the face-to-face teaching modality, it could be concluded that this factor is significant with a medium-effect size of d = .629, while

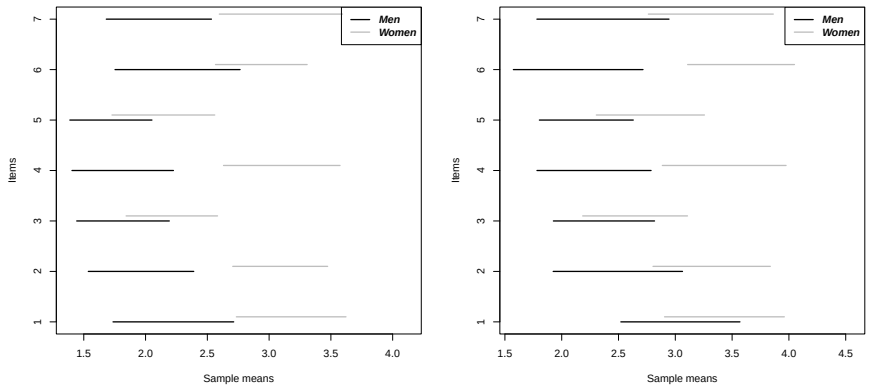

Fig. 2 Graphical sample means for men *vs.* women with face-to-face (left) and online (right) methodoly

it is not significant for online modality with a small-effect size of $d = .225$. Nevertheless, for question $I.6$, sex is significant when teaching modality is online and the effect size is large ($d = .837$) but it is not significant with the face-to-face modality with a small-effect size ($d = .392$).

5 Conclusions and Future Research

In this paper we have stated an extended version of Cohen's d index when the random experiment involve interval-valued data.

Besides, it would be desirable to study the properties of these sample measures estimating the corresponding population measure (like unbiasedness, consistency, and so on).

In a similar way, it is possible to calculate this index with fuzzy data obtained from the responses of fuzzy rating scales-based questions (Castaño et al., 2020). By extending the concept of effect size to more complex type of data, it will be possible to compare research results concerning both fuzzy or interval-valued data and real-valued data.

On the other hand, we are now studying other effect-size measures of difference on means and the extension of these concepts to other situations.

Acknowledgements The research in this paper has been partially supported by from Principality of Asturias Grant AYUD/2021/50897, and the Spanish Ministry of Economy and Business Grant PID2019-104486GB-I00. Their financial support is gratefully acknowledged. The authors would like to the reviewers for valuable and helpful comments to improve the quality of this work.

References

American Psychological Association. (2010). Publication manual of the American Psychological Association (6th ed.). APA, Washington, D.C.

Aumann, R.J. (1965). Integrals of set-valued functions. J. Math. Anal. Appl. **12** 1–12

Bertoluzza, C., Corral, N., Salas, A. (1995). On a new class of distances between fuzzy numbers. Mathware Soft Comput. **2**, 71–84

Castaño, A.M.; Lubiano, M.A.; García-Izquierdo, A.L. (2020). Gendered Beliefs in STEM Undergraduates: A Comparative Analysis of Fuzzy Rating versus Likert Scales. Sustainability **12**, 6227.

Cohen, J. (1965). Some statistical issues in psychological research. In: Wolman, B.B. (ed.) Handbook of clinical psychology, pp. 95–121. Academic Press, New York

Cohen, J. (1969). Statistical Power Analysis for the Behavioral Sciences, 1st ed. Academic Press, New York

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences, 2nd ed. L. Erlbaum Associates, Hillsdale, N.J.

Ellerby, Z., Wagner, C., Broomell, S.B. (2021). Capturing richer information: On establishing the validity of an interval-valued survey response mode. Behav. Res. (in press)

Gil, M.Á., Lubiano, M.A., Montenegro, M., & López, M.T. (2002). Least squares fitting of an affine function and strength of association for interval-valued data. Metrika **56**, 97-111

Hays, W.L. (1963). Statistics for psychologists. Holt, Rinehart and Winston, New York

Hedges, L.V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. J. Educ. Stat. **6**(2), 106–128

Körner, R. (1997). On the variance of fuzzy random variables. Fuzzy Sets Syst. **92**(1), 83–93

Likert, R. (1932). A technique for the measurement of attitudes. Arch. Psychol. **22**, 140–155

Lubiano, M.A., Gil, M.Á., López-Díaz, M., López, M.T. (2000). The $\overrightarrow{\lambda}$-mean squared dispersion associated with a fuzzy random variable. Fuzzy Sets Syst. **111**, 307–317

Lubiano, M.A., Montenegro, M., Sinova, B., De la Rosa de Sáa, S., Gil, M.Á. (2016). Hypothesis testing for means in connection with fuzzy rating scale-based data: algorithms and applications. Eur. J. Oper. Res. **251**, 918–929

Matheron, G. (1975). Random Sets and Integral Geometry. J. Wiley & Sons, New York

Minkowski, H. (1903). Vorlumen und Oberflache. Mathematische Annalen **57**, 447–495

Themistocleous, C., Pagiaslis, A., Smith, A., & Wagner, C. (2019). A comparison of scale attributes between interval-valued and semantic differential scales. Int. J. Market Res. **61**(4), 394–40

Vitale, R.A. (1985). L_p metrics for compact, convex sets. J. Approx. Theory **45**(3), 280–287

Wagner, C., Miller, S., Garibaldi, J.M., Anderson, D.T., Havens, T.C. (2015). From interval- valued data to general type-2 fuzzy sets. IEEE Trans. Fuzzy Syst. **23**, 248–269

Weinert, C., Maier, C., Laumer, S. (2015). Why are teleworkers stressed? An empirical analysis of the causes of telework-enabled stress. In: Thomas, O., Teuteberg, F. (eds) 12th international conference on Wirtschaftsinformatik, pp. 1407–1421. Available at `https://aisel.aisnet.org/wi2015/94`

Wilkinson, L., Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. American Psychologist **54**, 594–604