A Meta-analysis Approach for Estimating Salary Mean and Its Confidence Interval

Flora Ferreira[1(✉)], José Soares[1], Fernanda Sousa[2], Filipe Magalhães[2], Isabel Ribeiro[2], Dânia Pinto[1], and Pedro Pacheco[1,2]

[1] BERD - Bridge Engineering Research and Design, Matosinhos, Portugal
{flora.ferreira,jose.soares,dania.pinto,pedro.pacheco}@berd.eu
[2] Faculty of Engineering (FEUP), University of Porto, Porto, Portugal

Abstract. A meta-analysis is the statistical pooling of the summary statistics from several selected studies to estimate the outcome of interest. A job's salary estimate is important information for both job applicants and companies, that is reported on different websites. By combining data from different sources a mean estimate more representative of the target population can be obtained, especially when the data has high variability and dependence on different factors, as in the salary case. However, data are not reported in each source by the same statistics values. Sometimes, the data are summarized by reporting the sample median and one or both of (i) the minimum and maximum values and (ii) the first and third quartiles. Additionally, the sample size is not always reported. In this paper, we aim to provide a step-by-step process to estimate the salary mean and its confidence interval by combining the data from different sources. We illustrate the process with an example dataset of seven job titles. The performance of two different alternatives to estimate the sample mean is evaluated, and the variation in the outcomes between websites is discussed. A high range of other quantitative data could benefit from the proposed process to obtain an estimate representative of the target population.

Keywords: Confidence interval · Meta-analysis · Pool data · Salary estimate

1 Introduction

Meta-analysis is the statistical combination of results from a set of different studies to generate a summary (pooled) estimate. By combining the data from different sources, an outcome estimate more representative of the target population can be provided [3]. Websites like glassdoor.com, salary.com, and payscale.com

This article is a result of the project BEIS (Bridge Engineering Information System), supported by Operational Programme for Competitiveness and Internationalisation (COMPETE 2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF).

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
O. Gervasi et al. (Eds.): ICCSA 2022 Workshops, LNCS 13377, pp. 627–639, 2022.
https://doi.org/10.1007/978-3-031-10536-4_41

offer information about salaries, including an estimation of how much a person should be earning based on job title and location. A job's salary estimate is important information for both job applicants and companies [14]. When looking for new job opportunities, the likely salary range helps job seekers understand which professions and locations offer higher average salaries, and on the other hand, helps employers understand what the employee hopes to receive and in making recruitment-related decisions. Getting a salary range estimation by combining the data from different websites will provide a salary estimate more representative of the source population. Ideally, to provide a quantitative estimation of a continuous variable, the sample size, mean, and standard deviation are required [18]. However, sometimes, the statistics reported differs from source to source. Instead, the sample mean and standard deviation, the median, the minimum, and maximum values, and/or the first and third quartiles are reported. In order to pool the results reported in the different sources in a consistent format is important to estimate the sample mean and standard deviation. Different methods have been proposed for estimating the sample mean and the sample standard deviation from the median, range, and/or interquartile range [2,7,9,12,18]. The methods proposed by Luo et al. [9] and Want et al. [18] are the formula-based methods with the best performance in estimating the sample mean and standard deviation, respectively. In these methods, the sample size is used, and higher accuracy for the estimation of the sample mean and standard deviation is achieved compared with other methods, such as proposed in [2,7], where the sample size is not fully used or even ignored. However, in some cases, the sample size is not reported and in these cases, alternative estimators independent of the sample size are useful. The methods proposed in [2,7,9,18] are derived from the assumption that the outcome variable is normally distributed. To overcome the limitation of the normality assumption, in [12], two approaches, Box-Cox (BC) method and Quantile Estimation (QE) method, were proposed to estimate the sample mean and standard deviation when data are suspected to be non-normal. The results presented in [12] show that BC and QE methods often outperformed the methods developed by Luo et al. [9] and Want et al. [18] when applied to non-normal data.

In this paper, with the purpose to obtain a range estimation of a job's salary representative of the reality of the market, we present a step-by-step process to estimate the mean salary and its confidence interval by combining the data from different websites.

In the following section, we introduce the data reported on the websites considered in this study. Taking into account the different sets of summary statistics that are reported on the websites, in Sect. 3, we describe the existing methods for estimating the sample mean and standard deviation. The approaches to meta-analyze the mean and its confidence interval is also introduced in Sect. 3. We describe the proposed process in Sect. 4. The results obtained based on an example dataset of seven job titles are presented and discussed in Sect. 5. The conclusion and future work are presented in Sect. 6.

2 Case Study

Throughout the paper, we use the following notation for sample summary statistics:

a = the minimum value,
q_1 = the first quartile,
$\widetilde{x}$ = the median,
$\overline{x}$ = the mean,
q_3 = the third quartile,
b = the maximum value,
n = the sample size.

In this study, five salary research websites were considered. Table 1 gives a summary of all statistics reported on these websites. On glassdoor.com website the 10th and 90th percentiles as low and high values, respectively, and the sample size are reported. In this study, we assume the low and the high values as minimum and maximum values, respectively, of $n = 0.80\times$ sample size.

Table 1. Summary table of statistics estimates report in each website considered in this study

Website	$\overline{x}$	$\widetilde{x}$	a	q_1	q_3	b	n
averagesalarysurvey.com	X						X
glassdoor.com		X	X			X	X
payscale.com		X		X	X		X
salaryexplorer.com	X	X	X			X	
salaryexplorer.com	X	X	X	X	X	X	

3 Methods

3.1 Sample Mean and Standard Deviation Estimators

Let $\overline{X}$ and S denote estimates of the sample mean and the sample standard deviation, respectively. Based on previous studies [2,7,9,12,18] and the data reported in the different websites (Table 1), we consider the following sets of summary statistics denoted by Scenario 1 ($\mathcal{S}_1$), Scenario 2 ($\mathcal{S}_2$), Scenario 3 ($\mathcal{S}_3$), and Scenario 4 ($\mathcal{S}_4$):

$$\mathcal{S}_1 = \{\widetilde{x}, a, b, n\},$$

$$\mathcal{S}_2 = \{\widetilde{x}, q_1, q_3, n\},$$

$$\mathcal{S}_3 = \{\widetilde{x}, a, b\},$$

$$\mathcal{S}_4 = \{\widetilde{x}, a, q_1, q_3, b\}.$$

In scenarios $\mathcal{S}_1$ and $\mathcal{S}_2$, where the sample size is known, Luo et al. [9] proposed the following sample mean estimators, respectively:

$$\overline{X} = \left(\frac{4}{4+n^{0.75}}\right)\frac{a+b}{2} + \left(\frac{n^{0.75}}{4+n^{0.75}}\right)\widetilde{x}, \tag{1}$$

$$\overline{X} = \left(0.7+\frac{0.39}{n}\right)\frac{q_1+q_3}{2} + \left(0.3-\frac{0.39}{n}\right)\widetilde{x}. \tag{2}$$

For scenario $\mathcal{S}_3$ where the median, the minimum and maximum are reported, the following estimator for the sample mean proposed by Hozo et al. [7] can be used:

$$\overline{X} = \frac{a+2\widetilde{x}+b}{4}. \tag{3}$$

For scenarios $\mathcal{S}_4$, following the same idea in Hozo et al. [7], Bland [2] proposed the following estimator for the sample mean:

$$\overline{X} = \frac{a+2q_1+2\widetilde{x}+2q_3+b}{8}. \tag{4}$$

The sample standard deviation estimators in scenarios $\mathcal{S}_1$ and $\mathcal{S}_2$ are given by [18]:

$$S = \frac{b-a}{2\Phi^{-1}\left(\frac{n-0.375}{n+0.25}\right)}, \tag{5}$$

and

$$S = \frac{q_3-q_1}{2\Phi^{-1}\left(\frac{0.75n-0.125}{n+0.25}\right)}, \tag{6}$$

respectively, where $\Phi^{-1}(z)$ is the inverse function of $\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$, that is, the upper zth percentile of the standard normal distribution.

In scenario $\mathcal{S}_3$, the value of the sample size is unknown. In this case, the following range rule of thumb can be used to estimate the sample standard deviation:

$$S = \frac{b-a}{4}. \tag{7}$$

Hozo et al. [7] showed that the formula (7) is a good estimator when the sample size range between 15 and 70. For small ($n \leq 15$) and large ($n > 70$) sample sizes Hozo et al. [7] proposed other formulas. However, the lack of knowledge of the value of n makes it not possible to choose the formula based on n.

In the case of scenario S_4, Bland [2] considered the following estimator for the sample standard deviation:

$$S = \left(\tfrac{1}{16}(a^2+2q_1^2+2\widetilde{x}^2+2q_3^2+b^2) + \tfrac{1}{8}(aq_1+q_1\widetilde{x}+\widetilde{x}q_3+q_3b) - \tfrac{1}{64}(a+2q_1+2\widetilde{x}+2q_3+b)^2\right)^{\frac{1}{2}}. \tag{8}$$

Summarily, Table 2 presents the proposed estimators under different scenarios based on previous studies [2,7,9,18].

Table 2. Summary table for estimating the sample mean and standard deviation under different scenarios

	$\mathcal{S}_1$	$\mathcal{S}_2$	$\mathcal{S}_3$	$\mathcal{S}_4$
$\overline{X}$	Eq. (1)	Eq. (2)	Eq. (3)	Eq. (4)
S	Eq. (5)	Eq. (6)	Eq. (7)	Eq. (8)

3.2 Box-Cox Transformations

The above estimators perform well when the distribution of the outcome variable is symmetric (normal). McGrath et al. [12] proposed to apply Box-Cox transformations before employing the methods of Luo et al. [9] and Wan et al. [18] to transform a dataset into a more symmetric (normally-distributed) dataset. The proposed method, which McGrath et al. [12] denote by BC method, consists of the following four steps:

1. The power parameter λ is obtained through an optimization algorithm such that the distribution of the transformed data is most likely to be symmetric and therefore most normally distributed.
2. The quantiles of x are converted into the quantiles of $f_\lambda(x)$, where f_λ represent the Box-Cox transformation.
3. The mean and standard deviation of $f_\lambda(x)$ are estimated using the methods Luo et al. [9] and Wan et al. [18], respectively.
4. At last, the mean and standard deviation of $f_\lambda(x)$ are inversely converted into the mean and standard deviation of $x = f_\lambda^{-1}(y)$.

The optimization step consists, specifically, to find a finite value of λ such that:

$$f_\lambda(q_3) - f_\lambda(\widetilde{x}) = f_\lambda(\widetilde{x}) - f_\lambda(q_1) \tag{9}$$

in $\mathcal{S}_2$,

$$f_\lambda(b) - f_\lambda(\widetilde{x}) = f_\lambda(\widetilde{x}) - f_\lambda(a) \tag{10}$$

in $\mathcal{S}_3$, and

$$\arg\min_\lambda \left[\left(f_\lambda(q_3) - 2f_\lambda(\widetilde{x}) + f_\lambda(q_1)\right)^2 + \left(f_\lambda(b) - 2f_\lambda(\widetilde{x}) + f_\lambda(a)\right)^2 \right] \tag{11}$$

in $\mathcal{S}_4$.

The description of the optimization algorithm used to find λ can be found in Appendix B of [12].

Box-Cox transformation function f_λ and respective inverse Box-Cox transformation function f_λ^{-1} are defined as follows:

$$f_\lambda(x_i) = y_i = \begin{cases} \frac{x_i^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln(x_i) & \text{if } \lambda = 0 \end{cases}, \tag{12}$$

$$f_\lambda^{-1}(y_i) = x_i = \begin{cases} (\lambda y_i + 1)^{1/\lambda} & \text{if } \lambda \neq 0 \\ e_i^y & \text{if } \lambda = 0 \end{cases}. \tag{13}$$

As in [12], Monte-Carlo simulation was used to compute the mean and standard deviation of $x \sim f_\lambda^{-1}(N'(\mu, \sigma))$, where $N'(\mu, \sigma^2)$ is the symmetrically truncated normal distribution with mean μ and variance σ^2 bounded within support $[f_\lambda(0), 2\mu - f_\lambda(0)]$ (see [12] for more details).

3.3 Measuring Symmetry

To quantify the degree of symmetry of a distribution, different measures of skewness have been proposed [10] based on different statistics measures. Based on the extremes values (maximum and minimum) and the median Galip Altinay [1] proposed the following coefficient of skewness:

$$SK_G = \frac{b + a - 2\widetilde{x}}{b - a}. \tag{14}$$

Based on the first and third quartiles and the median, the Bowley's coefficient of skewness is given by:

$$SK_B = \frac{q_3 + q_1 - 2\widetilde{x}}{q_3 - q_1}. \tag{15}$$

3.4 Measuring Heterogeneity

Heterogeneity refers to the variation in study outcomes between studies. When the difference in results between studies is higher, the summary estimate may not be representative of individual studies. Then, the heterogeneity among studies considering the similarity of point estimates should be evaluated. The classical measure of heterogeneity/inconsistency is Cochran's Q, which is given by:

$$Q = \sum_{i=1}^{k} w_i \left(\overline{X}_i - \widehat{\mu}\right)^2, \tag{16}$$

where w_i is the weighting factor associated to the study i, $\overline{X}_i$ is the mean of study i, $\widehat{\mu}$ is the global estimated mean, and k the number of studies. The weights w_i are estimated by

$$w_i = \left(S_i^2 + \widehat{\tau}^2\right)^{-1}, \tag{17}$$

where S_i is the estimated standard deviation of ith study and $\widehat{\tau}^2$ is an estimate between-study variation [8]. One commonly used method for estimating is the method of moments estimator by DerSimonian and Laird [4]:

$$\widehat{\tau}^2_{DL} = \max\left\{0, \frac{\left[\sum_{i=1}^k (S_i)^{-2}(\overline{X}_i - \widehat{\mu}_0)^2\right] - (n-1)}{\sum_{i=1}^k (S_i)^{-2} - \sum_{i=1}^k (S_i)^{-4} / \sum_{i=1}^k (S_i)^{-2}}\right\}, \tag{18}$$

where $\widehat{\mu}_0$ is defined as

$$\widehat{\mu}_0 = \frac{\sum_{i=1}^{k}(S_i)^{-2}\overline{X}_i}{\sum_{i=1}^{k}(S_i)^{-2}}. \tag{19}$$

Statistic I^2 that is calculated from Cochran's Q the corresponding degrees of freedom $(k-1)$ is an intuitive and simple measure of heterogeneity among studies, usually used in meta-analysis [8] and defined by [6]:

$$I^2 = \begin{cases} 0 & \text{if } Q \leq k-1 \\ \frac{Q-(k-1)}{Q} \times 100\% & \text{if } Q > k-1 \end{cases}. \tag{20}$$

If the value of I^2 is between 0% to 40%, 30% to 60%, 50% to 90%, and 75% to 100%, then may represent low, moderate, substantial, and considerable heterogeneity, respectively [3].

3.5 Estimating the Mean and Its Confidence Interval

An estimator for μ is given by:

$$\widehat{\mu} = \frac{\sum_{i=1}^{k} w_i \overline{X}_i}{\sum_{j=1}^{k} w_i} \tag{21}$$

and a $100(1-\alpha)\%$ confidence interval (CI) for the mean, with $\alpha \in (0,1)$, can be calculated as [5]:

$$\left(\widehat{\mu} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{\sum_{j=1}^{k} w_j}}, \widehat{\mu} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{\sum_{j=1}^{k} w_j}}\right), \tag{22}$$

where $z_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$th quantile of the standard normal distribution and w_i is given by (17).

In the meta-analysis, there are two statistical procedures: fixed- and random-effects [5]. In the fixed-effects meta-analysis, it is assumed that the between-studies variance is zero ($\tau^2 = 0$), and only the internal sampling variability of each of the studies is considered. By contrast, in random-effects meta-analysis, between-study variance ($\tau^2 > 0$) is considered enabling the modeling of differences between studies. We estimate both a fixed-effect and random-effect pooled mean (and its 95% CI) denoted by $\widehat{\mu}_{FE}(CI_{FE})$ and $\widehat{\mu}_{RE}(CI_{RE})$, respectively. For fixed-effect case we assume $\widehat{\tau}^2 = 0$ and for random-effect case we estimate $\widehat{\tau}^2$ using the formula (18).

4 Proposed Process

The five main steps of the proposed process to estimate the sample mean and its confidence interval are briefly described in Fig. 1. First, a collection of the summary statistics should be collected from different sources (websites). Then,

according to the set of summary statistics, the identification of the scenario ($\mathcal{S}_1$, $\mathcal{S}_2$, $\mathcal{S}_3$, $\mathcal{S}_4$ or not defined) must be identified for each source. The relative precision of the estimators depends heavily on the skewness of the data [11,12]. While for symmetry data, the sample mean and standard deviation estimators based on raw quantiles perform best, for skewness data BC method allows to obtain estimates with lower relative error [12]. Therefore, we recommend selecting the estimators' data based on the skewness. Thus, the third step consists of the calculation of the coefficients of skewness. According to the scenario, the sample mean and standard deviation estimates are calculated in the fourth step. The quantiles (raw or BC transformed) associated with the lower coefficient skewness should be used to estimate the two statistics. In the fifth and last step, the mean and its 95% CI as well the heterogeneity measures are obtained.

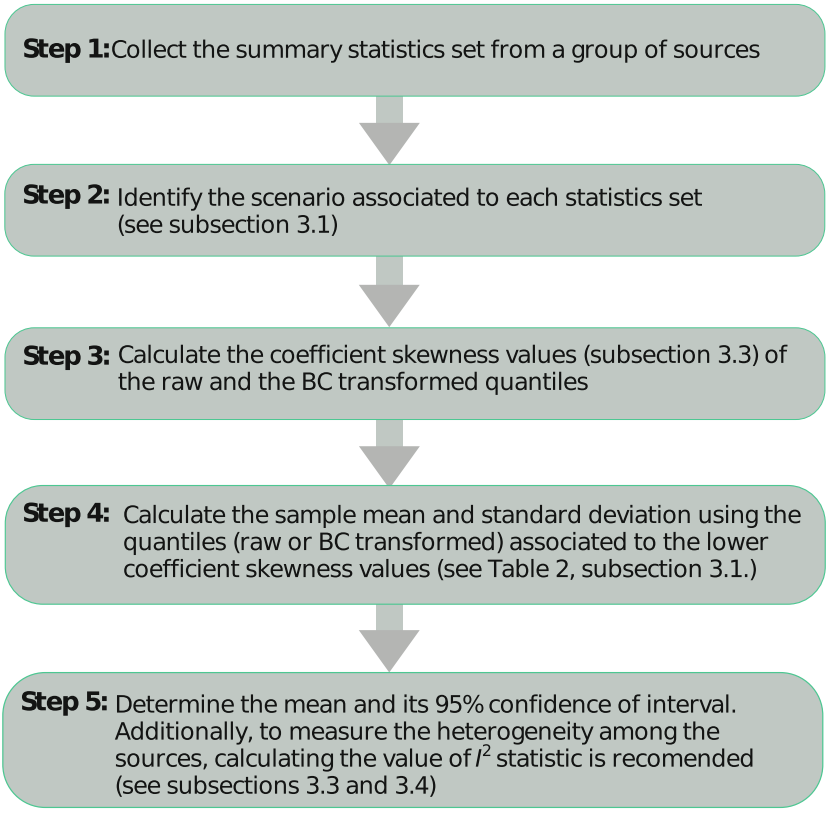

Fig. 1. Recommended steps to obtain the estimating of salary mean and its CI.

5 Results

For this study, a set of seven job titles was randomly chosen to implement the proposed process. The search of the annual salary in euros was done for a selected

country on 28 February 2022 (see Table 3). Since in averagesalarysurvey.com just the sample mean and sample size present in Table 3 are available, these data were not used to estimate the mean and the CI. This sample mean was used to verify if the estimated CI includes it (see Subsect. 5.2).

Table 3. Sample mean and sample size of the annual salary in euros reported in averagesalarysurvey.com

j	Job title	Country	$\overline{x}$	n
1	Account Manager	Italy	58437	14
2	Architect	France	37412	20
3	Civil Engineer	Spain	54461	15
4	Financial Analyst	Germany	63414	39
5	Marketing Manager	France	70922	25
6	Project Manager	Spain	51756	27
7	Quality Manager	Germany	57820	3

The Python programming language using open-source software libraries and tools, including Pandas [13], NumPy [15] and SciPy [17] were used to obtain all estimates presented in this section.

5.1 Sample Mean and Standard Deviation Estimation

In the case, of the websites salaryexplorer.com and salaryexplorer.com, as the sample mean value is known, just the standard deviation needs to be estimated. The relative error (RE) of the sample mean was used as a performance measure. The RE is given by:

$$RE(\widehat{x_j}) = \frac{|\overline{X}_j - \overline{x}_j|}{\overline{x}_j}, \tag{23}$$

where j correspond to the job enumerated as $j = 1, ..., 7$.

Table 4 reports the coefficients of skewness (Galip's coefficient, SK_G and/or Bowley's coefficient, SK_B) based on raw and BC transformed quantiles, and the RE values when the sample mean is known. For the data collected from the websites glassdoor.com, payscale.com, and salaryexplorer.com, $\mathcal{S}_1$, $\mathcal{S}_2$, and $\mathcal{S}_3$, the coefficients of skewness values are lower for the BC transformed quantiles, and as expected, in $\mathcal{S}_3$, we can confirm that the relative error for the mean estimated based on BC transformed quantiles (BC method) is lower compared with the mean estimated using directly the raw quantiles. Contrarily, in the case of the data collected from salaryexplorer.com, $\mathcal{S}_4$, the RE is lower when the mean is obtained using the raw quantiles. In fact, while SK_G values are lower for BC transformed quantiles, SK_B values are lower for raw quantiles.

Table 4. Coefficients of skewness based on raw and BC transformed quantiles, and the RE values when the sample mean is known

Website	Job	Raw quantiles			BC transformed quantiles		
(Scenario)	(j)	SK_G	SK_B	$RE(\widehat{x_j})$	SK_G	SK_B	$RE(\widehat{x_j})$
glassdoor.com	1	0.43	–	–	0.23	–	–
$(\mathcal{S}_1)$	2	0.13	–	–	<0.01	–	–
	3	0.15	–	–	<0.01	–	–
	4	0.15	–	–	<0.01	–	–
	5	0.24	–	–	0.06	–	–
	6	0.09	–	–	<0.01	–	–
	7	−0.13	–	–	<0.01	–	–
payscale.com	1	–	0.17	–	–	0.07	–
$(\mathcal{S}_2)$	2	–	0.31	–	–	0.21	–
	3	–	0.21	–	–	0.10	–
	4	–	0.28	–	–	0.21	–
	5	–	0.11	–	–	0.03	–
	6	–	−0.08	–	–	<0.01	–
	7	–	−0.08	–	–	<0.01	–
salaryexplorer.com	1	−0.18	–	0.052	<0.01	–	0.035
$(\mathcal{S}_3)$	2	−0.06	–	0.021	<0.01	–	0.017
	3	−0.06	–	0.021	<0.01	–	0.017
	4	−0.05	–	0.020	<0.01	–	0.017
	5	−0.18	–	0.053	<0.01	–	0.035
	6	−0.05	–	0.019	<0.01	–	0.016
	7	−0.06	–	0.021	<0.01	–	0.018
salaryexplorer.com	1	0.13	−0.11	0.039	0.05	−0.16	0.042
$(\mathcal{S}_4)$	2	0.22	−0.14	0.070	−0.22	−0.22	0.075
	3	0.17	−0.13	0.054	0.05	−0.19	0.058
	4	−0.10	−0.04	0.060	−0.01	0.02	0.065
	5	0.17	−0.12	0.053	0.05	−0.18	0.057
	6	−0.03	−0.06	0.027	0.02	−0.03	0.028
	7	−0.10	−0.10	0.060	−0.01	0.02	0.064

5.2 Salary Mean Estimate and Its CI

Table 5 presents the salary mean estimate and its CI obtained by undertaking both a fixed-effect and a random-effects meta-analysis for the seven job titles, as well as the estimated heterogeneity values. In the third case, jobs 2, 3, and 7, the statistics I^2 was 0%, indicating no heterogeneity across the sources. For these

Table 5. Salary mean estimate and its 95%CI

Job(j)	$\overline{x}(n)$[a]	$\widehat{\mu}_{FE}$	(CI_{FE})	I^2_{FE}	$\widehat{\mu}_{RE}$	(CI_{RE})	I^2_{RE}
1	58437(14)	40089	(32179,47999)	31%	42520	(31373,53667)	0%
2	37412(20)	59358	(49269,69448)	0%	59358	(49269,69448)	0%
3	54461(15)	39524	(32138,46910)	0%	39524	(32138,46910)	0%
4	63414(39)	63130	(53311,72948)	11%	63143	(52372,73915)	0%
5	70922(25)	54953	(45314,64592)	51%	60259	(42692,77827)	9%
6	51756(27)	44800	(37541,52059)	2%	44849	(37440,52258)	0%
7	57820(3)	65714	(55494,75934)	0%	65714	(55494,75934)	0%

[a]Sample mean ($\overline{x}$) and sample size(n) of the annual salary in euros reported in averagesalarysurvey.com (see Table 3).

cases, the obtain mean estimate and respective CI for FE and RE are equal. On the other hand, in the presence of heterogeneity ($I^2 > 0$), jobs 1,4,5, and 6, the confidence intervals CI_{RE} are wider than CI_{FE}, as expected.

Checking if the salary mean reported on averagesalarysurvey.com website (Table 5) that was not used to estimate the confidence of intervals belongs to the corresponding CIs, we can observe that in 3 of the 7 job titles, jobs 1,2, and 3, the salary mean does not belong to both estimated CIs. For job 5, the case with the highest I^2 values, the salary mean reported in averagesalarysurvey.com does not belong to CI_{FE} but belongs to CI_{RE}. In the remaining cases, jobs 4, 6, and 7, the salary mean belongs to both CI_{FE} and CI_{RE}.

6 Discussion and Future Work

In this paper, we proposed a step-by-step process for estimating the salary mean and its confidence interval by pooling means estimated via transformation-based approaches. The pooled salary estimate is more representative of the reality in the market than the salary estimate from an individual source. In our results, the confidence intervals were estimated based on data collected from 4 websites. Checking whether the salary mean reported on another website is not used to determine whether the CIs belongs (or not) to the CIs obtained (Table 5), we can observe that for 4 job titles the salary mean reported on averagesalarysurvey.com belongs to at least one of the estimated CI. However, in the remaining 3 job titles, this did not happen. A well-conducted meta-analysis based on data collected from a larger number of sources may provide a more valid and reliable estimate of the outcome of interest [16]. Then, for the salary estimate to be more representative of the salary paid a higher number of websites is recommended.

The salary mean and its confidence interval were obtained both around the random-effects summary estimate and around a fixed-effect summary estimate. When there is no heterogeneity among the studies, the fixed-effect method and the random-effects method will give identical results. When heterogeneity is present, a confidence interval around the random-effects and the fixed effect summary estimate is different, including a wider CI in the case of random-effects. The

decision between fixed- and random-effects meta-analyses should not be made on basis of heterogeneity measures, and there is no universal recommendation [3]. A fixed-effect meta-analysis would be preferred if the studies considered are truly identical. However, if the studies differ (e.g. different approaches to acquiring the data) a random-effects meta-analysis is usually more appropriate. In job's salary depends on many factors, including the level of experience, location, possessed skills, and others, we recommend the use of the random-effects analysis. For instance, the salary reported on two websites can represent a salary estimate of the two different groups of workers in terms of location or experience. The salary estimate was obtained based on the sample mean and standard deviation calculated from the median and measures of spread (range and/or interquartile range). Better is the estimate as closer to a symmetric distribution is the data. Following the idea of McGrath et al. [12] the incorporation of Box-Cox transformations into the methods for estimating the sample mean and standard deviation, is a possible solution to overcome the limitation of the assumption of normality when the data is non-normal. For example, in the case of the data collected from salaryexplorer.com, Scenario $\mathcal{S}_3$, the relative errors were all lower when the Box-Cox transformations were incorporated in the process of estimating the sample mean (Table 4). However, if the transformed are not normally distributed, then estimating the mean and standard deviation of the underlying distribution may not be appropriate. In fact, in the case of the data collected from salaryexplorer.com, scenario $\mathcal{S}_4$, the relative error does not decrease with the incorporation of the Box-Cox transformation. In the cases where the studies report medians and heavily skewed data, directly meta-analyzing medians may be better-suited [11]. In future work, we intend to investigate the performance of median-based approaches to estimating a salary range.

In this study, a limited number of sets of summary statistics (4 scenarios) were considered. However, on some sources, a different set of summary statistics including other statistics, such as 10th and 90th percentiles, could be collected. Therefore, in future research, the inclusion of other scenarios should be considered.

Despite some limitations presented, a high range of other quantitative data could benefit from the proposed process to obtain a mean value (and its confidence of interval) representative of the target population.

References

1. Altinay, G.: A simple class of measures of skewness (2016)
2. Bland, M.: Estimating mean and standard deviation from the sample size, three quartiles, minimum, and maximum. Int. J. Stat. Med. Res. **4**(1), 57–64 (2015)
3. Deeks, J.J., Higgins, J.P., Altman, D.G., Group, C.S.M.: Analysing data and undertaking meta-analyses. Cochrane handbook for systematic reviews of interventions, pp. 241–284 (2019)
4. DerSimonian, R., Kacker, R.: Random-effects model for meta-analysis of clinical trials: an update. Contemp. Clin. Trials **28**(2), 105–114 (2007)

5. Hedges, L.V., Vevea, J.L.: Fixed-and random-effects models in meta-analysis. Psychol. Methods **3**(4), 486 (1998)
6. Higgins, J.P., Thompson, S.G., Deeks, J.J., Altman, D.G.: Measuring inconsistency in meta-analyses. Bmj **327**(7414), 557–560 (2003)
7. Hozo, S.P., Djulbegovic, B., Hozo, I.: Estimating the mean and variance from the median, range, and the size of a sample. BMC Med. Res. Methodol. **5**(1), 1–10 (2005)
8. Huedo-Medina, T.B., Sánchez-Meca, J., Marin-Martinez, F., Botella, J.: Assessing heterogeneity in meta-analysis: Q statistic or i^2 index? Psychol. Methods **11**(2), 193 (2006)
9. Luo, D., Wan, X., Liu, J., Tong, T.: Optimally estimating the sample mean from the sample size, median, mid-range, and/or mid-quartile range. Stat. Methods Med. Res. **27**(6), 1785–1805 (2018)
10. Mahmoudi, M.R., Nasirzadeh, R., Baleanu, D., Pho, K.H.: The properties of a decile-based statistic to measure symmetry and asymmetry. Symmetry **12**(2), 296 (2020)
11. McGrath, S., Zhao, X., Qin, Z.Z., Steele, R., Benedetti, A.: One-sample aggregate data meta-analysis of medians. Stat. Med. **38**(6), 969–984 (2019)
12. McGrath, S., Zhao, X., Steele, R., Thombs, B.D., Benedetti, A., Collaboration, D.S.D.D.: Estimating the sample mean and standard deviation from commonly reported quantiles in meta-analysis. Stat. Methods Med. Res. **29**(9), 2520–2537 (2020)
13. McKinney, W., et al.: Data structures for statistical computing in python. In: Proceedings of the 9th Python in Science Conference, vol. 445, pp. 51–56. Austin, TX (2010)
14. Tanwar, K., Kumar, A.: Employer brand, person-organisation fit and employer of choice: Investigating the moderating effect of social media. Personnel Review (2019)
15. Van Der Walt, S., Colbert, S.C., Varoquaux, G.: The numpy array: a structure for efficient numerical computation. Comput. Sci. Eng. **13**(2), 22–30 (2011)
16. Vetter, T.R.: Systematic review and meta-analysis: sometimes bigger is indeed better. Anesthesia Analgesia **128**(3), 575–583 (2019)
17. Virtanen, P., et al.: Scipy 1.0: fundamental algorithms for scientific computing in python. Nature Methods **17**(3), 261–272 (2020)
18. Wan, X., Wang, W., Liu, J., Tong, T.: Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. BMC Med. Res. Methodol. **14**(1), 1–13 (2014)