On the efficacy of the wisdom of crowds to forecast economic indicators

Nilton S. Siqueira Neto\textsuperscript{a} and José F. Fontanari\textsuperscript{b} \textsuperscript{e}

Instituto de Física de São Carlos, Universidade de São Paulo, Caixa Postal 369, 13560-970 São Carlos, São Paulo, Brazil

Received 20 December 2022 / Accepted 9 January 2023 / Published online 19 January 2023
© The Author(s), under exclusive licence to EDP Sciences, SIF and Springer-Verlag GmbH Germany, part of Springer Nature 2023

Abstract. The interest in the wisdom of crowds stems mainly from the possibility of combining independent forecasts from experts in the hope that many expert minds are better than a few. Hence the relevant subject of study nowadays is the \textit{Vox Expertorum} rather than Galton’s original \textit{Vox Populi}. Here we use the Federal Reserve Bank of Philadelphia’s Survey of Professional Forecasters to analyze 15455 forecasting contests to predict a variety of economic indicators. We find that the median has advantages over the mean as a method to combine the experts’ estimates: the odds that the crowd beats all participants of a forecasting contest is 0.015 when the aggregation is given by the mean and 0.026 when it is given by the median. Both aggregation methods yield a 20\% error on the average, which must be contrasted with the expected error of a randomly selected forecaster, which is about 22\%. We conclude that selective attention is the most likely explanation for the mysterious high accuracy of the crowd reported in the literature. This conclusion is strengthened by the rebuff of the claim that this high accuracy results from the cancellation of the participants’ errors: we find no meaningful correlation between the asymmetry of the distribution of the individuals’ estimates and the collective error.

1 Introduction

The phenomenon known as wisdom of crowds or Vox Populi is more than a century old, being brought to light by Galton’s 1907 study of a contest to guess the weight of an ox at the West of England Fat Stock and Poultry Exhibition in Plymouth [1] (see [2] for a brief historical account). Galton used the median of the participants’ estimates as the collective estimate and found that it overestimated the weight of the ox by less than 1\% of the true weight [1,2]. The wisdom of crowds has been brought to the attention of a wider audience by a series of business bestsellers published in the early 2000s [3–5], but the remarkable accuracy of the estimate that results from the aggregation of many independent individuals’ judgments is still somewhat of a mystery today.

A difficulty to address the wisdom of crowds is that it seems to have distinct meanings for different researchers. On the one hand, some researchers view this phenomenon as the finding that the crowd can solve problems better than most individuals in it, including experts [3]. In this perspective, the actual accuracy of the crowd estimate is not relevant: what matters is that its estimate is better than the estimate of most of the participants of the forecasting contest. In other words, the crowd is viewed as another forecaster and the focus is on its rank among the participants. On the other hand, some researchers stress the accuracy of the crowd estimate, despite the large dispersal of the individuals’ judgments, without much regard to its rank [6]. This latter perspective seems fitter to describe Galton’s original reaction to the surprising trustworthiness of the popular judgment [1]. Here, we consider extensively these two perspectives.

It is interesting that both views of the wisdom of crowds can partly be endorsed depending on the method we choose to aggregate the independent individuals’ judgments. For instance, if, on the one hand, we adopt Galton’s suggestion and choose the median of the participants’ estimates as the collective estimate, then it follows that the crowd will always beat at least half of the individuals. In fact, assuming that the true value of the ox weight is to the right of the median then the median is closer to the true value than all the individuals’ estimates that are to the left of the median, which accounts for half of the total estimates. A similar logic applies to the case the true value is to the left of the median. On the other hand, if we adopt the mean of the individuals’ judgments as the collective estimate we can easily show that the error of the collective estimate can never be greater than the average individual error, which equals the expected error of the estimate of a randomly selected participant [5]. As an aside, we note that in Galton’s ox-weighing judgment the choice...
of the mean as the collective estimate yields zero error [2].

A natural explanation for the wisdom of crowds calls upon the well-known fact that the combination of unbiased and independent estimates guarantees the accuracy of the statistical average, provided the number of estimates is large [7]. In other words, if the judgments made by numerous different people scatter symmetrically around the truth, then the collective estimate is likely to be very accurate. Somewhat surprisingly, this explanation is rather popular and claims that “the wisdom of crowds effect works if estimation errors of individuals are large but unbiased such that they cancel each other out” abound in the literature [6]. It is of course unarguably true that if the individuals’ judgments are unbiased then the wisdom of crowds effect will hold. The trouble with this explanation is the assumption that the independent individuals’ judgments are unbiased, that is, that their means coincide with the true value of the unknown quantity so that there are no systematic errors on the individuals’ estimates. The typical data do not support this assumption. In fact, the pronounced asymmetry of the distribution of individuals’ judgments has not eluded Galton’s attention, who even tried to fit it with a two-piece distribution [2]. The nonzero skewness of the distribution of individual judgments offers strong evidence for biases on the individuals’ judgments [8,9]. In addition, since there is no meaningful correlation between the asymmetry of the distribution of the individuals’ judgments and the collective estimation error [10], the cancellation of the participants’ estimation errors cannot be the general explanation for the wisdom of crowds.

Another difficulty to understand the wisdom of crowds is that much of the evidence in favor of its existence has relied on anecdotes, such as the celebrated Galton’s ox-weighing contest [3]. This may lead to the suspicion that selective attention is at play, i.e., that prominence is given to the successful outcomes only. For instance, the analysis of three wisdom of crowds experiments, viz., the estimate of the number of candies in a jar, the estimate of the length of a paper strip and the estimate of the number of pages of a book, in which the crowd estimate was given by the arithmetic mean of the participants’ judgments, resulted in 16.5, 1.8, and 28.4% errors, respectively [11]. Nevertheless, one hears mostly of the high accuracy of the crowd estimates [3–5]. As a matter of fact, it is not possible to assess the accuracy of an estimate without considering some standard of comparison and in this paper we offer some proposals in that direction. Of course, a simple and natural comparison standard is the prediction of the winner of a forecasting contest. In particular, for the three wisdom of crowds experiments mentioned above, the percent errors of the winners’ predictions were 0.9, 0.4 and 2, respectively [11].

To avoid the selective attention fallacy and aim at statistically meaningful conclusions we need to consider very many independent wisdom of crowds experiments or forecasting contests. Each experiment is fully characterized by the distribution of individuals’ estimates. Here we accomplish this by using forecasts of economic indicators that are publicly available in the Federal Reserve Bank of Philadelphia’s Survey of Professional Forecasters (FRBP-SPF) database [12]. In particular, we consider quarterly projections of 20 economic indicators, which amounts to 15455 independent forecasting contests.

In this paper we expand on a previous study of the FRBP-SPF database [10] by almost doubling the number of forecasts considered and by comparing two methods to combine the individuals’ estimates, viz., the mean and the median. In addition, we assess the accuracy of the crowd’s performance by comparing it with the standard time series forecasting algorithm, the ARIMA (AutoRegressive Integrated Moving Average) model, as well as with the winners of the forecasting contests. Overall we find that the median is superior to the mean as a means to combine the individuals’ judgments, but both choices yield a 20% error on average. This is a great performance when compared with ARIMA’s that yields a 31% error on the average and it does not fare badly when compared with the 15% error of the contests’ winners. However, since the expected error of a randomly selected forecaster is about 22%, our conclusion is that selective attention is the most likely explanation for the mysterious accuracy of the crowd. This conclusion is strengthened by the rebuff of the popular explanation that the participants’ errors cancel each other out resulting in the high accuracy of the crowd estimate.

2 The FRBP-SPF database

The Federal Reserve Bank of Philadelphia’s Survey of Professional Forecasters (FRBP-SPF) offers quarterly projections for five quarters of a variety of economic indicators since 1968 [12] (see [13] for a recent review of the statistical literature based on that database). This survey provides forecasts of the current quarter (i.e., of the quarter in which the survey is held) and of each of the next four quarters. For instance, consider a survey of a particular variable, say NGDP, held in the first quarter of 2006 (or 2006:Q1 in the notation of the documentation of the FRBP-SPF database). The last known historical quarter at that date is 2005:Q4 and the quarterly observation dates forecast are 2006:Q1 (i.e., the current quarter), 2006:Q2, 2006:Q3, 2006:Q4 and 2007:Q1. We focus on the surveys of the 20 economic indicators or variables, viz., NGDP, PGDP, CPROF, UNEMP, EMP, INDPROD, HOUSING, TBILL, TBOND, RGDP, RCONSUM, RRESIN, RRESINV, RFEDGOV, RSLOV, REXPORT, CPI, CORECPI, PCE, and COREPCE from the date they entered the FRBP-SPF database until 2020:Q4, which means we need the quarterly historical values of those variables until 2021:Q4. The variables NGDP, PGDP, CPROF, UNEMP, INDPROD, HOUSING, and RGDP were the first to enter the FRBP-SPF database (1968:Q4) and CORECPI was the
last (2007:Q1). The meaning of these economic variables can be found in the documentation of the FRBP-SPF database [12]. The only reason we have listed them here is to allow the reproducibility of our findings, as we do not distinguish them in our analysis. All forecasters are select economists and the mean number of participants for survey is \( \langle n \rangle = 35 \), whereas the minimum number is \( n_{\text{min}} = 8 \) and the maximum \( n_{\text{max}} = 87 \).

At this stage, we can define a forecasting contest (or a wisdom of crowds experiment) in the framework of the FRBP-SPF database more precisely. We recall that, similarly to Galton’s ox-weighing contest, a forecasting contest involves a group of individuals making independent judgments about the value of an unknown quantity. For instance, consider the forecast of the variable NGDP for 2021:Q4. There are five distinct forecasting contests (one for each forecast horizon) to predict the value of that variable, viz., the surveys held at 2021:Q4, 2021:Q3, 2021:Q2, 2021:Q1 and 2020:Q4. Hence each survey of a particular economic variable can be seen as a forecasting contest. We note that not necessarily the same economists participate in all those surveys since they are held at different dates.

Since the forecasters are unaware of each other’s predictions, it is not too far-fetched to assume that their predictions are independent, which seems a necessary condition for the wisdom of crowds to work [6,14]. We clump together the 20 variables listed above as well as the 5 forecast horizons to form a large ensemble of forecasting contests. Since we need to use percentage errors to compare the forecasters’ performances in those different contests, we have eliminated the contests for which the true (historical) value of the variable is zero. Fortunately, only the variables CPI and PCE have historical values equal zero and those happened for 2013:Q2 only. As there are five forecasting contests associated to each of those values, namely, the contests held at 2013:Q2, 2013:Q1, 2012:Q4, 2012:Q3, and 2012:Q2, we have eliminated solely 10 forecasting contests in total. Overall the total number of forecasting contests in our study is 15455. This amount contrasts with previous (mostly anecdotal) analyses that considered only a few forecasting contests.

As pointed out, the FRBP database considered here consists of predictions of expert economists, which contrasts with the common, but essentially incorrect, view of the wisdom of crowds as the aggregation of predictions of ordinary people. Of course, the wisdom of crowds as a method of information aggregation depends on the presence of an expressive number of experts in the crowd as nothing good can come from the aggregation of random information [4]. In fact, the interest of business-minded researchers in the wisdom of crowds stems from the possibility of combining the forecasts of experts in the hope that many expert minds are better than a few. It is interesting that just after the publication of Galton’s paper on the ox-weighing contest, it was pointed out that Galton had reported the efficacy of Vox Expertorum rather than of Vox Populi, as the participants of the contest were most likely butchers and farmers whose livelihood depended on their ability to judge the weight of farm animals before trading [15]. Moreover, the 6 pence tickets probably deterred the entrance of dilettantes in the contest.

### 3 Mean versus median

The first question we explore is whether there is an advantage of choosing the mean over the median (and vice versa) to combine the independent participants’ judgments. We recall that in the ox-weighing contest, the median overestimated the weight by 0.8% whereas the mean yielded zero error [2].

Let \( g_i \) be the estimate of some unknown quantity (e.g., the weight of the ox in Galton’s contest) by individual \( i = 1, \ldots, n \), so the arithmetic mean of the individuals’ estimates is

\[
\langle g \rangle_n = \frac{1}{n} \sum_{i=1}^{n} g_i .
\]

If we denote by \( G \) the true value of the unknown quantity, then the collective error resulting from the use of the arithmetic mean to aggregate the individuals’ estimates is

\[
\gamma_{\text{mean}} = |\langle g \rangle_n - G| .
\]

Defining the average quadratic individual error as

\[
\epsilon_{\text{quad}} = \frac{1}{n} \sum_{i=1}^{n} (g_i - G)^2 ,
\]

and the diversity of the estimates as

\[
\delta = \frac{1}{n} \sum_{i=1}^{n} (g_i - \langle g \rangle_n)^2 ,
\]

we obtain the identity

\[
\gamma_{\text{mean}}^2 = \epsilon_{\text{quad}} - \delta ,
\]

which is Page’s diversity prediction theorem [5]. This theorem asserts that, for a given forecasting contest, the (quadratic) collective error equals the average (quadratic) individual error minus the prediction diversity. This result is sometimes viewed as an indication that the increase in the prediction diversity \( \delta \) results in the decrease of the quadratic collective error \( \gamma_{\text{mean}}^2 \). Of course, since \( \delta \) and \( \epsilon_{\text{quad}} \) cannot be varied independently of each other, this interpretation is not correct. In fact, there is no evidence of meaningful correlation between the diversity of the estimates and the collective error [10]. We note that \( \delta \) is known in the statistical literature as the precision of the estimates, that is, the closeness of repeated estimates (of the same quantity) to one another [16].
Fig. 1 Scatter plot of the relative collective errors $\gamma_{\text{mean}}/|G|$ and $\gamma_{\text{median}}/|G|$. The solid line is the diagonal $\gamma_{\text{mean}}/|G| = \gamma_{\text{median}}/|G|$. Pearson’s correlation coefficient is $r = 0.98$. Each symbol stands for one of the 15455 forecasting contests, but not all contests are shown: the upper ranges of the $x$ and $y$ axes were limited to 1 to aid visualization, so about 3% of the forecasting contests are not represented in the figure.

To introduce the median we need to define the empirical cumulative distribution function [17]

$$F_n(g) = \frac{1}{n} \sum_{i=1}^{n} I(g_i \leq g)$$

where the indicator function is $I(g_i \leq g) = 1$ if $g_i \leq g$ and 0 otherwise. The median is then given by

$$F_n^{-1}(1/2) = \inf \{ g : F_n(g) > 1/2 \}$$

where $F_n^{-1}(q)$ with $q \in [0,1]$ is the empirical quantile function, i.e., the inverse of the empirical distribution function. The collective error resulting from the use of the median to aggregate the individuals’ estimates is then

$$\gamma_{\text{median}} = |F_n^{-1}(1/2) - G|.$$  

Figure 1 shows the scatter plot of the relative collective errors calculated with the mean and the median, as defined in Eqs. (2) and (8), respectively. As expected, there is a high correlation between these two errors, viz., Pearson’s correlation coefficient is $r = 0.98$. The median beats the mean in 50.4% of the forecasting contests. Note the outliers with large $\gamma_{\text{mean}}$ but small $\gamma_{\text{median}}$ that exemplify the robustness of the median to extreme individual estimates.

To offer more quantitative information on the data shown in the scatter plot, Fig. 2 exhibits the histograms and the cumulative distributions of the number of forecasting contests with a given relative collective error. The histograms are practically identical for the two aggregation methods, median and mean. In fact, both methods yield relative errors less than 0.05 in 58% of the forecasting contests, but their relative errors are greater than 0.1 in 28% of the forecasting contests. In addition, the mean of the relative collective error calculated with the mean is $0.20 \pm 0.01$ and the mean of the relative collective error calculated with the median is $0.19 \pm 0.01$. The cumulative distributions exhibited in this figure show that about 3% of the forecasting contests have relative collective errors greater than 1 and are left out of the histograms (as well as of the scatter plot of Fig. 1) for the sake of visual clarity. We stress, however, that the forecasting contests left out of Figs. 1 and 2 are not influential in leading to better overall results for the median. In sum, whereas the crowd performance is clearly distinguishable since 58% of its predictions yielded an error of 5% or less, it falls short of being miraculous since it errs by more than 10% in 28% of the forecasting contests. In view of these findings, it seems likely that selective attention has played a considerable role to secure the popularity of the wisdom of crowds.

Next, we measure the fraction of individuals’ estimates that beat the collective estimate for each forecasting contest. The results are presented in form of histograms and cumulative distributions in Fig. 3, where the height of the bars is the proportion of forecasting contests for which that fraction equals $\xi \in [0,1]$. As already pointed out, the median estimate is always
better than the estimates of half of the participants at least, so there is no forecasting contest for which $\xi_{\text{median}} > 0.5$. However, the mean estimate beats the majority of the participants in 67% of the forecasting contests only. This number is obtained by evaluating the cumulative distribution function at $\xi_{\text{median}} = 0.5$. The fraction of forecasting contests for which $\xi_{\text{mean}} = 0$ is 0.015 and the fraction for which $\xi_{\text{median}} = 0$ is 0.026. This is the situation where the crowd beats all participants. Therefore, if the performance criterion is beating most or all participants then the median should definitely be preferred over the mean as an aggregation method since it is by construction guaranteed to beat at least half of the participants and has almost double the chances of beating all them as compared with the mean.

4 Testing the unbiased estimates assumption

As pointed out, a rather natural explanation for the wisdom of crowds assumes that the individuals’ estimates are unbiased, that is, that the errors spread in equal proportion around the correct value of the unknown quantity. Amusingly, if this assumption were correct one might harvest the benefits of the wisdom of crowds by asking a single individual to make several estimates at different times (see, e.g., [18]). Of course, the almost insurmountable difficulty to ensure that a same individual produces a large number of independent estimates of a same quantity makes the direct validation of the unbiased estimates assumption unfeasible. However, we can perform a much simpler and more informative test, viz., to check whether there is a meaningful correlation between the (unsigned) asymmetry of the distribution of individuals’ estimates and the collective error. In fact, according to the unbiased estimates assumption, the more asymmetric the distribution of estimates is, the more unlikely that the errors cancel each other out and hence the greater the collective error.

We recall that the skewness $\mu_3$ of a distribution is a dimensionless measure of its asymmetry defined as

$$\mu_3 = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{g_i - \langle g \rangle_n}{\delta^{1/2}} \right)^3, \quad (9)$$

where $\langle g \rangle_n$ is the mean of the individuals’ estimates defined in Eq. (1) and $\delta$ is the sample variance of those estimates defined in Eq. (4). A negative value of $\mu_3$ implies that the left tail of the distribution of individual estimates is longer than the right tail, whereas a positive $\mu_3$ indicates a right-tailed distribution. Since the skewness $\mu_3$ measures the asymmetry with respect to the mean, in this analysis we will only consider the case that the collective estimate is calculated using the mean of the individuals’ estimates.

In Fig. 4 we show the scatter plot of the relative collective error $\gamma_{\text{mean}}/|G|$ calculated with the mean and the unsigned skewness $|\mu_3|$. The solid line is the linear regression. The Pearson correlation coefficient is $r = 0.03$. The upper range of the $x$ axis was limited to 1 to aid visualization, but about 3 percent of the forecasting contests have relative errors greater than 1.
Fig. 5 Histogram of the skewness $\mu_3$ of the distributions of individuals’ estimates. The wisdom of crowds experiments produce left- and right-tailed distributions in approximately the same proportion.

no meaningful correlation between the asymmetry of the distribution of individuals’ estimates and the collective error. More pointedly, Pearson’s correlation coefficient for these two variables is $r = 0.03$ and the linear regression is $|\mu_3| \approx 0.81$. Particularly conspicuous are the great number of forecasting contests with high asymmetry and low prediction error, as well as with low asymmetry and high prediction error, that produce the triangle-shaped spreading of points depicted in Fig. 4. It is worth to mention here that if one were to expect that the estimates of a particular individual were unbiased, i.e., that they fluctuated symmetrically around the truth, that individual should most certainly be an expert on the matter in question. For this reason, we claim that the observed asymmetry of the distributions of the experts’ predictions in the FRBP database offers strong evidence to reject the unbiased estimates assumption as an explanation for the accuracy of the crowd’s predictions.

Figure 5 shows that the distribution of the skewness is approximately symmetric, i.e., the number of right- and left-tailed distributions of individuals’ estimates are roughly the same in the forecasting contests of the FRBP database. Hence the asymmetry of those distributions has no role in the wisdom of crowds phenomenon. In fact, a scatter plot of the (signed) skewness against $\gamma_{\text{mean}}/|G|$ yields an almost perfectly symmetric distribution of points around $\mu_3 = 0$ (data not shown).

5 Assessing the accuracy of the forecasters and the crowd

Our previous analysis yielded that the crowd errs by less than 5% in 58% of the forecasting contests, regardless of the method – mean or median – used to aggregate the participants’ estimates. Here we address the question: how good is this performance? Since the crowd beats all participants, which are all expert economists, in less than 3% of the forecasting contests, perhaps an error of 5% is nothing to boast about. Actually, we could as well ask: How good are the expert predictions, anyway? Of course, to answer those questions we must specify a standard of comparison. The seminal work of Meehl in the 1950s that made the astonishing claim that well-informed experts predicting outcomes were not as accurate as simple algorithms [19] (see also [20]) offers a clue. So here we use the basic ARIMA (AutoRegressive Integrated Moving Average) model, which is widely used in fitting and forecasting nonstationary time series [21, 22], as the comparison standard to assess the goodness of both forecasters’ and crowd’s predictions.

The time series for a fixed variable, say NGDP, is the list of its true values $G$ for each month since the date the variable was included in the FRBP survey. We used the 12 values of the variable before the moment of the prediction to set the optimum parameters of the ARIMA model. In other words, we train the ARIMA model using a 12-quarters moving window. Larger window sizes produced similar results. Once the ARIMA model...
model is trained, we use it to predict the values of the variable at the five quarters subsequent to the moment of the prediction, similarly to what is asked to the participants of the FRBP-SPF. Since a fraction of the data of the FRBP-SPF database is used to train the ARIMA model, the number of forecasts used in this section is reduced to 13893.

Let us first assess the accuracy of the individual forecasters. To achieve that we introduce the mean individual error

$$\epsilon = \frac{1}{n} \sum_{i=1}^{n} |g_i - G|,$$

which can be interpreted as the expected error of an expert selected at random among the $n$ participants of the forecasting contest. Figure 6 shows the scatter plot of the relative error of the ARIMA model $\gamma_{\text{ARIMA}} / |G|$ and the relative mean individual error $\epsilon / |G|$ for the 13893 forecasting contests. We find that the ARIMA model is more accurate than a randomly selected expert in 35% of the forecasting contests. The average relative mean individual error is $0.22 \pm 0.01$ and the mean relative error of the ARIMA model is $0.31 \pm 0.01$. Hence, at least in economics, the experts have a considerable advantage over simple time series forecasting algorithms.

Next, we compare the performance of the ARIMA model with the crowd’s performance, which for this analysis we define as $\gamma_{\text{best}} = \min\{\gamma_{\text{mean}}, \gamma_{\text{median}}\}$. The results are actually insensitive to this definition because the accuracies of the median and the mean are very close (see Figs. 1 and 2). Figure 7 shows the scatter plot of the relative error of the ARIMA model $\gamma_{\text{ARIMA}} / |G|$ and the relative collective error $\gamma_{\text{best}} / |G|$. The ARIMA model beats the crowd in 28% of the forecasting contests. We recall that the mean relative error of the crowd is $0.19 \pm 0.01$ and the mean relative error of the ARIMA model is $0.31 \pm 0.01$.

An alternative way to assess the goodness of the predictions of the crowd is to compare them with the individual predictions that were closest to the true values, viz., the winning predictions of the forecasting contests. We have already pointed out that the odds that the crowd beats the winner is 0.015 when the aggregate of the individuals’ estimates is the mean and 0.026 when it is the median. Here we add that the mean relative error of the winners is $0.15 \pm 0.01$, so the crowd performance as measured by its mean relative error is really not too far from the best the experts can offer.

6 Conclusion

The notion that a group of cooperating individuals can solve problems more efficiently than when those same individuals work in isolation is hardly contentious [23], although the factors that make cooperation effective still need some straightening out [24,25]. In fact, cooperation may well lead the group astray resulting in the so-called madness of crowds, as pointed out by MacKay almost two centuries ago [26]: “Men, it has been well said, think in herds; it will be seen that they go mad in herds, while they only recover their senses slowly, and one by one”. Less dramatically, cooperation may simply undermine the benefits of combining independent forecasts, which is known as the wisdom of crowds [6,14].

More precisely, the wisdom of crowds is the idea that the aggregation of the predictions of many independently deciding individuals is likely to be more accurate than the predictions of most or even all individuals within the crowd [3]. Here we use the Federal Reserve Bank of Philadelphia’s Survey of Professional Forecasters (FRBP-SPF) to offer an extensive statistical appraisal of the wisdom of crowds for the prediction of economic indicators. This approach is immune to the selective attention fallacy that has most likely influenced the anecdotal evidences for the wisdom of crowds that abound in the literature [11]. In addition, by focusing on the predictions of experts only we are in fact probing the so-called Vox Expertorum [15], which is the relevant voice when the forecasts address economic, social and political issues.

Our analysis of more than ten thousand forecasting contests reveals a noteworthy performance of the crowd, but that can hardly justify the ado about its wisdom. In particular, the odds that the crowd beats all participants of a forecasting contest is 0.015 when the mean is used to aggregate the individuals’ estimates and 0.026 when the aggregation is given by the median. This
means that the crowd’s performance in Galton’s weighing contest, for which the mean collective estimate has zero error [2], was quite atypical. The median has clear advantages over the mean as the method to combine the individuals’ estimates, specially when the crowd’s performance is compared with the participants’. In particular, the median is always guaranteed to beat at least half of the participants, whereas the mean beats the majority of the participants in 67 percent of the forecasts only (see Fig. 3) and, as just pointed out, the median has almost double the chances to beat all participants as compared to the mean.

Regarding the accuracy of the predictions, however, the advantage of the median is not so noticeable: it is more accurate than the mean in 50.4% of the forecasts. The mean relative errors of both aggregation methods are around 20%, which is not really bad considering that the mean relative error of the winners is 15% and of a standard prediction algorithm (ARIMA) is 31%. It is worth mentioning that the expected error of a randomly chosen forecaster is 22%, so the participants of the FRBP survey predict better than the ARIMA on average.

Another significant result of our extensive statistical analysis of the wisdom of crowds is the rejection of the unbiased estimates assumption, which purports to explain the accuracy of the crowd by assuming that the individuals’ judgments scatter symmetrically around the true value. If this were correct we should find a strong correlation between the asymmetry of the distribution of the individuals’ estimates and the error of the crowd’s estimate. We find no meaningful correlation between these two quantities (viz., Pearson correlation coefficient is $r = 0.03$) and, in fact, there is a large number of forecasting contests with high asymmetry and low collective error (see Fig. 4). Because of the large number of forecasting contests used in our analysis, the $p$-values associated to the correlation coefficients are extremely small, so all correlations reported in this paper are statistically significant. Their meaningfulness, however, is determined by the magnitude of the correlation coefficient $r$.

Our study of the wisdom of crowds dovetails with Quetelet’s original proposal of Social Physics as an empirical science [27,28], where the regularities of the patterns observed in human behavior and refined through a statistical analysis are interpreted as the laws of the society (see also [29]). Perhaps, the statistical features reported here for the prediction of economic indicators hold true for the prediction of other continuous-valued quantities as well and so they may shed light on how (expert) humans estimate unknown quantities.

**Acknowledgements** The research of JFF was supported in part by Grant No. 2020/03041-3, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) and by Grant No. 305620/2021-5, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

**Author contributions**

All the authors contributed equally to the paper.

**Data Availability Statement** This manuscript has associated data in a data repository. [Authors’ comment: The economic forecasting data considered in the manuscript are from the Federal Reserve Bank of Philadelphia’s Survey of Professional Forecasters and are available at https://www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters.]

**References**

1. F. Galton, Nature **75**, 450 (1907)
2. K.F. Wallis, Stat. Sci. **29**, 420 (2014)
3. J. Surowiecki, *The Wisdom of Crowds: Why the Many Are Smarter than the Few and How Collective Wisdom Shapes Business, Economics, Societies, and Nations* (Doubleday, New York, 2004)
4. C. Sunstein, *Infotopia: How many minds produce knowledge* (Oxford University Press, Oxford, 2006)
5. S.E. Page, *The difference: How the power of diversity creates better groups, firms, schools, and societies* (Princeton University Press, Princeton, 2007)
6. J. Lorenz, H. Rauhut, F. Schweitzer, D. Helbing, Proc. Natl. Acad. Sci. USA **108**, 9020 (2011)
7. J.M. Bates, C.W.J. Granger, Oper. Res. Q. **20**, 451 (1969)
8. U.W. Nash, PLoS ONE **9**, e112386 (2014)
9. U.W. Nash, J. Math. Psychol. **77**, 165 (2017)
10. S.M. Reia, J.F. Fontanari, J. Stat. Mech. **2021**, 053402 (2021)
11. D.A. Nobre, J.F. Fontanari, Complex. Syst. **29**, 861 (2020)
12. Federal Reserve Bank of Philadelphia. [https://www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters](https://www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters), (2022)
13. M. P. Clements, R. W. Rich, J. S. Tracy, Surveys of Professionals, Working Paper No. 22-13. Federal Reserve Bank of Cleveland, (2022). [https://doi.org/10.26509/frbc-wp-2022-13](https://doi.org/10.26509/frbc-wp-2022-13)
14. A.J. King, L. Cheng, S.D. Starke, J.P. Myatt, Biol. Lett. **8**, 197 (2011)
15. F.H. Perry-Coste, Nature **75**, 509 (1907)
16. P.-N. Tan, M. Steinbach, V. Kumar, *Introduction to data mining* (Pearson Education Limited, New York, 2014)
17. L. Wasserman, *All of statistics: a concise course in statistical inference* (Springer, New York, 2004)
18. E. Vul, H. Pashler, Psychol Sci. **19**, 645 (2008)
19. P. Meehl, *Clinical versus statistical prediction* (University of Minnesota Press, Minneapolis, 1954)
20. P.E. Tetlock, D. Gardner, *Superforecasting: the art and science of prediction* (Crown Publishing Group, New York, 2016)
21. C. Chatfield, *Time-series forecasting* (CRC Press, Boca Raton, 2000)
22. R.J. Hyndman, G. Athanasopoulos, *Forecasting: principles and practice* (OTexts, New York, 2018)
23. L. Rendell, R. Boyd, D. Cownden, M. Enquist, K. Eriksson, M.W. Feldman, L. Fogarty, S. Ghirlanda, T. Lillcrap, K.N. Laland, *Science* **328**, 208 (2010)
24. J.F. Fontanari, *PLoS ONE* **9**, e110517 (2014)
25. S.M. Reia, A.C. Amado, J.F. Fontanari, *Phys. Life Rev.* **31**, 320 (2019)
26. G. MacKay, *Extraordinary delusions and the madness of crowds* (Richard Bentley, London, 1841)
27. S.M. Stigler, *Statistics on the table: the history of statistical concepts and methods* (Harvard University Press, Cambridge, 2002)

28. K. Donnelly, *Adolphe quetelet, social physics and the average men of science, 1796–1874* (University of Pittsburgh Press, Pittsburgh, 2015)
29. M. Perc, *Sci. Rep.* **9**, 16549 (2019)

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.