EXTENSION OF A 2018 SAMPLE-BASED STUDY ON THE LEVEL OF AWARENESS REGARDING BIG DATA IN THE STATISTICS COMMUNITY OF PAKISTAN

Saleha Naghmi Habibullah
Department of Statistics, Kinnaird College For Women, Lahore, Pakistan
saleha.habibullah@kinnaird.edu.pk

Whereas statisticians of advanced countries are developing new methodologies to reveal patterns and trends inherent in extremely large data sets, the statistics communities of developing countries are deficient in this regard. In the year 2018, a sample survey was carried out in Pakistan to ascertain the level of awareness regarding big data among academics and practitioners of statistics in the country. The survey revealed that, for many terms related to big data, there did not exist much awareness among the statisticians of the country. This paper extends the 2018 study in terms of coverage, scope and depth of analysis. Results of the extended survey seem to confirm the findings of the previous year indicating that there is a need for multi-pronged strategies to create awareness in the statistical community of Pakistan regarding big data and its role in evidence-based decision-making conducive to development and progress of the country.

INTRODUCTION

With the widespread use of the internet, the world is experiencing an unending accrual of enormous amounts of data pertaining to multifarious disciplines. The term 'Big Data' relates to data-sets which are so massive and elaborate that they cannot be dealt with through usual methods. As such, new methodologies are being developed in the technologically advanced countries in order to deal with extremely large data sets encountered in various fields (see Keller et al., 2012; Varian, 2014; Torrecilla & Romo, 2018). The situation of the developing countries is markedly different and they need assistance in order to reap the benefits of Big Data Analytics (see Kshetri, 2014; Manske et al., 2015; Runde, 2017).

'Big Data' being a relatively new term in Pakistan, the author felt the need to conduct a survey in order to determine the extent to which the statisticians' community of the country is cognizant about this term and related concepts. As such, in the summer of 2018, a questionnaire-based sample survey was carried out in order to determine the existing level of awareness regarding big data among academics and practitioners of statistics (Habibullah, 2018). Although small-scale, the survey conducted by the author in 2018 can probably be regarded as the first of its kind in the history of Pakistan. Analysis of the data acquired from seventeen respondents indicated that, for a majority of the statisticians, there did not exist a clear understanding of various terms and concepts associated with Big Data.

The current paper can be regarded as an extension of the 2018 study as it includes PhD, MPhil, MSc and BS students of statistics in addition to faculty-members and professional statisticians. Recognizing that not much change would have occurred during a span of only nine months, the same questionnaire was sent by the author to a number of prospective respondents (including students) in Spring 2019. Replies received from the seventeen respondents of the 2018 survey were *put together* with the responses of the twenty-one persons who sent filled out questionnaires subsequent to the completion of the analysis of the 2018 survey data. As such, the results reported in the following sections of this paper pertain to a total of thirty-eight respondents.

OVERALL PROPORTIONS

First and foremost, we considered the overall proportions of respondents who indicated that they were aware of the meanings of the technical terms contained in Qs. 11, at least to some extent. For this purpose, (i) we merged the response category "I am very well aware of this concept" with the response category "I am aware of this concept but only to some extent" in order to obtain the category "Aware" and (ii) we merged the response category "I have heard this term but I am not aware of its meaning" with the response category "Never heard this term" in order to obtain the category "Not aware". Table 1 contains results obtained subsequent to this merging process.

In: S Budgett (Ed.), *Decision Making Based on Data*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
August 2019, Kuala Lumpur, Malaysia. iase-web.org/Conference_Proceedings.php

Table 1. Overall Proportions of Respondents contained in the 2018/2019 Combined Sample belonging to Category "Aware" with reference to each of the Twenty-Four Technical Terms given in Qs. 11

Sr. No.	Technical term	Total No. of Responses received	Number (proportion) belonging to Category "Aware"	Sr. No.	Technical term	Total No. of Responses received	Number (proportion) belonging to Category "Aware"
i	Data Analysis	36	35 (97.2%)	xiii	Artificial Intelligence	38	19 (50.0%)
ii	R	37	35 (94.6%)	xiv	Data-warehousing	37	16 (43.2%)
iii	Data Science	38	32 (84.2%)	xv	The Internet of Things	36	15 (41.7%)
iv	Data Mining	38	31 (81.6%)	xvi	Distributed processing	35	12 (34.3%)
v	Big Data	37	30 (81.1%)	xvii	Cloud computing	37	12 (32.4%)
vi	Bayesian Analysis	38	30 (78.9%)	xviii	Python	37	11 (29.7%)
vii	Algorithm	38	30 (78.9%)	xix	Exabyte	38	9 (23.7%)
viii	Data Analytics	37	28 (75.7%)	xx	Petabyte	38	8 (21.1%)
ix	Business Analytics	37	26 (70.3%)	xxi	Brontobyte	38	7 (18.4%)
x	Java	36	25 (69.4%)	xxii	Grid computing	36	5 (13.9%)
xi	Machine Learning	38	20 (52.6%)	xxiii	Hadoop	37	5 (13.5%)
xii	Data Engineering	37	19 (51.4%)	xxiv	Crowd-sourcing	36	3 (8.3%)

Remark: Belonging to Category "Aware" implies that these respondents reported that they were aware of the meanings of the technical term, at least to some extent; in the Table, proportions have been written in descending order.

From Table 1, we are able to make the following statements with reference to awareness/non-awareness among statisticians regarding the meanings of various technical terms associated with Big Data:

- "Data analysis" and "R" are the only two terms for each of which more than ninety percent of the respondents indicated that they were aware of the meaning of this term, either wholly or partially;
- "Data Science", "Data Mining" and "Big Data" are the three terms for each of which between eighty and eighty-five percent of the respondents indicated that they were aware of the meaning of this term, either very well aware or aware to some extent;
- "Bayesian Analysis", "Algorithm", "Data Analytics" and "Business Analytics" are the four terms for each of which between seventy and seventy-nine percent of the respondents indicated that they were aware of the meaning of this term, either fully or partly ;
- Of the remaining fifteen technical terms contained in Qs. 11, less than seventy percent of the respondents indicated that they were aware of the meanings of those terms; the six terms "Exabyte", "Petabyte", "Brontobyte", "Grid Computing", "Hadoop" and "Crowdsourcing" seem to be the ones that the statistics community is *least familiar* with.

Comparing the proportions of the 2018/2019 combined sample with the proportions of the 2018 sample, we found that, for a multitude of technical terms included in Qs. 11, the proportions of the 2018/2019 combined sample were not substantially different from those of the 2018 sample. This authenticates the view that not much change occurred in the awareness-level of the country's statisticians during a span of just nine months.

COMPARISONS BETWEEN VARIOUS CATEGORIES OF RESPONDENTS

Having looked at the overall picture, we focused on comparisons between various categories of respondents with reference to the awareness/non-awareness regarding big data and the related terminologies. We compared students' situation with that of professional statisticians, female statisticians' situation with that of their male counterparts and the situation of statisticians some portion of whose education has been in a foreign country with that of statisticians whose entire education had been in their home country.

Table 2. Proportions of Academic/Professional Statisticians and PhD/MPhil/MSc/BS students contained in the 2018/2019 Combined Sample belonging to Category "Aware" with reference to Selected Technical Terms out of the Twenty-Four given in Qs. 11

Sr. No.	Technical term	Academic/Professional Statisticians		PhD/MPhil/MSc/BS Students		Difference in Proportions (arranged in descending order)
		Total No. of Responses received	Number (proportion) belonging to Category "Aware"	Total No. of Responses received	Number (proportion) belonging to Category "Aware"	
1	Machine Learning	29	17 (58.6%)	9	3 (33.3%)	25.3%
2	Data Engineering	28	16 (57.1%)	9	3 (33.3%)	23.8%
3	Artificial Intelligence	29	16 (55.2%)	9	3 (33.3%)	21.9%
4	Business Analytics	28	21 (75.0%)	9	5 (55.6%)	19.4%
5	Big Data	28	24 (85.7%)	9	6 (66.7%)	19.0%
6	Hadoop	28	5 (17.9%)	9	0 (0.0%)	17.9%
7	Exabyte	29	8 (27.6%)	9	1 (11.1%)	16.5%
8	Bayesian Analysis	29	24 (82.8%)	9	6 (66.7%)	16.1%
9	Distributed Processing	26	7 (26.9%)	9	5 (55.6%)	-28.7%
10	Java	28	17 (60.7%)	8	8 (100.0%)	-39.3%

Remark 1: Table 2 contains only those terms for which a difference of 15% or more was observed in the proportions pertaining to the two categories of respondents.
Remark 2: Belonging to Category "Aware" implies that these respondents reported that they were aware of the meanings of the technical term, at least to some extent.

From Table 2, it is interesting to see that, whereas on the one hand, "Machine Learning", "Data Engineering" and "Artificial Intelligence" are the three terms for each of which the difference in proportions of academic/professional statisticians and PhD/MPhil/MSc/BS students falling in Category "Aware" exceeded 20%, on the other, for two of the technical terms, "Distributed Processing" and "Java", considerably larger proportions of students than academic/professional statisticians seemed to belong to in Category "Aware".

From Table 3, it is interesting to see that, whereas on the one hand, "Data-warehousing", "Hadoop" and "The Internet of Things", "Artificial Intelligence", "Cloud Computing" and "Machine Learning" are the six terms for each of which the difference in proportions of male and female statisticians (students included) falling in Category "Aware" exceeded 35%, on the other, for the basic technical term, "Big Data", a slightly larger proportion of females than males reported that they were aware of the meaning of this term, either wholly or partly.

From Table 4, it is easy to see that, for ten of the twenty-four technical terms given in Qs. 11, considerably higher proportions of statisticians some portion of whose education had been in a foreign country reported that they were aware of the meanings of these terms (at least to some extent) as compared with those whose entire education had been in their home country. For each of the three terms

"Cloud Computing", "Artificial Intelligence" and "The Internet of Things", the difference in proportions of respondents falling in Category "Aware" exceeded 40%.

Table 3. Proportions of Male and Female Statisticians (students included) contained in the 2018/2019 Combined Sample belonging to Category "Aware" with reference to Selected Technical Terms out of the Twenty-Four given in Qs. 11

Sr. No.	Technical term	Male Statisticians (students included)		Female Statisticians (students included)		Difference in Proportions (arranged in descending order)
		Total No. of Responses received	Number (proportion) belonging to Category "Aware"	Total No. of Responses received	Number (proportion) belonging to Category "Aware"	
1	Data-warehousing	9	8 (88.9%)	28	8 (28.6%)	60.3%
2	Hadoop	10	5 (50.0%)	27	0 (0.0%)	50.0%
3	The Internet of things	9	7 (77.8%)	27	8 (29.6%)	48.2%
4	Artificial Intelligence	10	8 (80.0%)	28	11 (39.3%)	40.7%
5	Cloud computing	10	6 (60.0%)	27	6 (22.2%)	37.8%
6	Machine Learning	10	8 (80.0%)	28	12 (42.9%)	37.1%
7	Brontobyte	10	4 (40.0%)	28	3 (10.7%)	29.3%
8	Python	10	5 (50.0%)	27	6 (22.2%)	27.8%
9	Grid Computing	9	3 (33.3%)	27	2 (7.4%)	25.9%
10	Petabyte	10	4 (40.0%)	28	4 (14.3%)	25.7%
11	Data Engineering	10	7 (70.0%)	27	12 (44.4%)	25.6%
12	Exabyte	10	4 (40.0%)	28	5 (17.9%)	22.1%
13	Crowd-sourcing	9	2 (22.2%)	27	1 (3.7%)	18.5%
14	Algorithm	10	9 (90.0%)	28	21 (75.0%)	15.0%
15	Big Data	10	7 (70.0%)	27	23 (85.2%)	-11.8%

Note: Remarks 1 and 2 immediately below Table 2 apply to Table 3 too.

DISCUSSION AND CONCLUDING REMARKS

During the past few years, 'Big Data' has become a buzzword among the statisticians of the developed world. However, to this date, it is almost an 'alien' term among the statistics communities of some of the developing countries. In 2018, a semi-structured questionnaire was devised by the author and administered on a sample of Pakistani statisticians in order to determine their knowledgeability regarding big data and the related concepts. Being probably the first of its kind in the country, it can be regarded as an exploratory survey to ascertain the factual situation in the country. The current paper is based on an extension of the 2018 survey in terms of coverage, scope and depth of analysis.

Table 4. Proportions of Statisticians (students included) contained in the 2018/2019 Combined Sample Some Portion of whose Education had/had not been in a Foreign Country belonging to Category "Aware" with reference to Selected Technical Terms out of the Twenty-Four given in Qs. 11

Sr. No.	Technical term	Statisticians (students included) some portion of whose education had been in a Foreign Country		Statisticians (students included) no portion of whose education had been in a Foreign Country		Difference in Proportions (arranged in descending order)
		Total No. of Responses received	Number (proportion) belonging to Category "Aware"	Total No. of Responses received	Number (proportion) belonging to Category "Aware"	
1	Cloud computing	8	6 (75.0%)	29	6 (20.7%)	54.3%
2	Artificial Intelligence	8	7 (87.5%)	30	12 (40.0%)	47.5%
3	The Internet of things	8	6 (75.0%)	28	9 (32.1%)	42.9%
4	Data Engineering	8	6 (75.0%)	29	13 (44.8%)	30.2%
5	Machine Learning	8	6 (75.0%)	30	14 (46.7%)	28.3%
6	Data-warehousing	8	5 (62.5%)	29	11 (37.9%)	24.6%
7	Java	8	7 (87.5%)	28	18 (64.3%)	23.2%
8	Business Analytics	8	7 (87.5%)	29	19 (65.5%)	22.0%
9	Distributed processing	8	4 (50.0%)	27	8 (29.6%)	20.4%
10	Data Analytics	8	7 (87.5%)	29	21 (72.4%)	15.1%

Note: Remarks 1 and 2 immediately below Table 2 apply to Table 4 too.

Results reported in this paper seem to *confirm* the findings of the previous year which pointed to a lack of awareness among the statistics community of Pakistan regarding big data and the related terminologies. Similar to the previous year, for a considerably large number of terms related to big data, there seems to be not much awareness. The ranking in Table 1, facilitates the identification of the technical terms where there is greatest need to educate the statisticians. As far as comparisons between various categories of respondents are concerned, it is only natural to expect academic/professional statisticians to be more knowledgeable than students. It is thus interesting to find that, for two of the technical terms, the students seem to have an edge over their seniors, and one wonders why? Results pertaining to comparisons between male and female statisticians with reference to awareness regarding various technical terms relating to big data are not surprising as, even in this modern era, the socio-cultural norms of the country facilitate a greater amount of 'exposure' to the men than the women. However, as far as awareness regarding the basic term, "Big Data" itself is concerned, the female statisticians do not seem to be handicapped as compared with their male counterparts. Results pertaining to comparisons between statisticians some portion of whose education had been in a foreign country with those whose entire education had been in their home country are as expected. For a number of technical terms, considerably higher proportions of respondents some portion of whose education had been in a foreign country were falling in Category "Aware" as compared with respondents whose entire education had been in their home country.

The world having entered an era in which we are witnessing an 'explosion' of digital data, developing countries cannot afford to 'sit back' and be oblivious of big data and the related challenges. As such, it is recommended that similar exploratory surveys be conducted in other developing countries so that the ground realities are unveiled. Identification of areas of weakness (with reference to knowledgeability regarding big data among statisticians) will facilitate launch of initiatives for the commencement of the process of capacity-building to be able to deal with the complexities of extremely

large data sets that cannot be handled by ordinary methods. Only then will the statisticians of these countries be able to play their role in the optimal utilization of big data for *evidence-based decision-making and policy formulation* conducive to progress and development.

ACKNOWLEDGMENTS

The author would like to acknowledge the contribution of her former student, Ms Kessica Xavier (MPhil Statistics) as well as of Ms Zulaikha Mashkoor (MPhil Statistics) both of whom assisted in the preparation of the data-sheet, analysis of the collected data and compilation of results.

REFERENCES

Habibullah, S. N. (2018). A sample survey on the current level of awareness regarding Big Data among academics and practitioners and statistics in Pakistan. Retrieved from https://www.bigsurv18.org/conf18/uploads/251/252/Saleha_N._Habibullah_Full_Paper_BigSurv_2018.pdf.

Keller, S. A., Koonin, S.E., & Shipp, S. (2012). Big data and city living – what can it do for us?. *Significance, 9*(4), 4–7.

Kshetri, N. (2014). The emerging role of Big Data in key development issues: Opportunities, challenges, and concerns. *Big Data & Society*, *1*(2).

Manske, J., Sangokoya, D., Pestre, G., & Letouzé, E. (2015). Opportunities and Requirements for Leveraging Big Data for Official Statistics and the Sustainable Development Goals in Latin America. *White Paper, Harvard Humanitarian Initiative, MIT Media Lab and Overseas Development Institute*, 15.

Runde, D. (2017). The data revolution in developing countries has a long way to go. *Forbes*, *February 25, 2017*.

Torrecilla, J. L., & Romo, J. (2018). Data learning from big data. *Statistics & Probability Letters*, *136*, 15–19.

Varian, H. R. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives, 28*(2), 3–28.