A Sample Survey on the Current Level of Awareness Regarding Big Data Among Academics and Practitioners of Statistics in Pakistan

by

Saleha Naghmi Habibullah
Department of Statistics
Kinnaird College For Women
Lahore, Pakistan

Abstract

In developing countries such as Pakistan, there is a fairly strong tradition of theoretical development as well as practical application of survey sampling. However, in these countries, a large number of academics and practitioners of Statistics are unfamiliar with the true meaning of terms such as Big Data, Exabyte, Petabyte, Brontobyte, Artificial Intelligence, Machine learning, Data mining, Data warehousing, Distributed processing, Grid computing, Cloud computing and the like. In this paper, we report the results of a survey carried out to ascertain the current level of awareness regarding Big Data among academics and practitioners of Statistics in Pakistan. Respondents to a questionnaire formulated for this purpose include lecturers, assistant professors, associate professors and professors of Statistics working in various universities and colleges of Pakistan, as well as statistical officers working at the Pakistan Bureau of Statistics, the provincial bureaus of Statistics and/or other data-collecting organizations of the country. Results of the survey seem to indicate that there is a need for multi-faceted efforts aimed at creating awareness regarding Big Data, the related technologies, challenges and future prospects among members of the statistical community of Pakistan.

Keywords: Big Data, Survey, Academics, Practitioners of Statistics.

1. Introduction:

With the generalized used of computers and the advent of the internet in particular, not only has the world witnessed a revolution in the field of communication, it is also experiencing an ongoing accumulation of huge amounts of data pertaining to diverse disciplines. Multifarious transactions and registries that are taking place on a continual basis have caused a 'deluge' of data. The term 'Big Data' pertains to these types of data-sets which are so huge and complex that they cannot be analyzed through ordinary methods. As such, the technologically advanced countries are developing specialized algorithms that are able to deal with massive data-sets measured in Exabytes, Petabytes, Brontobytes etc. Without exaggeration, it can be said that we have entered an age in which we are witnessing an 'explosion' of digital information.

The situation of some of the developing countries being substantially different from that of the Western world, to this date, there is very little awareness regarding Big Data among the statisticians working in these countries. This paper presents the results of a survey that has been carried out in Pakistan to determine the current level of awareness regarding Big Data among academics and practitioners of Statistics in the country. The following sections of this paper

present (i) a brief review of the literature on this topic, (ii) the methodology that was adopted to conduct the survey, (iii) analysis of the collected data and (iv) results and conclusions.

2. Literature Review:

In this section, we present a brief review of some of the papers that have emerged during the past few years.

Fisher et al. (2012) express the opinion that although universities are creating advanced degree programs in analytics, there is a shortfall of statisticians and, in the near future, demand will exceed supply. This shortfall enables data-savvy end users to do their own analyses, without expert supervision which exposes them to the pitfalls that scientists are trained to avoid. The authors provide examples of such pitfalls, including (i) being careless with missing data values; (ii) misapplying statistical tests and (iii) overfitting models.

Schenker et al. (2013) assert that, despite the enormous potential for contributions by statisticians, the statistics community has not been involved in BigData activities to the extent that is required. They opine that there are three reasons for this disconnect: (i) the media and public lack a general understanding of what statisticians contribute to society; (ii) only a few statistician are engaged in Big Data projects; (iii) the statistical community is disconnected from the new community of Data Scientist.

Wang et al. (2015) assert that the role of computational statisticians in scientific discoveries resulting from analyses of Big Data has been under recognized. They opine that Big Data present opportunities as well as challenges to statisticians. They go on to summarize recent methodological and software developments in statistics that address the challenges related to Big Data.

A number of other papers are also available in the scientific literature that indicate the deficiency of the statistics community with reference to Big Data.

3. The Survey:

Big Data being a relatively new term in Pakistan, the author felt the need to conduct a survey aimed at ascertaining the extent to which the statistics community in the country is familiar with this term and related concepts. During the month of August 2018, a questionnaire consisting of eighteen questions was devised by the author and emailed to 68 statisticians working in various cities and towns of Pakistan. (The author utilized (i) the mailing list of the Islamic countries Society of Statistical Sciences (ISOSS), the headquarter of which is located in Lahore, (ii) the social media group “ISOSS & PISTAR”, (iii) another social media group “Statistician Forum”, (iv) homepages of the Statistics Departments of various universities (and associated links), (v) homepages the Pakistan Bureau of Statistics and the provincial bureaus of Statistics (and associated links), and (vi) some statisticians personally known to the author.) In order to increase the probability of a higher response-rate, along with reminders, the author sent a number of emails that carried the request that the questionnaire may be forwarded to other statisticians and they be requested to email to the author the filled out questionnaire.

Despite considerable effort during the short time that was available to the author for the conduct of this survey, the total number of respondents until August 31, 2018 turned out to be 17. The questionnaire is given in the Appendix.

4. Main Results:

In this section, we present the main results obtained regarding the level of awareness pertaining to various technical terms Big Data among members of the Statistics community who belonged to our sample and responded to the questionnaire.

4.1 Overall proportions:

First and foremost, we present the overall proportions for the various technical terms contained in Qs. 11. Table 4.1 contains this information, the technical terms 'ranked' with respect to the proportion of respondents who answered that they have a good understanding of the topic.

Table 4.1
Overall Proportions for the Technical Terms of Qs. 11

Sr. No.	Technical term	I am very well aware of this concept	I am aware of this concept but only to some extent	I have heard this term but I am not aware of its meaning	Never heard this term	Non-response	Total No. of Responses received
i	R	11 64.7%	5 29.4%	0 0.0%	0 0.0%	1 5.9%	17 100.0%
ii	Data Analysis	10 58.8%	5 29.4%	0 0.0%	0 0.0%	2 11.8%	17 100.0%
iii	**Big Data**	**7** **41.2%**	**6** **35.3%**	**3** **17.6%**	**0** **0.0%**	**1** **5.9%**	**17** **100.0%**
iv	Bayesian Analysis	7 41.2%	9 52.9%	1 5.9%	0 0.0%	0 0.0%	17 100.0%
v	**Data Science**	**7** **41.2%**	**8** **47.1%**	**2** **11.8%**	**0** **0.0%**	**0** **0.0%**	**17** **100.0%**
vi	Data Mining	6 35.3%	8 47.1%	2 11.8%	1 5.9%	0 0.0%	17 100.0%
vii	Data Analytics	6 35.3%	6 35.3%	3 17.6%	1 5.9%	1 5.9%	17 100.0%
viii	Algorithm	5 29.4%	9 52.9%	1 5.9%	2 11.8%	0 0.0%	17 100.0%
ix	Artificial Intelligence	4 23.5%	9 52.9%	3 17.6%	1 5.9%	0 0.0%	17 100.0%
x	Machine Learning	4 23.5%	7 41.2%	5 29.4%	1 5.9%	0 0.0%	17 100.0%

Table 4.1 (continued)
Overall Proportions for the Technical Terms of Qs. 11

Sr. No.	Technical term	I am very well aware of this concept	I am aware of this concept but only to some extent	I have heard this term but I am not aware of its meaning	Never heard this term	Non-response	Total No. of Responses received
xi	Business Analytics	3 17.6%	10 58.8%	2 11.8%	1 5.9%	1 5.9%	17 100.0%
xii	The Internet of Things	3 17.6%	5 29.4%	1 5.9%	7 41.2%	1 5.9%	17 100.0%
xiii	Java	2 11.8%	9 52.9%	5 29.4%	0 0.0%	1 5.9%	17 100.0%
xiv	Cloud computing	2 11.8%	5 29.4%	5 29.4%	4 23.5%	1 5.9%	17 100.0%
xv	Exabyte	2 11.8%	4 23.5%	3 17.6%	8 47.1%	0 0.0%	17 100.0%
xvi	Petabyte	2 11.8%	4 23.5%	2 11.8%	9 52.9%	0 0.0%	17 100.0%
xvii	Brontobyte	2 11.8%	4 23.5%	1 5.9%	10 58.8%	0 0.0%	17 100.0%
xviii	Distributed processing	2 11.8%	3 17.6%	7 41.2%	3 17.6%	2 11.8%	17 100.0%
xix	Data Engineering	1 5.9%	5 29.4%	9 52.9%	1 5.9%	1 5.9%	17 100.0%
xx	Python	1 5.9%	5 29.4%	7 41.2%	3 17.6%	1 5.9%	17 100.0%
xxi	Grid computing	1 5.9%	2 11.8%	5 29.4%	8 47.1%	1 5.9%	17 100.0%
xxii	Crowdsourcing	1 5.9%	1 5.9%	4 23.5%	10 58.8%	1 5.9%	17 100.0%
xxiii	Data-warehousing	0 0.0%	8 47.1%	3 17.6%	5 29.4%	1 5.9%	17 100.0%
xxiv	Hadoop	0 0.0%	4 23.5%	2 11.8%	10 58.8%	1 5.9%	17 100.0%

It is interesting to note that, of the twenty-four technical terms, only nine attracted a response from each one of the seventeen respondents. As many as thirteen technical terms experienced one non-response each whereas two experienced two non-responses each.

From the proportions contained in Table 4.1, we are able to make the following statements with reference to the responses obtained from the respondents in our sample:

- As far as the technical term of primary interest to the researcher i.e. 'Big Data' is concerned, it is heartening to note that 7 of the 17 respondents (i.e. 41.2%) reported that they are well aware of the meaning of this term and 6 (i.e. 35.3%) indicated that they are aware of it to some extent. Only 3 out of 17 respondents (i.e. 17.6%) reported that they had heard this term but were not aware of its meaning, and not a single one stated that he or she had never heard this term.
- Similarly, for the technical term that is a 'buzz word' in the world today i.e. 'Data Science', 7 of the 17 respondents (i.e. 41.2%) reported that they are well aware of the meaning of this term and 8 (i.e. 47.1%) indicated that they are aware of it to some extent. Only 2 out of 17 respondents (i.e. 11.8%) reported that they had heard this term but were not aware of its meaning, and not a single one stated that he or she had never heard this term.
- The technical concepts regarding which ten or more respondents out of seventeen (i.e. more than 58%) indicated that they are very well aware of these concepts are 'R' and 'Data Analysis'; as well, for each of these two concepts, nearly 30% of the respondents indicated that they were aware of it to some extent.
- Between 35% and 42% of the respondents reported sound awareness of the concepts 'Bayesian Analysis', 'Data Mining' and 'Data Analytics', and, for each of these five concepts, between 35% and 53% indicated that they were aware of it to some extent.
- Between 17% and 30% of the respondents claimed that they were well aware of the meanings of the terms 'Algorithm', 'Artificial Intelligence', 'Machine Learning', 'Business Analytics' and 'The Internet of Things' and, for each of these five concepts, between 29% and 53% indicated that they were aware of it to some extent whereas between 17% and 48% indicated that either they had heard this term but were not aware of its meaning or they had never heard this term prior to this survey.
- For as many as twelve of the twenty-four technical concepts included in Qs. 11 of the questionnaire (i.e. 'Java', 'Cloud computing', 'Exabyte', 'Petabyte', 'Brontobyte', 'Distributed processing', 'Data Engineering', 'Python', 'Grid computing', 'Crowdsourcing', 'Data-warehousing' and 'Hadoop'), less than 12% of the respondents reported that they were well aware of the term. For six of these technical terms, between 47% and 59% of the respondents reported that they had never heard this term.

4.2 Cross-tabulations:

Interested in comparing the situation of the female respondents with that of the males, we performed a cross-tabulation of gender with each and every one of the technical terms given in Qs. 11. Although this sample of seventeen respondents is not random in the strict sense of the word, in each case, we applied the 2×4 extension of the Fisher's Exact Test / Freeman-Halton Test to obtain an indication of association between gender and awareness regarding the meaning of the technical term. Four of the twenty-four tables seemed to indicate an association, and each one of these is presented here along with the p-value of the test applied.

Interested in comparing the older statisticians with the younger ones, we divided the respondents into two categories i.e. (i) 45 years of age or younger and (ii) 46 years of age and above. Having done so, we performed cross-tabulations of age-bracket with awareness regarding various technical terms but did not detect any significant differences in proportions.

Similarly, inquisitive about any differences in the situation of the statisticians with higher educational qualifications (MPhil/PhD) with those having lower educational qualifications, we performed cross-tabulations of educational qualifications with awareness regarding various technical terms but did not detect any significant differences in proportions.

Table 4.2
Awareness Regarding 'Machine Learning' *

Response / Gender	I am very well aware of this concept	I am aware of this concept but only to some extent	I have heard this term but I am not aware of its meaning	Never heard this term	Non-response	Total
Male	3	4	0	1	0	8
Female	1	3	5	0	0	9
Total	4	7	5	1	0	17

***The p-value of the Fisher's exact test = 0.057 which is only very slightly bigger than 0.05. As such, we may conclude that there does exist a difference between the proportions of male and female statisticians in Pakistan who are well-aware of this concept.

Table 4.3
Awareness Regarding 'The Internet of Things'**

Response / Gender	I am very well aware of this concept	I am aware of this concept but only to some extent	I have heard this term but I am not aware of its meaning	Never heard this term	Non-response	Total
Male	3	4	0	1	0	8
Female	0	1	1	6	1	9
Total	3	5	1	7	1	17

**The p-value of the Fisher's exact test = 0.021 implying that there is an association between the gender of the person and his/her awareness regarding 'The Internet of Things'. As such, we may conclude that there does exist a difference between the proportions of male and female statisticians who are aware of the meaning of this term.

Table 4.4
Awareness Regarding 'Data Warehousing'***

Response / Gender	I am very well aware of this concept	I am aware of this concept but only to some extent	I have heard this term but I am not aware of its meaning	Never heard this term	Non-response	Total
Male	0	7	0	0	1	8
Female	0	1	3	5	0	9
Total	0	8	3	5	1	17

***The p-value is 0.001 implying that the computed value of the Fisher's exact test statistic is highly significant. As such, we may conclude that there does exist a difference between the proportions of male and female statisticians who are aware of the meaning of the term 'Data Warehousing'.

Table 4.5
Awareness Regarding 'Hadoop'****

Response / Gender	I am very well aware of this concept	I am aware of this concept but only to some extent	I have heard this term but I am not aware of its meaning	Never heard this term	Non-response	Total
Male	0	4	2	2	0	8
Female	0	0	0	8	1	9
Total	0	4	2	10	1	17

****The p-value is 0.004 implying that the computed value of the Fisher's exact test statistic is highly significant. As such, we may conclude that there does exist a difference between the proportions of male and female statisticians who are aware of the meaning of the term 'Hadoop'.

Last but not the least, in order to find out whether any differences exist between the situation of the academic statisticians and that of statisticians working in non-academic organizations, we performed cross-tabulations of the type of organization in which the respondent was employed with awareness regarding various technical terms but did not detect any significant differences in proportions.

5. Limitation of the Study:

As mentioned in Section 3, the survey questionnaire was emailed to 68 members of the statisticians' community in Pakistan. However, the author had only about two weeks for

compiling the first version of the results (those contained in this paper) and, in this limited time-period, despite reminders, responses were received from seventeen persons only. For this particular questionnaire, the low response rate can also be interpreted as an indicator of 'not caring' for this particular topic by members of the statistics community; it might also mean that people who are not familiar with the technical terms contained in the questionnaire are uncomfortable saying it. The author intends to continue the process of data-collection and to update the results subsequent to receipt of a reasonably large number of responses.

6. Concluding Remarks:

From the results presented in Section 4, it seems reasonable to conclude that the programming language 'R' is about the only 'entity' related to Big Data about which good awareness exists among a fairly large proportion of statisticians in Pakistan --- and the reason for this is not Big Data but (i) the fact that R comes in handy for small data-sets too (the types of data-sets that the academic statisticians of Pakistan are dealing with routinely), (ii) R has become pretty well-known in Pakistan during the past eight to ten years.

For a lot many terms related to Big Data, there seems to be not much awareness. In Table 4.1, the ranking with respect to the proportion of respondents who indicated that they have a good understanding of the technical term facilitates the identification of the technical terms that the statisticians feel less familiar with.

For a few terms related to Big Data, it appears that, in the country, the proportions of female statisticians having at least some amount of awareness regarding these terms is lower than the corresponding proportions of male statisticians. This is totally understandable given the fact that, even though we are eighteen years into the twenty-first century, the socio-cultural norms of many countries including Pakistan provide a greater amount of 'exposure' to the male half of the society than the female half.

The results of the survey seem to indicate that there is a need for multipronged efforts aimed at creating awareness regarding Big Data and related concepts as well as methodologies for dealing with Big Data among the statistician community of Pakistan. Capacity-building to be able to deal with Big Data is the need of the twenty-first century and the first phase --- awareness-building --- will be the forerunner of the era when the Pakistani statisticians will have a fairly good idea regarding the challenges surrounding Big Data Analytics as well as the future prospects. Only then will they be able to play their role with reference to the utilization of Big Data for evidence-based decision-making conducive to development in various sectors leading to the advancement and progress of the country.

7. Acknowledgments

The author would like to thank Dr. Zahoor Ahmad, Lahore Garrison University, Lahore, Dr. Zahid Asghar, Quaid-e-Azam University, Islamabad and Mr. Syed Waseem Abbas, Bureau of Statistics, Punjab, Lahore for their contribution in the finalization of the questionnaire that was devised in order to conduct this survey. As well, the author would like to acknowledge a great deal of contribution on the part of her former student, Miss Kessica Xavier, MPhil Statistics,

who assisted in the preparation of the data-sheet, the analysis of the collected data and the compilation of the results.
Sincere thanks and appreciation for Prof. Emer. Edith Seier, East Tennessee State University, USA and Prof. Andrea Blejec, National Institute if Biology, Slovenia on account of the pains that they took in order to review the paper and to provide invaluable suggestions.

8. References

- Chun Wang, Ming-Hui Chen, Elizabeth Schifano, Jing Wu, and Jun Yan October (2016) 'Statistical Methods and Computing for Big Data', Statistics and Its Interface, 399-414, 9.
- Schenker, N., Davidian, M., and Rodriguez, R. (2013), "The ASA and Big Data," Amstat News, 432, 3–4.
- Danyel Fisher, Rob DeLine, Mary Czerwinski, Steven Drucker (2012) "Interactions with Big Data Analytics", Magazine interactions, 50-59, 19(3)

APPENDIX
SURVEY QUESTIONNAIRE

Dear Colleague,

This questionnaire is being mailed to you in connection with a survey intended to ascertain the level of awareness regarding Big Data among the academics and practitioners of Statistics in Pakistan. I am optimistic that I will obtain your kind consent to act as one of the respondents. If so, you are requested to provide responses to each of the following questions to the best of your ability. Your genuine and well thought-out responses will go a long way in making this survey a success. Your cooperation in this regard will be much appreciated.

1. Your gender: _____
2. Your age: _____
3. Your educational qualification: _____
4. Has any portion of your education been in a foreign country? _____
5. Are you employed or have you retired ? _____
6. Your profession: _____
7. City/town of Pakistan in which you are/have been working_____
8. Is it one of the bigger cities or one of the smaller towns ?_____
9. Are you/have you been working in university/college or in a non-academic organization?_____
10. Are you/have you been working in public sector or private sector? _____
11. Please tick the appropriate option for each of the technical terms given in the following table:

12.

Serial No.	Technical term	I am very well aware of this concept	I am aware of this concept but only to some extent	I have heard this term but I am not aware of its meaning	Never heard this term
i	Big Data				
ii	Data Science				
iii	Machine Learning				
iv	Data Mining				
v	Business Analytics				
vi	Data Analytics				
vii	Data Analysis				
viii	Bayesian Analysis				
ix	Artificial Intelligence				
x	Data Engineering				
xi	Algorithm				
xii	Exabyte				
xiii	Petabyte				
xiv	Brontobyte				
xv	Python				
xvi	Java				
xvii	R				
xviii	Hadoop				
xix	Data-warehousing				
xx	Cloud computing				
xxi	Distributed processing				
xxii	Grid computing				
xxiii	Crowdsourcing				
xxiv	The Internet of Things				

13. For each of the technical terms of Question no. 11 for which you indicated that you have thorough knowledge / some amount of knowledge, please provide a definition of each one of those terms. (Please do not use Google etc for this purpose):

    Please type inside the (extendable) box.

Definition no. 1: Definition no. 2: Definition no. 3: And so on

14. In your opinion, does the subject of Statistics have any role to play in Data Science? If so, please explain in a few words.

15. Does Bayesian Analysis have any role to play in Data Science? If so, please explain in a few words.

16. Does Machine Learning have anything to do with Data Science? If so, please explain in a few words.

17. Can the term "Data Science" be applied in the case of a small data-set also? If so, please explain in a few words.

18. Your Name (optional): _____

19. Your Institution/Organization (optional): _____

Thank you very much for your time and kind cooperation.

For any query, comments or suggestions, please write to
Dr. Saleha Naghmi Habibullah, Professor of Statistics, Kinnaird College For Women, Lahore
(email: salehahabibullah@gmail.com).

==========
==========