# ENABLING STATISTICS UNDERGRADS OF DEVELOPING COUNTRIES TO TAKE THEIR FIRST STEPS TOWARD DATA SCIENCE THROUGH MOBILE APPLICATIONS

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

*Data Science is a catchword today. Whereas developed countries have advanced significantly in this field, the developing countries are far behind. Keeping in view the importance of this discipline in today's data-driven world, efforts are required to enable the statistics communities of developing countries to take their first steps toward data science. This paper focuses on undergraduate students and proposes the introduction of two 3 credit-hour courses on introductory data science in the BS Statistics programs of developing countries. As non-trivial proportions of undergraduate students in these countries do not possess laptops but do have access to smartphones and the internet, it is proposed that these courses use freely downloadable mobile applications that will need to be developed for this purpose. Such an initiative could be a significant first step toward the inculcation of data-science-related knowledge and insight in statistics undergrads of developing countries.*

## INTRODUCTION:

Enormous amounts of data are being accumulated in the world by the minute and this phenomenon has caused the emergence of a new discipline called Data Science which relies heavily on computer algorithms. Substantial advancement has occurred in this discipline during the past few years in the developed countries. The developing countries, however, are far behind in this discipline due to a variety of reasons including the enormous differences that exist between the socio-economic conditions of the developed and the developing countries. In particular, the statistics communities of these countries are unaware of the complexities of this field. Nonetheless, statisticians of developing countries cannot afford to remain oblivious of this highly important discipline in today's data-driven era. Efforts need to be made to enable practitioners, teachers and students of Statistics living and working in developing countries to take their *first steps* toward the acquisition of knowledge and insight with regard to introductory-level data science.

This paper focuses on undergraduate students of Statistics in developing countries who need to be exposed to the basics of data science through a pair of 3 credit-hour courses 'Introductory Data Science I' and 'Introductory Data Science II' that will need to be made an *integral component* of the BS Statistics program. Keeping in view the fact that non-trivial proportions of undergraduate (teenage) students in the developing countries do not possess laptops but do have access to smartphones and to the internet, it is proposed that BS Statistics students be exposed to introductory-level data science through *freely downloadable mobile applications* designed for handling modest/moderate amounts of real data.

In a 2018 Consensus Study Report of the National Academies of Sciences, Engineering, and Medicine (NASEM) "Data Science for Undergraduates: Opportunities and Options", it has been indicated that data science is a broad concept involving principles for data collection, storage, integration, analysis, inference, communication, and ethics appropriate for this new data-driven era. As described in the Report, there exist some key concepts that can be regarded as *precursors* for being able to deal with medium-sized datasets. In addition to mathematical and statistical foundations, these include computational foundations, data management, data description and visualization, data modeling and assessment as well as workflow and reproducibility, the basics of each of which will need to be included in the 3 credit hour course entitled 'Introductory Data Science I' very importantly.

One of the key concerns when designing a course is the set of *learning outcomes* or, in other words, the competencies that the students will be expected to leave with. The other equally important matter to be considered is the set of mechanisms by which the learning outcomes will be *assessed* by the instructors. Each of these will need to be given top priority when designing the pair of 3 credit hour courses on data science.

In: P Arnold (Ed.), *New Skills in the Changing World of Statistics Education*
*Proceedings of the Roundtable conference of the International Association for Statistical Education (IASE),*
July 2020, Online. Voorborg, The Netherlands. ©2020 ISI/IASE iase-web.org/Conference_Proceedings.php

Notwithstanding the various challenges that it entails, this initiative may prove to be an effective way of exposing undergraduate students of statistics in developing countries to the basics of data science.

## LITERATURE REVIEW

In this section, we present an overview of a few papers that have emerged during the past few years discussing the utilization of mobile applications that have been designed and developed for educational purposes.

Allen et al. (2016) assert that, although quantitative research methods are a known area of weakness for many undergraduate psychology students, emerging research suggests that mobile technologies offer many possibilities for facilitating learning. The authors introduce StatHand, a free cross-platform application designed to support students' statistical decision making. They present an overview of the rationale behind StatHand and describe the feature set of the application. Guidelines for integrating StatHand into the research methods curriculum are also provided.

Ling et al. (2014) investigate whether the use of mobile phone apps have an impact on students' learning of new statistical concepts. The authors compared a control group consisting of twelve students with a group consisting of thirteen students that used a statistical mobile app during a simulated Statistics lecture and the results provided evidence in favour of the app group. Members of this group felt strongly that mobile applications helped them understand the new concepts more clearly and quickly.

Falloon (2013) asserts that, although the past years have witnessed a range of new technological gadgets emerging on the education scene, the excitement surrounding these technological innovations has failed to match the reality of their performance. The author presents the results of a study on students' interaction with iPad apps, the purpose being to highlight factors that affect the students' learning pathways. The study focuses specifically on design and content features of apps selected for literacy, numeracy and problem-solving capabilities of 5-year-old students. The author suggests that researchers, teachers and developers should work together on the utilization of methodologies introduced by the author to collect data on the reality of student engagement with digital devices as this will contribute toward a significant improvement in the design of mobile apps used by young students for learning purposes.

Even though Falloon (2013)'s study focuses on 5-year-old students, a considerable amount can be learnt from the suggestions given in this paper when designing mobile applications for undergraduate students of statistics.

## A PAIR OF COURSES ON DATA SCIENCE THROUGH MOBILE APPLICATIONS

It is proposed that a pair of 3 credit-hour courses entitled 'Introductory Data Science I' and 'Introductory Data Science II' be introduced in the BS Statistics Programs of universities located in developing countries and be made an *integral component* of each such Program. The optimal time for exposing the students to these two courses seems to be semesters V and VI as, by this time, the students will have already studied the basics of mathematics and statistics. The ground reality of developing countries is that unignorable proportions of undergraduate (teenage) students in these countries do not possess laptops but many of them do have access to smartphones and the internet (at least some of the time), and this fact can be exploited advantageously for exposing those who are enrolled in BS Statistics programs to introductory-level data science even if they do not possess laptops/personal computers. This will be achievable through the creation of freely downloadable mobile applications especially designed for handling modest/moderate amounts of real data.

Another point to be noted here is that, in any university of the developing world, if the courses 'Introductory Data Science I' and 'Introductory Data Science II' are to be made an *integral component* of the BS Statistics Program (as proposed), then, evidently, two optional courses of the existing Program will need to be replaced by these two courses. Keeping in view the importance of this new discipline in today's data-driven world, universities of developing countries may not have much difficulty in letting go of some of the optional courses.

PRECURSORS

The NASEM Report (2018) asserts that data science is a new field that is emerging out of various established fields, including mathematics, statistics, computer science, information technology (IT), operations management, and business analytics. However, core data science concepts involving these principles are not being covered by mainstream training in any one of these fields due to the fact that data science is *not reducible* to any of the pre-existing fields. As such, it is important to realize that, prior to hands-on experience with the handling of medium-sized datasets on smartphones/laptops, the students of developing countries enrolled in the BS Statistics Programs will need to be equipped with some knowledge of *precursors* including (i) computational foundations, (ii) data management and curation, (iii) data description and visualization, (iv) data modeling and assessment, and (v) workflow and reproducibility in addition to mathematical and statistical foundations that they will have already been exposed to during the first four semesters. We present an outline of each of these below, as mentioned in the NASEM Report (2018). It must be kept in mind, however, that, in the 3-credit hour course "Introductory Data Science I", the instructor will be able to cover only the very basics of each of these concepts.

*Computational Foundations*

As stated in the NASEM Report (2018), a data science student needs to be prepared to work with data as they are commonly found in the workplace and research laboratories. Working with data requires *extensive* computing skills that enable one to access and organize data in databases, scrape data from websites, process text into data that can be analyzed, ensure secure data storage, and protect confidentiality. Along with extensive coursework in computer science, students need to be competent in dealing with professional statistical analysis software packages and need to understand the computational and algorithmic problem-solving principles that underlie these packages. Students will also be provided instruction in algorithms, state of the art information technology, data structures, object-oriented programs, and workflow. New pathways may be needed to establish appropriate depth in algorithmic thinking and abstraction in a streamlined manner. Computational problem-solving skills will recur throughout the data scientist's workflow.

*Data Management and Curation*

As indicated in the NASEM Report (2018), for a typical data analysis project, a substantial amount of effort goes into cleaning, merging, and organizing data. With the availability of large databanks accessible to the public, it has become imperative to teach students about the many features of data-management so that they become comfortable with data of different types such as relational data, text, images, etc. According to the Report, 'data provenance', 'data preparation/transformation', 'data management', 'record retention policies', 'data subject privacy', 'missing and conflicting data', 'modern databases', etc. are some of the key concepts/skills that students of data science need for succeeding in their profession.

*Data Description and Visualization*

Students of data science need to learn about basic descriptive statistics as well as about customary graphics such as histograms, scatterplots, historigrams and the like in order to be able to discern the essence of a data-set. Next, students will need to be instructed upon how to use simple graphics to check data for various types of flaws, glitches and inconsistencies. Having acquired basic proficiency in this area, the students will be ready to indulge in exploratory data analysis (EDA). According to the NASEM Report (2018), "Data visualization is at the core of data science insight extraction, communication with others, and quality assurance." The report asserts that all students of data science should be exposed to the concepts of data consistency checking, exploratory data analysis, grammar of graphics, static and dynamic visualizations, and dashboards.

*Data Modeling and Assessment*

The NASEM Report (2018) asserts that data scientists enjoy the liberty of applying a wide variety of models and methods but the question is how to identify which models are most appropriate in a given situation. A related challenge is how to assess whether the assumptions and conditions required for applying that method are credible. Data modeling and assessment concepts and skill play a central role in being able to do data science. Of particular importance are (i) *machine learning,* (ii) *multivariate modeling and supervised learning,* (iii) *dimension reduction techniques* and *unsupervised learning,* (iv) *deep learning, model assessment and sensitivity analysis*, and also (v) *model interpretation*.

*Workflow and Reproducibility*

As indicated in the NASEM Report (2018), the term 'workflows' refers to pipelines of processes the objective of which is to combine various steps to accomplish the target at hand. Workflows need to be documented, improved, shared and generalized due to the fact that various tasks performed by a data scientist need to be *reproducible* and *replicable*. This enables others to understand the various steps involved in the data analysis process which, in turn, increases confidence in the results and facilitates *re-use* of analyses or results in a meaningful way. Needless to say, students of data science need to be exposed to the concept and methodology of workflows through programming languages such as R and Python. Key workflow and reproducibility concepts/skills that are important for all students of data science include *workflows and workflow systems*, *documentation and code standards*, *source code control systems*, *reproducible analysis*, and *collaboration.*

These five precursors will be the subject of 'Introductory Data Science I' wherein students will be learning the very basics of these concepts (to the extent that will be manageable through the functionalities of the mobile applications that will need to be created for this purpose). Once this is accomplished, the students will be ready for handling modest/moderate amounts of real data through mobile applications which will be the main theme of 'Introductory Data Science II'.

## MOBILE APPLICATIONS FOR DATA SCIENCE

Mobile applications are becoming increasingly popular all over the world and the speed with which apps pertaining to a wide variety of disciplines are being developed is truly remarkable! However, data science being a relatively new field, mobile apps designed for learning/executing various techniques of data analysis are yet limited in number. A few examples of these types of applications that have already been developed are presented below.

- *Basic Statistics:* This app is for beginners in data science/data analytics. It can be regarded as a refresher on various statistical measures. It includes a wide variety of topics including frequency distribution, graphs, data description, probability, distributions, estimation, hypothesis testing and others. (See Analytics Vidhya, 2015).
- *Statistics and Sample Size:* It is the statistical calculator for android devices and helps in calculating various statistical metrics such as sample size, statistical distribution table, statistical analysis and many others. According to Analytics Vidhya (2015), through the utilization of this app, one can save much time to be spent on doing calculations and is a handy tool for basic as well as scientific statistics. A limitation is that it supports only .csv files but this is not too big an impediment for dealing with moderate amounts of data.
- *Excel Tutorial:* This app enables the user to learn Excel on an android device. This app consists of tutorials covering a wide variety of Excel topics including sorting, filtering, pivot tables, keyboard shortcuts, what if analyses, etc. According to Analytics Vidhya (2015), this provides one of the best Excel tutorials on Google Play Store.
- *R Programming:* This app introduces the basics of R Programming which can be recommended for complete beginners. In addition to vectors, functions, matrices and factors, it contains data frames, lists and other entities related to data analysis. (See Analytics Vidhya, 2015).
- *QPython:* According to Analytics Vidhya (2015), this app enables one to run Python scripts and projects on one's android device; it consists of Python interpreter, console, editor, and many useful libraries; Python code & files can also be executed from QR codes. Thus, it is a useful app for doing data science through Python.
- *Tableau Mobile:* This app provides a simple and fast way for searching and exploring dashboards and workbooks. Users can browse the project hierarchy and see search results presented with clear structure,

just like they would on Tableau Online or Tableau Server. Users will spend less time looking for dashboards and more time analyzing the insights within them. Users can access the application even if the device is offline (See Cox, 2019.)

- *QlikView on Mobile:* This set of purpose-built apps enables one to instantly access and explore business intelligence data through just touch, pinch, zoom, and swipe actions. It thus enables one to stay in touch with a complete set of live data while one is on the run. Consequently, one is able to see what is happening with one's business, and is able to make smart and quick decisions on the go. In fact, one gets a wholesome business discovery experience including interactive analysis, rich visualization, and associative search. It is equipped with browser-based client technology and server-side security and manageability. (See QlikView on Mobile - Get in Touch With Your Data.)

With reference to the mobile applications indicated above, it is relevant to note that, at the present time, there exist a large number of freely available online tools for descriptive statistics and statistical inference that can easily be found through search engines (on personal computers, laptops and/or mobile phones) as long as one is connected to the internet. A distinctive feature of *mobile apps,* however, is that a number of them are developed in such a way that once downloaded and installed on the phone, they can be used even when one is offline. This is definitely an advantage in poor countries that are faced with shortages of electricity along with other challenges.

Although a few mobile apps are already in the market that have been designed for initial-level data science (some of which have been outlined above), *many more* will need to be developed in order to cater for the requirements of teaching and learning introductory-level data science. Once created and introduced in the market, these apps will provide ample opportunities to the students to practice various aspects of data science such as data cleaning, data analysis, inference, etc. It should be kept in mind that, in a group of students enrolled in these two courses, some might be having laptops/personal computers in their homes while others might not. As such, the mobile apps will need to be designed in such a way that they can be run on smartphones, laptops and personal computers, and work equally well on each device.

## LEARNING OUTCOMES AND ASSESSMENT

One of the foremost considerations in designing and developing a course of study is to define the set of *learning outcomes* that the students will be expected to have acquired upon completion of the course. As far as designing a pair of courses on data science based on *mobile apps* is concerned, it is obvious that the learning outcomes will be closely linked with the capabilities of the apps that are to be used. With reference to the apps that are already in the market, the following learning outcomes come to mind:

- students will be able to describe the data at hand through numerical descriptive measures and graphs,
- students will be able to calculate statistical metrics such as sample size,
- students will be able to perform a variety of data-related tasks on Excel such as sorting, filtering, pivot tables, etc.,
- students will be able to write basic R code (involving vectors, functions, matrices, factors, data frames, lists and other entities) for data analysis
- students will be able to run Python scripts and projects
- students will be able to interact with data through Tableau dashboards
- students will be able to indulge in a wholesome business discovery experience encompassing interactive analysis, rich visualization, and associative search.

Considering the rapidity with which technology is advancing and the speed with which mobile apps are being developed for a wide variety of purposes, it is obvious that the coming years will witness an ongoing series of innovations in mobile app creation for data science and, as such, learning outcomes for app-based courses in data science will need to be *continually updated* in order to keep pace with the new developments.

Another highly important consideration in designing a course is the set of mechanisms by which the learning outcomes will be *assessed* by the instructors. Obviously, this matter will need to be given as much importance and attention as the set of learning outcomes. Needless to say, similar to the learning outcomes, the mechanisms for assessment will need to be continually revised/updated in accordance with the developments that will continue to occur in mobile app creation for data science.

## CONCLUDING REMARKS

Data Science has become a buzzword in the present day and age where a multitude of data are being accumulated perpetually, and many undergraduates in the developed countries aspire to become data scientists. As far as undergrads of developing nations are concerned, they are as dynamic, enthusiastic and passionate as their developed-world counterparts but, many a time, face difficulties in achieving their goals due to non-trivial socio-cultural and economic constraints. The desire to study new, emerging disciplines such as data science is hindered due to lack of state-of-the-art computing resources (networks, servers, storage, applications, etc.) in educational institutions coupled with non-availability of laptops/personal computers in the students' homes. Nonetheless, keeping in view the importance of this discipline in today's data-driven world, statisticians of these countries cannot afford to 'sit back and relax'. Consolidated efforts are required to enable the statistics communities of developing countries to take their first steps toward the acquisition of knowledge and insight in the discipline of data science. The pair of 3 credit-hour courses on introductory data science based on freely downloadable mobile applications proposed in this paper as an integral component of the BS Statistics Programs of developing countries may appear to be an impracticable idea to some, but caters to the ground realities of undergraduate students of these countries. Challenging and daunting as it may appear to be today, this pair of app-based courses may turn to be the harbinger of full-fledged BS Programs on Data Science in the universities of developing countries in the years to come.

## ACKNOWLEDGMENT

The author is indebted to the reviewers for valuable comments and suggestions that have significantly enhanced the quality of the paper.

## REFERENCES

Allen, P. J., Roberts, L. D., Baughman, F. D., Loxton, N. J., Van Rooy, D., Rock, A. J., & Finlay, J. (2016). Introducing StatHand: A cross-platform mobile application to support students' statistical decision making. *Frontiers in psychology*, *7*, 288.

Analytics Vidhya. (2015). 18 Useful Mobile Apps for Data Scientist / Data Analysts. Retrieved February 12, 2020. www.analyticsvidhya.com/blog/2015/12/18-mobile-apps-data-scientist-data-analysts/

Cox, J. (2019, February 28). Introducing the new Tableau Mobile: Redesigned browsing experience and offline access to data, Retrieved from https://www.tableau.com/about/blog/2019/2/introducing-new-tableau-mobile#:~:text=When%20you%20are%20connected%20to,if%20the%20device%20is%20offline.&text=However%2C%20more%20complex%20actions%20like,not%20possible%20in%20Interactive%20Previews.

Falloon, G. (2013). Young students using iPads: App design and content influences on their learning pathways. *Computers & Education*, 68, 505-521.

Ling, C., Harnish, D., & Shehab, R. (2014). Educational apps: using mobile applications to enhance student learning of statistical concepts. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *24*(5), 532-543.

National Academies of Sciences, Engineering, and Medicine (2018). *Data Science for Undergraduates: Opportunities and Options.* Washington, DC: The National Academies Press.

QlikView on Mobile - Get in Touch With Your Data. Retrieved from https://rightqlik.com/qlikview-on-mobile.php . Retrieved on October 31, 2020.