The Biostatistics Master of Public Health in the USA: Practical Information for Other Countries

Kendra K. Schmid, Gleb R. Haynatzki*

Biostatistics Department
College of Public Health
984375 Nebraska Medical Center
ghaynatzki@unmc.edu

Abstract

This article considers the general approach to education leading to the Master of Public Health (MPH) in Biostatistics degree in the USA. The discussion covers the (bio)statistics education that a typical US student receives from kindergarten to graduate school, as well as the role of mathematics education in the MPH of Biostatistics curriculum. The hope is that the US experience could be of help to colleagues from other countries in designing their Master's-level Biostatistics programs (MPH, MSc or others).

I. Introduction

In the modern world availability of data is increasing much faster than our capabilities to process them and extract valuable information. Two areas generating large amounts of data are medicine and public/population health. Biostatistical methods are currently the most reliable and widely accepted approach to analyze biomedical and public health data.

Biostatistics and the biostatistics education in the USA have a long tradition that follows that of the British school, which is much more applied than other major European schools, e.g. the German, French, and Russian schools of probability and statistics. Currently, the USA, Great Britain and Europe have statistics education systems that are converging at some level, usually at the doctoral level. However, it is interesting to evaluate the prevalent approach taken in the USA towards the education of Master of Public Health (MPH) biostatisticians, whose realization on the job market is very practice-oriented. The insights from our study could be helpful in the development of similar Master's-level Biostatistics programs (MPH, MSc or others) around the world, including Bulgaria.

II. The K-12 Years

In theory, the mandatory statistics (and biostatistics) education in the USA begins in kindergarten (K) and continues through high school (Grade 12), i.e. K-12, following certain standards formulated by the NCTM (National Council of Teachers of Mathematics) since statistics is taught as part of the mathematics curriculum. A set of well-defined but rudimentary statistics skills are sometimes developed during K-5, depending on how much time a mathematics teacher spends on statistics exercises and problem solving. For example, the most recent set of expectations from all students in grades K-2 are formulated by the NCTM and listed in [1]. The respective NCTM expectations from all students in grades 3-5 are listed in [2], and we give them here as an example:

*We are greatly indebted to our colleagues from the biostatistics department and the MPH program for their substantial support in developing the Biostatistics MPH Concentration.

- design investigations to address a question and consider how data-collection methods affect the nature of the data set;
- collect data using observations, surveys, and experiments;
- represent data using tables and graphs such as line plots, bar graphs, and line graphs;
- recognize the differences in representing categorical and numerical data;
- describe the shape and important features of a set of data and compare related data sets, with an emphasis on how the data are distributed;
- use measures of center, focusing on the median, and understand what each does and does not indicate about the data set;
- compare different representations of the same data and evaluate how well each representation shows important aspects of the data;
- propose and justify conclusions and predictions that are based on data and design studies to further investigate the conclusions or predictions;
- describe events as likely or unlikely and discuss the degree of likelihood using such words as certain, equally likely, and impossible;
- predict the probability of outcomes of simple experiments and test the predictions; and
- understand that the measure of the likelihood of an event can be represented by a number from 0 to 1.

The above expectations seem well-intended and important but, as mentioned above, the respective skills are rarely developed at a satisfactory level because of the insufficient number of hours spent on the material, lack of mathematical maturity, and lack of real datasets and projects to work on. The NCTM expectations from students in Grades 6-8 [3] and Grades 9-12 [4] sound even more comprehensive and exciting than those for lower grades, but the results are similar, i.e. students rarely develop any good statistical skills in regular mathematics classes. Statistics skills may be sometimes touched upon in science classes (physics, chemistry, biology, ecology, etc) and social studies classes (history, geography, economics, etc) in middle and high school, but the direct connection between statistics, and its' use for data collection and analysis in other fields is not realized. Also, science fair projects (on natural or mathematical phenomena) for middle and high school students are another opportunity at working on statistical and probabilistic ideas.

The first serious statistics class available to K-12 students is the AP Statistics (AP = Advanced Placement, cf. [5]), usually available to high school students in Grades 10-12. It is of the same difficulty and follows identical syllabus (including sometimes the textbook) as the Beginner's Statistics Course at college/university undergraduate (Bachelor's) level.

III. Undergraduate (Bachelor's) Degree Years

The AP/Beginner's Statistics Course is a very important introductory statistics class. In particular, it provides an opportunity to high school students to obtain college/university credit, after passing the respective national AP Statistics exam with grade at least 3, from range 1-5. A good example of a widely used textbook in this course is "Introduction to Practice of Statistics" by David S. Moore and George P. McCabe [6]. It does not require Calculus as a prerequisite and is therefore a truly applied statistics class. It covers the following material (for more details, check the AP Statistics website, [5]):

- data types: continuous/numerical, categorical/nominal, ordered categories / ranks;
- measures of central tendency for continuous data: mean, median, mode;
- measures of spread for continuous data: standard deviation (SD), range, quartiles, percentiles, inter-quartile range (IQR);
- data summaries: tables, charts, plots;
- probability laws;
- binomial experiment and binomial distribution;
- normal distribution and normal approximation to the binomial distribution;
- Central Limit Theorem (CLT);
- sampling and experimentation: planning and conducting a study;
- simple random samples and the resulting sampling distributions, use of CLT;
- confidence intervals (CIs) for population means – continuous data case;
- 1- and 2-sample t-tests, including the matched pairs t-test;
- one-way Analysis of Variance (ANOVA);
- sometimes, nonparametric tests;
- correlation and simple/univariate linear regression;
- population proportions: CIs, 1- and 2-sample z-tests;
- contingency tables (one- and two-way) and the Chi-square test for dichotomous data from two or more groups, for goodness of fit, homogeneity of proportions, and independence;
- usually an advanced graphing calculator, with statistical functions is used in high school, whereas at college level a computer with a statistics software package is used to store and retrieve data, and to apply methods of analysis discussed in class to a dataset;
- interpretation of the results from statistical analysis.

These concepts are worked out on lots of exercises and problems taken usually from published studies or publicly available data sets, and students can get a firm grasp of the concepts and their applications. However, the textbook datasets are of limited size and often statistics calculations could be accomplished by hand. Software (e.g. Minitab, SPSS, SAS, R) is often used for the sake of learning it, and the courses are often taught by individuals with no statistics background. The AP/Beginner's Statistics Course does not have Calculus or Linear/Matrix Algebra as prerequisites, but a college student who has taken these classes may also take the first Probability/Mathematical Statistics class, which class essentially justifies mathematically the methods used in the AP/Beginner's Statistics course. There is a wide variety of options for undergraduate students to expand their knowledge of statistics within their Bachelor's degree curriculum, including obtaining a major or a minor in statistics, which are offered at most universities. Other opportunities include intermediate-level applied statistics classes offered by psychology, economics, and other social sciences departments.

IV. The Master of PH Program – Competencies

The Biostatistics MPH programs in the USA look very much alike as long as competencies and specialized biostatistics courses are concerned, regardless the institutions offering them. To be more specific, we will give as an example the Biostatistics Concentration of the Master of Public Health (MPH) Program at the College of Public Health in the University of Nebraska Medical Center (UNMC), Omaha, in whose development the authors of this article participated very actively. Although a graduate with a MPH degree will be primarily prepared to work in the area of biomedicine or public health, often on interdisciplinary teams, s/he could equally successfully work in the area of, say, banking or whenever statistical modeling is a major expectation. A Biostatistics MPH program focuses on the following competencies which could be easily reformulated or translated for areas other than biomedicine or public health.

1. Statistical Considerations in Study Design

   A Formulate pertinent research questions and hypotheses in statistical terms

   B Identify strengths and weaknesses of study designs and implement scientifically and statistically sound design strategies.

   C Select variables relevant to a specific public health or biomedical problem for utilization in statistical design and analysis.

   D Recognize sources of bias and confounding in study design.

   E Determine statistical power and sample size needed for future public health and biomedical studies.

2. Perform Statistical Analysis of Data

   A Apply appropriate statistical methods for estimation and inference, including univariate and multivariate methods appropriate for continuous, categorical, and time-to-event data.

   B Utilize a software package for data management, statistical analyses and data presentation.

   C Apply statistical methods for quality control and data cleaning to already collected data, before the actual statistical analysis.

   D Verify assumptions of statistical tests and models and implement appropriate methods to address observed violations of the assumptions.

   E Apply basic measures to account for confounding factors in the analysis of public health and biomedical studies, including matching, and multivariable analysis.

   F Evaluate the strengths and limitations of statistical analyses of public health and biomedical studies.

3. Interpretation and Dissemination of Results

   A Develop written and oral presentations based on statistical findings for both public health professionals and lay audiences

4. Ethical/Legal Treatment of Human Subjects

   A Be familiar with the Institutional Review Board (IRB) research requirements and processes .

V. The Biostatistics MPH Program – Course Work

The prerequisites for entering a Biostatistics MPH program are Differential and Integral Calculus and an undergraduate statistics course, whereas a Linear/Matrix Algebra course is recommended. In reality, only a few college students take any of these classes (and even fewer students take both classes), and therefore only few students are ready to enroll in a Biostatistics MPH program. Once the prerequisites are met and a student is admitted to the MPH program in Biostatistics, they take the following courses.

1. Biostatistics I, which is an enhanced version of the AP/Beginner's Statistics Course, e.g. with nonparametric tests and a statistical package such as SAS® (SAS Inst, Inc, Cary, NC, USA) or IBM SPSS®, among others. A good example of a widely used textbook for this course is "Fundamentals of Biostatistics" by Bernard Rosner [7].

2. Applied Linear Statistical Models is a rigorous applied statistics course which is the fundamental course of the Biostatistics MPH program, and all other specialized classes depend on it and often have it as a prerequisite. This course aims at preparing the student to analyze continuous response data and interpret results using methods of multiple linear regression and analysis of variance (ANOVA). The major topics to be covered include simple and multiple linear regression model specification and assumptions, specification of covariates, confounding and interactive factors, model building, transformations, ANOVA model specification and assumptions, analysis of covariance (ANCOVA), multiple comparisons and methods of adjustment, fixed and random effect specification, nested and repeated measures designs and models, and diagnostic methods to assess model assumptions. A good example of a widely used textbook for this course is "Applied Linear Statistical Models" by Michael H. Kutner, Christopher J. Nachtsheim, John Neter, and William Li [8].

3. Categorical Data Analysis is a specialized course focused on the methods for analysis of categorical response and count data. The major topics covered include proportions and odds ratios, multi-way contingency tables, generalized linear models, logistic regression for binary response, models for multiple response categories, loglinear models, and simple mixture models for categorical data. A good example of a widely used textbook for this course is "An Introduction to Categorical Data Analysis" by Alan Agresti [9]. Sometimes, another more software oriented textbook is added, e.g. "Categorical Data Analysis Using The SAS System" by Maura E. Stokes, Charles S. Davis, and Gary G. Koch [10].

4. Survival Data Analysis teaches the basic methods of statistical survival analysis used in clinical and public health research. The major topics covered include the Kaplan-Meier product-limit estimation, log-rank and related tests, and the Cox proportional hazards regression model, including with time-dependent covariates. A good example of a widely used textbook for this course is "Survival Analysis: A Self-Learning Text" by David G. Kleinbaum and Mitchel Klein [11]. Sometimes, another more software oriented textbook is added, e.g. "Survival Analysis Using the SAS System: A Practical Guide" by Paul D. Allison [12].

5. Correlated/Repeated Measures Data Analysis focuses on the methods for analysis of correlated continuous, binary, and count (response) data. The major topics covered include linear models for longitudinal continuous data, generalized estimating equations, generalized linear mixed models, impact of missing data, and design of longitudinal and clustered studies. A good example of a widely used textbook for this course is "Applied Longitudinal Analysis" by Garrett M. Fitzmaurice, Nan M. Laird, and James H. Ware [13]. Sometimes, another textbook is added, e.g. "Analysis of Longitudinal Data" by Peter J. Diggle, Patrick Heagerty, Kung-Yee Liang, and Scott L. Zeger [14].

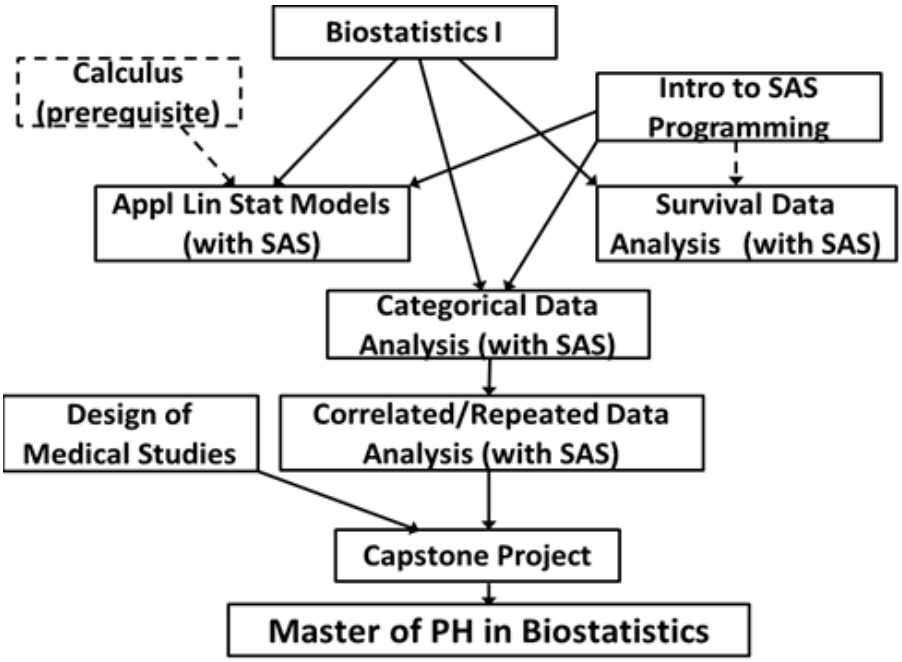

Figure 1: *A flowchart showing the required courses for the MPH in Biostatistics degree. The arrows indicate which course is a prerequisite for which. Note that the course Applied Linear Statistical Models could be used as a prerequisite for the course Categorical Data Analysis.*

6. Design of Medical Studies aims at preparing the Biostatistics MPH student to understand and apply principles and methods in the design of biomedical and public health studies, with a particular emphasis on randomized, controlled clinical trials. The major design topics covered include sample selection, selecting a comparison group, eliminating bias, need for and processes of randomization, reducing variability, choosing endpoints, intent-to-treat analyses, sample size justification, adherence issues, longitudinal follow-up, interim monitoring, research ethics, and non-inferiority and equivalence hypotheses. A good example of a widely used textbook for this course is "Fundamentals of Clinical Trials" by Lawrence Friedman, Curt D. Furberg, and David L. DeMets [15].

7. Elected courses, e.g. Introduction to SAS Programming, Advanced Epidemiology.

8. Service Learning/Capstone Experience project on real data. This is an internship experience that prepares the soon-to-be MPH in Biostatistics graduates for the realities of their future job and workplace. It is conducted at an institution other than the College, e.g. a research hospital, a city/state health department, or a community organization that are substantially involved in public health. The student gives a presentation on their project shortly before graduation, and may or may not continue working at that institution after graduation.

Next, a flowchart of the different courses included in a Biostatistics MPH Program is shown. Note that, in addition to the above mentioned required specialized biostatistics classes and electives, there are also some other core, non-biostatistics courses (e.g. The US Healthcare System), which will not be discussed in detail here.

VI. Mode of Delivery

In determining the mode of delivery, consideration should be given to the target audience of the degree program and what may best fit those students. In the MPH program at UNMC, efforts were made to deliver instruction in ways that would attract students who may otherwise be unable to attend courses due to scheduling conflicts or distance from the campus. Three main modes of delivery have been utilized and are briefly discussed.

First is the traditional classroom delivery that is mostly lecture based with assignments, projects and/or exams as assessment. Second, a video streaming method was used for students located away from the main campus to gain access to course lectures. This method required students to login to the live class session and participate in real-time. It also required students to submit assignments via electronic mail instead of on paper and to arrange for exam proctors at their location(s), but still did not allow for accommodation of schedules. Third is an asynchronous online delivery. With this method, students can work asynchronously on their own time within given deadlines from the course instructor. The online delivery requires more initial preparation, but allows for accommodation of students and instructors in different locations and with different schedules as it does not require any real-time interaction. The MPH in Biostatistics at UNMC can be completed either entirely in class, entirely online, or in combination for added flexibility for the students.

VII. Discussion

The MPH in Biostatistics degree awarded in the USA is a very applications-oriented degree aiming at preparing students to learn how to analyze real data sets, interpret the obtained results, and communicate them to clients. It does not aim at teaching them how to develop new statistical methods or do rigorous mathematical proofs, which would be the purview of a Ph.D. Program in Biostatistics. Thus, no Probability/Mathematical Statistics course is required as a prerequisite or as part of the MPH in Biostatistics, but having taken such a course would considerably strengthen the application of a student who has applied to a MPH in Biostatistics program, since every (bio)statistics teacher prefers students with strong quantitative background.

A biostatistician with a Biostatistics MPH degree must also have excellent skills with a statistical software package since no real data set could be efficiently managed and analyzed without a computer. Usually, SAS® (SAS Inst, Inc, Cary, NC, USA) is the most adequate for the purpose, and is used at numerous companies and universities in the USA. However, SAS is high priced which may be prohibitive for institutions or companies not having sufficient funds to invest in it. Another statistical software package, R, which is free, is becoming very popular around the world, including at companies and universities.

On real datasets/projects, a MPH biostatistician should be ready to work alone or team up with a Ph.D. biostatistician, and should have good communication skills, in order to collaborate efficiently with clients on their projects.

The road to Ph.D. in Biostatistics may or may not go through a MPH degree in Biostatistics, and if it does, may require additional theoretical coursework to supplement the MPH coursework. However, in our experience, many students enrolled in a Ph.D. in Statistics program wish to take biostatistics classes, since that expertise makes them more marketable after graduation.

It seems that mathematics and statistics departments in different countries offering a MPH degree in Biostatistics currently struggle with the issue of how much mathematical preparation a MPH student needs. It is obvious that if such a graduate will continue into a Ph.D. program in Biostatistics, s/he will need better mathematical preparation than someone who will join, e.g., industry. However, if the mathematical preparation of the majority of students entering a MPH program is high already, the issue is resolved by itself since in this case more mathematical rigor could be incorporated into the MPH program.

Finally, holders of a MPH in Biostatistics degree have been very successful at working as data analysts in industry and at universities in the USA, and therefore the US experience could be of help to other countries around the world in designing and establishing their own MPH in Biostatistics programs.

References

1. http://standards.nctm.org/document/chapter4/data.htm
2. http://standards.nctm.org/document/chapter5/data.htm
3. http://standards.nctm.org/document/chapter6/data.htm
4. http://standards.nctm.org/document/chapter7/data.htm
5. http://www.collegeboard.com/student/testing/ap/sub_stats.html
6. D. S. Moore, G. P. McCabe G. P. Introduction to the Practice of Statistics, 7th ed., W.H. Freeman and Company, 2012.
7. B. Rosner. Fundamentals of Biostatistics, 7th ed., Brooks/Cole, 2011.
8. M. H. Kutner, C. J. Nachtsheim, J. Neter, W. Li. Applied Linear Statistical Models, 5th ed., McGraw-Hill Education. 2004.
9. A. Agresti. An Introduction to Categorical Data Analysis, 2nd ed., Wiley-Interscience, 2007.
10. M. E. Stokes, C. S. Davis, C. S., G. G. Koch. Categorical Data Analysis Using The SAS System, 3rd ed., SAS Institute, 2012.
11. D. G. Kleinbaum, M. Klein. Survival Analysis: A Self-Learning Text, 3rd ed., Springer, 2011.
12. P. Allison. (2010) Survival Analysis Using the SAS System: A Practical Guide, 2nd. ed., SAS Institute.
13. G. Fitzmaurice, N. Laird, J. H. Ware. Applied Longitudinal Analysis, 2nd ed., Wiley, 2011.
14. P. J. Diggle, P. Heagerty, K.-Y. Liang, S. Zeger. Analysis of Longitudinal Data, 2nd. ed., Oxford Statistical Science, 2013.
15. L. Friedman, C. D. Furberg, D. L. DeMets. Fundamentals of Clinical Trials, 4th ed., Springer, 2010.