TEACHING STATISTICS IN THE CONTEXT OF BIOLOGY: THE SYMBIOSIS EXPERIENCE

Edith Seier[1] and Karl H. Joplin[2]
[1]Department of Mathematics and Statistics, East Tennessee State University, United States of America
[2]Department of Biological Sciences, East Tennessee State University, United States of America
seier@etsu.edu

We are part of a team that designed and taught a sequence of three integrated courses combining biology, mathematics, and statistics, at the freshman/sophomore level (SYMBIOSIS project, HHMI grant # 52005872). This presentation describes the statistical content of the integrated courses and the stand-alone introductory statistics course that emerged from it. The focus of the paper is on the questions that arose in the design of the courses and the way they were addressed when preparing the syllabus and the teaching material.

INTRODUCTION

Research in biology has been the motivation for the development of many statistical methods. Biology also provides a very good context for the teaching of statistics. However, questions and challenges arise when deciding what statistical topics to teach to biology students and how to teach them. Research in biology in general and genetics in particular, has become increasingly quantitative. The *Bio2010* initiative (National Research Council, 2003) and the *Scientific Foundations for Future Physicians* (AAMC & HHMI, 2009), encourage a more quantitative preparation of future researchers in biology and medicine. Efforts have been made to implement these suggestions at several universities. However, the new courses and research experiences on quantitative methods for biology target mostly graduate or advanced undergraduate students. At ETSU the decision was made to focus first on incoming freshmen.

Faculty from the Department of Biological Sciences and the Department of Mathematics and Statistics at ETSU developed and co-taught a sequence of three double credit courses at the freshman/sophomore level integrating biology, mathematics and statistics (Joplin et al., 2009). The modules or chapters have biological names and the mathematics and statistics topics are introduced in those specific contexts. Based on that integrated experience, new biology, calculus, and introductory statistics courses and integrated modules emerged. This presentation focuses on the statistical component of the integrated courses, the resulting new introductory statistics course and some of the additional modules. Teaching statistics totally immersed in the biological context posed some challenges and the decisions made for these integrated courses definitely impacted the design of the stand-alone course *Introduction to Statistics in a Biological Context*.

Some of the questions addressed in this presentation are:

- Generic course or focused course?
- How to introduce statistical inference during the first weeks of class?
- Randomization based inference only, or also normal based inference?
- Less or more probability?
- Should the students be introduced to the concept of power?
- What to teach in the first course?
- Some topics for later.
- What software to use?

GENERIC COURSE OR FOCUSED COURSE?

At our university, the introductory statistics course is a general education requirement but students in the Bachelor of Science program in biology can substitute it for a calculus course. As a result of the *SYMBIOSIS* experience, the mathematics department is creating a version of the introductory statistics course focused on applications to biology that contains some topics not included in the general course. The biology department plans to require both statistics and calculus as co-requisites to their introductory biology courses.

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php

To illustrate the integration between biology and statistics we will mention one simple example. In the second module ('The Cell') of Symbiosis I, statistical graphs and descriptive statistics are covered including correlation. In a data base of mammalian red blood cells/ml for different species, two species immediately stand out because of the large number of red blood cells: the vicuña and the llama. We ask the students: 'What is different for these two species from the other species in the data set? One suggestion is the altitude at which they generally live. Does something similar happen to humans? Fortunately, there is a published data set (Spector, 1959) with the altitude of residence and the number of red blood cells for 17 individuals. This example is used to introduce the correlation coefficient and, from there, the discussion of the biological principle of adaptation follows.

INTRODUCING STATISTICAL INFERENCE EARLY IN THE SEMESTER

The first module in Symbiosis I is called 'The Scientific Method', thus the issue of testing hypotheses is discussed early in the semester. Two cases of hypotheses testing are introduced during the first weeks of class and then treated with more detail later. The first one corresponds to a case common in scientific research: the comparison of treatment vs. control in an experiment. A hands-on activity with plastic chips is done in class to introduce the randomization test. A program in R that mimics the hands-on activity is used to get a large number of re-groupings of the data to calculate an approximated p-value. The basic concepts of probability and the binomial distribution are also introduced rather early in the semester and the exact test for proportions is discussed as an exercise on the binomial distribution.

The early introduction of hypothesis testing has been a positive experience since students have the opportunity throughout the course of practicing the writing of statistical hypotheses, calculating an exact or approximated p-value, and deciding whether to reject or not reject the null hypothesis. In the stand-alone introductory statistics course, the notion of hypotheses testing is also introduced during the first week. The randomization test, bootstrapping, and the exact test with the binomial are introduced after a chapter on data production and descriptive statistics,

RANDOMIZATION METHODS ONLY?

Randomization methods are excellent to introduce ideas of statistical inference very early in the semester because they need few pre-requisites. These methods are steadily making their way into introductory statistics courses (Lock & Lock, 2008). In traditional introductory statistics courses, inference is studied toward the end of the semester after the pre-requisites for normal based inference have been covered. The question is: Should we teach the whole introductory statistics course using only randomization methods and forfeit the normal based inference? It was decided to cover both, the randomization methods in the first part of the course and the traditional normal based inference at the end of the semester since the students will later take biology courses where the t-test is a standard procedure. In order to give a sense of unity to the different hypotheses testing methods, students are reminded that a key question is 'how likely is it to get this result or a more extreme one when the null hypothesis is true?', and that there are several ways of calculating or approximating that probability depending on the context.

LESS OR MORE PROBABILITY?

A current point of discussion in statistical education forums is whether the probability content of introductory statistics courses should be reduced. In the case of biology and health science students we feel they need more, not less probability. Exposure to more probability distributions and some applications of probability to genetics might be convenient. In the *SYMBIOSIS* courses and the introductory statistics course that emerged from them, all the probability exercises come from a biological/health sciences context. One of these applications is conditional probability in the context of medical diagnosis. The Bayes rule is taught using two-way tables (Rossman & Short, 1995) and probability trees, before even looking at the formula. A coin model and probability trees are used to understand genotypes, phenotypes, and the passing of information from parents to offspring in Mendelian genetics. This is followed by the goodness of fit test using Mendel's data for seven characteristics of peas. Probability exercises related to the inheritance of diseases associated with either recessive or dominant alleles of genes are included.

DNA is presented as a sequence of 4 letters (A,C,G,T) and simple exercises to calculate the probability of palindromes (associated with restriction enzymes that cut the DNA strand), and other specific sequences are discussed.

SHOULD THE STUDENTS BE INTRODUCED TO THE CONCEPT OF POWER?

Power of statistical tests is one of those topics that are frequently omitted in introductory statistics courses due to lack of time. The understanding of power is not easy for introductory statistics students. The sample size problem is usually treated in an introductory statistics course solely in the context of confidence interval estimation. However, experimentation and hypotheses testing are common in the biological/medical literature. The idea of power is introduced in the context of the exact test for proportions using a binomial table and a ruler. Later, power curves in the normal based inference context are shown. The learning goal is that students understand the notion of power and make sense of a power curve plot in relation to sample size.

WHAT TO TEACH IN THE FIRST COURSE?

In the first course of the integrated sequence, *SYMBIOSIS I*, the modules or chapters are: The Scientific Method, The Cell, Size and Scale, Mendelian Genetics, DNA Genetics, and Evolution. The statistical topics are introduced in each specific context.

The chapters in the one-semester stand-alone introductory statistics course, which emerged from the symbiosis experience, are: The Scientific Method, What Are the Data Telling Us? (statistical graphs including bivariate and multivariate ones, and descriptive statistics including correlation), Inference by Randomization, Probability and the Binomial Distribution, Testing Hypotheses using the Binomial Distribution, Conditional Probability, Discrete Distributions, Applications of Probability to Mendelian and DNA Genetics, Continuous Distributions, Checking Models, Normal-based Inference, Regression Models and More on Design of Surveys and Experiments. The basics on random sampling and experimental design are introduced at the beginning and reinforced throughout the course.

SOME TOPICS FOR LATER

The statistical topics in the *SYMBIOSIS II* and *III* courses have also become part of stand-alone modules and tutorials that can be used either inside a course (biology, math, or statistics) or in integrated workshops. Those modules will be made available in the internet, some of them are described next.

The Chronobiology module (from *SYMBIOSIS* II) has four sections (the third one is the statistics component of the module): rhythms in nature, harmonics, the periodogram, and circadian oscillators and entrainment. Another statistical topic from *SYMBIOSIS* II is a brief introduction to non-linear regression in the context of enzyme kinetics.

There are two modules in SYMBIOSIS III that have a strong statistical component: Developmental Biology and Introduction to Bioinformatics. They are taught one after the other and they form a unit. The first part of the statistical material consists of 11 one-hour lessons covering the following topics: microarray data files, normalization, ANOVA, finding the most differentially expressed genes (3), matrices, principal components (2), clustering (2). After finishing the lectures, students were assigned a project to analyze gene expression through the developmental stages of *Drosophila melanogaster* (embryo, larva, metamorphosis and adult) using data from the NCBI data base. The second module ends with a brief introduction to the comparison of DNA sequences, for which the NCBI data base and software (BLAST and Clustal-W) are used.

WHAT SOFTWARE TO USE?

In class we mainly use R, but Minitab was also used and our students work in the lab with what they feel more comfortable, with the exception of the topics for which R is definitely used (bootstrapping, randomization test, star plots, applying t-test and ANOVA to hundreds of genes, etcetera). Students are introduced to R not only because it is powerful and freeware, but because it is extensively used in bioinformatics. Programming in R is not a learning goal, rather the commands are given to the students and they replace the data in the examples by their own data.

CONCLUSION

To include hypothesis testing and estimation from the beginning of the semester via randomization methods is useful not only because students are motivated by seeing the role of statistics in the scientific method, but also because repeating the concepts related to inference several times during the semester helps the students to become more familiar with them.

Teaching in the context of a given field, biology in our case, helps us to select which topics to teach or emphasize in a course where the generic version is traditionally crowded with many topics. It is also easier to find motivating examples and exercises for the students.

We find the experience of teaching biology, mathematics, and statistics in an integrated way to be very interesting. For a statistician, to co-teach a course in context with the specialist in an area of application of statistics can be beneficial even if it is done for a limited time. What we learned from the experience definitely influenced the way the introductory statistics course for biology and pre-professional majors was designed. We feel that we benefited more from this integrated experience than if we had just tried to add biological examples to a standard introductory statistics course. That feeling was shared by the participating mathematicians. The material prepared for the integrated course was easily transformed into material for stand-alone introductory statistics and calculus courses. Modules with the more advanced material can be used in courses in one of the three fields or in integrated courses or workshops.

During the years in which we were designing and teaching the integrated courses, we realized that a good level of communication between the participating biologists and statisticians had been achieved, and also between mathematicians and statisticians.

Working in the design of the integrated courses there was the pressure of having to teach statistical topics embedded into, and in agreement with, the biological topics that were being covered. That forced us to make several decisions about what to cover and emphasize, such as the extent of probability content and applications, and the location of inference in the course. Even when that pressure is off in a stand-alone introductory statistics course, some of the decisions are maintained, smoothing out the transition from topic to topic and completing details that would make the stand-alone course more cohesive from the statistical point of view. The integrated experience and day to day collaboration was an invaluable source of inspiration for the material of the separated new courses. The evaluations of the students with respect to the statistical component were good. One of the main satisfactions was to teach topics such as microarray data analysis to sophomore students, and that the students had the background to be able to cope with that material.

REFERENCES

National Research Council (2003). *Bio 2010 Transforming Undergraduate Education for Future. Research Biologists*. Washington DC: The National Academies Press.

Joplin, K., Seier, E., Helfgott, M., Karsai, S., Knisley, J., Moore, D., & Miller, H. (2009). SYMBIOSIS An Integration of Biology and Statistics at the Freshman Level: Walking Together Instead of on Opposite Sides of the Street. *MAA notes on Undergraduate Mathematics for the Life Sciences: Process, Models, Assessment, and Directions (forthcoming).*

Lock, R. H., & Lock, P. F. (2008). Introducing Statistical Inference to Biology Students Through Bootsrapping and Randomization. *Primus*, *18*, 39-48.

Rossman, A. J., & Short, T. H. (1995). Conditional Probability and Education Reform: Are They Compatible. *Journal of Statistics Education*, *3*(2).

Spector, W. S. (1956). *Handbook of Biological Data*. Philadelphia: Saunders.

AAMC & HHMI (2009). Scientific Foundations for Future Physicians report. Online: www.hhmi.org/grants/sffp.html.