AN EARLY START ON INFERENCE

Edith Seier
East Tennessee State University, Johnson, Tennessee, USA
seier@etsu.edu

Starting to teach inference early in an introductory statistics course means that the students have more time to assimilate the new concepts involved in estimation and hypothesis testing, especially if they are exposed to them in a sustained way throughout the semester. I teach a special section of the algebra-based course in which we start writing statistical hypotheses during the first week. We tell the students the general idea about testing hypotheses and that the details on how to calculate or approximate the p-value will depend on the context and the tools available. Randomization methods (permutation tests and bootstrapping) are introduced first because they require less background. After covering the basics of probability, the binomial distribution is used to do inference about one proportion. The classical methods using the normal, t-student and Chi-square distributions are studied at the end of the semester after these distributions have been introduced.

INTRODUCTION

In traditional one-semester algebra-based introductory statistics courses, statistical inference is usually introduced in the last third part of the semester. The reason is that prerequisites such as the normal distribution need to be covered first. For many students, it is difficult to grasp the main ideas of inference in a short time when they are already overwhelmed and close to the end of the semester. They have to face, at the same time, new important concepts and numerous formulas. Starting to introduce the language and ideas of statistical inference much earlier allows us to capture the attention of the students when they still have more time and energy. It also allows the students to see the concepts again and again during the semester, thus acquiring more familiarity with them. Each one of us might think on a different way of approaching the early introduction of inference in an intro-stats course. This paper relates my own experience on designing and teaching a course that introduces inference quite early. We first developed and taught the material for two years as part of a special project that integrated biology, mathematics and statistics at the freshman level (Joplin et al., 2013). The statistical component later evolved into the statistics course that is today and that we have taught for the last three years as a special section of the general education introductory statistics course. The integrated course started with a discussion of the scientific method and that explains in part why we talk about hypotheses right at the beginning of the semester in the stand-alone stats course. One of the main characteristics of the course is that instead of teaching hypotheses testing using a single approach (either normal-based inference or randomization), we adopt an eclectic approach using different methods around the central idea that what we need is to calculate or approximate the p-value, and that there is more than one way of doing this. Seier and Joplin (2011) is used as the textbook for this course.

WRITING STATISTICAL HYPOTHESES IN THE FIRST WEEK

In the first week of classes the concepts of population, individuals, samples, variables, parameters and statistics are introduced. Parameters and statistics are presented as summaries of the values of the variables at the population and sample levels respectively, and the symbols for population and sample mean and proportion are given. We also talk about hypotheses people make, especially in science, and how we translate either scientific hypotheses or research questions into two contrasting ideas that are called statistical hypotheses, and that are written in terms of parameters. We practice writing H_o and H_a either in terms of μ or p for a given story depending on the nature (quantitative or categorical) of the variable. One-group and two-group examples are considered. Students are told that decisions are made with regard to the null hypothesis (reject or not reject) based on data and that we need to collect data through an observational study or an experiment and learn how to display and summarize the data. We tell the students that after we get the information from the survey or experiment we need to evaluate how likely it is that such an outcome, or a more extreme one, happens when the null hypothesis is true. If such outcome would only very rarely or almost never happen when the null hypothesis is true, then the data might be

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

giving us evidence against the null hypothesis. Even when we have not introduced probabilities formally yet, students usually have an intuitive understanding of the word *probability*. We tell them that the probability of getting the outcome we got, or a more extreme one, receives a special name: *p-value*. In order to calculate or at least approximate the p-value we need to have an idea of what can happen when the null hypothesis is true and that there are several ways of getting such knowledge (a distribution of reference) depending on the context (the nature of the null hypothesis) and the tools we have available. We also tell them that they will see several concrete examples throughout the course. Thus, when we later teach an actual testing hypothesis method, the students are already familiar with the notion of statistical hypotheses and they know how to write H_o and H_1 for a given story. They also know that they will need to produce, in some way, a distribution under H_o to gauge how likely the outcome of the survey or experiment is to happen when H_o is true.

After this first chapter about statistical hypotheses, we discuss the methods of producing data by observation and experimentation. The chapter on statistical plots and descriptive statistics for quantitative and categorical variables comes next. The cases of one and two variables are covered; correlation, odds ratio and relative risk are included.

RANDOMIZATION TESTS AND BOOTSTRAPPING

In recent years randomization methods are being included in some introductory statistics courses such as Rossman and Chance (2008), Tingle et al. (2011) and Lock et al. (2012). We introduce *randomization methods*, right after the descriptive statistics chapter. We begin with the comparison of two groups because it is a more common research question in science than the one-group case and examples tend to be more exciting. We also find that the randomization test to compare two groups is easier to explain than bootstrapping. The randomization test is introduced with a tactile experience and a small data set involving treatment and control. The idea is to test H_o:$\mu_1=\mu_2$ against H_a:$\mu_1>\mu_2$. The means of treatment and control groups are calculated and their difference $d = \bar{x}_1 - \bar{x}_2$ is marked on a horizontal axis drawn on the board. Students are organized in teams and each team receives a set of plastic chips (n_1 of one color, for treatment, and n_2 of another color for the control group) in which the observed values of the response variable have been written. They are asked to mix the chips and randomly select n_1 chips to form *random group 1*and the remaining will form *random group 2*. They calculate the mean of each randomly formed group and report the difference of means, which we plot on the axis on the board. Soon an empirical distribution starts forming and the difference reported for the treatment and control group starts to look quite extreme to have happened just by chance. In order to have an empirical distribution with more random regroupings, code that mimics the tactile experience is given and pasted into R obtaining a dotplot with the differences of means for thousands of randomly formed groups. The approximated p-value is calculated as the proportion of differences of means that are equal or exceed the value of *d*. Knowledge of R is not a pre-requisite for the course; we do a small introduction in the course. Code in R is available from http://faculty.etsu.edu/seier/RcommCh3.txt
The randomization test for paired data is studied using the example with Darwin's maize data found in Fisher (1966).

Bootstrapping is used to do estimation with percentile confidence intervals. A tactile experience is used to grasp the notion of bootstrap samples. The code in R mimics the experience, displays the bootstrap replicates of the sample mean in a dotplot, and finds the central 95% or 90% part of the empirical distribution to produce the confidence interval. The confidence interval is used to test hypotheses in the one group case, H_o: $\mu=\mu_o$.

Thus, at the end of that chapter the students know how to estimate parameters using confidence intervals and how to test hypotheses for the one-population, paired data and the two-populations case when working with quantitative variables using randomization methods. By now they have revised the notion of statistical hypotheses, and have a more concrete idea of the p-value and how to make a decision about H_o based on it. The concept of Type I and II errors have been introduced as well. They also have experimented with the effect of confidence on the width of the confidence interval while looking at the empirical distribution of the bootstrap replicates.

USING THE BINOMIAL DISTRIBUTION TO DO INFERENCE

Immediately after the basic concepts of probability are introduced, the binomial distribution is studied. After doing some simple exercises, the distribution is used to do hypothesis testing when H_o: $p=p_o$. Each student is given a two-color plastic chip (red and yellow) and asked to test the hypothesis H_o: $p=1/2$ where p is the probability of red vs. H_a: $p>1/2$. They are asked to flip the chip 10 times and keep track of the number of times in which the red color comes on top. They are then asked to calculate the probability of their outcome, or a more extreme one, happening when the null hypothesis is true. We then see other examples, varying the value of p_o and with alternative hypotheses of the type $<$ and $\neq$. Students learn how to use a probability distribution to calculate a p-value and review how to use the p-value to make a decision about H_o. The concepts of Type I and Type II errors are reviewed as well. They learn the exact test for one proportion, which is omitted in many introductory statistics courses where students are usually warned against using the normal approximation when samples are small but are not told what to use instead. This test could be even used to introduce the ideas on hypotheses testing *from scratch,* as explained in Seier and Robe (2002). The binomial table provides an easy way of calculating the probability of Type I error (α) and Type II error (β) without using formulas, and of introducing the concept of power as explained in Seier and Liu (2013).

NORMAL-BASED INFERENCE

After briefly talking about the idea of discrete and continuous probability distributions, we cover the Normal distribution and mention the Chi-square distribution in relation to the normal. The idea of sampling distribution of the sample mean and the t-Student distribution are introduced. The usual confidence interval and testing hypotheses topics for means and one proportion of any algebra-based introductory statistics course are also studied. Due to time constraints we choose to focus on the t-tests when dealing with inference about means. The confidence interval for the mean, when the variance is known, is only addressed to discuss sample size calculation in the context of estimation. The Chi-square tests of goodness-of-fit and independence are covered as well. Students are already very familiar with the concepts of statistical hypothesis, p-values and confidence, so they can focus on the methods and don't have to face at once a lot of new concepts *and* formulas.

OTHER TOPICS COVERED

We learn how to calculate conditional probabilities from two-way tables but also using probability trees to apply Bayes Rule. We practice with the vocabulary of sensitivity, specificity, false positive, false negative, positive predictive value and negative predictive value in the context of medical diagnosis. The last topic of the semester is regression.

ANCILLIARY MATERIAL AND EVALUATION

The course is taught twice a week, one in a regular classroom and one in a computer lab, with the same instructor. Students have access to Minitab and R. PowerPoint presentations, other reading material, assignments and data files are available from a platform (D2L). We give several multiple choice quizzes especially in the first part of the semester. Practices quizzes are made available on-line. There are two partial exams and a comprehensive final. Four labs are also assigned; three of them (one on descriptive statistics and two on inference) require the use of computers. A final data analysis project including descriptive statistics and inference is also part of the evaluation. In the data analysis project we include exercises about exploring data and looking for patterns so that students don't get the idea that statistics is only about testing hypotheses.

PERFORMANCE OF THE STUDENTS

As in any other course, the performance of students depends mainly on study habits, background, motivation, attendance, and the time and effort they dedicate to the course. We have taught the course 3 times as a stand-alone statistics course. The composition of the students in the course has been different each year due to the fact that it has taken some time for the recruitment process to become more efficient and to enroll more of the students who are the target population for this section. The first year, the students came to this section in a random way. They were not in

any sort of special program and were no different than the students in any other section of the regular course in a regional university with a non-selective admission process. There was variability in their scores as in any freshman statistics course, but they did not do worse than in the regular course. The second year, almost one third of the students were from different fields but they were enrolled in the honors program and two-thirds were students who were not in any special program and registered in this section just because the time was convenient. The performance of the honors students was clearly better than in the non-honors group. The third time we offered the course (Fall 2013) we had a very interesting combination of ten students who were in the honors program but who were mainly not science oriented, and nineteen students some of whom are interested in health or science oriented careers but who are not in the honors program. The honors students were from areas as diverse as art, business, psychology, mass communication, history, and political science; they tended to be very dedicated students and all perform extremely well in the course. We noticed more variability among the non-honors students. In the first exam, which included several inference questions, the scores for the non-honors group this time were in the range 64-100 with a mean of 83 and median 85, which was better than the performance of the non-honors students in the past two years (means 72 and 68, medians 77 and 65 in the first exam, respectively). This year, the final exam the scores were in the range 57-95 (median=81.25%) and 82-100 (median = 95%) for the regular and honor students respectively. Only one student of the twenty-nine who took the first exam did not show up for the final exam. Students in general get to be more comfortable with the language and ideas of inference than what they usually are in the regular course and tend to answer correctly some conceptual questions on statistical inference that we always include.

CONCLUSION

During the last five weeks of the semester we cover the inference for means using the t-distribution, inference for proportions using the normal approximation for large samples, and the Chi-square tests for goodness of fit and independence. However, we also discuss statistical inference in three previous sections of the course: writing statistical hypothesis in the first week, randomization tests and bootstrapping in week 4, and the exact test for a proportion in week 6. This allows us to progressively introduce the ideas of hypothesis testing and estimation and sustain their discussion throughout the semester. This enables us to focus on some important inference concepts before dealing with the numerous formulas of the normal-based inference section. Students have more time to feel comfortable with the basic ideas of inference because they see them several times. It is an approach that has worked for us. Each instructor could think of different ways of increasing the points of contact with the ideas of inference throughout the semester.

REFERENCES

Fisher, R.A. (1966). *The design of experiments* (8th ed.). London: Oliver & Boyd.

Joplin, K., Karsai, I. Moore, D., Miller, H., Seier, E., Godbole, A., & Helfgott, M. (2013). SYMBIOSIS: An integration of biology, math and statistics at the freshman level: Walking together instead of on opposite sides of the street. In *Undergraduate mathematics for the life sciences: Models, processes, and directions, MAA Notes, Vol. 81* (pp. 97-103). Washington DC: Mathematical Association of America.

Lock, R.H., Lock, P.F., Lock, K., Lock, E.F., & Lock, D.F. (2012). *Unlocking the power of statistics.* New York: Wiley.

Rossman, A. & Chance, B. (2008). Concepts of statistical inference: A randomization based curriculum. *Available at: http://statweb.calpoly.edu/csi*

Seier, E., & Robe, C. (2002). Ducks and green – an introduction to the ideas of hypothesis testing. *Teaching Statistics, 24*(3), 82-86.

Seier, E., & Liu, Y (2013) An exercise to introduce power. *Teaching Statistics*, *35*(1), 53-56.

Seier, E., & Joplin, K. (2011). *Introduction to statistics in a biological context*. Charleston: Createspace.

Tintle, N., Van der Stoep, J., Holmes, V.L., Quisenberry, B., & Swanson, T. (2011). Development and assessment of a preliminary randomization-based introductory statistics curriculum. *Journal of Statistics Education, 19 (1).* www.amstat.org/publications/jse/v19n1/tintle.pdf