AN INTERACTIVE METHODOLOGY TO TEACH SAMPLING AND STATISTICAL INFERENCE

SUNDEFELD, Maria Lucia Marçal Mazza
School of Dentistry of Araçatuba, UNESP- University of São Paulo State, Brazil.
mlsundef@foa.unesp.br

PERRI, Sílvia Helena Venturoli
Veterinary Medicine School of Araçatuba, UNESP- University of São Paulo State, Brazil.

RODRIGUES, Marco Aurélio Borella
UNESP- University of São Paulo State, Brazil.

ABSTRACT

In the course of our experience teaching statistics to students in biological fields, we noticed that they have great difficulty understanding the statistical inference process and sampling theory. These students should be prepared to understand scientific papers, to develop research in health areas, and to establish a good relationship with patients when explaining about diseases and treatment decisions. This paper aims to test a new methodology to teach Biostatistics by developing the statistical thinking to understand inference. Two classes of biological courses at the University of São Paulo State in Araçatuba, Brazil were chosen to be the experimental and control group. In the first group, a new methodology was developed. At the end of the year, both classes were assessed on the same test as to the understanding of the statistical inference.

INTRODUCTION

In Brazilian schools of health sciences, biostatistics is taught in the first year of the course. Unfortunately, students are not motivated to study any discipline related to mathematical calculation, only biological subjects. However, the students need to learn biostatistics in order to be prepared to understand scientific papers, to develop research, to help improve their knowledge in health areas, and to establish a good relationship with patients when explaining about diseases and treatment decisions (Gigerenzer, 2003). Some of these students will have opportunity to participate in a health care survey, and need to be prepared to plan it.

In the course of our 26-year experience teaching statistics to students in biological fields, we noticed that the opinion of the students is the same as that pointed out by Neville Davies (2006), that "statistics is hard, useless, and boring". They have great difficulty understanding the statistics, particularly inference processes and sampling theory. So, this research aims to test a new approach to teach Biostatistics, by developing statistical thinking to understand the reasoning of inference.

METHODOLOGY

Two 2010 classes of biological science students at the University of São Paulo State (UNESP) in Araçatuba, Brazil, were chosen to be the experimental and control group. Both were comprised of freshmen, the experimental group consisting of 80 Dentistry students and the control group with n=40 Veterinary Medicine students. The students were selected by the same college entrance examination and showed the same general performance. Two professors taught these courses, one of them responsible for each course and the other as a collaborator. These two professors have had many years of experience working together.

In the control group, in the School of Veterinary Medicine, the conventional approach was maintained. The course began with descriptive statistics, followed by basic notions of probability, basic notions of sampling, hypothesis tests, association and correlation tests, regression and analysis of variance, with health coefficients encompassing theory and exercises extracted from text books. (Parsen, 1960; Zar, 1999; Bussab & Morettin, 2002; Bolfarine & Bussab, 2005).

In the experimental group, in the School of Dentistry, a new methodology was developed. The course syllabus started with notions of sampling theory by introducing probabilistic and non-probabilistic sampling, including different sampling designs such as simple random sampling (SRS), stratified sampling, and cluster sampling. It also included different types of observation drawing, including random and systematic draws (Zar, 1999; Sundfeld, Silva, Frei & Correa, 2002; Bussab & Morettin, 2002; Bolfarine & Bussab, 2005). At the beginning of the experimental course, each student drew one or more samples using simple random sampling. The sample size was specified in the classroom by using the mathematical expression for sample size, $n = \frac{\sigma^2}{(d/1.96)^2}$, with μ and σ^2 known and $d = x - \mu$, the acceptable difference between the sample mean and the population mean (Kish, 1965).

These samples were used to teach descriptive statistical subjects, such as mean, variance, standard deviation, skewness, kurtosis, tables, and graphs. Each student calculated the values in his or her own sample. Next, an Excel database was built containing all individual samples. From this database, the sample statistics of different samples were calculated and the distribution of sample means was built. Through fitting the distribution of sample means to the normal distribution (Cochran, 1977), the students could visualize the Central Limit Theorem: "In a finite population or sampling with replacement, the values of the sample can be considered as independent random variables with the same probability of distribution as the population. With the set of the infinite sample determined by one drawing procedure, we can build the sampling distribution of the means that theoretically will approximate the normal distribution with all the means of samples obtained by this procedure". When we draw an SRS of size n from a population, we want to obtain an estimator for the unknown population mean or proportion. Ignoring measurement or observation error, the standard deviation (called standard error) calculated by $s / \sqrt{n}$ quantifies the random error inherent in the sampling procedure. The result obtained from the sample is called the point estimate. Information for a population mean is estimated by the 95% confidence interval, when ($n \geq 30$), as $[\bar{x} \pm 1.96 \frac{s}{\sqrt{n}}]$, using $\alpha = 0.05$ in the normal distribution (Kalton, 1983; Silva, 2001). Although the Central Limit Theorem result described is based on sampling with replacement, in-class practice carried out the sampling without replacement. The Kolmogorov-Smirnov test, Lilliefors test, and Shapiro-Wilk test were used to verify the fit of the distribution of sample means to the normal distribution. Afterwards, hypothesis tests and other issues that required notions of inference were developed.

At the end of the year, both courses were assessed by asking questions about understanding, interpretation, and development of statistical thinking about inference and the Central Limit Theorem. The two classes were compared using the two-proportion test, at the α = 5% significance level.

RESULTS

In the experimental group, databases from dentistry were used. One example is the discrete numerical database expressing dental caries, the DMF (Decay/Missing/Filled) index, N=8052. The students drew 98 samples with n=114. The class built the distribution of the data from one sample, and the distribution of sampling means with all 98 sample means. Figures 1 and 2 show the distributions and p-values of the three fitting tests.

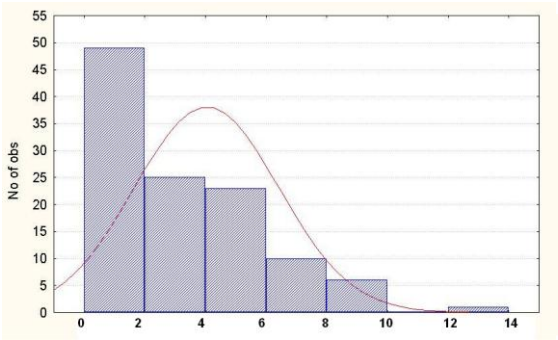

Figure 1. DMF Index of one sample
K-S: p<.05 Lilliefors: p<.05 S.Wilk: p<.001

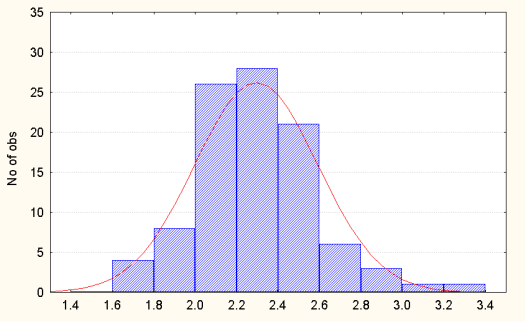

Figure 2. DMF index of the Distribution of Sampling Means K-S: p>.20; Lilliefors: p>.20 S.Wilk: p<.6370

In the database used in this example, the parameters were, $\bar{x} = 2.20$; s= 2.68; $\frac{s}{\sqrt{n}} = 0.25$ and 95% CI = [1.72; 2.69]. There are different means, different standard deviations, and different intervals, so we asked the students: "What do you think about it? Which is the correct one?"

Each student typed the results from his or her sample into the common Excel file, and then confidence intervals were built for all drawn samples. Each student computed an individual confidence interval. Eventually comparing all intervals built by the class, the students verified the intersection of most of them. Each sample led to the same conclusion: 95% of the samples will have the population mean inside the confidence interval. Table 1 presents 13 out of the 98 confidence intervals, just to show what the students built in the classroom.

Table 1. Some Confidence Intervals of 95% built in class by the students.

Student	***LowerLimit***	***CI(95%) representation***	***UpperLimit***
44	**1.4882**	_____	**2.4066**
54	**1.9603**	_____	**3.1099**
52	**1.6248**	_____	**2.7612**
56	**1.6187**	_____	**2.5215**
48	**1.7927**	_____	**2.7511**
67	**1.9963**	_____	**3.0915**
22	**1.8628**	_____	**2.8740**
45	**1.7464**	_____	**2.6745**
40	**1.8571**	_____	**2.8973**
21	**1.7165**	_____	**2.6871**
64	**1.7680**	_____	**2.8108**
8	**1.4176**	_____	**2.3016**
38	**1.9620**	_____	**3.0906**

Students noticed that the result obtained in the sample is a point estimate, while the information of the population mean is estimated from the 95% confidence interval. The distribution of the DMF measures in the samples is asymmetric, but the distribution of the sampling means is symmetric. The key for understanding the statistical interpretation is that 95% of the drawn samples have the means inside the confidence interval at 95% confidence, i.e., the drawn samples have the means between 2.46 and 3.54.

Using results from software for the Kolmogorov-Smirnov test, Lilliefors test, and Shapiro-Wilk test, we can show that the distribution of the sampling means is always fitted to the

Normal Distribution. Therefore, the experimental group verified the Central Limit Theorem empirically, a topic very difficult to understand theoretically by students of non-mathematical service courses. In the control group, the Central Limit Theorem was taught through examples of small populations usually presented in textbooks. (Zar, 1999; Bussab & Moretin, 2002)

At the end of the year, both classes took the same test about understanding sampling and statistical inference. As this assessment was taken without prior notice, the percentage of the School of Dentistry students who were present was 86.25%, 69 of the 80 students; in Veterinary Medicine, 33 of 40 (82.5%) students registered at the beginning of the year took the test.

COMPARING EXPERIMENTAL AND CONTROL CLASSES

It was verified that the understanding of the subject was deeper in the experimental group, rather than the control group, through the following eight questions.

1. How many samples of size n can we get from a population of size N?
2. If you have a simple random sample (SRS) from a population, will the mean of this sample be the exact mean of the population?
3. Explain the answer to the question above.
4. What is the distribution of sample means?
5. What does the 95% confidence interval mean?
6. If you have several samples of the same population, will the average and standard deviation of each of them be the same as the other samples?
7. Explain the answer to the question above.
8. What parameters form the basis for statistical inference?

In Question 1, our concern was verifying whether the students noticed that it is possible to obtain a great number of samples of size n from a population N, as students in a biological course are likely to think that only one sample can be obtained from a population. Therefore, the expected correct answer in this question would not be the mathematical formula, but an awareness that there is a ¨great number of possible samples¨, not only one.

Concerning Question 2, the expected correct answer is ¨No¨. This answer would be justified in question 3 with the statement that 95% of the samples drawn have the population mean inside this 95% confidence interval, so the population mean would not be a point estimate but it would be inside this interval.

The answer to Question 4 is that the distribution of sampling means is a symmetrical distribution fit to the normal distribution, built from the means of each sample of size n obtained from the same population.

In Question 5, students are expected to answer that 95% of the samples of size n obtained from the same population will have means inside $\left[\bar{x} \pm 1.96 \frac{s}{\sqrt{n}}\right]$ with $\bar{x}$ the mean of one sample and s the standard deviation. The value 1.96 defines the 95% interval in the standard normal N(0, 1) distribution.

In Question 6, the expected correct answer is “No”, and in Question 7 it would be justified that the likelihood of different samples having the same statistics is very slim.

Concerning Question 8, students are expected to mention the mean, standard deviation, standard error, and/or 95% confidence interval.

Table 2 shows the count and percentage of correct answers for both groups. The two groups were compared with the two-proportion test with $\alpha = 5\%$ significance. Only question 2 has no significant difference. We can verify that the experimental Dentistry School students showed significantly better results in all the questions except question 2.

Table 2. The number and the percentage of correct answers in the assessment of 2010 Dentistry and Veterinary Course students at UNESP, Araçatuba, SP, Brazil.

Question	Dentistry	Veterinary	p-value

1	66 (95.65%)	25 (75.76%)	0.0072*
2	65 (94.20%)	28 (84.85%)	0.2359
3	57 (82.61%)	21 (63.64%)	0.0345*
4	35 (50.72%)	6 (18.18%)	<0.0017*
5	27 (39.13%)	0 (0.00%)	<0.0001*
6	69 (100.00%)	27 (81.81%)	0.0014*
7	69 (100.00%)	27 (81.81%)	0.0014*
8	54 (78.26%)	0 (0.00%)	<0.0001*

* Statistically significant difference at 5% level

CONCLUSION

As the results of the assessment favored the experimental group, we propose an interactive methodology in order to understand theoretical concepts of statistical inference. Each student should draw a sample to be used during a Biostatistics course, compare the values obtained with other samples, and built the distribution of sampling means with all the means of drawn samples. We suggest that the Biostatistics course should begin with sampling, and develop descriptive statistics in different samples drawn by the students. By teaching service course statistics through applying related issues, doing and learning step by step, and visualizing the problem the students will slowly absorb the sense of statistics even without the mathematical concepts.

REFERENCES

Bolfarine, H., & Bussab, W. O. (2005). *Elementos de Amostragem.* São Paulo: Edgard Blücher.

Bussab, W. O., & Morettin, P. A. (2002). *Estatística básica.(5ª ed.).* São Paulo: Saraiva.

Cochran, W. G. (1977). *Sampling Techiques (3rd ed.).* New York: John Wiley & Sons.

Cox, B.G. & Cohen, S. B. (1985). *Methodoligical Issues for health care survey.* New York: Marcel Dekker.

Davies, N. (2006) Real data, real learning and the London Olympics. Significance, jun.

Gigerenzer G, Edwards A. (2003). Simple tools for understanding risks: from innumeracy to insight. *BMJ,* 327:741-4.

Kalton, G. (1983). *Introduction to survey sampling*. Beverly Hills: Sage Publications.

Kish L. (1965). *Survey sampling.* New York: John Wiley & Sons.

Parzen, E. (1960). *Modern probability theory.* New York: John Wiley & Sons.

Silva, N. N. (2001). *Amostragem probabilística.* São Paul: Editora da Universidade de São Paulo.

Sundefeld, M. L. M. M., Silva, N. N., Frei, F. & Correa, D. C. (2002) A tool to learn sampling: doing and learning through an interactive software. In: *XXIst International Biometric Conference Freiburg. International Biometric Society*.

Zar, J.H. (1999). *Biostatistical analysis (4th ed.).*New Jersey: Prentice-Hall.