THE INTERACTIVE SUBMARINE: USING BOXPLOTS AS A LIKELIHOOD APPROACH

Pedro Campos
LIAAD INESC TEC and FEP, University of Porto, Portugal
pcampos@fep.up.pt

In this work, we explore the capacity of students and professionals with different education levels to cope with boxplots and to draw conclusions. An experiment has been conducted where specimens of a subspecies of a reptile are distinguished by weight, age and size. Participants compare sample data with population data in order to identify the subspecies that have been collected. Young students seem to be more prepared to analyze and interpret boxplots than older students and adults which reinforces the need for the training of statistical reasoning in adult ages. The submarine simulator (or Interactive Submarine) is included in Exploristica (www.exploristica.com), an itinerant exploratory exhibition consisting of various interactive modules with the aim of bringing the fundamentals of Statistics and Probability to basic and secondary schools, conveying the statistical concepts in a practical and experimental way.

STATISTICAL LIKELIHOOD

In statistical inference, the likelihood function indicates how likely a particular population is to produce an observed sample. The function $L(\theta \mid Xo) = P(Xo \mid \theta)$ is called a likelihood function given that Xo is the observed realization of vector X and θ is the corresponding parameter under study. One of the well-known uses of the word "likelihood" in statistics is the maximum likelihood estimation method, which the general idea is to find the population that is more likely than any other to produce the observed data Xo for estimating the parameter θ. Because of its versatility and favorable large sample asymptotic properties, the method of maximum likelihood (ML) is probably the most widely used method of estimation for parametric statistical models.

Based on this idea, several authors aimed at teaching the concept of likelihood in many different ways. Hawkins and Hawkins (1998), explored whether lawyers can understand and apply the criminal burden of proof, and whether they are able to make the sort of likelihood estimates that the expert witness suggested would help the jurors in some legal cases in court. Meeker and Escobar (2005) developed a way of teaching ML by presenting, to students, confidence regions/intervals based on ML estimation. Rohde (2003) presented an extended discussion of the likelihood concept, by introducing the work of Royall (1977, cf. Rohde, 2003), that contains links between data and beliefs. Royall states that observations can lead to answering one of three questions (i) what do I believe?; (ii) what do I do? and (iii) what does the data say? The first requires input of prior beliefs; the second requires both prior beliefs and statements of consequences of actions (losses); the third is connected with the law of likelihood and is called interpretation of statistical evidence provided by data. This discussion also leads to the prior information that may exist when one observes data, which occurs within the Bayesian approach in statistics. Bolstad (2010) presented a graphical approach for teaching the difference between the two approaches (likelihood inference and Bayesian inference). What we can judge from all these approaches is that the concept of likelihood is hard to teach for levels of education where the concepts of probability are still not acquired by students. That is why likelihood is usually taught in university levels and not in basic or secondary schools. However, boxplots (a convenient way of graphically depicting groups of quantitative data), may constitute a good approach of transmitting likelihood to students in lower levels of education, as we can see in the following sections.

BOXPLOTS AS A LIKELIHOOD APPROACH

Boxplots (or box plots) are graphical methods of representing important characteristics of data, based on the five-number summary of the data: minimum, maximum, 1st quartile, median and 3rd quartile (Everit, 1998). The box corresponds to the inter-quartile range, which is the difference between the 3rd quartile and the 1st quartile. Boxplots may also have lines extending vertically from the boxes to include all observations from the minimum until the maximum value of the dataset. A slight different type of graphs contains "whiskers" that are extended to include all but outside

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

observations (outliers), these being indicated separately. The whiskers may also be useful for indicating variability outside the upper and lower quartiles. When these lines exist, these types of graphics are named box-and-whiskers plots (see Figure 1).

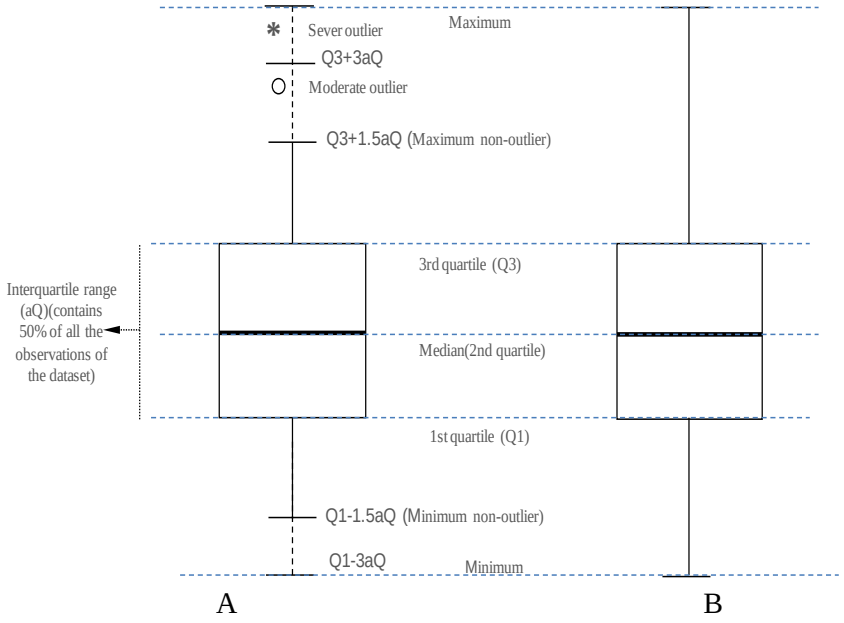

Figure 1 – Box-and-whisker plot (A) and boxplot (B) compared for the same dataset.

Boxplots focus on three main features of the data: 1) Center, which is determined by the median, i.e., the point where about half of the observations are on either side; 2) Spread, that refers to the variability of the data; and 3) Shape of a distribution, which is described by symmetry. These graphical representations are also known as Tukey box graphs, after the name of the statistician that stands behind many works in exploratory data analysis (Tukey, 1977). Boxplots and box-and-whisker plots, (henceforward, just boxplots) have many strengths: one is that the chosen quartiles can be compared effectively. Few arguments exist against the use of boxplots in the teaching and learning of statistics. However, according to Bakker, Biehler and Konold (2004), some of the features of boxplots make them particularly difficult for young students to use in authentic contexts. These difficulties include: (i) boxplots generally do not allow perceiving individual cases; (ii) boxplots operate differently than other displays students encounter; (iii) the median is not as intuitive to students as we once suspected; (iv) quartiles divide the data into groups in ways that few students few students (or even teachers) really understand. These authors recommend that educators consider these features as they determine whether, how, and when to introduce boxplots to students.

THE INTERACTIVE SUBMARINE

The Interactive Submarine is one of the game-type modules of Exploristica (see Figure 2), where a news species of a reptile is being studied in a lake. Participants start by seeing a movie in the right part of the module where they are trained about the goal of the game, and the statistical concepts being learned. Then, they have to capture the reptiles, using the joystick in the left part of the module. A balance, a ruler and a camera are ready to measure weight, age, size in the laboratory at the right part of the module. A kind of 4D (augmented reality) device is installed in the module so that participants can rotate a small piece of vinyl and see the image of the reptile “alive” on the screen.

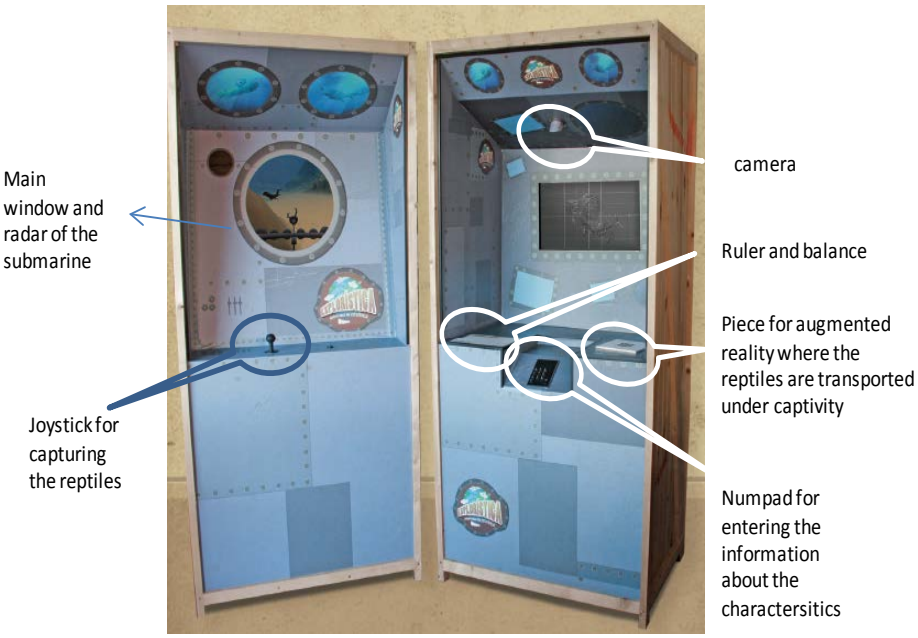

Figure 2. Left part (main submarine window) and right part (laboratory) of the Interactive Submarine

Taking into account the existence of three different subspecies of the reptile in the same lake, which are distinguished by their main characteristics (weight, age, and size), participants have to produce boxplots of these variables and compare them with the boxplots of the population to identify the subspecies that have been collected. A similar game containing the same goals of the module (except that participants would not have to capture and measure the reptiles), has been made in a survey. A Google Drive's form was sent by email to different target groups in order to measure the ability of people with different skills to use boxplots. A sample of 74 observations was collected under a convenience sampling scheme: students of basic and secondary school are from Escola Tomaz Pelayo (9th to 12th grades – corresponding to ages 14 to 17), as well as the teachers. Students of college /master belong to University of Porto/Faculty of Economics. Professional/Statisticians work in public administration services. Data was collected during the first semester of 2014. A summary of the obtained stratified sample is presented in Table 1.

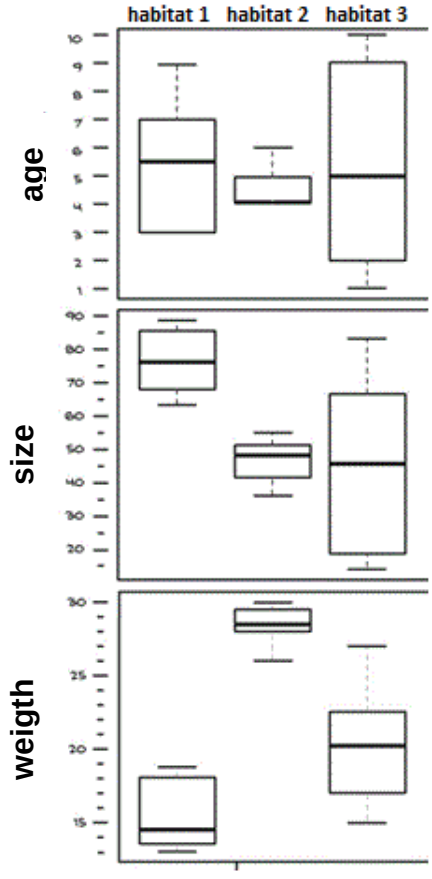

Figure 3. Boxplots with information for variables weight, size and age, based on subpopulation data

Strata	Percent (%)
Student (basic school)	34,2
Student (secondary school)	17,8
Student (college /master)	28,8
Teacher	1,4
Professional (Statistician)	16,4
Total	100,0

Table 1 – Details of the collected sample of the participants in the survey. Ages in different school levels are the following: 9-14: basic school; 15-17: Secondary school; College /master: over 17.

In this survey-version of the game, all participants were given a fixed sample of reptiles. Only 5 reptiles have been caught with the following characteristics: weight: 10, 12, 15, 20, 12; size: 72, 85, 74, 73, 78; age: 3, 5, 6, 4, 3. Participants were asked four questions concerning the collected data and the boxplots of the habitats for each subpopulation (real data) being depicted in Figure 3.: 1) What was the habitat from where data was collected? 2) What is the habitat with many young and old specimens? 3) What is the habitat with many heavy-weight individuals as well as with more homogenous ages? 4) What is the habitat with both thin and tall individuals?

To answer the survey, we have to compare the data for the 5 reptiles, with the corresponding habitats and boxplots for each subpopulation depicted in Figure 3. It seems clear that the answers are: 1) habitat 1; 2) habitat 3; 3) habitat 2; 4) habitat 1.

One of the aspects that it is worth noting is the difference between the correct answers for the different types of targets (here seen as strata). Curiously, the percentage of correct answers is higher for the students of basic school. Besides, the highest percentage of incorrect answers occurs in professionals. What kind of conclusions can we draw from here?

Strata	Answers			Total
	Habitat 1	Habitat 2	Habitat 3	
Student (basic school)	**87,5%**	0%	12,5%	100,0%
Student (secondary school)	**76,9%**	7,7%	15,3%	100,0%
Student (college /master)	**66,7%**	9,5%	23,8%	100,0%
Teacher	**0%**	100,0%	0%	100,0%
Professional (Statistician)	**55,0%**	27,0%	18,0%	100,0%

Table 2 – Question 1: description of answers according to the strata of the sample (correct answer is Habitat 1).

CONCLUSIONS

We observed that younger individuals seem to be more prepared to analyze and interpret boxplots. We recall that boxplots were not built with sample data, but with data for a particular subpopulation. Maybe it caused some confusion in the respondents trying to find an exact resemblance between the boxplots and the sample data. Furthermore, it is also important to note that boxplots are taught at the basic and secondary school in Portugal. On the other hand, we may observe that it is strange that such rather simple graphical representations like boxplots are not so well interpreted by people in university and professionals boxplots should be capable of explaining the data in a simple way, because they do not involve complex concepts of statistics. Therefore, one should encourage statistical training to reinforce statistical reasoning.

REFERENCES:

Bakker, A., Biehler, R., & Konold, C. (2004). *Should young students learn about box plots?* Paper presented at the IASE Roundtable, Curricular Development in Statistics Education, Sweden.

Bolstad, W. (2010). *Comparing the Bayesian and likelihood approaches to inference: A graphical approach.* Paper presented at the Eighth International Conference on Teaching Statistics, Ljubljana, Slovenia.

Everit, B., S. (1998). *The Cambridge dictionary of statistics*. Cambridge University Press.

Hawkins, P., & Hawkins, A. (1998). *Lawyers' likelihoods*. Paper presented at the Fifth International Conference on Teaching Statistics, Singapore.

Meeker, W. Q., & Escobar, L. A. (1995). Teaching about approximate confidence regions based on maximum likelihood estimation. *The American Statistician, 49*(1), 48-53.

Rohde, C. A. (2003). *Teaching the likelihood paradigm in biostatistics*. Paper presented at the 54th Session of the International Statistical Institute, Berlin.

Royall, R. M. (1977). *Statistical evidence: A likelihood paradigm*. Chapman and Hall.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley Publishing Co.