Genuine or impostor: using a biometric recognition system to describe the hypothesis test methodology

Jorge Domínguez-Domínguez† & Jorge Axel Domínguez-López‡

Mathematics Research Center, Aguascalientes, Mexico†
Conteck, Houston, Texas, United States of America‡

ABSTRACT: The methodology for the statistical hypothesis test is presented using the simulation of a security system. In order to authenticate users, the system uses biometric information from the user, particularly, the iris. The method shows by way of animation a process to contrast a hypothesis; therefore, offering an alternative approach for the teaching and learning of this subject. Several statistical and probability concepts are required to understand and perform a hypothesis test. For instance, it is vital to comprehend random variables, statistics, density functions and probability distribution of a random variable. These subjects are explained in detail using animations. Then, the user will be able to describe the procedure of the statistical hypothesis test. This method is also utilised in similar cases where the user performs statistical inference of a parameter and/or contrast of parameters from information about a random variable. This is illustrated using a visual approach in order to assist in the solid understanding of the procedures used to perform hypothesis test on one or two populations.

INTRODUCTION

Statistics is the process used to discover or provide an answer about the real world using data that have been collected, analysed and interpreted. Hence, statistical studies are presented as search procedures. First, a problem is defined, and from there, a number of questions is generated. The authors explain and answer those questions using an appropriate method of data collection and analysis. Initially, students tackle this situation using the concepts of descriptive statistics.

Many syllabi and statistics textbooks illustrate the parameters estimation procedure using confidence intervals and hypothesis tests. These concepts make up the core of statistical inference. The hypothesis test is a standard procedure, which is commonly utilised in a great diversity of professional fields. Accordingly, the hypothesis test has applications in many areas of knowledge in addition to statistics. Consequently, it is vital to illustrate the main ideas and their characteristics related to statistical inference.

This article focuses on describing a didactic development, which illustrates the statistical process to test a hypothesis using animations. This development has a visual approach and is based on the concept of learning by playing. The system simulates the identification of a person based on the iris. During the animations and simulations, the different basic statistical concepts used in the hypothesis test methodology are described. Moreover, the idea behind using an advanced technological system is to motivate the students to learn more about the problem, as well as the techniques utilised to solve it.

Statistical inference is based on concepts of probability theory. Thus, probability distributions play an important role. Consequently, it is necessary for students to understand and master the notions of distributions. In order to make these more appealing and easier to understand, part of the didactic development includes a module for the density function and the cumulative distribution functions for several distributions. This module illustrates the main ideas for the calculation of the probabilities. The user captures the input parameters for the desired distribution.

The distribution is represented graphically and the user can adjust the threshold(s) to obtain the area to the left, to the right or between the values. The concepts of probability distribution are utilised to teach the essential ideas to test a hypothesis graphically and by using animation.

The didactic material has been incorporated into software, which includes calculations, plotting, simulations and animations. This project is named CalEst. Only the modules used to illustrate the statistics concepts needed for hypothesis testing are described in this article. To learn more about the CalEst project's view refer to *A Visual Approach to Teaching and Statistics* [1].

This pedagogical proposition has been shown to be useful as it provides instructors with an extra tool with which to teach statistics using an interactive and visually attractive approach. Likewise, students have an opportunity to learn a range of statistical techniques by experimenting with the software. The idea of learning by playing, arises because the system allows students to explore different scenarios, and learn the concepts of statistics and probability using the different animations and simulations, as well as the tools to perform calculations and graphics. This didactic software has been employed in a range of workshops with teachers and students with positive results, which are described below.

THE PRACTICAL PROBLEM: IRIS RECONGNITION

How can one be sure a person is who he or she says he or she is? If one does not know the person, one asks for a form of ID. People have several unique characteristics they can use to identify themselves. One can use fingerprints, voice, the iris and retina, the ear, DNA, even the way one walks. Each feature has its pros and cons. One can chose to use the iris as it has a great mathematical advantage in that its pattern has high variability between different persons [2]. Actually, the iris from the left eye is completely different from the right side. In order to validate a person from his biometric information, it is necessary to perform a hypothesis test. There are two different approaches to doing the identification: when one knows the person and when one does not. For the former, one compares the person's iris with the stored information, while for the latter one compares the person's iris against all irises stored in the system.

METHODOLOGY

The main goal of the hypothesis test is to make possible an adequate selection between two hypotheses. When one wants to measure the quality of sample parameters and assume the data follow a known distribution (commonly the normal distribution), then, one uses a parametric hypothesis test.

For iris identification, one wants to make a decision to either accept or reject a person. One simulates a person security system, which has to grant or deny access to a person. People present an iris and the system compares it against irises stored in the database [3]. During the comparison of both irises, one obtains a Hamming distance, which is a value between 0 and 100 (%). The Hamming distance is the random variable X and it is assumed to have a normal distribution with mean μ and standard deviation σ.

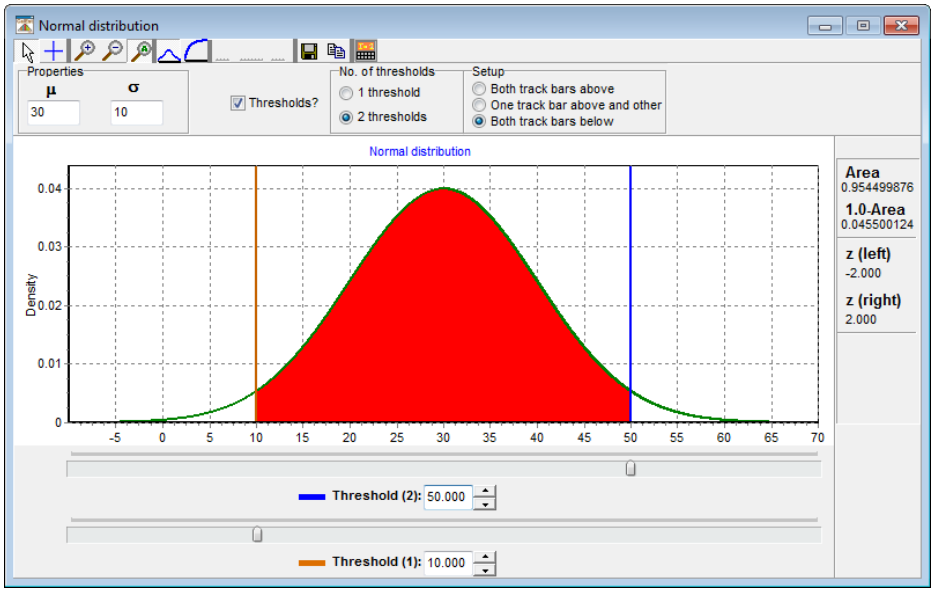

Figure 1: Probability between two values (thresholds) of a variable.

Bearing in mind the characteristics of the normal distribution, one can calculate the probabilities of matching two irises. In order to illustrate the use of the didactic material, one must start considering the following case: a normal distribution with $\mu = 30$ and $\sigma = 10$. Figure 1 illustrates this scenario for the particular case, where two thresholds have been selected. Calculating the probability using a threshold, X_c, is used as reference to accept or reject the null hypothesis. Using the tool, it is possible to obtain the probability in three different cases of the iris matching: the distance is equal or less than 10 ($P(X \leq 10)$), greater or equal to 50 ($P(X \geq 50)$), or it is between 10 and 50 ($P(10 \leq X \leq 50)$).

Hypothesis Test for the Iris Identification: Genuine or Impostor

As mentioned above, a security system has been developed in order to illustrate the process of hypothesis testing. This simulation is shown in Figure 2. These elements describe the concepts related to the procedure of hypothesis testing. By clicking on Read Iris, the system simulates the process of capturing an image of the user, performing the image

processing needed to locate the iris and extract its pattern. The system then determines if the user is genuine or is an impostor.

Hence, the input is the iris of the user and the random variable, X, is the index of how much the input iris and the stored iris are alike. For the example depicted in Figure 2, the impostor (i.e. two different eyes) has a mean $\mu = 30$ and a standard deviation $\sigma = 10$, while for the genuine (i.e. both iris belong to the same person) has a mean $\mu = 68$ and a standard deviation $\sigma = 10$. The x-axis represents the level of similarity and is expressed as a percentage.

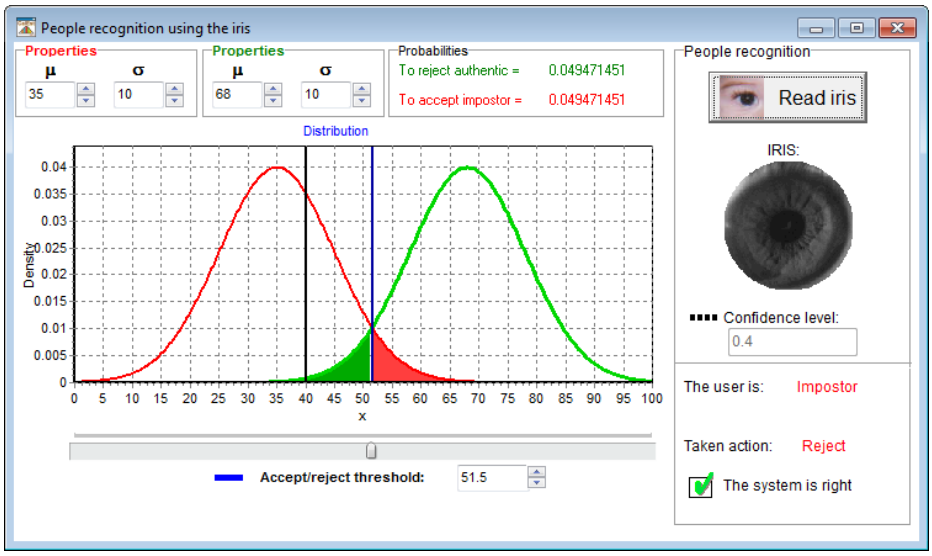

Figure 2: Didactic description of the hypothesis test used to simulate an identification system.

Every testing compares two hypotheses within the context of the information provided by the sample. Two hypotheses are proposed, which are the null hypothesis (H_0) and the alternative hypothesis (H_1):

H_0 : The system recognise an impostor user;

H_1 : The system recognise a genuine user.

The hypothesis testing methods allow a suitable decision to be made between these hypotheses, with a certain level of error. There is always a possibility of making an error at the moment of choosing. Error Type I occurs when the null hypothesis is rejected but it was true. Error Type II happens when one is wrong in accepting the null hypothesis. These both errors are expressed as probabilities. Consequently, the need for probability calculations arises. For Error Type I, it is the probability of rejecting H_0 when H_0 was true and is given by $\alpha = P(reject\ H_0 | H_0\ is\ true)$.

In the iris identification, Error Type I occurs when the system does not reject the user and he was an impostor. Error Type II is given by the probability of not rejecting H_0 when H_0 was false. Error Type II happens when the system rejects the user when he was genuine. The probability of Error Type II is $\beta = P(no\ reject\ H_0 | H_0\ is\ false)$. In summary, Error Type I is the probability of accepting an impostor and Error Type II is the probability of rejecting a genuine person.

The didactic material shown in Figure 2 illustrates the four cases one can have, when one performs a hypothesis test. The system simulates the identification and hypothesis testing, as well as showing the actual decision made by the system (i.e. accepting or rejecting) and the actual status of the user (i.e. genuine or impostor). These four cases are summarised in Table 1.

Table 1: Hypothesis test for the iris identification.

	H_0 is true The user is impostor	H_0 is false The user is genuine
Accept Null Hypothesis Reject access to the user	Right decision	Wrong Decision Error Type II
Reject Null Hypothesis Grant access to the user	Wrong Decision Error Type I	Right decision

Different exercises can be done using the didactic in order to describe the probabilities of both error types and their complements. If one simulates several iris readings, one can calculate the proportions of the errors and, hence, estimate

approximately their probabilities. Also, it is possible to create different scenarios by changing the means and standard deviations. This is useful for illustrating the role that each parameter plays in the hypothesis testing.

Formalising Hypothesis Testing Using Random Samples

This is the method used in the process of teaching statistics to explain the concepts of hypothesis testing. By the reading of several irises, one can take random samples to verify if the system is recognising properly both the genuine and impostor users. Consider the following case, where the iris index is greater than 36 and has a standard deviation of 27, where σ is assumed to be known and the data follow a normal distribution. The hypothesis for this scenario is expressed as follows:

$$H_0 : \mu = 36$$
$$H_1 : \mu > 36 \ (\mu = 65) \tag{1}$$

The steps followed to perform the analysis consist of selecting a random sample from a population of possible users. A sample size of $n = 9$ users is proposed. The iris reading indexes (multiplied by 100) are: 24, 32, 28, 66, 73, 69, 85, 54 and 27. The sample mean is $\bar{X} = 50.8$.

Similar to the iris system, CalEst Project has a module to illustrate graphically the concepts of hypothesis testing. Figure 3 shows the details of the hypothesis test described by Equation (1).

It is important to bear in mind that hypothesis testing is based on the information given by the sample. So that, if a null hypothesis is rejected, the sample data offer evidence to not support the hypothesis. The didactic material grabs this idea and allows the user to experiment with the threshold. By moving the threshold, one can see how the probabilities for Error Type I and II change.

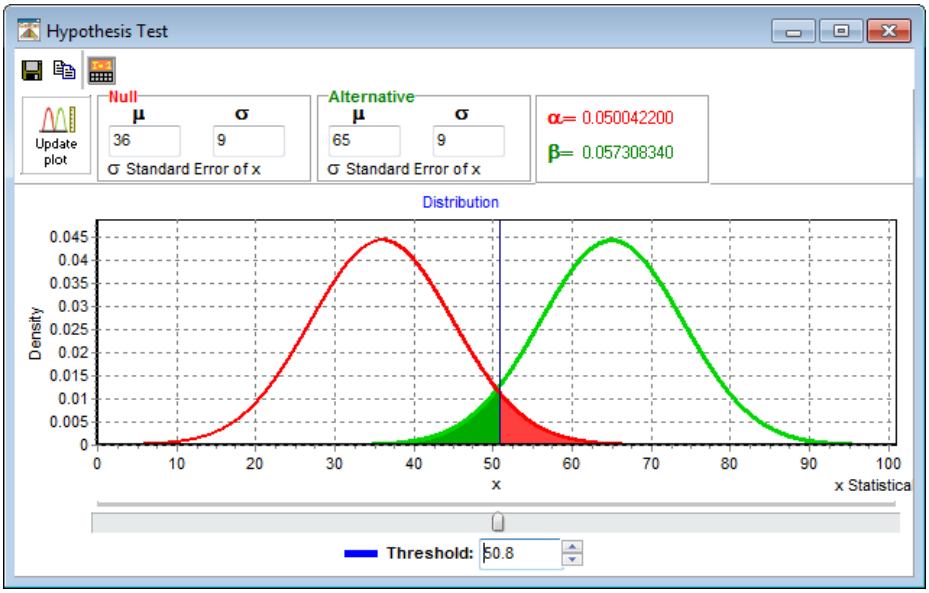

Figure 3: Description of the approach of a hypothesis test.

RESULTS

Three different options to illustrate the concepts behind hypothesis test have been described. The probability distributions are basic to understanding the concepts of statistical inference, such as significance level, confidence intervals and types of errors [5]. Utilising the normal distribution, students learn to calculate probabilities from real problems without the distraction of standardisation. Students also understand the notions of density functions, probability distributions and the relationship between them.

Moreover, rather than using tables, students can use this tool to solve the exercises and problems given in textbooks. The experience gained from using this material allows its usefulness to be observed. After each workshop, each student is asked to complete a questionnaire, which measures their level of understanding, how much they liked the tool, etc. Results show that students learnt to set out abstract expressions. The same ideas used for the normal distribution can also be used in other probability distributions used in statistical inference, such as t-Student, F, Chi-squared, as well as for other distributions used in other fields of knowledge.

By using hypothesis tests to solve the problem of whether a subject is genuine or is an impostor, the users develop skills in the use of the concepts of hypothesis testing. The acquired knowledge can be utilised in similar cases.

Finally, the case given by Figure 3 allows for generalising the notions of the hypothesis testing with two parameters, under the assumption of normality. Similarly, this can be applied in the cases of hypothesis testing for a single proportion or two proportions. Likewise, the standard normal distribution of the data is assumed. Once the student has command of these concepts, it is easy to move forward to other hypothesis tests, which involve parametric significance levels (α) and descriptive (p). In this case, it is necessary to turn to the suitable distribution probability.

CONCLUSIONS

The proposed development is presented as a suite, as it integrates calculations and graphics with a collection of simulations and animations. Thus, the system becomes an innovative, visually attractive tool, which can be utilised to assist in the teaching/learning of statistics and probability concepts. On one hand, it gives resources to the instructor to explain different matters in a more enjoyable and easy to understand way.

Furthermore, it allows the teacher to provide a deeper exposition of the concepts. Its friendly user-interface and animations also encourage students to explore and learn by themselves. CalEst helps students to understand the concepts. In addition, it motivates them to learn more about statistics and probability and how to apply that knowledge to solve real-life problems. The suite can also be used to solve problems and exercises from a range of high school and college level textbooks.

REFERENCES

1. Domínguez, D.J. and Domínguez-L. J.A., *CalEst: A Visual Approach to Teaching and Statistics*. Mexico: Conteck (2010).
2. Daugman, J., How iris recognition works. *IEEE Transactions on Circuits and Systems for Video Technol.*, 14, **1**, 21-30 (2003).
3. Daugman, J., The importance of being random: statistical principles of iris recignition. *Pattern Recognition*, 36, 279-291 (2003).
4. Domínguez D.J. and Domínguez-L. J.A., *Statistics and Probability: The World of Data and Random*. Mexico: Oxford University Press (2006).
5. Johnson, R. and Kuby, P., *Elementary Statistics*. Boston: Duxbury Press (2011).