VISUALIZING the *p*-VALUE and UNDERSTANDING HYPOTHESIS TESTING CONCEPTS USING SIMULATION in R

Leslie Chandrakantha
lchandra@jjay.cuny.edu
Department of Mathematics & Computer Science
John Jay College of Criminal Justice of CUNY
USA

ABSTRACT: *Hypothesis testing is used to make inferences about population parameters. Hypothesis testing is among the most challenging topics at the conceptual level for many students. Research has shown that the use of computer simulation methods as an alternative to traditional methods enhances the understanding of the concepts [36]. R, a free software environment, has many capabilities that are ideal for performing computer simulations. In this paper, we describe how to use simulations in R that place the concept of p–value in a central role to understand the hypothesis testing in introductory level. The p-value distribution and the relationship between the p-value and the evidence against the null hypothesis are shown graphically to enhance the understanding of the concepts.*

1. INTRODUCTION

An understanding of the concepts of statistical inference is critical in order to work with data. We are living in an era where the world's most valuable resource is no longer oil, but data [17]. To deal with this new mandate, it is important for students to have a better understanding of data analysis methods. Many college majors require at least one statistics course. Fundamental statistical concepts such as sampling distributions, the central limit theorem, confidence intervals, hypothesis testing, and *p*-values are vital at the introductory level and beyond. We will focus on the need to provide a more secure grasp of the concepts of statistical hypothesis tests and *p*-values. Hypothesis testing is considered one of the first statistical inference methods and it is widely used to this day [22].

Hypothesis testing is a procedure for testing a claim or hypothesis about a population parameter. Even though the original form of hypothesis testing goes back to the work of John Arbuthnot in 1710 [15], the modern forms of hypothesis testing was introduced by Ronald Fisher, Jerzy Neyman and Egon Pearson in the 1920s [11-13, 27-28]. NHST (Null Hypothesis Significance Testing) is the most commonly used procedure for testing data. NHST is a combination of the Fisher's and Neyman-Pearson's approaches [29]. In NHST, a null hypothesis is posed and the related data is generated. Then the evidence against the null hypothesis is assessed using a statistical estimate. Applications of hypothesis testing can be found in many areas including, sciences, engineering, business, finance, psychology and social sciences [21]. The *p*-value is used to assess the evidence against the null hypothesis in the hypothesis testing framework [2]. In other words, *p*-values tell you how surprising data is, assuming there is no effect. A solid understanding of the *p*-value concept is essential in performing hypothesis testing and making the correct decision. Since all statistical software packages calculate *p*-values, now more and more researchers and instructors use the *p*-value approach to make decisions.

Due to the abstract nature of the concepts, traditional methods for teaching hypothesis testing may not lead to a good understanding for students who do not have strong mathematical skills. Many students in introductory statistics are pursuing non-quantitative majors and hence they may have difficulties in grasping concepts when traditional algorithmic approaches to instruction are used as opposed to newer approaches. Due to this lack of understanding, students may not be able to come to correct conclusions and may not be able to apply the testing approach to new contexts [4]. The use

of computer simulation methods to mimic the real situations helps with understanding abstract concepts [35]. Cobb [7] noted that incorporating computer simulation techniques to illustrate key concepts allows students to discover important principles themselves. Mills [24] has given a comprehensive review of the literature of computer simulation methods used in all areas of statistics to help students understand difficult concepts. Lock et al. [23] has noted that randomization-based simulation methods make the fundamental concepts of statistical inference more visual and intuitive, and help students see connections that are otherwise lost with numerous different formulae. Tintle et al. [31] developed a randomization based curriculum to teach introductory statistics. Their assessment showed that the students learned significantly more about statistical inference using the new curriculum.

Almost all statistical software packages offer ways to perform simulation. In this paper, we describe how to use R to perform hypothesis testing using repeated simulated sampling. R is a free, powerful, and flexible statistical programming language and computing environment. It has become very popular among educators and statisticians. R runs on all of the commonly used computer platforms including Windows, Unix/Linux, and the Macintosh operating system (Mac OS). Statistics instructors of all levels are now using R to teach and perform statistical data analysis. Several educators including Verzani [33] and Zhang [36] have proposed R as a tool in teaching statistics at introductory level. R can easily generate random samples from many data sets and a variety of probability distributions. Hallgren [16] has used R and noted that simulation methods are flexible and can be applied to a number of problems to obtain answers that may not be possible to derive through other approaches. The Mosaic package in R is intended to support teaching statistical concepts in introductory level [20]. This package includes the commands to carry-out randomization based statistical inferences including confidence intervals, testing for proportions and testing means for two groups. However, although the Mosaic package includes the procedure for a traditional one-sample *t*-test, it does not include the randomization procedure that would allow creation a sampling distribution from the original sample or a one sample *t*-test based on that distribution. In this paper, we illustrate this procedure and the connection between the *p*-value and the evidence against the null hypothesis.

We describe how to use simulation in R to enhance the understanding of hypothesis testing at the introductory level. We generate random samples from a given sample from a population and compute the empirical sampling distribution of a test statistic. This sampling distribution is known as the randomization distribution [9]. Then we use this sampling distribution to show the *p*-value graphically. The *p*-value distribution and the relationship between the *p*-value and the evidence against the null hypothesis are shown graphically to enhance understanding of the concepts.

This paper is organized in 5 sections. Section 2 gives a brief overview of hypothesis testing and the *p*-value. Section 3 provides a quick look at R and simulation. In section 4, we demonstrate the simulation of hypothesis testing and *p*-values. Section 5 ends the paper with some concluding remarks.

2. AN OVERVIEW OF HYPOTHESIS TESTING

Hypothesis testing is a common procedure that uses statistical evidence from a sample to draw a conclusion about a hypothesis. The hypothesis is a quantitative statement formulated about a population characteristic. There are a number of different types of hypothesis tests useful for different hypothesis scenarios and data samples. In this discussion, we consider one sample hypothesis testing. *Figure 1* gives an example for a basic idea of hypothesis testing. In this example, we consider the population of all men. A sample is randomly drawn from the population. To compare the sample and the population, we introduce a quantification that maps the sample to a number. This number could

represent a sample characteristic such as mean height, mean weight or mean age of the sample. This mapping is called a test statistic.

A hypothesis could be that the mean weight of all men equals 172 pounds. Such a hypothesis is called a null hypothesis and it is denoted by H_0. The idea behind hypothesis testing is the same as the idea behind a criminal trial. The person is presumed to be not guilty (assumed null hypothesis) until there is sufficient evidence to declare the person guilty. This population consists of a larger number of (may be infinite) adult men and the weight can follow a probability distribution with the hypothesized mean of 172 pounds. If we take many samples of the same size from the population and compute the corresponding sample characteristic (in this case sample mean weight), that forms a probability distribution called the sampling distribution for mean weight. In order to evaluate the validity of the null hypothesis, we can compare the value of our sample characteristic (obtained from the sample we took) with the hypothesized population characteristic. From this comparison, we compute another numerical value, called the *p*-value, which provides the evidence against the null hypothesis being true. Finally, we make the conclusion based on this *p*-value. This basic idea of one sample hypothesis tests can be extend to other hypothesis testing scenarios.

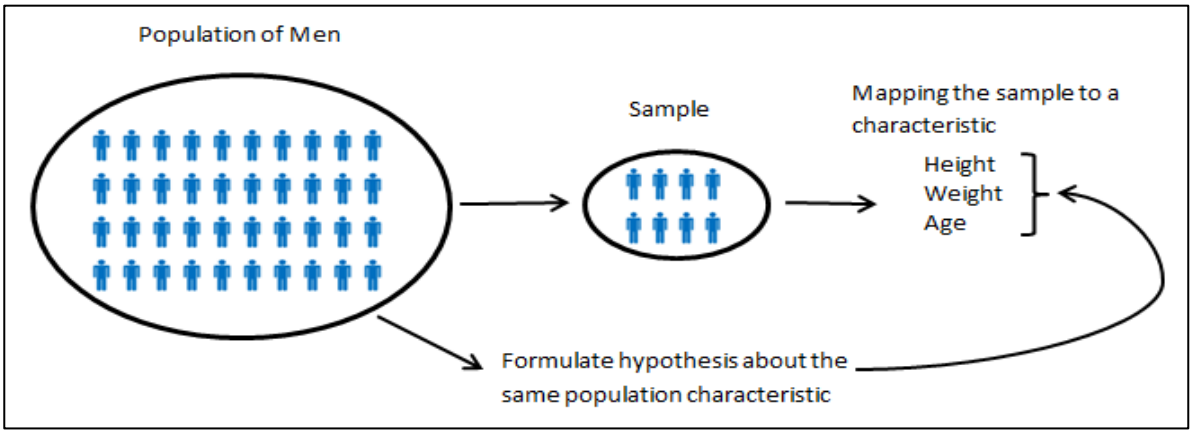

Figure 1: The diagram of the example explains the basic idea of hypothesis testing

2.1 HYPOTHESIS TESTING PROCEDURE

The basic idea of hypothesis testing is shown in the diagram above. Now we give some more details of the key parts of hypothesis testing. The following steps are used in hypothesis testing:

1. Define the null hypothesis (H_0) and the alternative hypothesis (H_1).
2. Choose the significance level α.
3. Select the appropriate test statistic and sampling distribution when H_0 is true.
4. Select a sample and compute the value of the test statistic.
5. Determine the critical value(s) and specify the critical region **or** compute the *p*-value
6. Make the decision (fail to reject H_0 or reject H_0).

Step 1: Null Hypothesis H_0 and Alternative Hypothesis H_1

In this step, we define two hypotheses: the null hypothesis (H_0) and the alternative hypothesis (H_1). Both hypotheses express statements about a population parameter. The null hypothesis is a statement that is assumed to be true at the beginning. The alternative hypothesis is a statement that is contradictory to the null hypothesis. The null hypothesis states that the population parameter is equal to a hypothesized value, but depending on the situation the alternative hypothesis can be one of the three forms: greater, less, or not equal. If our population parameter is the mean μ, then H_0 will be of the form $\mu = \mu_0$ where μ_0 is the null hypothesized value. Then H_1 can be $\mu > \mu_0$, $\mu < \mu_0$ or $\mu \neq \mu_0$.

The alternative hypothesis H_1: $\mu > \mu_0$ is called the right-sided hypothesis because it assumes values greater than μ_0. The alternative hypothesis H_1: $\mu < \mu_0$ is called the left-sided alternative and H_1: $\mu \neq \mu_0$ is called the two-sided alternative.

Step 2: Choose the significance level α

In making a decision in a hypothesis test, two errors could happen: rejecting the null hypothesis when it is actually true and failing to reject the null hypothesis when it is actually false. The first one is called the Type 1 error and the second one is called the Type 2 error. The significance level, denoted by α, is the probability of making the Type 1 error. This value will be used in making the decision of rejecting the null hypothesis or not. This procedure will be explained in step 5. When conducting a hypothesis test, we have the freedom to choose the value of α. Even though the common choice of α is 0.05, one needs to be aware of the potential consequences when selecting this value. If the possible consequences of making the Type 1 error are great, it is advisable to use a lower significance level.

Step 3: Select the appropriate test statistic and sampling distribution when H_0 is true

The proper definition of a statistic is a characteristic of a sample. In hypothesis testing, we compute the value of a test statistic assuming the null hypothesis is true. In this way the statistic measures the degree of agreement between the sample data and the null hypothesis. Before we substitute values of the sample data, a test statistic is a random variable that follows a certain probability distribution. This probability distribution is called the sampling distribution of that statistic. This sampling distribution represents the values of the test statistic assuming the null hypothesis is true. Common examples of statistics are sample mean, sample proportion and sample variance. Each of these statistics has their own probability distribution under certain conditions. For example, the sample mean $\bar{X}$ has a normal distribution if the original population is normally distributed or an approximate normal distribution for larger samples if the original population is not normally distributed. This fact is given in the Central Limit Theorem [5]. If we standardize $\bar{X}$, the resulting test statistic will have either a standard normal distribution or a *t* distribution depending on whether the population standard deviation is known or not. The test statistic plays a major role in hypothesis testing. Which test statistic to use will depend on the parameters that appear in the hypotheses we formulate. For example, if the test is about the population mean, we use the Z or *t*-test depending on whether the standard deviation is known or not. For testing for variance, we use the Chi–square test.

Step 4: Select a sample and compute the value of the test statistic.

As we noted in step 3 above, before we compute the value of the test statistic based on a random sample, it is a random variable. To assess the evidence against the null hypothesis, we need to take a random sample from the population and compute the value of the test statistic assuming the null hypothesis is true. As given in steps 5 and 6, the value of the test statistic along with the relevant sampling distribution will be used to make the conclusion.

Step 5: Determine the critical value(s) and specify the critical region **or** compute the *p*-value

Once the value of the test statistic is computed as indicated in step 4, one can use either the critical value approach or the *p*-value approach to make the decision.

In the critical value approach (known as the traditional method), the observed value of the test statistic is compared to the critical value. The critical value is a cutoff value to determine whether or not the observed test statistic value is more extreme than what would be expected if the null hypothesis is true. The critical value is computed based on the assumed significance level α and the sampling (probability) distribution of the test statistic. The critical value divides the area under the sampling distribution curve into two regions: the critical (rejection) region and the noncritical (non-rejection) region.

In the *p*-value approach, the evidence against the null hypothesis is true is compared with the significance level α. The *p*-value is the probability, (assuming the null hypothesis is true) that the test statistic would take a value as extreme or more extreme than what was actually observed [8]. Since all statistical software calculate *p*-values, more and more researchers and instructors are using the *p*-value approach to make decisions. The smaller the *p*-value, the stronger the evidence is against the null hypothesis. The final decision of rejecting or failing to reject the null hypothesis will be based on the calculated *p*-value.

The idea of the *p*-value is somewhat misunderstood and one needs a good understanding of it to make a correct decision in hypothesis testing. There are several misconceptions associated with the interpretation of a *p*-value [14]. One of the most common ones is that the *p*-value is the probability that the null hypothesis is true. As mentioned in the previous paragraph, as the *p*-value is calculated under the assumption that the null hypothesis is true, it does not provide information regarding whether the null hypothesis is actually true. Similarly, the *p*-value cannot be interpreted as the probability that the alternative hypothesis is true. The American Statistical Association (ASA) believed that the scientific community could benefit from a formal statement clarifying several widely agreed upon principles underlying the proper use and interpretation of the *p*-value. The ASA Statement on *p*-Values by Wasserstein and Lazar [34] gives the description on this statement. Recently, Benjamin & Berger [1] have given three recommendations for improving the use of *p*-values.

A critical value and a *p*-value of a hypothesis test with a right-sided alternative hypothesis are shown in *Figure 2*.

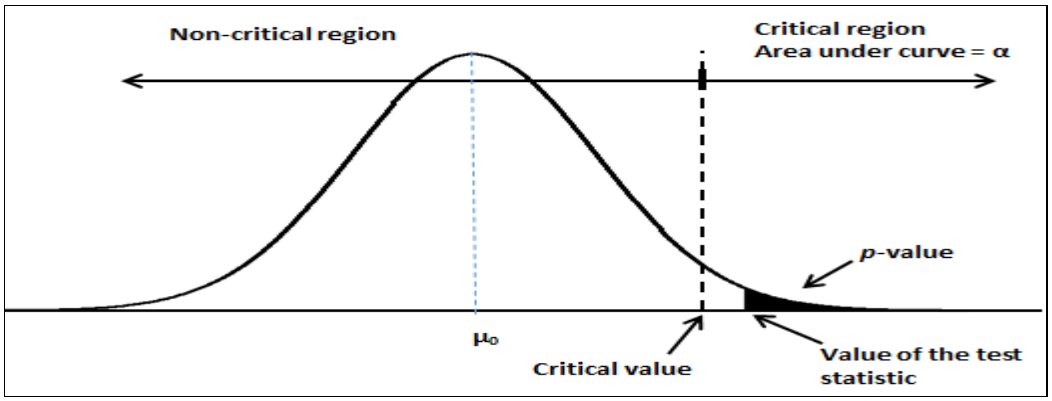

Figure 2. Critical region and *p*-value of sampling distribution of test statistic.

Step 6: Make the decision (fail to reject H_0 or reject H_0).

In this final step, we decide about the null hypothesis. In step 5, we have given two approaches. In the critical value approach, if the observed value of the test statistic falls within the critical region, the null hypothesis is rejected. Otherwise we fail to reject the null hypothesis. In the *p*-value approach, if the *p*-value is less than the assumed significance level α, the null hypothesis is rejected. If this does not hold true, we fail to reject the null hypothesis.

3. A QUICK LOOK AT R AND SIMULATION

Although it is not feasible to provide a general introduction to R in this paper, we will provide enough background to understand the remainder of this paper. R is a programing language used to study statistics problems and to conduct research [30]. R is a relatively simple syntax-driven and case-sensitive language. Even though the syntax for writing instructions may be somewhat difficult in the beginning for students with little or no prior programming experience, most students have become comfortable using R [16]. R is an object-oriented program that works with data structures

such as vectors (one-dimensional array) and data frames (two-dimensional arrays). A vector contains a list of values. When the R program is started and after it prints an introductory message, the R interpreter prompts for input with > (the greater-than sign). The interpreter executes expressions that are typed at the command prompt. For example:

```
> 2 + 4*5
[1] 22
> 1:5
[1] 1 2 3 4 5
> 1:5 + c(2,5,-4,6,0)
[1]  3  7 -1 10  5
>  x  <- 1:5
> 2*x
[1]  2  4  6  8 10
```

Most of the above R statements are self-explanatory except for the following:

- Simple (vector) output is prefixed by [1]. If the output extends over several lines, the index number of the first element in each line appears in square brackets at the beginning of the line.
- c() function combines its arguments to create a vector. The arguments are specified within parentheses and separated by commas.
- <- is the assignment operator. The equal sign (=) may also be used for assignment purposes. Variables are created and memory is allocated to them dynamically. Variable names can consist of any combination of lower and upper case letters, numerals, periods, and underscores, but cannot begin with a numeral or an underscore. R is case sensitive and there is no limit on the number of characters in a name.

Once we have a vector of numbers, we can apply built-in functions to get useful statistical summaries and visual displays. R also provides functions for generating random samples from various probability distributions.

sample **function**

The *sample* function generates a random sample of specified size from a set of values with or without replacement. Let us suppose values are stored in a vector named x. To take a random sample of size *n* without replacement from the set x, we use the following R commands:

```
>  sample(x, n)    or  > sample(x, n, replace = FALSE)
```

To obtain a sample of size *n* with replacement, we use the following command:

```
>  sample(x, n, replace = TRUE)
```

We can use the *sample* function to obtain a random number from a set of numbers, say 1 through 10, in the following way:

```
> sample(1:10,1)
[1] 8
```

Control Structures

R has the standard control structures such as *if*, *while*, and *for*. These can be used to control the flow of an R code. We will demonstrate the use of control structures in R using the following code segment. Let us assume that we have stored 1000 numbers in the vector named x. The following code will compute the average of the non-negative numbers in vector x. The symbol # is used to write comments.

```
> total <- 0              # variable total initialized to 0
> count <- 0              # variable count initialized to 0
> for(i in 1:1000)
+ {
+   if (x[i] >= 0)
+   {
+     count <- count +1    # count the non-negative values
+     total <- total + x[i]   # add the non-negative values
+   }
+ }
> average <- total/count    # compute the average
```

In the above code segment, the *for* loop iterates 1000 times, selecting only nonnegative numbers using the *if* statement. It computes the average as well. The variable named *count* counts the number of nonnegative numbers stored in x. We will use the control structures when we discuss simulations in this paper.

Simulation Methods

Simulation methods combine both mathematical and logical concepts that try to imitate the operations of a real-life process or system through the use of computer software. The computer simulation approach has been used in mathematical modeling and to solve complex problems [32]. In recent times, many instructors use simulation techniques in their classrooms to explain difficult concepts [6]. In certain situations, relevant data is not readily available or cannot be obtained. In these situations, simulation modeling provides ways to study such problems. In the simulation process, a computer program or a software application is used to generate steps that are formulated to model and solve the given problem. Based on the assumptions and the rules of the problem, outcomes or the values of the variables of the model are calculated iteratively until desired results are obtained.

Even though many statistical questions can be answered using mathematical analysis, the complexity of some statistical questions makes them more easily answered through simulation methods [26]. In these cases, simulation may be used to generate datasets that conform to a set of known parameters such as the mean or the variance specified by the researcher. There are many technological tools and computer programming languages (such as R) available to perform such simulations.

3. HYPOTHESIS TESTING EXAMPLE

Now we use an example to perform a one sample *t*-test and simulation to compute the *p*-value. Data for this example are obtained from [25].

Does the use of fancy type fonts slow down the reading of text on a computer screen? Adults can read four paragraphs of text in an average time of 23 seconds in the common Times New Roman font. 25 adults were asked to read this text in the ornate font named Gigi. Here are their times:
23.2, 21.2, 28.9, 27.7, 23.4, 27.3, 16.1, 22.6, 25.6, 32.6, 23.9, 26.8, 18.9, 27.8, 21.4, 30.7, 21.5, 30.6, 31.5, 24.6, 23.0, 28.6, 24.4, 28.1, 18.4.

Suppose that reading times are normally distributed. Is there good evidence that the mean reading time for Gigi fonts is greater than 23 seconds? In other words, is μ greater than 23 seconds for Gigi fonts?

The null and alternative hypotheses are: H_0: μ = 23 seconds and H_1: μ > 23 seconds. The test statistic used for this test is $(\bar{X} - \mu)/(s/\sqrt{n})$ and that follows the *t*-distribution with 24 degrees of freedom under the assumption that H_0 is true. *s* is the sample standard deviation.

The *t.test* command in R performs the *t*-test and produces the *p*-value. *Figure 3* shows the R output. The data are stored in a data frame named *times*. One can access data by saving them in an R, csv or text file for the *t.test* command.

```
> times <- c(23.2, 21.2, 28.9, 27.7, 23.4, 27.3, 16.1, 22.6, 25.6, 32.6, 23.9, 26.8, 18.9, 27.8, 21.4, +
30.7, 21.5, 30.6, 31.5, 24.6, 23.0, 28.6, 24.4, 28.1, 18.4)
> t.test(times, mu = 23, alternative = 'greater')

      One Sample t-test

data:  times
t = 2.5073, df = 24, p-value = 0.009669
alternative hypothesis: true mean is greater than 23
95 percent confidence interval:
 23.68356      Inf
sample estimates:
mean of x
  25.152
```

Figure 3. R output of one sample *t*-test.

To perform a left tailed or two tailed test, 'greater' should be replaced with 'less' or 'two.sided' respectively. Since the *p*-value (0.0097) is less than the significance level α (say 0.05), we have sufficient evidence to reject the null hypothesis and conclude that the true mean reading time is greater than 23 seconds.

We notice that the implementation of hypothesis testing in R is fairly simple. One can use the *t.test* command to obtain the *p*-value without understanding the concepts given in section 2. Those concepts are hidden behind the *t.test* command. We need to caution that a thorough understanding of hypothesis testing cannot be obtained by simply executing it in R or any software package. One needs to understand the concepts before executing it in software. Software is just a tool to get the output and it saves time and the effort of doing the tedious calculations.

To get a better understanding of the *p*-value and to assess the evidence against the null hypothesis based on this sample, we generate new samples that are consistent with the null hypothesis, H_0: μ = 23 seconds. We use randomization testing procedures in resampling techniques to construct a sampling distribution that can be used to make inferences about the population [9]. The randomization distribution is constructed under the assumption that the null hypothesis is true. That means the randomization distribution is centered on the value in the null hypothesis. The original sample is shifted so that the sample mean equals the hypothesized population mean (i.e., the value in the null hypothesis). Samples of the same size as the original sample are drawn with

replacement from the shifted distribution and the mean of each randomization sample is calculated. This distribution is called the randomization distribution of the test statistic. This resampling procedure is similar to the bootstrap procedures [10, 18]. In bootstrap procedures, we take random samples of the same size with replacement from an original sample and calculate the value of a test statistic. This sampling distribution is called the bootstrap distribution and that can be used to make inferences about the original population. What makes a randomization distribution different from bootstrap distribution is that it is constructed given that the null hypothesis is true.

In our example, the sample mean of the original sample is 25.152. To ensure that the null hypothesis ($\mu = 23$) is satisfied, we subtract 2.152 from each reading time to produce a new set of times with a mean exactly equal to 23. To generate a randomization distribution of sample means while assuming that the null hypothesis is true, we select samples of size 25 at a time (with replacement) from the modified data and compute the mean of each sample. A set of sample means generated by this process will be a randomization distribution of values produced at random under the null hypothesis that $\mu = 23$. The R code below simulates this process and generates 10,000 sample means. A larger number of samples are needed to get a proper and precise impression on the complete distribution of the statistic. Hesterberg [18] noted that 1000 samples are needed for rough approximation of the distribution and 10,000 or more are needed for better accuracy.

```
> newTimes <- times - 2.152
> means <- c()
> for(i in 1: 10000)
+ {
+     x <- sample(newTimes, 25, replace = TRUE)
+    means[i] <- mean(x)
   }
```

Figure 4 shows a dot plot of the sample means generated by this randomization process. As expected, the distribution is centered at the null hypothesized value of 23.

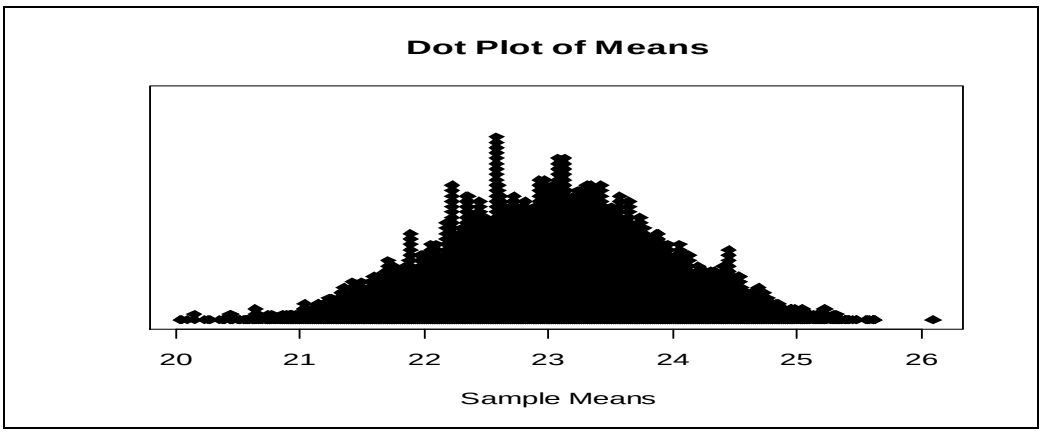

Figure 4. Dot Plot of 10,000 sample means.

To compute the empirical *p*-value, we compute the proportion of sample means that are as large (or larger) than the original sample mean 25.152. The following command computes this proportion.

```
> mean(means >= 25.152)
 [1] 0.0062
```

This empirical *p*-value is off by 0.003 in comparison to the value given in the *t.test* command. *Figure* 5 shows the probability histogram of sample means generated from the simulation process. The *p*–value is visualized in the right tail of the histogram. While the *t.test* command allows students to perform the hypothesis test, the above described simulation process provides a better understanding of the *p*-value.

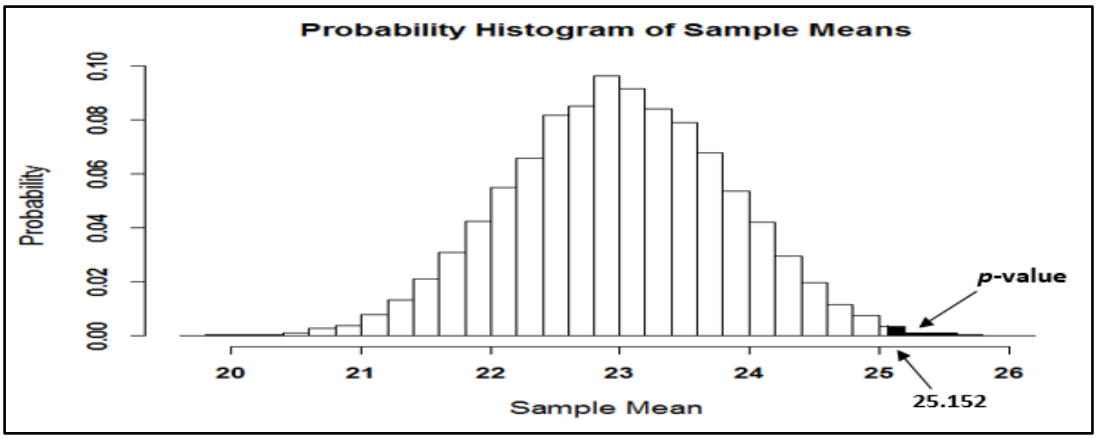

Figure 5. Probability histogram of sample means with *p*-value in right tail.

Now we compute the transformed *t*-test statistic values for simulated samples to observe the relationship between the value of the test statistic and the *p*-value. Because the test statistic varies from one sample to another, the *p*-value will also vary from one sample to another. This result makes the test statistic a random variable, and so the *p*-value will also be a random variable. The sampling distribution of the transformed test statistic $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ is a *t* distribution with *n*-1 degrees of freedom where $\bar{x}$ is the sample mean, *s* is the sample standard deviation, μ is the null hypothesized population mean, and *n* is the sample size. The *p*-value for each sample is calculated by finding the area under the *t* distribution curve that falls above the *t*-test statistic. The following R code segment simulates this process.

```
> tStats <- c()
> means <- c()
> pValues <- c()
> for(i in 1:10000)
+ {
+    x <- sample(newTimes, 25, replace = TRUE)
+   means[i] <- mean(x)
+   tStats[i] <- (mean(x) - mean(newTimes))/(sd(x)/sqrt(25))
+   pValues[i] <- 1 - pt(tStats[i],24)
+ }
```

As we noted earlier, the theoretical sampling distribution of the test statistic is a *t* distribution with 24 degrees of freedom. We compare the approximate sampling distribution that we obtained using the above R code with the theoretical distribution in *Figure 6* by overlaying the *t* distribution

on the histogram. We observe that the empirical distribution with sample sizes of 25 quite reasonably agrees with the theoretical distribution.

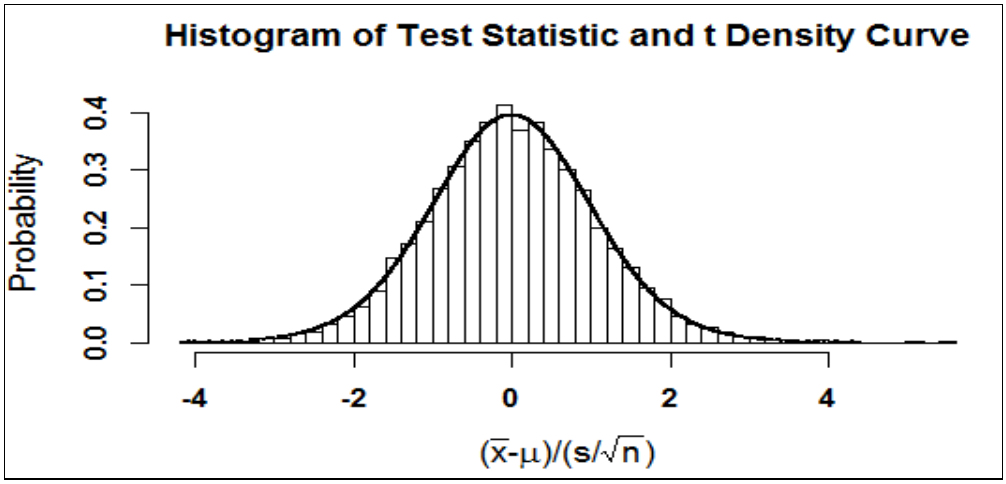

Figure 6. Empirical and theoretical sampling distribution of *t*-test statistic.

Now we plot the *p*-values against the corresponding sample means. The central theme of hypothesis testing is that the sample evidence is used for not supporting the null hypothesis. If the sample results are compatible with the null hypothesis, we fail to reject the null hypothesis. As the sample mean moves away from the null hypothesized mean in the direction of the alternative hypothesis, the strength of the support for the alternative hypothesis increases. Since the *p*-value is the likelihood of observing more extreme evidence than what the sample has produced assuming the null hypothesis, for larger sample means, *p*-value decreases. A visual image would be a good way to gain a better understanding of this concept. *Figure 7* gives a plot of *p*-values against sample means. In *Figure 7*, we have plotted 10,000 simulated sample mean *p*-value pairs. As the sample mean increases from the null hypothesized mean of 23, the support for the alternative hypothesis increases. From *Figure 7* one can notice that as support for alternative hypothesis increases, the *p*-value decreases. The smaller the *p*-value, the stronger the evidence is against the null hypothesis.

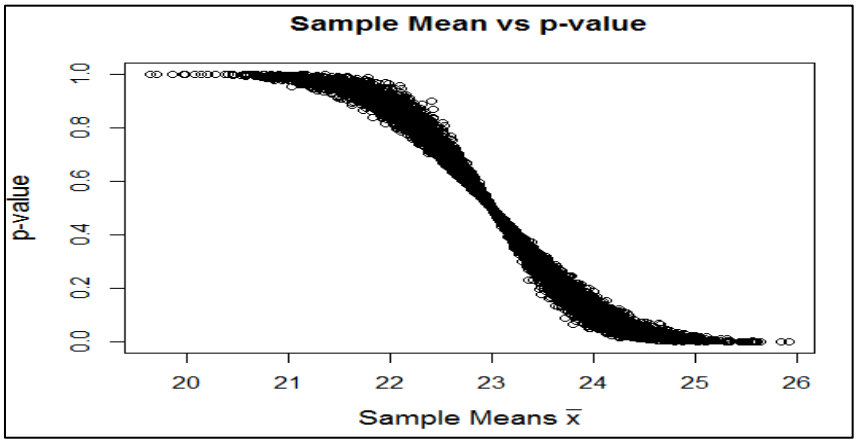

Figure 7. Plot of *p*-values against sample means.

p -Value Distribution

Simulation methods can be used to observe the *p*-value distribution and how the *p*-value behaves under the null and alternative hypotheses. As we noted earlier, the *p*-value is a random variable. It

varies as the test statistic varies. Under the null hypothesis, the *p*-value follows a uniform distribution on the interval from 0 to 1 [3,19]. The *Figure 8* shows the histogram of the 10,000 *p*-values from previous simulations assuming the null hypothesis.

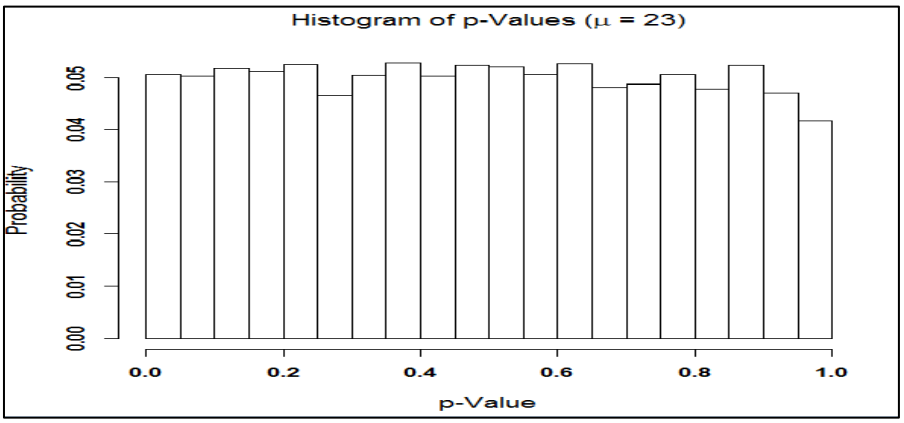

Figure 8. Histogram of *p*-Values under H_0.

The histogram in *Figure 8* shows that the *p*-values appear to be relatively uniform. About 5% of these *p*-values are in the first class interval from 0 to 0.05. What does this tell us? If we use the significance level α as 5%, we will reject the null hypothesis 5% of the times when the null hypothesis is indeed true. This is the probability of making a Type 1 error. From our simulation experiment, the first class interval in the histogram shows this empirically. The proportion of *p*-values from 0 to 0.05 is about 5%.

Now we observe the *p*-value distribution when we consider several values of the mean μ when the null hypothesis is false. The null hypothesized value of the mean μ is 23. We use the same simulation procedure used earlier to generate 10,000 random samples from the randomization distribution with mean μ = 24 and μ = 25 and compute the value of the test statistic $t = \frac{(\bar{x}-23)}{s/\sqrt{25}}$ and the corresponding *p*-value for each sample. *Figure 9* shows the histograms of *p*-values for both cases.

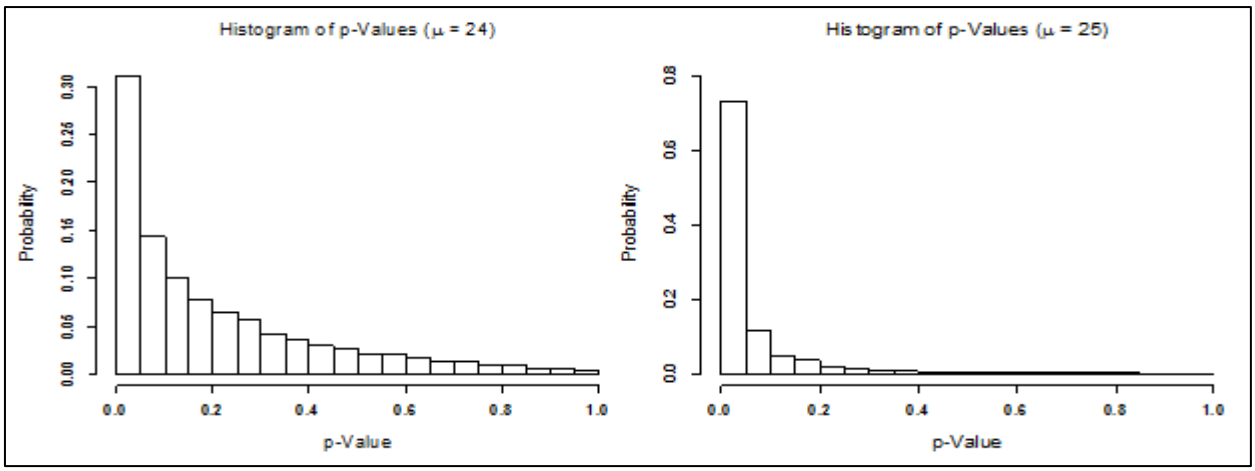

Figure 9. Histograms of *p*-values under H_1.

The shape of the *p*-value histograms in *Figure 9* is quite different from the one in *Figure 8*. More and more *p*-values are concentrating about 0 as μ increases. In the case of μ = 24, little more than 30% of the *p*-values are less than 0.05. Assuming a 5% significance level, about 30% of the 10,000 samples (tests) correctly reject the null hypothesis. In the case of μ = 25, about 75% of the 10,000

samples (tests) correctly reject the null hypothesis. We notice that the farther the actual value of the mean (parameter) is from the null hypothesized value, the more the distribution of *p*-value will be closer to 0. This indicates that there is a greater chance that the test will correctly reject the null hypothesis.

5. CONCLUSION

In this paper we showed how to use R and simulation to understand the hypothesis testing concepts. Due to the abstract nature of the concepts of hypothesis testing at an introductory level, using R with a simulation approach benefits students. Previous research has shown that the traditional procedures for teaching hypothesis testing are ineffective [8]. We used the randomization and simulation of these distributions and their visualization to introduce the fundamental concepts of hypothesis testing, *p*-values, and decision-making. By generating random samples from an original sample, we computed the empirical sampling distribution of the test statistic assuming the null hypothesis is true. Then we showed the *p*-value distribution and the relationship between the *p*-values and the evidence against the null hypothesis graphically. From this visual display, one can observe that as evidence against the null hypothesis increases, the *p*-value decreases. Furthermore we showed that the farther the actual parameter from the null hypothesized value, the more the *p*-values concentrate about 0. These are the key ideas that we want our students to understand. These simulation methods are comprehensible to students with varying levels of mathematics knowledge and experience.

REFERENCES

[1] Benjamin, D. J. & Berger, J. O. (2019). Three Recommendations for Improving the Use of *p*-Values. *The American Statistician*, 73: sup1, 186 –191. https://doi.org/10.1080/00031305.2018.1543135

[2] Biau, D. J., Jolles, B. M., & Porcher, R. (2010). P Value and the Theory of Hypothesis Testing: An Explanation for New Researchers. *Clinical Orthopedics and Related Research*, 468(3), 885–892.

[3] Breheny, P,, Stromberg, A,, & Lambert, J. (2018). *p*-Value Histograms: Inference and Diagnostics. *High-Throughput*, 7(3), 23, doi: 10.3390/ht7030023

[4] Chandrakantha, L. (2015). Simulating One-Way ANOVA Using Resampling. *The Electronic Journal of Mathematics and Technology (EJMT)*, 9(4), 281-296.

[5] Chandrakantha, L. (2018). Simulating Sampling Distribution of the Mean in R. *The Electronic Journal of Mathematics and Technology (EJMT)*. 12(2), 309-321.

[6] Chandrakantha, L. (2014). Visualizing and Understanding Confidence Intervals and Hypothesis Testing Using Excel Simulation. *The Electronic Journal of Mathematics and Technology (EJMT)*, 8(3), 212-221.

[7] Cobb, P. (1994). Where is the Mind? Constructivist and Sociocultural Perspectives on Mathematical Development. *Educational Researcher*, 23, 13-20.

[8] Dambolen, I.G., Eriksen, S. E., & Kopsco, D. P. (2009). An Intuitive Introduction to Hypothesis Testing. *INFORMS. Transaction of Education*, 9(2), 53-62.

[9] Edgington, E. & Onghena, P. (2007). *Randomization tests*, Chapman & Hall: New York.

[10] Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York, USA.

[11] Fisher, R. A. (1925). *Statistical Methods for Research Workers*; Genesis Publishing Pvt Ltd.: Guilford, UK.

[12] Fisher, R. A. (1992). The Arrangement of Field Experiments (1926), *In Breakthroughs in Statistics*; Springer: Berlin, Germany, 82-91.

[13] Fisher, R. A. (1929). The Statistical Methods in Physical Research. *Proc. Soc. Psych. Res.*, 39, 189-192.

[14] Goodman, S. (2008). A dirty dozen: twelve p-value misconceptions. *Seminars in Hematology*, 45(3), 135-140.

[15] Hacking, I. (2016). *Logic of Statistical Inference*; Cambridge University Press: Cambridge, UK.

[16] Hallgren, K. A. (2013). Conducting Simulation Studies in the R Programming Environment. *Tutorial in Quantitative Methods for Psychology*, 9(2), 43-60.

[17] Helbing, D. (2015). The Automation of Society Is Next: How to Survive the Digital Revolution. https://ssrn.com/abstract=2694312.

[18] Hesterberg, T.C. (2015). What Teachers should know about Bootstrap: Resampling in the Undergraduate Statistics Curriculum. *The American Statistician*, 15(4), 371-386. https://doi.org/10.1080/00031310.2015.1089789.

[19] Hung, H.M. J., O'Neill, R. T., Bauer, P., & Kohne, K. (1997). The Behavior of the P-Value When the Alternative Hypothesis is True. *Biometrics*, 53(1), 11-22.

[20] Kaplan, D., Horton, N. J., & Pruim, R. (2019). Randomization-based inferences using mosaic package, 2019, https://cran.r- roject.org/web/packages/mosaic/vignettes/Resampling.pdf.

[21] Kim, J, H., & Robinson, A. P. (2019). Interval-Based Hypothesis Testing and Its Applications to Economics and Finance, *Econometrics MDPI*, 2-22.

[22] Lehman, E. (2005). *Testing Statistical Hypotheses*; Springer: New York, NY, USA.

[23] Lock, K.L., Lock, R. H., Lock, P. F., Lock, E. F., & Lock, D. F. (2014). STATKEY: Online Tools for Bootstrap Intervals and Randomization Tests. *Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9)*. https://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_9B2_LOCKMORGAN.pdf

[24] Mills, J. D. (2002).Using Computer Simulation Methods to Teach Statistics: A Review of the Literature. *Journal of Statistics Education (Online)*, 10 (1). http://www.amstat.org/publications/jse/v10n1/mills.html .

[25] Moore, D. S. (1996). *Essential Statistics.* W. H. Freeman & Company, New York, USA.

[26] Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine.* 38(11), 2074-2102. https://doi.org/10.1002/sim.8086.

[27] Neyman, J., & Pearson, E.S. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London*, 231, 289-337.

[28] Neyman, J., & Pearson, E.S. (1967). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika.* 20,1-2.

[29] Perezgonzalez, J. D. (2015). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Frontiers in Psychology*. 6, 223. https://doi.org/10.3389/fpsyg.2015.00223.

[30] R Development Core Team. (2008). *R: A language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, ISBN 3-900051-07-0.

[31] Tintle, N., VanderStoep, J., Holmes, V., Quisenberry, B., & Swanson, T. (2011). Development and Assessment of a Preliminary Randomization-Based Introductory Statistics Curriculum. *Journal of Statistics Education*. 19(1), DOI: 10.1080/10691898.2011.11889599.

[32] Veltern, K. (2009). *Mathematical Modeling and Simulation, Introduction to Scientist and Engineers. John* Wiley & Sons, New York.

[33] Verzani, J. (2004). Using R for Introductory statistics. *Chapman & Hall/CRC Press*.

[34] Wasserstein, R. L., & Lazar, N. A.(2016). The ASA Statement on *p*-Values: Context, Process, and Purpose. *The American Statistician*. 70(2), 129-133. https://doi.org/10.1080/00031305.2016.1154108

[35] Widiyatmoko, A. (2018). The Effectiveness of Simulation in Science Learning on Conceptual Understanding: A Literature Review. *Journal of International Development and Cooperation*, 24 (1), 35-43.

[36] Zhang, M. & Mass, Z. (2019). Using R as a Simulation Tool in Teaching Introductory Statistics. *International Electronic Journal of Mathematics Education (IEJME).* 14(3), 599-610. https://doi.org/10.29333/iejme/5773