A Concrete Strategy for Teaching Hypothesis Testing

Franz LOOSEN

This paper describes a physical device that can be used as a teaching aid for hypothesis testing instruction. The accompanying verbal commentary is sketched, and advantages over traditional teaching methods are discussed.

KEY WORDS: Hypothesis testing; Statistical teaching aids.

1. INTRODUCTION

This paper presents a simple device (hereafter referred to as the demonstrator) that can be used as a teaching aid for introducing the basic concepts in the "classical" approaches to hypothesis testing in a coherent way. The demonstrator can be used to illustrate Fisher's (1935, 1970) procedure of testing null hypotheses and Neyman–Pearson's (1928, 1933) procedure where one is forced to choose between two rival hypotheses and the concept of power is introduced. The demonstrator is not intended to illustrate the Bayesian inferential procedures (Lindley 1965) nor the decision theory approach originated by Wald (1950).

2. DIFFICULTIES IN UNDERSTANDING THE SUBJECT

In my experience over many years of teaching an introductory course in statistics attended by undergraduate students in the behavioral sciences at the University of Leuven, almost all difficulties with the comprehension of the theory of hypothesis testing result from the following two points:

1. The fact that the subject is traditionally approached in an interweaved way from three perspectives:

   a. the perspective of the ignorant statistician who does not know "the state of nature" (the real world) and can only hypothesize and reason about it in a conditional form,
   b. the perspective of the state of nature, and
   c. the perspective of the student who is studying the subject, but who is at the same time also an omniscient observer because he is informed about the state of nature and the way of operating of the statistician.

2. The usual instructional practice of putting distributions referring to different cases (i.e., "H_0 is true" and "H_0 is false") on the same set of axes. For most students this combined presentation conflicts with the knowledge that only one of the distributions fits the case in any concrete situation. A related confusing practice is to omit explicit plotting of the distribution representing the state of nature, and to refer to it by the distribution used for testing a true hypothesis. Hence the conceptual difference between a distribution that refers to the real world and one that refers to a hypothetical construction is faded.

In order to avoid confusion of perspectives and ambiguity in graphical representations, the present instructional method proposes a three-level frame of reference in conducting hypothesis testing. Separate sets of axes are reserved for plotting: (1) the sampling distribution in the state of nature, (2) the sampling distribution representing the case where H_0 is true, and (3) the sampling distribution representing the case where H_0 is false.

The three-level frame is based on the conviction that the common practice to represent distributions under different states on the same set of axes, although being perfectly appropriate for treating the problem from the point of view of the statistician who cannot distinguish at any stage between testing a true or false hypothesis, is ill-suited for educational purposes. Students are always informed about the state of nature. Consequently, they spontaneously contrast testing a true versus a false hypothesis. Hence it seems natural to use a representational format that fits in with this mental approach by assigning separate sets of axes to distributions under different states. The demonstrator was designed to materialize this idea.

Of course, the simple use of the demonstrator does not lead to insight automatically. The accompanying verbal commentary that will be outlined in Section 4 is equally important. Because the theory of hypothesis testing can be found in all introductory textbooks, the commentary in this paper will be limited to specifying the way in which the concepts are presented in dialogue with the apparatus.

3. DESCRIPTION OF THE APPARATUS

The demonstrator (see Fig. 1) is simple and inexpensive to construct. (A copy of the demonstrator can be purchased at production cost (approximately $150) from the author.) Basically, it consists of a free-standing transportable wooden frame about 1 m wide and 1 m high. The rear of the frame is covered with dark green baize. In the upper surface of the three horizontal bars that are marked with a scale (of 5–39) two parallel grooves are cut out over their entire length. Inside these grooves wooden curves of the normal distribution can be placed and moved over the entire range of the scale. (In the present case all curves represent sampling distributions of the mean. Five curves are white and five others are yellow. In the white curves the standard error of the mean is 3 (referring to the case where $\mathbf{X} \sim N(\mu, 15^2)$ and the sample size is $n = 25$). In the yellow curves the standard error of the mean is 1.6 ($n = 88$). In the set of white curves one is completely white (both at the

Franz Loosen is Professor, Department of Psychology, University of Leuven, B-3000 Leuven, Belgium.

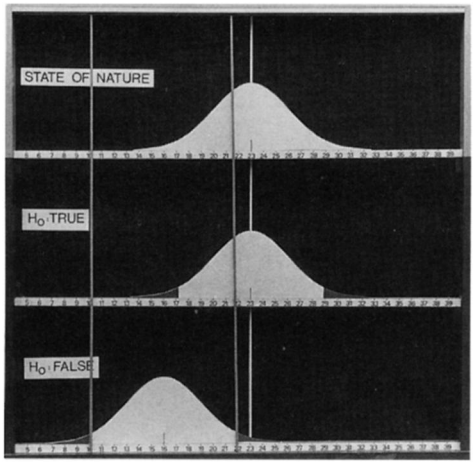

Figure 1. The Demonstrator.

front and at the rear), two have their upper and lower 2.5% tails marked in red paint, and the remaining curves have either their lower 5% tail or their upper 5% tail painted in red. The yellow curves are marked analogously to the white curves. Of course, other types of distributions (i.e., uniform and triangular distributions) and other significance regions (i.e., 1% and 10%) can be used.) There are also two red metal rods suspended in rails underneath the upper bar of the frame. These rods can move in a vertical position over the entire range of the scale in front of the curves. An analog white rod can move behind the curves.

The upper one-third part of the frame (referred to as the upper section) is painted in sky blue. The rest of the frame has the same color as the baize. The upper section is reserved for the sampling distribution in the state of nature. A curtain in sky-blue material can cover the upper section (see Fig. 2). The middle and lower sections are reserved for the cases "H_0 is true" and "H_0 is false," respectively.

4. INTRODUCTORY INSTRUCTION

In practice, I always start with a numerical example that is progressively built up on the board and culminates in an outline that fits the three-level frame of the demonstrator. Subsequently, the explanation is reworded, elaborated, and illustrated on the demonstrator. The numerical example describes the "classic" case where the hypothesis is tested that the mean μ of a normal distributed variable $\mathbf{X}$ is equal to a specified value when the standard deviation σ is known.

First, the distribution representing the state of nature is specified. Next, the hypothetical distributions used for testing a true and a false hypothesis are introduced. Finally, the viewpoint of the ignorant statistician is highlighted.

Step 1: The State of Nature. In Step 1 the state of nature is introduced. The concept is linked with a fictive omniscient external observer (for example, "God the Father" for people with Christian background). The practice of "personifying" the state of nature may seem odd, but in my experience it is an extremely efficient manner in concretizing the concept. Indeed, in this way the statistician, the student, and the omniscient can be contrasted in a lively "dialogical" style.

First, a random variable $\mathbf{X}$ is introduced. The variable is defined by $\mathbf{X} \sim N(23, 15^2)$. Next, the distribution of the sample mean $\bar{\mathbf{X}}$ is introduced in the following manner: "Statistics shows and the omniscient who is also an expert in statistics knows that in the case of a random variable $\mathbf{X}$ that is defined by $\mathbf{X} \sim N(23, 15^2)$ the sampling distribution of $\bar{\mathbf{X}}$ is defined by $\bar{\mathbf{X}} \sim N(23, 3^2)$ for samples of $n = 25$." This sampling distribution is drawn on the upper section of the board that is now labeled "state of nature" (see Fig. 3).

Step 2: Testing a TRUE Hypothesis. Step 2 approaches the subject from the point of view of the omniscient watching the statistician who is testing a true hypothesis. The middle section of the board is used for this purpose, and is labeled "H_0 is TRUE." The concepts of null hypothesis (H_0) and alternative hypothesis (H_1) are now introduced. The null hypothesis is defined as the exact hypothesis under test, and the alternative hypothesis is introduced as the "complement" of the null hypothesis. Thus when $\mu = 23$ and H_0 is true, then H_0: $\mu = 23$ and H_1: $\mu \neq 23$. In later examples it is shown on the demonstrator that any admissible hypothesis alternative to the one under test can be used as an alternative hypothesis. Moreover, the demonstrator also may be used to illustrate Neyman–Pearson's procedure

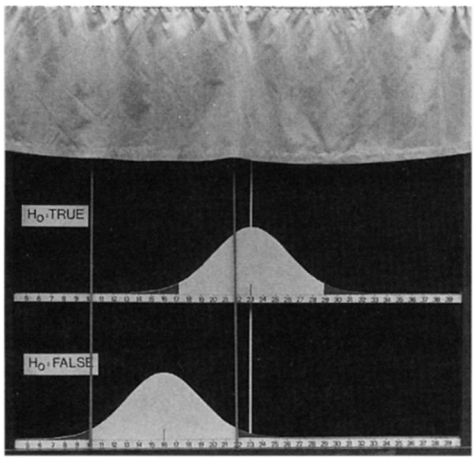

Figure 2. The Demonstrator Suggesting that the Sampling Distribution of the Mean is Unknown for the Statistician.

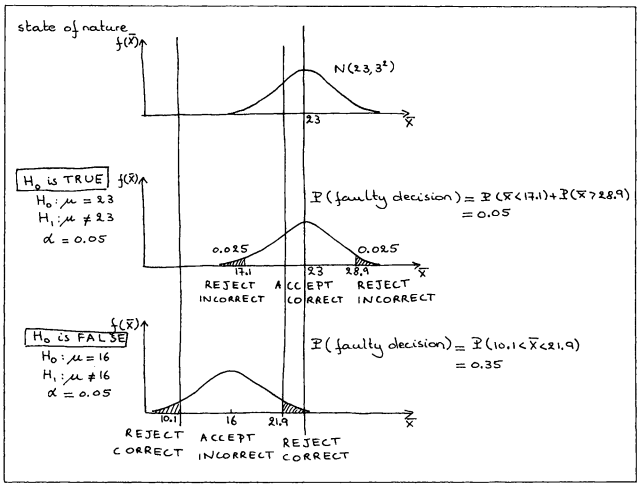

Figure 3. Outline on the Board.

where one is forced to choose between two rival hypotheses without referring to the state of nature.

In order to test H_0 empirically the statistician gathers data. In the present example a sample of 25 observations is drawn. If the sample data support H_0, then by default the statistician will continue to assert what is stated in H_0. On the other hand, if the data do not support H_0, the statistician will investigate whether the size of the departure from H_0 is so large that H_0 will have to be rejected as untenable. If H_0 is rejected, H_1 is accepted.

What exactly is meant by "the data do not support H_0"? This is defined in terms of chances of occurrence of what would be expected under H_0. Conventionally, the 5% (or 1%) of samples that deviate the most from what is stated in H_0 oppose H_0. At this point it is explained that the procedure for deciding whether to reject H_0 is based upon a probability model for observations. In the present case the sampling distribution of the mean is the appropriate model. This distribution is now drawn on the middle section of the board. In accordance with the null hypothesis the curve is centered on $\mu = 23$. Moreover, the curve is drawn in such a manner that corresponding abscissas in the upper and middle sections are vertically underneath one another.

Next, the critical points, the rejection regions (in red) and the acceptance region, are marked. It is explained that the choice of the rejection region(s) depends on the way in which H_0 can be false, and at the same time the term of level of significance (α) is introduced. In this context some explanation regarding the choice of reject and accept regions with respect to the level of significance should also be provided. For example, it should be pointed out that the rejection region is chosen so that the values that it contains have a total probability that is low on the null hypothesis, but that are better explained by the alternative hypothesis. Attention is also drawn to the fact that the actual values of α (.05, .01, etc.) are arbitrary. At this point the statistician's decisions are written underneath the abscissas of the curve: REJECT (for $\bar{\mathbf{X}} < 17.1$), ACCEPT (for $17.1 \leq \bar{\mathbf{X}} \leq 28.9$), and REJECT (for $\bar{\mathbf{X}} > 28.9$). It is specified that REJECT stands for "H_0 is rejected at the level of significance α," and "ACCEPT" must be interpreted as "H_0 is not rejected because there is no enough evidence to reject." The affirmative term "accept" is preferred to the negative expression "not rejected" because psychological studies have shown that negative statements are more difficult to comprehend than affirmative statements [see, for example, Carroll (1986)]. Subsequently, the judgments made by the omniscient about these decisions are "proclaimed" with a bass voice (simulating the voice of the omniscient) and written as "INCORRECT," "CORRECT," and "INCORRECT" underneath the respective decisions "REJECT," "ACCEPT," and "REJECT." Hence the concept of a wrong decision (Type I error) is introduced in a dialogical style.

It is explained that the fact that the decision procedure may lead to a wrong decision is not too serious if the risk associated with making a wrong decision is known. In the current context

$$\mathbf{P}(\text{wrong decision}) = \mathbf{P}(\bar{\mathbf{X}} < 17.1) + \mathbf{P}(\bar{\mathbf{X}} > 28.9).$$

At this point it is of the utmost importance that students realize that this probability, a probability that refers to the occurrence of an event, must be determined from a distribution that represents the real world, and not from a distribution that "exists merely in the mind of the statistician" as a hypothetical construction. Hence the student must realize that the probabilities $\mathbf{P}(\bar{\mathbf{X}} < 17.1)$ and $\mathbf{P}(\bar{\mathbf{X}} > 28.9)$ must be determined from the distribution in the upper section, and not from the distribution in the middle section of the board. In order to emphasize this and to avoid ambiguity in the computation of the respective probabilities the probability $\mathbf{P}$ is denoted hereafter by $\mathbf{P}_{\mu=23}$. Hence a notation that avoids confusion is

$$\mathbf{P}(\text{wrong decision}|H_0 \text{ is true}) = \mathbf{P}_{\mu=23}(\bar{\mathbf{X}} < 17.1) + \mathbf{P}_{\mu=23}(\bar{\mathbf{X}} > 28.9).$$

In order to delimit the areas representing $\mathbf{P}_{\mu=23}(\bar{\mathbf{X}} < 17.1) + \mathbf{P}_{\mu=23}(\bar{\mathbf{X}} > 28.9)$ in the distribution of the upper section, (red) vertical lines are drawn from the critical points of the hypothetical distribution in the middle section up to the x axis of the distribution in the upper section. It is explained that the statistician cannot calculate the size of the resulting surfaces in the distribution of the upper section. Indeed, this distribution is unknown to the statistician. However, the statistician knows that, if the null hypothesis is true, then the distributions in the upper and middle sections are identical. Hence in the case of a correct null hypothesis, the statistician can deduce from the known distribution of the middle section that

$$\mathbf{P}_{\mu=23}(\bar{\mathbf{X}} < 17.1) + \mathbf{P}_{\mu=23}(\bar{\mathbf{X}} > 28.9) = .05$$

in the unknown distribution of the upper section.

Step 3: Testing a FALSE Hypothesis. The case where the omniscient is watching the statistician who is testing a false hypothesis is now considered. For example, H_0: $\mu = 16$ (versus H_1: $\mu \neq 16$). The hypothetical sampling distribution of the mean on which the decision procedure is based is drawn in the lower section of the board. In accordance with the null hypothesis the curve is centered on $\mu = 16$. This curve is drawn in such a manner that corresponding abscissas in the different sections are vertically underneath one another. As in Step 2 the decisions of the statistician are written: REJECT (for $\bar{\mathbf{X}} < 10.1$), ACCEPT (for $10.1 \leq \bar{\mathbf{X}} \leq 21.9$), and REJECT (for $\bar{\mathbf{X}} > 21.9$). Next, the judgments made by the omniscient are pronounced with a bass voice and are written as "CORRECT," "INCORRECT," and "CORRECT," respectively. In this case the probability of a wrong decision (Type II error) is given by

$$\mathbf{P}(\text{wrong decision}|H_0 \text{ is false}) = \mathbf{P}_{\mu=23}(10.1 \leq \bar{\mathbf{X}} \leq 21.9).$$

As in Step 2 it is crucial that students realize at this point that this probability (i.e., β) cannot be determined from the distribution in the lower section (because this distribution exists only "in the mind of the statistician"), but from the one in the upper section because this distribution displays the probability distribution of the mean "in the real world." In the present example a simple calculation shows that

$$\mathbf{P}_{\mu=23}(10.1 \leq \bar{\mathbf{X}} \leq 21.9) = .35.$$

Step 4: The Viewpoint of the Ignorant Statistician. Step 4 reflects upon the previous steps from the point of view of the ignorant statistician. It is emphasized that the statistician will never know the state of nature. To make concrete this limitation the graph in the upper section is now erased in a theatrical manner. Later on when the demonstrator is being used the curtain is placed in front of the distribution in the upper section (see Fig. 2). So the students are confronted in a manifest manner with the fact that the distribution in the upper section exists (the distribution is physically behind the curtain), but is unknown to the statistician.

Subsequently, it is explained that although the statistician will never know whether or not the hypothesis is true or false, he/she is not prevented from reasoning in a conditional form and considering explicitly the implications of both eventualities. The statistician can therefore always declare that the decision process will take place at either the middle or the lower section of the outline on the board. Hence the statistician can reason in the following conditional way:

1. "**If H_0 is actually TRUE**, then the middle section represents the decision procedure. As a result of the known identity between the known hypothetical distribution in the middle section and the unknown real distribution in the upper section, I (the statistician) can calculate probabilities under the (unknown) real distribution in the upper section via the (known) hypothetical distribution in the middle section. Thus by means of the known hypothetical distribution, probabilities in the unknown real distribution can be deduced. Or, more concretely, the probability of a wrong decision is known if H_0 is true."
2. "**If H_0 is actually FALSE**, then the lower section represents the decision procedure. As a result of the known difference between the known hypothetical distribution in the lower section and the unknown real distribution in the upper section, I (the statistician) cannot know probabilities in the unknown real distribution in the upper section via the known hypothetical distribution in the lower section. Thus if H_0 is false, the probability of a wrong decision remains unknown."

On comparing 1 and 2 it is easy to comprehend that a statistician prefers rather to conclude "Reject H_0" than to conclude "Accept H_0." The probability of making an incorrect decision in case of the former is indeed known and can be "controlled" by using a smaller rejection region, whereas in case of the latter the probability of making an incorrect decision is unknown. Therefore, H_0 is always formulated with the intention of rejecting it. Attention is also drawn to the fact that in practical situations H_0 is virtually always false.

5. USING THE DEMONSTRATOR

Following the explanation on the board the same numerical example is recapitulated on the demonstrator. Prior to this it is advisable to focus the students' attention on the fact that most difficulties with the comprehension of the logic of hypothesis testing result from: (1) the conditional reasoning involved, and (2) the fact that the subject is approached from three points of view that are easily confused. The first warning is explained by pointing out that the statistician never knows whether the hypothesis is TRUE or FALSE, and is therefore always forced to consider the consequences of both cases. Hence the statistician will reason in the form of: "If the hypothesis is TRUE, then ...," and "If the hypothesis is FALSE and the state of nature is ..., then ...." The second warning is illustrated by commenting briefly on Figure 4 which is sketched on the board.

At the end of the recapitulation on the demonstrator the three distributions are located as shown in Figure 1. The dark (i.e., red) vertical bars are positioned for delineating the size of β (an area in the distribution of the upper section). The white vertical bar indicates the location of μ in the state of nature. In Figure 2 the blue curtain suggests that: "only the omniscient who sits high in the blue(!) sky knows the state of nature."

To consolidate knowledge the entire reasoning is repeated for different scenarios. For instance, for the same state of nature ($\mu = 23$) other null hypotheses are tested. For example:

1. In order to demonstrate that the probability for making a Type I error is maximum α for a one-tailed test (but exactly α for a two-tailed test), H_0: $\mu \geq 20$ is tested versus H_1: $\mu < 20$ (see Fig. 5) and H_0: $\mu \geq 18$ is tested versus H_1: $\mu < 18$. In case of H_0: $\mu \geq 20$ (versus H_1: $\mu < 20$) it is shown why the sampling distribution used for the decision process is centered on $\mu = 20$, and not on 21, 22, 23, etc.
2. H_0: $\mu \leq 21$ (versus H_1: $\mu > 21$) and H_0: $\mu \leq 14$ (versus H_1: $\mu > 14$) to demonstrate the impact of the value of μ in H_0 on β.
3. H_0: $\mu \geq 25$ (a false H_0) versus H_1: $\mu < 25$.

At this point it is also pointed out that in practice the significance test for the mean is not carried out on the probability distribution of the mean, but on the probability distribution of a function of the mean (i.e., the z or t score for the mean whose distribution is known when H_0 is true). Hence it is shown that the initial (unrealistic) assumption that the standard deviation σ is known does not create any difficulties in practice.

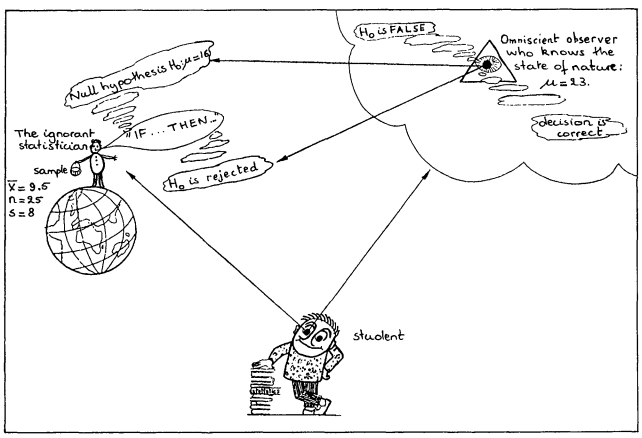

Figure 4. The Three Perspectives from which Hypothesis Testing can be Approached.

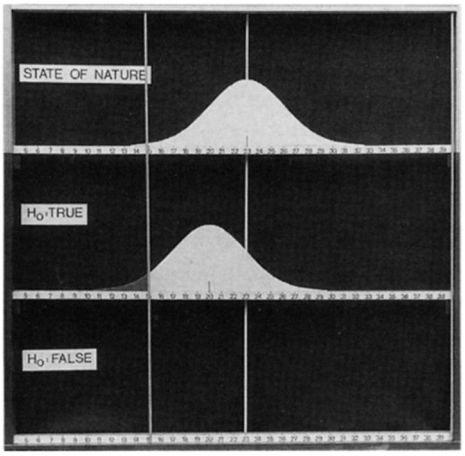

Figure 5. The Case where H_0: $\mu \geq 20$ is Tested versus H_1: $\mu < 20$. The state of nature is given by $\bar{X} \sim N(23, 3^2)$.

Finally, each student is provided with a personal mini-demonstrator in the shape of three pages of cardboard (DIN A4): (1) a green page whereupon three scales are printed representing the frame and the scales of the original demonstrator, (2) a white page with a dozen normal distributions (with $\sigma = 3$ and approximately 8.5 cm wide and 3 cm high) that are required to be cut out, and (3) an analogous yellow page whereupon normal distributions are printed with $\sigma = 1.6$ (approximately 4.5 cm wide and 5.5 cm high). Hence the students are able to experiment at home.

6. FURTHER APPLICATIONS

The demonstrator can also be used to show the impact of the sample size on the power of the test $(1-\beta)$. This can be done by using sample distributions with a smaller standard deviation. Figure 6 contrasts the cases where (for $\mu = 23$ in the state of nature) H_0: $\mu = 16$ is tested versus H_1: $\mu \neq 16$ for sample sizes of $n = 25$ (the clear distributions) and $n = 88$ (the shaded distributions). For $n = 25$, $1 - \beta = .35$, whereas for $n = 88$, $1 - \beta = .01$.

The demonstrator is also exceptionally well suited to introduce the concepts of operating characteristic curve and power curve in a concrete manner. This may be done by starting from the original numerical example. The distribution in the lower section remains fixed at $\mu = 16$. In the meantime the distribution in the upper section is progressively moved over the entire range of the scale. For several values of μ in the upper section the sum of the areas under the curve in the upper section (i.e., $1 - \beta$) falling "outside" the two red bars (which are positioned in the critical points of the distribution of the lower section) is computed. A graph of these sums against the corresponding values of μ in the upper section is plotted. It should be noticed that the general trend of the operating characteristic curve and the power curve can easily be sketched without effectively computing the probabilities. Indeed, the size of the relevant relative areas can roughly be estimated at sight.

The presence of the curtain enables the students to see that the statistician, in any case, can never know the power of his/her decision procedure. On the other hand, the student can also experience that the curtain does not prevent the statistician from plotting the power curve because this curve is merely the result of a reasoning process that considers the implications of "all possible" eventualities in the real world. Indeed, the power curve result from a conditional reasoning process ("If $\mu = \ldots$ in the state of nature, then $1 - \beta = \ldots$") that can be made concrete on the demonstrator by progressively moving the distribution in the upper section over the entire range of the x axis for some fixed distribution in the lower section.

Furthermore, the demonstrator can also be used to show how the requirements of high power and a small level of significance are conflicting, and how power can be increased without increasing the level of significance by increasing sample size. Finally, the concept of a p value (the probability of obtaining a value of the test statistic at least as extreme as the result observed, given that H_0 is true) can also be visualized by the demonstrator because the probability p is directly associated with an area under the curve in the middle section.

7. FINAL COMMENT

I have found that students taught with the demonstrator are enthusiastic about the subject and motivated by the challenging explorations that the device provides. More important, the students understand the basic concepts quite easily, and can successfully apply them to solve classic problems. The fact that the distributions are real objects that can be moved in any position and the students see at once the effects of the explorations on β and the power of the test $(1 - \beta)$, as well as the fact that all of these notions are introduced in a "tangible" manner by using the suggestive curtain and the movable vertical bars, seems to be of crucial importance to enhance understanding.

At present, there is a wide choice of valuable computer software available that illustrates the theory of hypothesis testing. However, in a first collective course of the basic principles the demonstrator presented here has several advantages over computer illustrations:

- The demonstrator is based on a special representational format that avoids most of the embarrassments induced by the computer software currently in use.

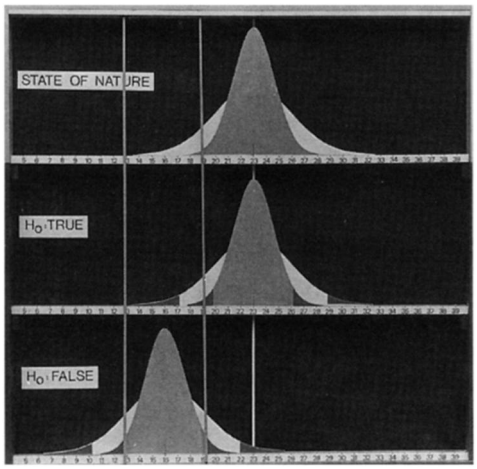

Figure 6. Impact of the Sample Size on the Probability of Making a Type II error.

- The demonstrator is flexible; one can easily concoct examples that clarify specific problems that arise in the course of the instruction and are not dealt with in standard texts.
- A demonstration with real objects is always more appealing and more suited to stimulate discussions than standardized computer illustrations on a monitor.
- The individual mini-demonstrators can be used without wasting time and attention in software problems at the expense of the comprehension of the subject.

Of course, it is always desirable to supplement the instructional strategy presented here further with computer explorations using graphical representation. The demonstrator is recommended primarily as an in-class interactive instructional aid prior to the use of other more advanced tools. I strongly urge interested teachers to build a demonstrator and experiment with it themselves to assess its effectiveness: the educational merits of the demonstrator can hardly be described; they must be experienced by using it. Perhaps readers who try it might comment on this issue in future letters.

[Received July 1994. Revised April 1996.]

REFERENCES

Carroll, D. W. (1985), *Psychology of Language*, Monterey, CA: Brooks-Cole.

Fisher, R. A. (1970), *Statistical Methods for Research Workers* (14th ed., revised and enlarged), New York: Hafner; Edinburgh: Oliver & Boyd.

——— (1935), *The Design of Experiments* (9th ed.), New York: Hafner; Edinburgh: Oliver & Boyd.

Lindley, D. V. (1965), *Introduction to Probability and Statistics from a Bayesian Viewpoint, Part 1: Probability; Part 2: Inference*, Cambridge: Cambridge University Press.

Neyman, J., and Pearson, E. S. (1928), "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference," *Biometrika*, 20A, 175–240, 263–294.

——— (1933), "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Philosophical Transactions of the Royal Society of London*, A 231, 289–337.

Wald, A. (1950), *Statistical Decision Functions*, New York: Wiley.