INFORMS Transactions on Education

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

An Intuitive Introduction to Hypothesis Testing

Ismael G. Dambolena, Steven E. Eriksen, David P. Kopcso,

To cite this article:
Ismael G. Dambolena, Steven E. Eriksen, David P. Kopcso, (2009) An Intuitive Introduction to Hypothesis Testing. INFORMS Transactions on Education 9(2):53-62. https://doi.org/10.1287/ited.1080.0019

Vol. 9, No. 2, January 2009, pp. 53–62
ISSN 1532-0545 | 09 | 0902 | 0053
INFORMS Transactions on Education
DOI 10.1287/ited.1080.0019

An Intuitive Introduction to Hypothesis Testing

Ismael G. Dambolena, Steven E. Eriksen, David P. Kopcso
Mathematics and Science Division, Babson College, Wellesley, Massachusetts 02457
{dambolena@babson.edu, eriksen@babson.edu, kopcso@babson.edu}

The traditional approach to teaching hypothesis testing, based on test statistics, is often perceived as lengthy and convoluted. This perception is of particular concern in business schools where the main focus of statistics education should be on providing practical decision-making tools to future managers. This paper discusses the results of a two-year experiment incorporating a more intuitive graph-based introduction to hypothesis testing that places the concept of *p*-value in a central role. Using this innovative approach at our institution we decreased class coverage time and improved students' understanding and retention with excellent results.

Key words: graphs; hypothesis testing; intuition; *p*-value; teaching statistics
History: Received: August 2007; accepted: August 2008. This paper was with the authors 5 months for 1 revision.

1. Introduction

A problem that continues to confront business school faculty teaching introductory statistics is the lack of understanding and appreciation of hypothesis testing by students. The introduction to hypothesis testing usually takes place early in the coverage of inferences. As such, this interaction sets a tone for the perceived usefulness of inferential statistics in general. We believe a clear explanation of hypothesis testing that appeals to intuition, thereby allowing the procedure to be internalized, is essential to promoting the benefits of decision making based on data.

In this paper we review the literature on intuitive approaches to hypothesis testing, and then describe the particular incident that piqued our curiosity about how we introduced hypothesis testing in our required undergraduate business statistics course. We then describe the traditional approaches, give a genealogy of when popular statistics textbooks started to emphasize the *p*-value approach, and trace hypothesis testing to early work by Laplace (Stigler 1986). The main contribution of this research is not the promotion of the *p*-value approach per se; some form of this approach has been used in the teaching of business statistics for many years. Instead, the main contribution is the delineation of a procedure whereby students arrive at an intuitive understanding of the strength of supporting evidence for an alternative hypothesis by examining data in a graphical format for several samples. Building on this visual image, students are then exposed to computer output that includes a numeric *p*-value for each of these samples. Finally, the relation between intuitive strength of evidence and *p*-values is graphically examined. We conclude our paper with strong statistical evidence of the improvement in student retention and understanding derived from this graphically reinforced approach to hypothesis testing. Based on these results we have continued to use this approach, and in this paper we supply sufficient detail to allow replication by any instructor who would like to try it.

2. Literature Review

Much discussion has taken place over the last 20 years on how to reform statistics education to enhance its effectiveness. Prominent in these efforts have been the Making Statistics More Effective in Schools of Business (MSMESB) conferences, which started in 1986. Love and Hildebrand (2002) give a brief history of these conferences and discuss their impact on courses, textbooks, and software. The Special Interest Group on Statistics Education of the Mathematical Association of America (SIGMAA Stat-Ed) has also been active in promoting improvements in the teaching of statistics and provides a forum for those interested in statistics education. Information on annual meetings and special sessions is available at http://www.pasles.com/sigmasastat.

Moore (1997, p. 123), who has published widely on the subject of statistical education, points out that it "is influenced by a movement to reform the teaching of mathematical sciences in general" and also

postulates that the "spirit of contemporary introductions to statistics should be very different from the traditional emphasis on lectures and on probability and inference." In his keynote address to a 2001 symposium on statistics education sponsored by the American Statistical Association, Moore (2001, p. 5) stated that teachers often imagine that they can gain by "presenting general principles or structures first, followed by special cases. This does not work. Few people learn from basic principles down to special cases. . . . Theory first in basic statistics is destined to fail—students have no idea what this is the theory of. . . . Statistics in practice has moved away from mathematics. So have the interests of students."

Based on a series of interviews with professional statisticians, Pfannkuch and Wild (2000, p. 132) observe that computer technology is enabling us to "downplay instruction in mechanical procedures and shift emphasis towards teaching the 'art' of statistics." Moore (2001, p. 5) postulates: "Drill only teaches drilling. Procedures and understanding are separate domains. Drill on procedures is not for this reason unimportant, but we should not be under the illusion that doing a procedure many times helps students understand it."

If one were to bring about reforms by first concentrating on concepts that are difficult to teach, hypothesis testing would be a prime starting point. The ideas of hypothesis testing are not easy to convey to students. A survey found that of a list of 30 core concepts in applied statistics, hypothesis testing and sampling distributions were considered the most difficult concepts to teach (McKenzie 2004). (This does not necessarily mean they are the most difficult to learn.) Informal discussions we have had with many colleagues also bear this out.

Intuition plays different roles in statistical thinking and statistics education. It is often used to enhance student understanding, sometimes in quite sophisticated ways. Franklin (1992) presents an exercise designed to help students better comprehend, through intuition and graphs, several multiple regression concepts with which they usually have difficulty. Many examples that apply to more elementary topics are given in Scheaffer et al. (1996). Several of them deal with hypothesis testing, and two among these ("Introduction to Hypothesis Testing" and "Coins on Edge") are focused on the development of key concepts. Other intuition-based exercises in this book are designed to illustrate situations where properly applied statistical inference produces better results than guesswork. An excellent example of this is "Random Rectangles," where the average area of a set of 100 rectangles is first judgmentally estimated by students and several of these estimates are recorded. Then several estimates are made using random sampling and these estimates are also recorded. Finally the data are graphed and the true average is produced, thus revealing a bias in the judgmental method.

Garfield (2005) includes a collection of fourteen articles in which instructors of innovative statistics courses describe examples of actual classroom practices. Although this book does not address the innovation we present in this article, it should be of interest to our readers.

Chatfield (1985) argues that intuition should play a heavier role in research. He points to a disturbing tendency: the use of high-level statistical techniques by people who do not fully understand them. He feels not only that more exploratory data analysis (EDA) is necessary, but also that EDA alone is often all that is required. He gives examples from the literature that illustrate his point and show that in many cases a more sophisticated analysis is incorrect because the required assumptions are not met.

Cobb and Moore (1997, pp. 815–816) amplify Chatfield's argument: "students like exploratory analysis and find that they can do it, a substantial bonus when teaching a subject feared by many. Engaging them early on in the interpretation of results, before the harder ideas come along . . . can help establish good habits that pay dividends when you get to inference."

It has long been accepted that graphical displays of data yield insights (see Tufte 1983 and Wainer 2005). In the preface of their text Moen et al. (1991, p. 5) describe one of its distinctive attributes as "the almost exclusive use of graphical methods for analysis of data from experiments." As De Veaux and Hand (2005, p. 236) point out, the use of "simple plots such as bar charts, histograms, scatterplots and time series plots can be invaluable . . . since the human eye has evolved to select anomalies . . . ." Martin (2003) suggests that analogies, although often not perfect, also appeal to students' intuition.

Various student-based activities, often using coins and dice (for example, see Gelman and Nolan 2002), have been used to teach hypothesis testing in introductory statistics courses. Eckert (1994) presents a demonstration using playing cards that has been useful in teaching basic concepts of hypothesis testing such as the formulation of a null hypothesis and using data to determine the strength of the evidence against the null hypothesis.

Lane-Getaz (2005) provides a compilation of 13 p-value misconceptions documented in empirical studies. She points out that some of these misunderstandings and misinterpretations appear not only to be common among statistics students but that they also occur among experienced researchers.

We believe that our approach is novel and useful. Furthermore, it is in agreement with the recommendations for the teaching of statistics provided in the Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report (American Statistical Association 2005). The GAISE project was funded by the American Statistical Association and the report was written by two groups of highly regarded teachers of statistics. One team considered elementary through high school statistics education and the other studied the current state of college statistical pedagogy. The report outlines the current state of statistics education and recommends future actions to make the teaching of statistics more effective.

3. The Symptom

Our required undergraduate statistics course, like most business schools, includes an introduction to hypothesis testing that is focused on z-tests and t-tests of a population mean. In our case this introduction extends over at least three 75-minute class periods. Below we discuss the results of presentations of this material in two consecutive years. In the first year one of us used the traditional test-statistic/rejection-region approach in each of two sections of the course. The computation and interpretation of p-values, examples contrasting one-sided and two-sided tests, and the use of t-tests followed. A two-sample test for the difference between means, a test for normality, the chi-square test of independence, and the t-test for the slope in simple regression were also covered as part of the course. Figure 1 shows Question 9 in the final exam. This question was worth four points. The combined results for the two sections are provided in Table 1. Twenty-nine of the 36 students who earned no points answered "yes" and explained that the sample mean (28.66 minutes) was less than the hypothesized value of μ (30 minutes) without reference to the p-value or variability. The remaining seven students made other serious conceptual mistakes.

Figure 1 Question 9 of the Final Exam

The director of quality of a large health maintenance organization wants to evaluate waiting time at a local facility. A random sample of 30 patients was selected from the appointment book and their waiting times in minutes were recorded. The results of a Minitab test using these data are shown below. Is there evidence that the average patient waiting time at the local HMO facility is less than 30 minutes?

Answer: —— Yes —— No (check one) Briefly explain your answer:

One-Sample T: Time

Test of mu = 30.00 vs mu < 30.00

Variable	N	Mean	Stdev	SE mean	T	P
Time	30	28.66	13.09	2.39	−0.56	0.29

Note. This question was included in the final examination to assess the students' understanding of the concept of statistical significance as measured using a p-value.

Table 1 Frequency Distribution of Points Earned by Students on Question 9

	Points					
	0	1	2	3	4	Total
Frequency	36	2	4	0	19	61

Note. Four was the highest score and zero the lowest.

We were very surprised by these results. A comprehensive examination one month after coverage of this relatively complex topic seemed to expose severe problems with understanding or retention of fundamental hypothesis testing concepts. We therefore decided to develop class materials to find out whether an innovative approach grounded on intuition and using graphs might help students better understand key hypothesis testing concepts and retain them longer.

4. Traditional Approaches

4.1. The Test-Statistic/Rejection-Region Approach

In their introduction to hypothesis testing many business statistics texts (for example, see Groebner et al. 2005, Newbold et al. 2007, Weiers 2006) start with a discussion of the basic ideas and terminology that includes concepts such as the null and alternative hypotheses, one-sided and two-sided tests, Type I and Type II errors, and levels of significance. They then describe and apply a structured procedure for performing a test using the test-statistic/rejection-region (TS/RR) approach. This procedure consists of the following steps (with minor variations between texts):

1. Formulate the null and alternative hypotheses.
2. Specify the level of significance and select the sample size.
3. Select the test statistic to be used and determine its distribution under the assumption that the null hypothesis is true.
4. Collect the sample and compute the value of the test statistic.
5. Determine the critical value(s) and specify the rejection region.
6. Reject the null hypothesis if the test statistic falls in the rejection region.
7. Interpret this decision in the context of the problem.

4.2. The Test-Statistic/p-Value Approach

After demonstrating the TS/RR procedure these texts discuss the computation and use of p-values as an

alternative approach. In the last few years some texts (Anderson et al. 2005, Berenson et al. 2006) have pointed out that, after the fourth step in the above procedure, one has a choice between using the TS/RR approach or the test-statistic/p-value (TS/PV) approach. Using the TS/PV approach only steps 5 and 6 of the procedure change:

5. Use the value of the test statistic to compute the p-value.

6. Reject the null hypothesis if the p-value is less than or equal to the level of significance.

The TS/RR approach, nevertheless, still tends to be emphasized. In this paper we will refer to TS/RR and TS/PV as "traditional approaches."

4.3. Shortcomings of the Traditional Approaches

We feel the traditional approaches have shortcomings when used in the introduction to hypothesis testing for students without a strong mathematical inclination. These approaches are lengthy, and although well founded in logic, are not intuitive to many students and have negative consequences, including the following:

- Students have a tendency simply to compare the sample mean to the hypothesized value of μ for a one-tailed test.
- Students do not understand the basic ideas of hypothesis testing or cannot retain their learning.
- Students cannot make sound managerial decisions based on their conclusions.
- Students cannot apply the testing approach to new contexts.

These shortcomings are of particular concern in business schools, where the main focus of statistics education should, in our view, be to provide practical decision-making tools to future managers. We have found that we can substantially mitigate these shortcomings and thus better serve our students with a more intuitive, graph-based approach using p-values. We hope that other teachers will benefit from this approach.

5. A Brief History of P-Values

The use of p-values in business statistics education started moving into the mainstream only about 25 years ago with the advent of widespread personal computing. In an article that lists first printed occurrences of terms used in mathematical statistics, David (1995) points out that while the *concept* of p-value goes back at least to Pierre Simon Laplace (for details see the next paragraph), the *term* is fairly new. In tracing the term back to Brownlee (1960), David (1995, p. 122) states that "numerous alternative terms have been used and to some extent are still being used: probability level, sample level of significance, observed significance level, significance probability, descriptive level of significance, critical level, significance level, probvalue, and associated probability." He also points out that Pearson (1900) already used P in this context. In a subsequent article (David 1998), he traces the term back to Deming (1943). More than a generation later there still seemed to be some ambiguity about the terminology: Freedman et al. (1978, p. 442) refer to the "observed significance level" of a test (a term that is still used on occasion) and write that "it is usually denoted P, for probability, and it is often called the P-value of the test." In their text for Continental Classroom, the first television course in statistics aired nationally on NBC, Mosteller et al. (1961, p. 304) referred to the p-value as the "descriptive level of significance." On the same page they also used the term "significance level" for the probability of a Type I error. An article by Gibbons and Pratt (1975), listed as the reference on p-values by *The Oxford Dictionary of Statistical Terms* (Dodge 2003), provides a comprehensive discussion. Dallal (2000) states that "as computers became readily available, it became common practice to report the observed significance level (or P value)."

Stigler (1986) describes in detail how Laplace, whose work linked the ocean tides to the moon's gravitation, tried to establish in 1823 whether the moon might also have an effect on the earth's atmospheric pressure. Based on historical data (three daily barometric readings over eight years), he tested the hypothesis to no effect. He determined that under this hypothesis, by chance alone the probability of a sample result being no more supportive of the alternative than the one he obtained was only 0.843, and felt that this probability was not large enough to establish that the moon's gravitation has an influence on atmospheric pressure. Moreover, his comments imply that he would have considered the results significant only if the observed probability had been in excess of 0.99. In today's terminology Laplace found a p-value of 0.157 and, as a consequence, dismissed the results as not statistically significant because he had selected a level of significance of 0.01.

6. The Intuitive Graph-Based Approach Using P-Values

One year after those surprising results from Question 9 of the final exam, the same instructor delivered the material on hypothesis testing using a graph-based/p-value (GB/PV) approach designed to appeal to a student's intuition. This approach consists of two components: a discussion that establishes the link between intuition and p-values, and a procedure for testing that is akin to the procedures for the traditional approaches.

The GB/PV approach has now been used several times and has undergone a few minor changes. Question 9 was included in the final exam every year the new approach was used (both in our undergraduate program and in the required introductory statistics course for our MBA students) and the results have always been much better than with the old approach. Furthermore, colleagues who did not use the new approach have included Question 9 in their final exams with results similar to those we obtained with the old approach.

6.1. Discussion

Immediately after briefly introducing the basic concepts and terminology of hypothesis testing, approximately 45 minutes are spent heavily involving the class in an interactive discussion that is key to our approach. The goal is to provide students with an intuitive understanding of the relationship between *p*-values and the degree of support for the alternative hypothesis of a test. This discussion takes place once and consists of the following steps (explained in detail later in this section):

(a) Formulate the null and alternative hypotheses.

(b) Show several samples with their dotplots and ask students to intuitively assess support of the alternative for each sample.

(c) Use software to produce *p*-values for each sample.

(d) Graphically link *p*-values to support for the alternative.

(e) Discuss these links and summarize them by using a rule of thumb.

6.2. Procedure for Testing

Once an intuitive understanding of the concept of a *p*-value is gained from the discussion involving steps A to E above, it is easy to use *p*-values to conduct tests in a variety of hypothesis testing situations. A structured procedure for performing a test using the GB/PV approach is similar to that of the traditional approaches but uses computer output and does not include manual computation of a test statistic. It could be a variation of the following:

1. Formulate the null and alternative hypotheses.
2. Specify the level of significance and select the sample size.
3. Collect the sample, develop a graphical or tabular display (see examples in Table 2), and intuitively make an assessment of the support for H_1.
4. Obtain the *p*-value from statistical software.
5. Reject the null hypothesis if the *p*-value does not exceed the level of significance.
6. Interpret this decision in the context of the problem.

Table 2 Some Appropriate Graphic and Tabular Displays of Data

Hypothesis test	Graphic or tabular display
Mean	Dotplot
Difference of means or analysis of variance	Grouped box plots
Chi-square test of independence	Cross-tabulation table
Slope in regression	Scatterplot

This procedure is used whenever a test is performed, and as noted the *p*-value is obtained from software. Valuable class time can be spent on the managerial implications of the statistical decision that has been so commonly reduced to a simple rule such as *if the p-value is less than or equal to alpha, reject* H_0.

6.3. Detailed Explanation of the Discussion

The class presentation was as follows: First, as had been done in the prior year with the TS/RR approach, basic concepts and terminology were briefly introduced, emphasizing that using sample evidence one must make a choice between the null Hypothesis (H_0) and the alternative Hypothesis (H_1). Also as before, we stressed that one lets H_0 stand unless the sample evidence clearly supports H_1. The introductory example for testing about μ, adapted from Anderson et al. (2005, p. 349), also remained the same:

> The United States Golf Association (USGA) stipulates that golf balls of any given brand, when tested at a special machine in the USGA headquarters, must not cover an average distance in carry and roll exceeding 280 yards. Manufacturers, therefore, want their balls to cover an average distance of 280 yards. On the one hand the USGA doesn't allow the average ball flight to be in excess of 280 yards, and on the other hand, for marketing reasons, manufacturers don't want the average ball flight to be less than 280 yards. We know from historical data that the distance covered by golf balls tested on the USGA apparatus is normally distributed. We will run tests to help manufacturers determine whether or not their balls cover 280 yards.

After agreeing that they would be testing H_0: $\mu = 280$ versus H_1: $\mu \neq 280$, students were shown samples from five different manufacturers and their respective dotplots (see Figure 2), and they were told that each of the five samples would be used for a separate test. They were also told that H_0 and H_1 would remain the same throughout the five tests, and that for each of the five tests their job was to intuitively assess how much evidence that sample provides in support of H_1. The possible choices were "none," "little," "quite a bit," and "a lot." These phrases were intentionally undefined. The purpose of this exercise was to help introduce the concept of hypothesis testing to business students in a manner that would help

them acquire an intuitive understanding of the underlying concepts and, consequently, be able to apply the concepts to future situations. The close association of the semantics familiar to business-school students with the unfamiliar computation of the *p*-value was meant to have a lasting impression on the students. As stressed in the introduction to the basic terminology and concepts of hypothesis testing, one lets H_0 stand unless the sample evidence clearly supports H_1, so "none" and "little" would connote a lack of support for the alternative whereas "quite a bit" and "a lot" would connote support. For each sample, after students voiced their opinions, it was fairly easy to get a consensus. They agreed that sample 1 in Figure 2 provided no evidence for H_1, samples 2 and 3 provided a lot of evidence, sample 4 provided quite a bit of evidence, and sample 5 provided little evidence. A record was kept of these intuitive assessments. An associate editor and referee of an earlier version of this paper suggested replacing "none," "little," "quite a bit," and "a lot." We have done so and have had good results. Consequently, in the future we will use "none," "not much," "a fair amount," and "a lot" as possible responses in our classes.

The statistical software package used in this course is Minitab. Students were shown how to perform a one-sample *t*-test for the mean using the first sample. It was noted that the Minitab output provides a *p*-value, a simple and effective way to test hypotheses. Test results for the five samples were given to the students and are included in Figure 3.

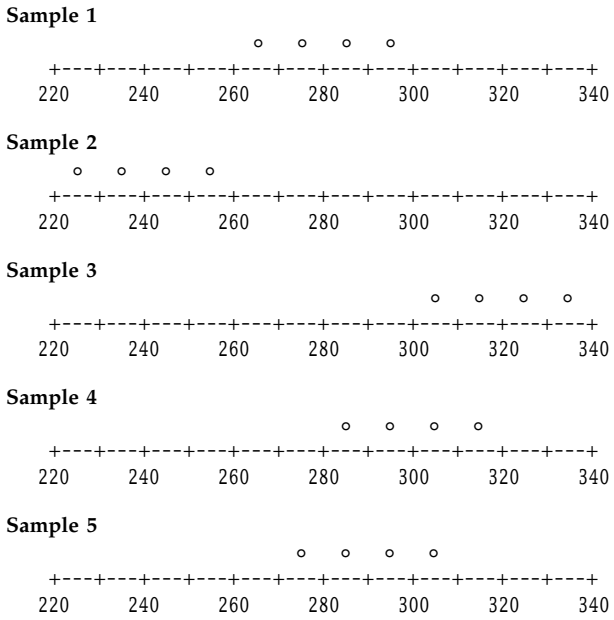
By this point considerable information had been collected on *p*-values as well as assessments of support for H_1 for several samples. The next question was whether one could come up with a way to put all this information together and make it meaningful. Could they somehow display this information graphically?

Eventually the graph in Figure 4 was developed. Here students could discern a clear relation between *p*-values and support for H_1, and they could also see that

- As the *p*-values decrease, the support for H_1 increases.
- If there is no support for H_1, then the *p*-value is one.
- If there is a lot of support for H_1, then the *p*-value is near zero.
- Last, and most important, the *p*-value has to be quite small in order for the support for H_1 to be substantial.

Figure 2 Five Samples and Their Dotplots

Sample number	Sample values			
1	265	275	285	295
2	225	235	245	255
3	305	315	325	335
4	285	295	305	315
5	275	285	295	305



Notes. The dotplots represent distances in yards covered by golf balls when tested at a special machine in the USGA headquarters. For each sample students must intuitively assess how much support it provides for the alternative hypothesis H_1: $\mu \neq 280$.

Figure 3 Minitab Tests of H_0: $\mu = 280$ vs. H_1: $\mu \neq 280$ Using the Five Samples in Figure 2

T-Test of the Mean
Test of mu = 280.00 vs mu not = 280.00

Variable	N	Mean	Stdev	SE mean	T	P
Sample1	4	280.00	12.91	6.45	0.00	1.000

T-Test of the Mean
Test of mu = 280.00 vs mu not = 280.00

Variable	N	Mean	Stdev	SE mean	T	P
Sample2	4	240.00	12.91	6.45	−6.20	0.008

T-Test of the Mean
Test of mu = 280.00 vs mu not = 280.00

Variable	N	Mean	Stdev	SE mean	T	P
Sample3	4	320.00	12.91	6.45	6.20	0.008

T-Test of the Mean
Test of mu = 280.00 vs mu not = 280.00

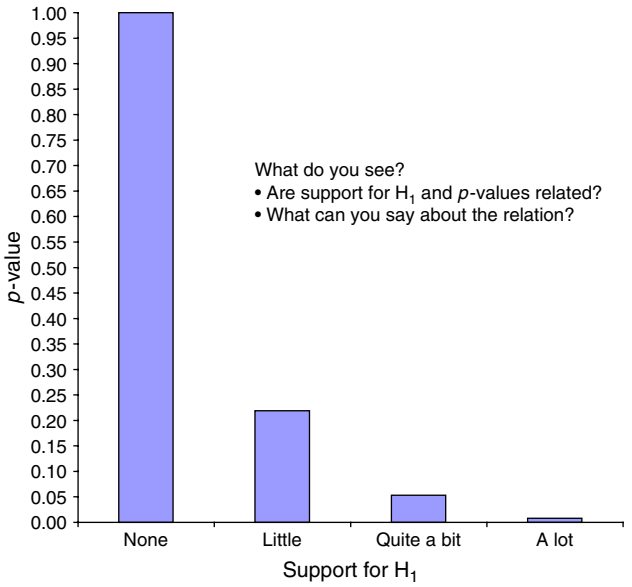
Variable	N	Mean	Stdev	SE mean	T	P
Sample4	4	300.00	12.91	6.45	3.10	0.053

T-Test of the Mean
Test of mu = 280.00 vs mu not = 280.00

Variable	N	Mean	Stdev	SE mean	T	P
Sample5	4	290.00	12.91	6.45	1.55	0.219

Note. We are interested in detecting a relationship, if it exists, between the sample *p*-value and the students' intuitive assessment of the support that the sample provides for the alternative hypothesis of the test.

Figure 4 **Bar Chart Demonstrating the Relation Between *p*-Values and Support for H_1**



Note. As the *p*-values decrease the support for H_1 increases, if there is no support for H_1 then the *p*-value is one, and if there is a lot of support for H_1 then the *p*-value is near zero. The chart also indicates that the *p*-value has to be quite small in order for the support for H_1 to be substantial.

This concept, based on their intuition, contradicts the natural inclination of many students to simply compare the sample mean to the hypothesized value of μ for a one-sided test, something we want them to avoid. We believe this conceptual mistake is the most common reason that about half of the prior year's students incorrectly answered "yes" on Question 9 of the final exam and explained that the sample mean was less than the hypothesized population mean.

These developments were followed by a discussion of *p*-values. It was explained that one may think of the *p*-value for a test as the probability of getting sample results as supportive of H_1 as those observed or even more so given that the null hypothesis is true. A formal definition based on test statistics, which at this point had not yet been discussed, was introduced later. Students were also told that *p*-values are routinely reported by statistical software, and that they could use the framework in Table 3 as a reasonable rule of thumb relating *p*-values to evidence in support of H_1.

Table 3 **Suggested Reasonable Guidelines for Converting a *p*-Value into an Inference Regarding the Alternative Hypothesis**

p-value ranges	Interpretation
p-value ≤ 0.01	Very strong evidence for H_1
$0.01 < p$-value ≤ 0.05	Rather strong evidence for H_1
$0.05 < p$-value ≤ 0.10	Rather weak evidence for H_1
p-value > 0.10	Weak or no evidence for H_1

This framework has obvious limitations. Is there a substantial difference between $p = 0.099$ and $p = 0.101$, or between $p = 0.049$ and $p = 0.051$? Furthermore, one could argue that from a practical perspective this representation may be misconstrued. For example, a *p*-value of 0.25 implies that if the null hypothesis is true the probability of obtaining a result at least as extreme as in the current sample when using an equivalent sampling process with the same population is 0.25. This result implies that the null hypothesis is plausible, but it does not imply that the alternative hypothesis is implausible.

However, because a required introductory business course is designed to teach the students to be good consumers of statistics and not practicing statisticians, we have adopted this simple set of rules in our classes. We tell our students they can use these rules when a problem statement does not specify a significance level and they wish to have a guideline for establishing their own significance level value. The choice of this set of guidelines is consistent with many introductory business statistics texts such as Moore et al. (2009) and Albright et al. (2006). These recommendations also address the concern raised by Gauvreau and Pagano (1994) that what is already a context-based rare event may not necessitate the prior specification of an alpha cutoff value.

In the discussion of levels of significance, we tell our students that the decision maker, prior to sampling, should select a value for alpha. Traditionally alpha is set at 0.01, 0.05, or 0.10, but there are reasons that the decision maker may select a different value. We explain to our students that for the purpose of our class, and hence in exams, unless one explicitly specifies an alpha greater than 0.10, *p*-values greater than 0.10 indicate that there is not statistically significant support for the alternative hypothesis.

7. Results

The identical Question 9 was included in the final examination the following year. A comparison of the new results with those of the first year appears in Table 4. The mean on Question 9 was 1.41 in the first year and 3.34 in the following year.

It is immediately clear from perusal of Table 4, as well as a comparison of the means, that there was a

Table 4 **Frequency Distributions for the Results on Question 9 for the Initial Year (as Shown in Table 1) and the Following Year**

	Points					
	0	1	2	3	4	Total
Frequency						
First year	36	2	4	0	19	61
Following year	8	1	2	0	48	59

substantial improvement in performance in the second year. For confirmation, we conducted a standard parametric one-tailed test for the difference between means. The p-value (on the order of 10^{-9}) indicated, as expected, strong support for the alternative hypothesis that there was an improvement in the mean score for Question 9 using the GB/PV method.

While the above results are convincing, an additional statistical test was performed to address the possibility that the test for the difference between means is inappropriate with this data because of a violation of the normality assumption. A one-tailed test for the difference between two proportions (the proportion of fours in the two samples) was significant with a p-value of the same order of magnitude as the above test for the difference between means.

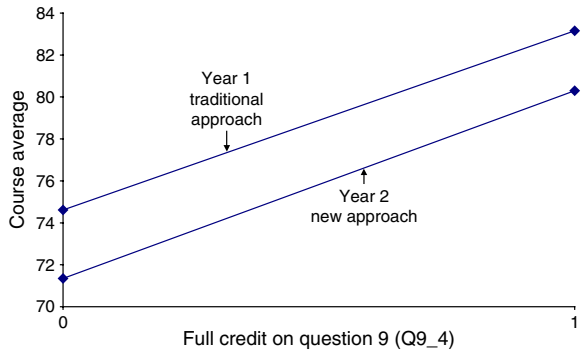
One might ask whether the observed improvement in performance on Question 9 during the following year could stem from factors other than exposure to the new GB/PV approach, such as a superior group of students. We examined this possibility in two ways. First, Figure 5 shows an interaction plot of the overall course average against an indicator variable, Q9_4, whose value is one if the student received full credit (all four points) on Question 9 and zero otherwise for each of the two years for which we collected data. The near parallelism of the lines shows that there is no interaction between course average and whether the teaching was done using the traditional approach or the new approach. Moreover, as the figure shows, overall class averages for the second year were lower than those for the first year.

For confirmatory purposes the binary logistic regression model in Figure 6 was built using as its response the indicator variable Q9_4. Selected as explanatory variables were the overall numeric course average (TermGrade); an indicator variable that took the value one if the student was exposed to the GP/PV approach and zero otherwise (GP/PV_Ind); and Interaction, the interaction of the course average with the GP/PV indicator. As the results in Figure 6 show, the p-values for TermGrade, the GP/PV indicator, and the interaction variable in this regression were 0.010, 0.662, and 0.313, respectively. Consequently, there was no significant effect of the interaction of the course average with the GP/PV indicator of performance on Question 9 when controlling for course average and for whether the GP/PV approach was used.

One might also ask whether students who took the final exam the following year could have seen Question 9 in the first year's final exam. The likelihood of this occurrence is negligible. Rules in our school specify that instructors must not return final exams but students may come to review them (although they do this very infrequently).

Note that we have continued to include Question 9 in final exams for sections taught using the GP/PV approach, and have also asked colleagues who do not use this approach to include Question 9 in their exams. Results have consistently shown a much better performance on Question 9 under the new approach than under the old approach.

Figure 5 **Interaction Plot of Overall Course Average Against Performance on Question 9**



Notes. The plotted lines are for the first year (Year 1, when hypothesis testing was taught using the traditional approach) and the following year (Year 2, when the new intuitive approach was used). Because the lines are essentially parallel, the graph shows no interaction between course average and teaching method. In fact the overall class averages for Year 1 were consistently lower than those for Year 2.

8. Conclusions

As business school faculty who teach statistics, we believe that our role is not to prepare a new generation of practicing statisticians. Our function, instead, is to provide practical decision-making tools to future managers.

We conclude that the traditional procedures for teaching hypothesis testing to our audience are ineffective. The advancements of data-analysis technology, which we have brought into the classroom over the last two decades, have presented us with a wonderful set of pedagogical opportunities. We believe that a combination of visual and graphical data presentations, along with the p-value approach to hypothesis testing, removes much of the "smoke and mirrors" attitude that business students attribute to statistical reasoning.

We suggest an approach to hypothesis testing in which the first step is to introduce a problem statement and then look at the data in graphical or tabular form. At this point the students make a preliminary decision based on what they observe in the histogram, boxplot, scatter plot, etc. The students should then use the software to generate the p-value and make a decision. This decision should not be as abstract as "Reject H_0" or "Do not reject H_0," but should answer the question "What do you conclude?" The final and

Figure 6 Binary Logistic Regression Model in Which the Response Is a Binary Variable that Was Assigned the Value One if the Student Earned a 4 on Question 9 and Zero Otherwise

Binary Logistic Regression: Q9_4 versus TermGrade, GB/PV_Ind, Interaction

Link Function: Logit

Response Information

Variable	Value	Count	
Q9_4	1	67	(Event)
	0	53	
	Total	120	

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Constant	−7.03496	2.47837	−2.84	0.005			
TermGrade	0.0789464	0.0305216	2.59	0.010	1.08	1.02	1.15
GB/PV_Ind	−1.93771	4.42706	−0.44	0.662	0.14	0.00	844.98
Interaction	0.0583670	0.0578361	1.01	0.313	1.06	0.95	1.19

Log-Likelihood = −57.015

Test that all slopes are zero: G = 50.688, DF = 3, P-Value = 0.000

Goodness-of-Fit Tests

Method	Chi-Square	DF	P
Pearson	98.577	104	0.632
Deviance	101.893	104	0.540
Hosmer-Lemeshow	7.615	8	0.472

Table of Observed and Expected Frequencies: (See Hosmer-Lemeshow Test for the Pearson Chi-Square Statistic)

		Group									
Value		1	2	3	4	5	6	7	8	9	10	Total
1												
	Obs	1	3	2	5	6	7	11	8	12	12	67
	Exp	1.3	2.3	3.2	4.6	5.9	7.4	9.5	10.3	11.0	11.6	
0												
	Obs	11	9	10	7	6	5	1	4	0	0	53
	Exp	10.7	9.7	8.8	7.4	6.1	4.6	2.5	1.7	1.0	0.4	
Total		12	12	12	12	12	12	12	12	12	12	120

Measures of Association: (Between the Response Variable and Predicted Probabilities)

Pairs	Number	Percent	Summary Measures	
Concordant	3011	84.8	Somers' D	0.70
Discordant	531	15.0	Goodman-Kruskal Gamma	0.70
Ties	9	0.3	Kendall's Tau-a	0.35
Total	3551	100.0		

Note. Explanatory variables are overall numeric course average (TermGrade), a GP/PV indicator variable (one if the student was exposed to the GP/PV approach and zero otherwise), and Interaction, the interaction of the course average with the GP/PV indicator.

most important step is for the students to interpret, in the context of the problem, what the hypothesis test result suggests about the situation at hand.

Acknowledgments

The authors thank Professor John D. McKenzie, Jr., of Babson College and an anonymous referee for their helpful comments and contributions.

References

Albright, S. C., W. L. Winston, C. J. Zappe. 2006. *Data Analysis and Decision Making*, 3rd ed. Thomson South-Western, Mason, OH.

American Statistical Association. 2005. The guidelines for assessment and instruction in statistical education (GAISE) college reports. American Statistical Association, Alexandria, VA. http://www.amstat.org/education/gaisel.

Anderson, D. R., D. J. Sweeney, T. A. Williams. 2005. *Statistics for Business and Economics*, 9th ed. Thomson South-Western, Mason, OH.

Berenson, M. L., D. M. Levine, T. C. Krehbiel. 2006. *Basic Business Statistics: Concepts and Applications*, 10th ed. Prentice Hall, Upper Saddle River, NJ.

Brownlee, K. A. 1960. *Statistical Theory and Methodology in Science and Engineering*. John Wiley, New York.

Chatfield, C. 1985. The initial examination of data. *J. Roy. Statist. Soc. Ser. A* (*General*) **148**(3) 214–253.

Cobb, G. W., D. S. Moore. 1997. Mathematics, statistics and teaching. *Amer. Math. Monthly* **104**(9) 801–823.

Dallal, G. E. 2000. *P* values. http://www.jerrydallal.com/LHSP/pval.htm. Last accessed on September 7, 2008.

David, H. A. 1995. First (?) occurrence of common terms in mathematical statistics. *Amer. Statistician* **49**(2) 121–133.

David, H. A. 1998. First (?) occurrence of common terms in mathematical statistics—A second list, with corrections. *Amer. Statistician* **52**(1) 36–40.

Deming, W. E. 1943. *Statistical Adjustment of Data*. John Wiley, New York.

De Veaux, R. D., D. J. Hand. 2005. How to lie with bad data. *Statist. Sci.* **20**(3) 231–238.

Dodge, Y., ed. 2003. *The Oxford Dictionary of Statistical Terms*. Oxford University Press, Oxford.

Eckert, S. 1994. Teaching hypothesis testing with playing cards: A demonstration. *J. Statist. Ed.* **2**(1). Available online at http://www.amstat.org/PUBLICATIONS/JSE/v2n1/eckert.html.

Franklin, L. A. 1992. Graphical insight into multiple regression concepts. *Amer. Statistician* **46**(4) 284–288.

Freedman, D., R. Pisani, R. Purvis. 1978. *Statistics*. W. W. Norton, New York.

Garfield, J., ed. 2005. *Innovations in Teaching Statistics*. The Mathematical Association of America, Washington, D.C.

Gauvreau, K., M. Pagano. 1994. Why 5%? *Nutrition* **10**(1) 93–94.

Gelman, A., D. Nolan. 2002. You can load a die but you can't bias a coin. *Amer. Statistician* **56**(4) 308–311.

Gibbons, J. D., J. W. Pratt. 1975. *P*-values: Interpretation and methodology. *Amer. Statistician* **29**(1) 20–25.

Groebner, D. F., P. W. Shannon, P. C. Fry, K. D. Smith. 2005. *Business Statistics: A Decision Making Approach*, 6th ed. Prentice Hall, Upper Saddle River, NJ.

Lane-Getaz, S. J. 2005. Summary of *p*-value survey research. *United States Conf. Teaching Statist.* (*USCOTS*). Ohio State University, Columbus, 13–17.

Love, T. E., D. K. Hildebrand. 2002. Statistics education and the Making Statistics More Effective in Schools of Business conferences. *Amer. Statistician* **56**(2) 107–112.

Martin, M. A. 2003. "It's like…you know": The use of analogies and heuristics in teaching introductory statistical methods. *J. Statist. Ed.* **11**(2). http://www.amstat.org/publications/jse.

McKenzie, J. D. 2004. Conveying the core concepts. *Amer. Statist. Assoc. Proc. Joint Statist. Meetings*, 2755–2757.

Moen, R. D., T. W. Nolan, L. P. Provost. 1991. *Improving Quality Through Planned Experimentation*. McGraw-Hill, New York.

Moore, D. S. 1997. New pedagogy and new content: The case of statistics. *Internat. Statist. Rev.* **65**(2) 123–165.

Moore, D. S. 2001. Undergraduate programs and the future of academic statistics. *Amer. Statistician* **55**(1) 1–6.

Moore, D. S., G. P. McCabe, W. M. Duckworth, S. L. Sclove. 2009. *The Practice of Business Statistics: Using Data for Decisions*. W. H. Freeman, New York.

Mosteller, F., R. E. K. Rourke, G. B. Thomas. 1961. *Probability and Statistics*. Addison-Wesley, Reading, MA.

Newbold, P., W. L. Carlson, B. Thorne. 2007. *Statistics for Business and Economics*, 6th ed. Prentice-Hall, Upper Saddle River, NJ.

Pearson, K. 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philos. Magazine, 5th Ser.* **50** 157–175.

Pfannkuch, M., C. J. Wild. 2000. Statistical thinking and statistical practice: Themes gleaned from professional statisticians. *Statist. Sci.* **15**(2) 132–152.

Scheaffer, R. L., M. Gnanadesikan, A. Watkins, J. A. Witmer. 1996. *Activity-Based Statistics: Instructor Resources*. Springer, New York.

Stigler, S. M. 1986. *The History of Statistics: The Measurement of Uncertainty Before 1900*. The Belknap Press of Harvard University Press, Cambridge, MA.

Tufte, E. R. 1983. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT.

Wainer. H. 2005. *A Trout in the Milk and Other Visual Adventures*. Princeton University Press, Princeton, NJ.

Weiers, R. M. 2006. *Introduction to Business Statistics*, 5th ed. Duxbury, Belmont, CA.