



INFORMS Transactions on Education

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

Identifying Addressable Impediments to Student Learning in an Introductory Statistics Course

Scott P. Stevens, Susan W. Palocsay,



To cite this article:

Scott P. Stevens, Susan W. Palocsay, (2012) Identifying Addressable Impediments to Student Learning in an Introductory Statistics Course. INFORMS Transactions on Education 12(3):124-139. <https://doi.org/10.1287/ited.1120.0085>

Full terms and conditions of use: <http://pubsonline.informs.org/page/terms-and-conditions>

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

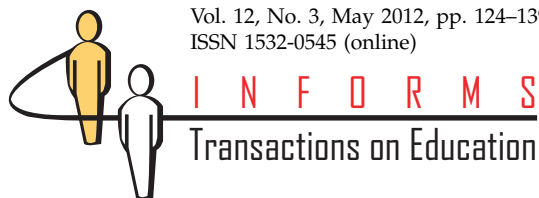
Copyright © 2012, INFORMS

Please scroll down for article—it is on subsequent pages



INFORMS is the largest professional society in the world for professionals in the fields of operations research, management science, and analytics.

For more information on INFORMS, its publications, membership, or meetings visit <http://www.informs.org>



Identifying Addressable Impediments to Student Learning in an Introductory Statistics Course

Scott P. Stevens, Susan W. Palocsay

Computer Information Systems and Management Science Department, James Madison University, Harrisonburg, Virginia, 22807
{stevensp@jmu.edu, palocssw@jmu.edu}

Using performance data on a common test instrument administered to more than 4,000 business statistics students studying under 16 different instructors over a period of 6.5 years, we identify areas of consistent student weakness. In addition to verifying difficulties found by earlier researchers, we document a fundamental problem in students' ability to reason with cumulative probabilities, a problem with implications for solving problems with both discrete and continuous distributions. Our performance data also suggest that it may be useful to view statistics problems using a taxonomy based more on solution procedure than on traditional statistics topics. We employ a new methodology based on cluster analysis to identify areas where the observed student difficulties may be particularly addressable. Finally, we offer some preliminary findings on how these problems may be addressed with practical teaching suggestions based on the approaches employed by our more successful instructors.

Key words: statistics education; assessment; statistical reasoning; problem-solving skills

History: Received: September 2010; accepted: January 2012

1. Introduction

The importance of *introductory statistics*, defined as the "non-calculus based, often terminal, introductory applied statistics course" (Garfield 2000, p. 2), is recognized by its required status for students pursuing studies in a wide range of disciplines. Concern about the effectiveness of teaching and learning of statistics at the introductory level has been a major focus of educational research and conferences for more than two decades. This concern is reflected in discussions of the status of and need for improvements in statistical education appearing throughout the literature, from the early 1970s (Brightman 1977, Brightman and Broida 1975, Hogg 1972) to the present (see references in Garfield et al. 2002b, Hall and Rowell 2008, Harrington and Schibik 2004, Holmes 2003). Zieffler et al. (2008) review a sample of published studies related to teaching and learning statistics at the introductory college level. Business statistics has received special attention since 1986 in the Making Statistics More Effective in Schools of Business conferences (Love and Hildebrand 2002, Ord 2010). Recently, Clayton and Sankar (2009) linked spreadsheet use to improved attitude and motivation of students in an introductory business statistics course.

If we are to improve statistics education, a number of questions naturally springs to mind. What are the common problems that students encounter? How far-reaching are their consequences? Can these problems be effectively addressed by changes in the pedagogy? And if so, what kinds of changes appear to be effective? This paper is intended to provide some additional insight into the answers to these questions.

As background, we briefly review the literature on assessment for statistics education and research into student misconceptions of statistics and probability. Then we describe our assessment instrument, a 63-question test administered to more than 4,000 students of introductory business statistics, and analyze the results to identify areas of common student difficulty. Our findings corroborate those of earlier researchers, and we identify a fundamental and previously unidentified weakness in student performance on problems concerning cumulative distributions. We also see evidence that the level of difficulty of a question appears to have more to do with the problem-solving skills that it requires than with the particular statistical topic that it addresses. Further, there is evidence that problems drawing on more than one skill show a considerable increase in the challenge that they present to the student.

We then take up the question of the addressability of a problem area by employing a new methodology. We use cluster analysis to partition our 16 introductory statistics teachers into clusters, grouping together teachers whose students tend to do well (or poorly) on the same sets of questions. Two natural clusters, which we refer to as Cluster A and Cluster B, emerge. Although there is no discernable difference between the clusters in some problem areas, the difference in others is dramatic. Where there is a substantial difference, Cluster A teachers always have the higher success rates. We believe that this difference indicates an area where pedagogical choices may significantly impact student learning.

There can be many effective ways to teach quantitative subject matter. Still, we hope that the practices of our Cluster A teachers, particularly in the areas where their students considerably outperform their Cluster B counterparts, may prove useful to other teachers of introductory statistics. With that in mind, we look into the pedagogical activities of the Cluster A teachers in these problem areas and report some specific techniques used by them. We conclude with a brief summary and a discussion of the limitations of this study as well as the work that remains to be done.

2. Background

When statistics instructors began implementing reform recommendations in the early 1990s, they recognized a need to study the effects of these changes on students' understanding of statistical concepts. This research was hampered by the lack of an appropriate assessment instrument for evaluating application of statistical knowledge rather than definitions and calculation procedures. This deficiency was initially targeted by Konold (1990) and Garfield (1991) in a National Science Foundation (NSF)-funded project on the effectiveness of a high school statistics curriculum where they developed the Statistical Reasoning Assessment (SRA) test. It provided a standard instrument at the pre-university level, consisting of 20 multiple-choice questions on probability and statistics concepts.

Identification of limitations in the scope of the SRA led to further development of assessment items and instruments in the NSF Assessment Resource Tools for Improving Statistical Thinking (ARTIST) project (Garfield et al. 2002a). The most significant outcome from this effort was the Comprehensive Assessment of Outcomes in Statistics (CAOS), constructed with a primary goal of providing a reliable instrument for assessment of student statistical reasoning and literacy across disciplines (delMas et al. 2007). It has recently become publicly available (<https://app.gen.umn.edu/artist>).

A secondary goal of CAOS was to support research for investigating areas where students show (or do not show) significantly improved performance from the beginning to the end of an introductory statistics course. In a sample of 763 students from 20 different institutions, delMas et al. (2007) reported that the average percentage of CAOS items correctly answered on the post-test was only 54%, representing a small (9% average) increase from the pre-test scores. This study provides recent evidence that reform efforts have not made a significant difference in student ability in statistical reasoning. Analysis of CAOS test results also showed statistically significant differences among instructors.

It has long been known that students struggle with statistical concepts and suffer from intuitions that conflict with statistical theory (Garfield and Ahlgren 1988). Numerous studies document misconceptions about averages (e.g., Watson 2007); chance events (e.g., Albert 2003, Garfield 2002, Hirsch and O'Donnell 2001); measures of variability (e.g., Ben-Zvi and Garfield 2004, Watson and Kelly 2007); sampling distributions (e.g., Chance et al. 2004); confidence intervals (e.g., Grant and Nathan 2008); and hypothesis testing (e.g., Dambolena et al. 2009, Haller and Krauss 2002, Vallecillos 2002). Additional references going back to the late 1970s are provided in delMas et al. (2007) and, for engineering disciplines, in Evans et al. (2003).

However, in a recent review of articles on students' misunderstandings in the area of statistical inference, Sotos et al. (2007) state that "the literature on education, and particularly publications providing empirical evidence of misconceptions in statistics, is sparse" (p. 100). They call for more *empirical* studies with structured methodologies to complement the copious theoretical discussions and descriptions of classroom experiences already published. In this paper, we address the need for further experimental work by reviewing six years of assessment data from an introductory undergraduate statistics course. From our study, we gain insight into the nature and characteristics of student errors within a framework of instructional differences, laying a foundation for identification of effective pedagogical approaches.

3. Test Instrument and Data Collection

The assessment instrument for our introductory business statistics course was developed in 2002–2003 through an iterative process that began with the identification of key learning objectives articulating the knowledge and skills students were expected to acquire. Test items were written to cover these objectives and reviewed by instructional faculty to ensure appropriate content and valid item construction. Several rounds of evaluation and feedback followed, with

adjustments being made to the learning objectives as needed.

The result was a test instrument consisting of 63 questions designed to be administered as a common final examination. Computation is kept to a minimum on the test. This computational simplicity is consistent with our heavy use of spreadsheet software to perform straightforward calculations. The test places more weight on conceptual and interpretation issues (37 questions) than computational ones (26 questions). For example, we ask students to identify which graphical display technique would be most appropriate for a given set of data (conceptual question) rather than having them draw a specific graph (computational question).

Students have two hours to complete the exam corresponding to less than two minutes on average per question. The multiple-choice instrument assures consistency of measurement across teachers, sections, and semesters and supported detailed item analyses. Students are permitted to bring an 8.5" × 11" formula sheet to the exam. Question reliability is evaluated using internal consistency measures and by comparing means for multiple sections taught by the same teacher in the same semester. Validity is assessed by comparing students' performance on the test instrument to their performance in the rest of the course.

We base our study on data for students in the semesters of spring 2003 through spring 2009, during which time 4,727 students completed the course and took the final exam. Extensive precautions were taken to ensure that the contents of the test remained uncompromised from semester to semester. The course and the management science course for which it is a prerequisite must both be completed before business students may begin their junior level course work. Consequently, the students in this study were overwhelmingly freshmen or first-semester sophomores planning to major in business.

4. Cluster Analysis of Assessment Results

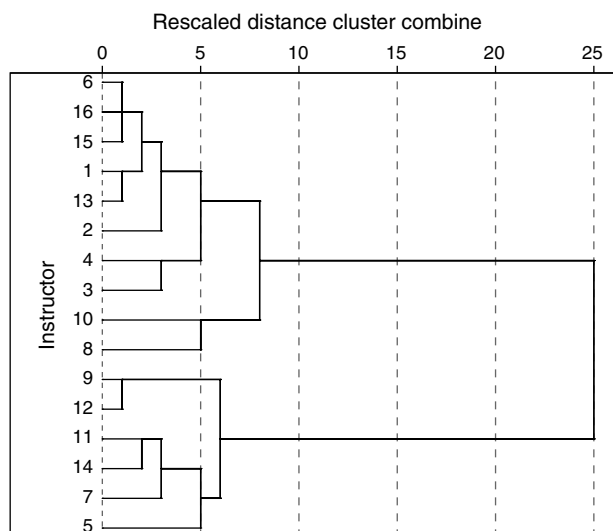
The success rate on a statistical topic suggests its overall level of difficulty. As teachers, however, our desire was to identify areas where changes in our course delivery could significantly improve student performance. Simply looking at which teacher's students did best on a particular question—or set of questions—was not particularly useful. Exceptional performance on a question by a class in a given semester might have many explanations such as a statistical fluke, the result of a nearly identical question being worked in class, or additional time being spent on the topic by a particular instructor. We decided on a broader approach to see if we could sensibly

partition our teachers into subgroups whose students showed similar strengths and weaknesses. Our hope was that differences between these groups could then help us identify not only those aspects of statistics that are difficult but those difficulties that might be addressable by teaching choices.

To this end, we computed a "performance profile" for each of our 16 teachers of introductory business statistics, consisting of the mean success rate of that teacher's students on each of the 63 assessment questions. We then performed a hierarchical cluster analysis (using Ward's method with a squared Euclidean distance function; Ward 1963), grouping together teachers who had similar performance profiles; that is, teachers whose students showed similar patterns of strength and weakness across the set of questions. The results appear in Figure 1.

The dendrogram shows, for example, that teachers 6, 16, and 15 all have similar performance profiles, as do teachers 1 and 13 and teachers 9 and 12. The first cluster (6, 16, 15) bears considerable similarity to the second (1, 13); these two clusters were then merged in the second stage of the cluster analysis. The horizontal axis in the dendrogram indicates the distance between two clusters that are being merged at a given stage in this process, standardized so that the distance between the final two clusters is always 25. In our case, the first cluster consisted of the first 10 teachers on the dendrogram (referred to here as Cluster B) and the 6 teachers in the lower portion of the dendrogram formed the other (referred to here as Cluster A). The long horizontal bars linked to these clusters indicate that the two groups show substantial differences from one another in performance profiles (Kaufman and Rousseeuw 1990).

Figure 1 Clustering of Teachers by Performance Profile



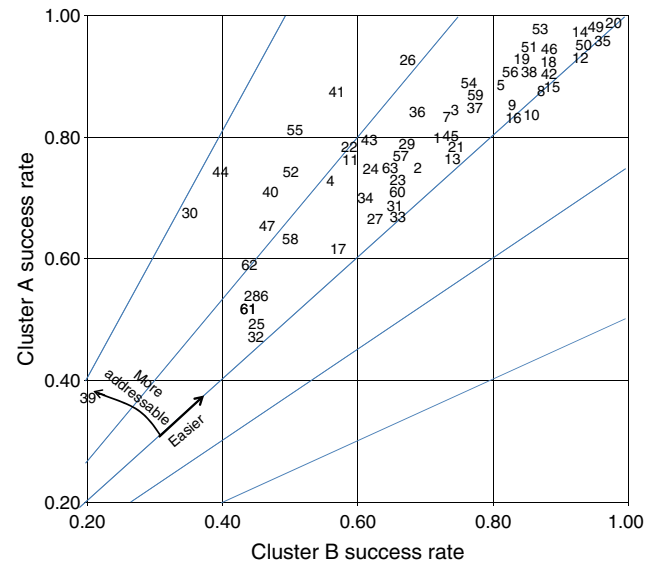
These two groups show some interesting differences in demographics: all six Cluster A teachers are full-time faculty who hold doctoral degrees, and five out of the six have taught at the university level for at least 10 years. The single exception was Teacher 9, who was mentored by Teacher 12 in teaching statistics during his first year at the university and who had previous teaching experience as a graduate student. All but one of the Cluster A teachers have advanced degrees in technical disciplines (mathematics, operations research/management science, engineering, etc.). Cluster B consists of one full-time and six adjunct teachers who are not doctorally qualified, two full-time doctorally qualified faculty members in technical disciplines, and one full-time doctorally qualified faculty/administrator in a nontechnical discipline. The teaching experience of Cluster B members varies considerably, from no teaching experience prior to teaching our course to more than 20 years in the classroom. Cluster B teachers instructed 57% of the students in this study (2,688), and Cluster A teachers taught the remaining 43% (2,039).

5. Identifying Addressability by Comparison of Clusters

We define the success rate of a cluster on a given question as the fraction of students in that cluster who answered the question correctly. A scatter plot of the success rate of Cluster B versus Cluster A on the 63 individual questions in the test instrument is shown in Figure 2. A point's height above the main diagonal line indicates the amount by which the success rate of Cluster A students exceeded the success rate of Cluster B on that question. The oblique lines on the plot show lines of constant relative success ratio defined as *Cluster B success rate* divided by *Cluster A success rate*. A relative success ratio of 0.77, for example, indicates a problem on which a Cluster B student is only 77% as likely to answer correctly as a Cluster A student. Relative success ratios significantly below one thus indicate questions where the Cluster A students showed markedly superior success rates. For this reason, we take the relative success ratio as a measure of the relative addressability of the difficulties a problem presents.

Several insights can be gained from examining the Figure 2 plot. Questions falling in the upper-right corner of the graph are ones on which students from both clusters do well; effective instruction does not appear to be a problem with these. Questions whose numbers fall near the 45-degree line of the graph show little difference in performance between the two clusters—both groups, for example, had about a 98% success rate on question 20 and about a 46% success rate on question 32. The plot also allows us to identify questions in which the performance between the

Figure 2 Comparison of Success Rates for Clusters A and B on all 63 Questions



two clusters was notably different. In this regard it is interesting to observe that Cluster A had the higher success rate on all but one of the questions (question 10) and on that question the success rates were within 2% of one another.

Although it is obvious that the students of Cluster A teachers perform better on average than their Cluster B counterparts, this is not in itself surprising. Random fluctuations will result in the students of some teachers doing better than those of others and if those more successful teachers are clustered together, it is clear that their students would on average outperform students from the other cluster. But the pattern seen here suggests more structure. Teachers in both clusters have the same number of contact hours with their students but Cluster A performance equals or exceeds that of Cluster B on virtually every question, sometimes dramatically. This pattern holds semester after semester. In the following section, we examine the issues of both difficulty and addressability for the learning objectives measured by our assessment test questions.

6. Student Performance on the Assessment Test

The most common approach used in studies of student difficulties with statistics is to partition the subject matter by course topic corresponding to standard textbook chapters (data presentation, probability, hypothesis testing, and so on). We take a different perspective, focusing on commonalities in the underlying problem-solving mechanisms. We characterize each of the 63 questions on the instrument as

falling into one of seven categories: rote calculation, application of definition, experiential validation, distribution identification, cumulative distributions and cutoffs, interpretation of probability density functions (pdfs), and inferential interpretation.

Each of the following subsections presents our results for one of these categories. The table in each subsection gives a brief description of the learning objective for each question in that category along with the question's aggregate and cluster-specific success rates. We also report each question's relative success ratio. The shading of the table entries indicates questions with overall success rate of 65% or below (dark gray), 66 to 80% (light gray), and above 80% (no shading). We refer to these categories as "hard," "intermediate," and "easy," respectively. Examining relative success ratios in conjunction with overall success rates provides guidance on specific issues where instructional methods can positively affect student learning. We do not include a subsection on distribution identification because our instrument included only two questions of this kind. Further, the results for question 17 were omitted because the question contained an ambiguity that may have substantially degraded student performance.

6.1. Rote Calculation

By "rote calculation," we mean determination of answers by application of a simple algorithm. As can be seen in Table 1, students in both clusters demonstrate competence in rote calculation and the success rates for the two clusters do not differ substantially. Success rates for both clusters are high, with an overall success rate of 88% and an average success ratio of 0.95. An algorithm becomes more complex by the introduction of conditional branching or the need to compute intermediate quantities as part of the solution process. We would expect performance to degrade as algorithm complexity increases or when more than one algorithm has to be applied, and our results are consistent with this.

The simplest questions, such as finding means, relative frequencies, or z scores when given μ and σ , require no branching and no computation of intermediate quantities. For such problems (questions 8, 12, 18, 35, 38, 46, 49, 50, 51, 53, and 56), overall success rates were 86% or better. But performance suffered when two skills were combined in a single problem: 88% of students could compute the relative frequency of a range of classes in a frequency histogram in question 15; 84% successfully computed the median in question 9; but only 74% of students could identify the category in the histogram that contained the median observation for question 16. Solving that problem required the combination of finding the location of the median value in an ordered data set, a simple branching based on parity, with that of reading a frequency

histogram. Almost all students missing this question chose the middle class, the third of five, ignoring the information actually contained in the histogram.

6.2. Application of Definition

Performance results for questions involving definitions are reported in Table 2. Because we are not interested in students merely reciting memorized definitions, our basic definitional problems fall into two categories. In questions 1, 2, 5, 6, 7, and 14, a quantity is specified in a scenario and the student provides the appropriate terminology (e.g., continuous data). In questions 3, 4, 13, 31, 37, and 40, a term is given (e.g., categorical variable) and the student applies its meaning in the provided scenario. The overall success rate on application of definition problems was 73% and the average relative success ratio was 0.89. Definition questions involving only one concept had success ratios between 0.84 and 0.92, indicating that Cluster A performance was noticeably but not dramatically better than that of Cluster B. Three questions showed an exception to this pattern: questions 13 and 14 where the two clusters were approximately equal in performance, and 40 where Cluster A considerably outscored Cluster B.

Although half of the definitional questions fall into the intermediate difficulty category, three are in the hard group with relative success ratios between 0.66 and 0.84. As with rote calculation, when a question involves the combination of two or more definitions, student performance drops substantially and the gap between the clusters widens. Compare the tabulated results for question 3 (dealing with a single variable type) and question 4 (dealing with two types). Note, too, that the concepts in questions 6 and 40 hinge on multi-stage processes.

6.3. Experiential Validation

Students find it very challenging to understand abstract principles (Willingham 2009). Many students also find it difficult to bring their experience and practical knowledge to bear on mathematical problems. We categorized theoretical problems with answers that can be confirmed via logical reasoning as experiential validation questions. Performance results for these questions are reported in Table 3. With the exception of question 36, all of these questions were scenario-based word problems stated in English without mathematical terms and notation. Most of these problems were of intermediate difficulty. The overall success rate on experiential validation questions was 74% and the average relative success ratio was 0.89.

For questions 23 and 24, which dealt with basic probabilities, our expectation was that students would be able to reason that the occurrence of one event does not change the probability of the other event when

Table 1 Results for Questions on Rote Calculation

Question number	Measured learning objective	Success rate (overall) (%)	Success rate (Cluster A) (%)	Success rate (Cluster B) (%)	Relative success ratio (B/A)
8	Compute relative frequency	87	87	87	0.99
9	Compute median of a data set ($n = 30$)	84	85	83	0.97
12	Relate impact of outliers on mean compared to median	93	93	93	0.99
15	Compute relative frequency of $s \geq i$ from a histogram	88	88	88	1.00
16	Interpret histogram to find median class	74	76	72	0.96
18	Apply the empirical rule to a normal distribution	90	93	88	0.95
19	Compute s in Excel	88	93	84	0.90
35	Estimate the mean of a normal variable from its pdf	95	96	95	0.99
38	Compute z score of an observation	87	91	85	0.94
46	Conditions for approximate normality of sampling distribution of the mean	90	93	88	0.95
49	Use z for confidence interval of one mean with σ known	97	98	96	0.98
50	Use t for confidence interval of one mean with σ unknown	94	95	93	0.98
51	Realize that proportion hypothesis tests use z	89	94	85	0.90
53	Choose correct degrees of freedom for a t distribution (Excel)	91	98	87	0.89
54	Compute standard error of the mean with σ unknown	81	88	76	0.87
56	Relate α , p -value, and hypothesis test conclusion	86	91	83	0.91

Table 2 Results for Questions on Applying Definitions

Question number	Measured learning objective	Success rate (overall) (%)	Success rate (Cluster A) (%)	Success rate (Cluster B) (%)	Relative success ratio (B/A)
1	Distinguish between a sample and a population	75	80	72	0.90
2	Recognize data as continuous	71	75	69	0.92
3	Identify categorical variables	78	84	74	0.88
4	Distinguish between nominal and ordinal variables	63	73	56	0.77
5	Distinguish a parameter from a statistic	84	88	81	0.92
6	Recognize a sampling process as stratified sampling	48	53	45	0.84
7	Identify incorrect data as measurement error	78	83	74	0.88
13	Box and whisker plot does not give mean	75	76	74	0.96
14	Recognize a skew distribution	95	97	94	0.97
31	Generate a discrete probability distribution	67	69	65	0.95
37	Characterize the standard normal distribution (μ and σ)	81	85	77	0.90
40	Explain meaning of “standard error”	57	71	47	0.66

Table 3 Results for Questions on Experiential Validation

Question number	Measured learning objective	Success rate (overall) (%)	Success rate (Cluster A) (%)	Success rate (Cluster B) (%)	Relative success ratio (B/A)
10	Choose chart type for univariate categorical data	85	83	85	1.02
11	Choose scatter plot to explore relationships in numeric bivariate data	66	76	59	0.77
17	Recognize that one can cannot compute μ or σ from categorical data	59	61	57	0.92
20	Compute marginal probability from a contingency table	98	99	97	0.99
21	Compute $P(A B)$ from a contingency table	76	78	74	0.94
22	Compute $P(A B)$ when A and B are mutually exclusive	67	78	58	0.75
23	Compute $P(A B)$ when A and B are independent	69	73	66	0.90
24	Compute $P(A \text{ and } B)$ when A and B are independent	67	75	62	0.83
36	Estimate the standard deviation of a normal variable from its pdf	75	84	68	0.82

the two events are independent. Independence should also make the joint probability easy to compute. But more than 90% of the students who answered question 24 incorrectly said that the joint probability was greater than or equal to the probability of one of the

events occurring, i.e., the conjunction fallacy. Mistakes in other questions in this category, such as failing to recognize that the probability of two mutually exclusive events occurring is zero in question 22 and choosing either an ogive or a histogram to explore the rela-

Table 4 Results for Questions on Cumulative Distributions and Cutoffs

Question number	Measured learning objective	Success rate (overall) (%)	Success rate (Cluster A) (%)	Success rate (Cluster B) (%)	Relative success ratio (B/A)
27	Compute $P(s > i)$ given cumulative distribution of a discrete variable	64	66	62	0.94
28	Compute $P(s = i)$ given cumulative distribution of a discrete variable	48	53	44	0.83
30	Compute $P(s \geq i)$ for a binomial variable (Excel)	49	68	34	0.50
52	Compute critical z for two-tailed test of one mean (Excel)	60	74	50	0.67

tionship between two ratio-scaled variables in question 11, showed a similar lack of rationality that was more prevalent in Cluster B. Sections A.1 and A.2 of Appendix A provide more detailed descriptions of these questions.

6.4. Cumulative Distributions and Cutoffs for Discrete Variables

It is often necessary to compute the probability that the value of a random variable lies within a specified range, and cumulative distributions play an important role in these calculations. Given the probability mass function or the relative frequency histogram for a discrete random variable, it is easy to construct the cumulative distribution as a running total of probabilities. We found that confusion often occurs, however, when the task is reversed and a student needs to recover the probability mass function from the cumulative distribution. All of the numeric discrete distributions on our test instrument are for variables that take on only nonnegative integer values. We refer to such variables as “counting variables.”

Our results in this category (Table 4) revealed a fundamental and pervasive weakness of students in working with cumulative distributions of discrete random variables, and this weakness was independent of the particular distribution in question. All four discrete cumulative/cutoff questions were in the “difficult” group of questions. The overall success rate on these questions was 55%, and the average relative success ratio was 0.73.

Questions 27 and 28 on our instrument were preceded by a brief scenario explicitly stating the probabilities that a specific counting variable took on values that were less than or equal to 4, 5, and 6. The variable was one that students were familiar with, such as the number of courses a student takes in a semester. The weakness suggested by the low success rates for these questions in both clusters is underscored by the choices made by those students answering incorrectly. Fully 43% of students computed $P(X = 6)$ as $1 - P(X < 6)$. Similarly, $P(X > 5)$ was computed as $1 - P(X < 6)$ by 16% of students and as $P(X < 6) - P(X < 5)$ by 13%.

As in previous categories, we saw evidence that calling on more than one skill set degrades performance and decreases the relative success ratio between the clusters. In question 30, students were given a variable X that is binomially distributed with specified values of n and p . They were then asked for the Excel expression that would give $P(X > 10)$. All choices were expressed in terms of the Excel BINOMDIST function and the meanings of the arguments in the BINOMDIST function were given. This question had only a 49% success rate with a huge gap between the groups. Some of this difference can be explained by a limited mastery of Excel probability functions. Still, 23% of students chose the formula for $1 - P(X < 10)$ rather than $1 - P(X < 9)$, and this choice was almost equally common in the two clusters. This problem of determining the correct cutoffs when working with a discrete cumulative distribution was widespread and pernicious in both A and B clusters.

6.5. Interpreting Probability Density Functions for Continuous Variables

For continuous variables, computation of probabilities requires use of the cumulative distribution. If the cumulative distribution function (cdf) for X is given by $C(x) = P(X \leq x)$, then $P(a \leq X \leq b)$ is simply calculated as $C(b) - C(a)$. In contrast to discrete random variables, this result holds even if one or both of the strict inequalities are replaced with weak inequalities. Unfortunately, as can be seen in Table 5, this simplification in determining cutoff points is more than counterbalanced by students' difficulty in understanding the nature and representation of continuous distributions. Five of the six pdf questions are classed in the hard question group. None of the pdf questions had a success rate above 66%. The overall success rate was 51% and the average relative success ratio was 0.84.

A discrete variable's distribution is usually specified by listing each possible value that the variable can take on, along with the probability that each of these values occurs. Most students readily understand

Table 5 Results for Questions on Probability Density Functions

Question number	Measured learning objective	Success rate (overall) (%)	Success rate (Cluster A) (%)	Success rate (Cluster B) (%)	Relative success ratio (B/A)
25	Recognize a binomial variable and compute its mean	47	49	45	0.91
32	Area corresponds to probability for continuous pdfs	46	47	45	0.95
33	Recognize that largest data values occur to the right of the pdf graph for a continuous variable	66	67	66	0.98
34	Compare μ and σ from the pdfs of two normal variables	65	70	61	0.87
39	Compute $P(0 \leq z \leq k)$, given $P(z \leq k)$	28	37	20	0.56
47	Interpret = NORMSDIST(k) (Excel)	54	65	46	0.71

this information when it is graphically represented as a bar chart. In contrast, we saw in §6.4 that students experience considerable trouble when the same information is conveyed by a cumulative distribution. For continuous variables, the cdf is an essential and unavoidable tool for computing probabilities. When the cdf is not expressible in closed form, a table or computer software is needed to find its values. For graphical representation, common practice is to focus on the graph of the pdf and draw on its parallels with a histogram rather than focusing on the graph of the cdf and its parallels to the ogive. The histogram/pdf correspondence may seem natural to those with experience with Riemann sums in integral calculus but is far from obvious to the uninitiated. Even the units on the vertical axis of a pdf graph are unclear to many, and pdfs are usually drawn without any vertical scale shown at all.

Our research indicates that many students are unable to correctly interpret the graph of the pdf. In questions 32 and 33 students were shown a graph of the pdf for the salary of a newly hired employee (Appendix A.3). The horizontal axis was partitioned into three salary ranges and students were asked which range was most likely to contain the salary of a randomly selected newly-hired employee. Only 46% correctly chose the range for which the area under the curve was the largest. More than a third of students (35%) appear to have applied the notion that probability is proportional to height and so chose the range including the mode. The remaining 19% chose the widest of the three ranges, a strategy that ignores the pdf altogether.

Students were then asked which range corresponded to the highest starting salaries. Although two-thirds of the students answered this correctly, 31% chose the range containing the mode, presumably because the pdf has the greatest height in this range. We saw similar performance in question 25 when asking for the mean of a binomial distribution

with $n = 15$ and $p = 0.2$. Fully 43% of students said the mean was 7.5, the midpoint of 0 and 15. All of these errors rates were largely independent of cluster.

In question 34, students were shown two normal curves on the same set of axes, one with a high mean and a low standard deviation and one with a low mean and high standard deviation (Appendix A.4). Although 91% of students realized that the wider curve had the higher standard deviation, only 72% correctly identified which curve had the higher mean. Many students seem to have associated the *height* of the peak of the normal curve with its mean rather than the value on the x -axis corresponding to that peak. Because the narrower curve is also the taller curve, this reasoning would lead to the wrong answer. This misconception occurred even though 95% could correctly report the value of the mean of a single normal distribution in rote calculation question 35 (§6.1). Because question 34 asked for a comparison of both means and standard deviations, the actual success rate for this question was even lower. Once again, the requirement to perform more than one task before reaching the final answer appears to contribute to the students' troubles, and we see a lower relative success ratio in such problems.

An extreme example of applying multiple algorithms is question 39, in which the student is given $P(z \leq 0.84) = 0.8$ and asked for $P(0 \leq z \leq 0.84)$ where z is a standard normal variable. The 28% success rate was the second worst on the test and had a 17% gap between clusters. We expected students to draw the standard normal curve, know that the total area under the curve is 1, and realize that the half of the curve to the left of 0 has area 0.5. Then, because area corresponds to probability, the answer must be $0.8 - 0.5 = 0.3$. Instead, 44% of students said the answer was 0.8, which is equivalent to saying that z can never be negative. This, even though on question 37, 81% of students knew that the z distribution has a mean of 0 and standard deviation of 1.

Table 6 Results for Questions on Inferential Interpretation

Question number	Measured learning objective	Success rate (overall) (%)	Success rate (Cluster A) (%)	Success rate (Cluster B) (%)	Relative success ratio (B/A)
41	Identify s in a sampling word problem	70	87	57	0.65
42	Realize that s estimates σ	88	90	87	0.97
43	Compute standard error of the mean	69	79	61	0.77
44	Population mean = mean of sampling distribution of the mean	54	74	39	0.53
45	Construct confidence interval for the mean given critical t value	76	79	74	0.93
48	Interpret a confidence interval	22	27	17	0.63
55	Conditions for approximate normality of sampling distribution of the proportion (population proportion unknown)	64	81	50	0.62
57	Smaller α means less chance of rejecting true H_0	71	77	66	0.86
58	Appropriate null hypothesis for a one-tailed test of two means	55	63	50	0.79
59	Distinguish a two-means test from a paired difference test	81	86	77	0.90
60	Recognize that H_0 : " μ_1 no larger than μ_2 " is a one-tailed test	68	72	66	0.91
61	Possible conclusions when H_0 is an equality	47	52	43	0.84
62	Relevance of equal variance assumption for two population test of means	50	59	44	0.74
63	Relate meaning of the Central Limit Theorem	69	75	64	0.86

6.6. Inferential Interpretation

Success in inferential statistics very much depends upon mastery of both concepts and calculations. Students did fairly well identifying a problem as requiring a difference-of-two-means test rather than a paired-differences test. With this exception, though, inferential concepts and interpretation were considerably more difficult for students than was rote calculation (see Table 6). The overall success rate on the inferential interpretation questions was 62%, and the overall success ratio was 0.78.

Confounding the notations for sample standard deviation (s), population standard deviation (σ), and the standard error of the mean ($\sigma_{\bar{x}}$) was also common. In a narrative on the confidence interval of the mean giving the value of the sample standard deviation and the sample size in the text, the lower success rate on question 41 in comparison to question 42 suggested that many students thought they were given σ although they were given s . When asked to compute the quantity $\sigma_{\bar{x}}$ in question 43, 16% confused $\sigma_{\bar{x}}$ with either σ or s . This was unexpected because students have an explicit formula for the calculation of this quantity. In question 44, almost half of the students did not realize that $\mu = \mu_{\bar{x}}$, i.e., the mean of the population was equal to the mean of the sampling distribution of the mean.

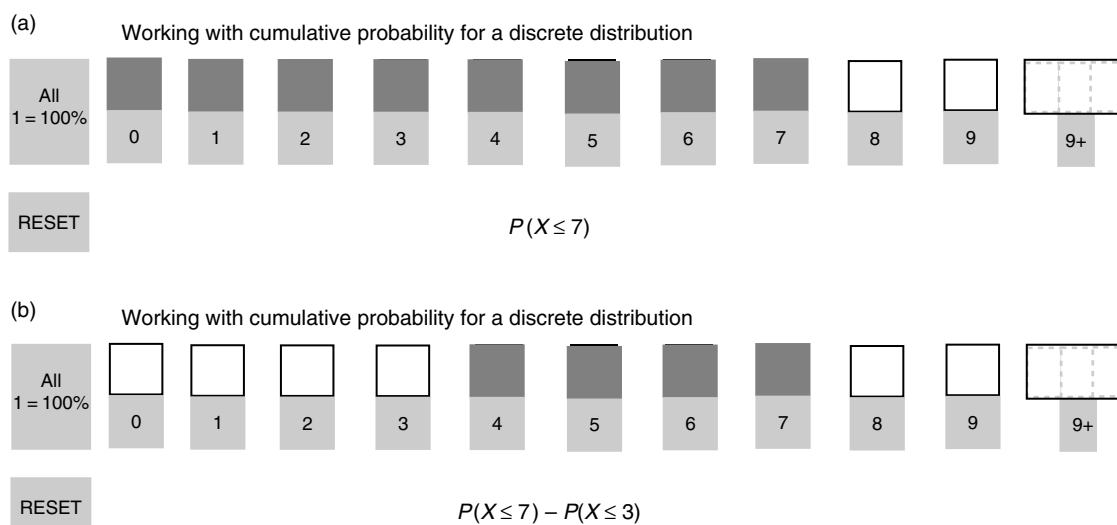
Computing a confidence interval (question 45) was within the grasp of most students; interpreting a confidence interval (question 48) was not. The most frequently selected interpretation, chosen by 26% of students, was that about 90% of the population means would fall within a 90% confidence interval. This choice is interesting because a population has only one mean. Another 22% thought that about 90% of the population values would fall within the interval.

And although a majority (86%) of students internalized the mechanical rule to reject the null hypothesis ($p\text{-value} < \alpha$) in rote calculation question 56, 29% could not go on in question 57 to conclude that a small value of α reduces the chance of rejecting a true null hypothesis.

Other common misunderstandings of inferential concepts included choosing H_0 : $\mu_1 \geq \mu_2$ over H_0 : $\mu_1 \leq \mu_2$ when asked for a null hypothesis that could lead to the conclusion that μ_1 is greater than μ_2 (question 58), not recognizing a one-tail test (question 60), believing that a hypothesis test can end in the conclusion that the null hypothesis is true (question 61), and concluding that the Central Limit Theorem states that all sufficiently large samples are normal (question 63). Additional answers suggested that many students have no clear idea of proper methodology when doing inferential statistics. For example, in question 62 students were asked what issue they would have to consider *before* deciding which of the two "difference of two mean" hypothesis tests they should conduct. They split equally between choosing "whether the two population variances can be considered as equal" and the nonsensical "whether the two population means can be considered equal."

The weak overall performance seen on inferential statistics in both clusters was not unanticipated. Analysis of data from the CAOS test (delMas et al. 2007) demonstrated that many students fail to understand confidence interval limits and significance tests after completing an introductory statistics course. In a recent review, Sotos et al. (2007) provide a detailed summary and classification of misconceptions of sampling distributions, confidence intervals, and hypothesis tests cited in both theoretical and empirical-based studies over the past two decades. Although the performance on these last questions was extremely poor,

Figure 3 Excel Spreadsheet for the Light Bulb Thought Experiment



their low success ratios suggest that these problems may be addressable to some degree.

7. Suggestions for Addressing Observed Problem Areas

Although there is no panacea for the difficulties students experience with introductory statistics, differences in the performance of the two clusters highlight areas where instructional practices may be an important factor. Using the cluster results as a guide, we have begun investigating pedagogical methods and activities that target specific types of errors and misunderstandings discovered in our analysis. Rote calculation shows high success in both clusters with little difference between them. Definition and validation problems show mostly intermediate success rates but limited difference between clusters: the median success rate was 75% and a median relative success ratio was a relatively high 0.90. The remaining three categories are areas of considerable student difficulty and relatively low relative success ratios, and so we focus on these. In this section, we discuss our preliminary findings on the approaches that Cluster A teachers take to the observed problems with cumulative distributions, probability density functions, and conceptual reasoning about inferential statistics.

7.1. Addressing Problems with Discrete Cumulative Distributions and Cutoffs

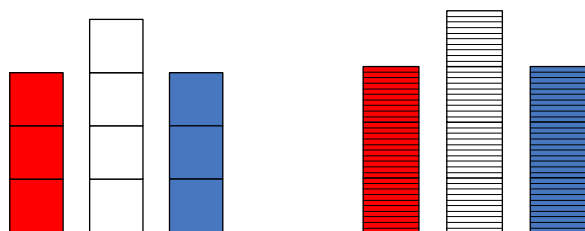
The challenge is to develop a solid intuition of the connection between the cumulative distribution of a discrete variable and the probability that its value falls in a specified range. We do this by having the student imagine a machine consisting of a row of numbered

pushbuttons beneath a row of light bulbs similar to that shown in Figure 3. Originally, all of the lights on the machine are off. Pressing the button of an *unlit* number causes its light to illuminate as well as all of the lights to its left. Pressing a *lit* number causes that light to turn off as well as all of the lights to its left. Saying it succinctly, pressing a button toggles its light and all of the lights to its left. We also have one more button labeled "1 = 100%" that illuminates all of the lights.

Knowing this, students are given a range of numbers and challenged to press at most two buttons so as to illuminate only the lights corresponding to that range. Figure 3 shows the solution process when the desired range runs from 4 to 7 inclusive. The student first presses the 7 (Figure 3(a)) and then the 3 (Figure 3(b)).

Most students have little trouble determining what numbers should be pushed, but this task is entirely equivalent to that of determining the correct cutoffs to use with the cumulative distribution for discrete probability calculations. If $C(x) = P(X \leq x)$ is the cdf of X , then the probability that X takes on a value in the lit range is simply $C(\text{ON}) - C(\text{OFF})$, where ON is the first button pushed and OFF is the second button pushed (if any). Our spreadsheet implementation of the machine, shown in Figure 3 and available with this paper (supplemental file *Discrete_Distribution_Lights.xls*)¹, displays this probability calculation underneath the pushbuttons. Although the buttons above assume that X is a counting variable, any discrete variable with a different range can

¹ Files that accompany this paper can be found and downloaded from <http://dx.doi.org/10.1287/ited.1120.0085>.

Figure 4 Visualizing the Meaning of a Bar Graph or Histogram ($N = 10$ and $N = 100$)

be handled in the same way by renumbering the buttons. For an upper tail probability, the student begins by pressing the 1 = 100% button to turn on all of the lights and then presses the highest numbered button not in the desired range.

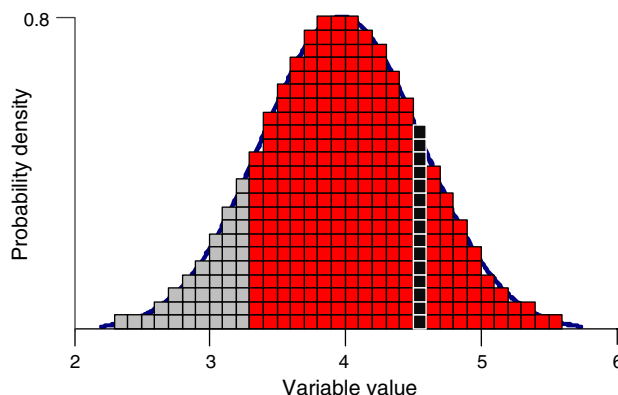
This approach makes clear that determining these cutoff points is more of an exercise in logical thinking than in statistical reasoning. We are in the process of collecting data on using this spreadsheet in the classroom and making it available to students for hands-on experimentation. Preliminary evidence points to the possibility that the cutoff problem may occur in part because students have difficulty interpreting English expressions such as “more than” and “at least.” This issue does not appear to have been documented in statistical education research to date and merits additional study.

7.2. Addressing Problems with Interpreting Probability Density Functions

If we are told that in a collection of objects, 30% are red, 40% are white, and 30% are blue, we may conceive of the bar graph of this population by imagining each item being represented by a block of some specific dimensions and then stacking them appropriately as shown in Figure 4. There is a natural correspondence connecting a relative frequency, a probability, and physically observable quantity—the height of the bar.

If we have the same proportions of colors in an infinite population, we draw the same bar chart but the relative frequency interpretation is now meaningless. Certainly saying that there are as many red objects as blue ones is not informative because both quantities are infinite. Intuition still serves, though—height in the new graph is proportional to probability and if the student envisions a population where N is very large (rather than actually infinite in order to apply the relative frequency interpretation) little practical harm is done at the introductory level. Many introductory statistics texts assume that the student can extend this work to the case of continuous variables with little difficulty. Our research suggests otherwise.

Because our goal is not to teach calculus or advanced probability theory to introductory statistics

Figure 5 A Finite Population Approximating the Continuous Distribution

students, how can we help students to develop an intuition comparable to the “blocks” model for discrete variables? Our solution, which we call “the sand interpretation,” is an attempt to provide a useful and simple “block-like” model for continuous variables. Preliminary evidence suggests that students taught in this way show considerable improvement in their understanding of pdfs. On the five pdf problems that caused students trouble (questions 32, 33, 34, 36, and 39), the average of the 484 students taught with the sand interpretation was 68% versus 54% for those who were taught without it. The average relative success ratio was 0.79.

For the sand interpretation, we begin with a pdf such as the normal curve shown in Figure 5 and wish to generate a finite population from it with a distribution close to that of the population represented by the pdf. We can do this by covering the pdf with a very fine mesh that cuts it into very small squares. We label each square with a number on the x -axis over which it rests. For example, all of the black squares in Figure 5 get values between 4.5 and 4.6. Then this finite collection of values has a distribution that approximates that of the underlying continuous population and this approximation will improve with a finer and finer mesh.

A “coarse mesh” picture like the one above allows a student to easily develop the appropriate intuitions about pdfs without knowledge of integral calculus. Probability can still be thought of in terms of relative frequency, which is proportional to the number of squares in a region. The number of squares, in turn, is proportional to the area that they occupy in the picture. It is clear, therefore, that the probability of getting a value less than 3.3 (the gray squares in Figure 5) is larger than the probability of getting a value between 4.5 and 4.6 (the black squares)—the black region is taller, but the gray region has more area.

We refer to this as the sand interpretation because we encourage students to next think of a very fine mesh with squares that are nearly invisible grains of sand. It is easy for students to visualize both the observation (the grain) and the population (the sand pile) at the same time. They can also understand that the mean of a distribution is the point along the x -axis upon which the sand dune would balance perfectly because the area of a region of the dune is proportional to its mass.

This interpretation is especially useful for a discussion of the sampling distribution of the mean. For $n = 5$, randomly select five grains of sand from the population pdf, read the numbers on those grains, and average them to get the sample mean. Write this average on a new grain of sand and begin to build a new sand pile on a new x -axis. Return the original five grains to the original pdf and repeat this for every possible set of five grains of sand. When you are done, the new sand pile is the pdf of the sampling distribution. Students can see how each “grain” corresponds to a sample mean rather than a single observation from the original population. Interactive software that approximates a real-time simulation of the sampling process is available from a number of sources (see Dinov et al. 2008, Mills 2002, Tsai and Wardell 2006).

7.3. Addressing Problems with Inferential Interpretation

Although the issues in these areas are unlikely to be addressed by a single remedy, interviews with the Cluster A faculty reveal that they generally follow a classic, systematic approach to characterizing a problem before applying a statistical formula or algorithm: understand the meaning of the question, record problem values in correct notation, and identify the quantity sought in the problem using appropriate notation. Equally important, they conduct a high-level analysis of the problem, considering which techniques or facts

among those studied might be relevant. After calculations are completed, they interpret their final answer in the context of the problem to see if it is reasonable.

From a student perspective, the algorithms for identifying and solving problems that involve probability distributions and inferential statistics are complex. They involve branching, checks of assumptions, and calculation of intermediate values. Two of our Cluster A teachers have adopted the use of flowcharts to provide students with a systematic way of identifying problem type and obtaining a solution. Example flowcharts are provided in Appendix B. Many students reported finding the flowcharts very helpful in structuring the material and preliminary results over a 2.5-year study period appear promising (see Table 7). During this time, 489 students were taught with flowcharts and 1,173 were taught without them. Note that flowcharts were not available to students during exams.

We have not conducted designed experiments on the extent to which flowchart use may improve student performance in this area, so the results seen above may confound the utility of flowcharts with other factors. Still, the notable difference between the two groups suggests that flowcharts may prove a useful tool in the introductory statistics teacher’s arsenal. Previous research has also shown that students’ analytical reasoning skills can be enhanced by the inclusion of more projects and cases that present statistical analysis as a process of inquiry with practical applications (see, e.g., Bell 2001, Cochran 2010).

8. Conclusions, Limitations, and Directions for Future Research

We examined longitudinal data on test scores in an introductory business statistics course and identified addressable difficulties through the use of cluster analysis. Rather than using the conventional statistical topic-based performance in the clusters based

Table 7 Success Rates With and Without Flowcharts

Question	Number	Success rate with flowcharts (%)	Success rate without flowcharts (%)	Relative success ratio (without/with) (%)
Identifying a binomial variable	Q26	93	71	0.76
Identifying a Poisson variable	Q29	85	73	0.86
Computing σ_x in a word problem where s is given	Q41	74	61	0.82
Constructing the confidence interval for μ with σ unknown	Q45	77	64	0.83
Checking normality of the \bar{x} distribution	Q46	96	81	0.84
Using z rather than t when constructing a confidence interval for μ with σ known	Q49	98	86	0.88
Using t rather than z when conducting a hypothesis test for μ with σ unknown	Q50	96	85	0.89
Using z rather than t for a hypothesis test of π	Q51	95	79	0.83
Checking normality of the p distribution	Q55	80	57	0.71

on categories constructed from shared characteristics in problem-solving complexity. This approach provided additional insights into statistical misconceptions and errors and revealed a previously undocumented difficulty in working with cumulative distributions and determining cutoffs for discrete distribution problems. It also brought attention to underlying problem-solving skills necessary for proper application of combinations of definitions, techniques, and concepts. Further research from this viewpoint may be helpful in understanding why reform efforts to date have not resulted in significant gains in students' statistical reasoning abilities.

The results of this study also indicate that teaching practices appear to play an important role in student achievement in introductory statistics. A preliminary review of instructors in Cluster A determined that they conduct well-organized classes, have high expectations of their students, give frequent and informative feedback, and generally have considerable teaching experience as well as in-depth knowledge of the subject area. In addition to a problem-solving approach, we also found that Cluster A instructors heavily emphasize mastery of the individual concepts and component tasks required to solve statistical problems. Although this emphasis seems to make little difference on the simplest problems, the gap between Cluster A and Cluster B mastery tended to widen (as evidenced by a lower relative success ratio) when tasks drew on two or more fundamental skills. Cluster A teachers also used more exercises that require students to determine *which* statistical tool or technique is most appropriate to a particular situation. These observations are consistent with the major conclusions in a study of outstanding college teaching described in Bain (2004).

Further investigation into the pedagogies and teaching philosophies of Cluster A faculty members is clearly warranted to better illuminate differences in cluster performance. In addition, our results should be viewed in light of the data-driven nature of the study, where patterns were located without benefit of formal experimental design and testing. It is possible that factors other than teacher (such as time of day the course is taken or the term in which the course is completed) might be confounded with the effect of the teacher, and this provides an avenue of future research. Preliminary work suggests that although there is a variation between fall and spring semesters (with somewhat stronger performance by students in fall), the difference in performance between students in one cluster and those in the other is fairly consistent.

We believe, however, that our empirical data analysis, in conjunction with an initial exploration of Cluster A instruction, demonstrates a novel way of identifying weaknesses in course content and delivery

and providing feedback on student learning outcomes in the context of a credible assessment process.

Supplementary Material

Files that accompany this paper can be found and downloaded from <http://dx.doi.org/10.1287/ited.1120.0085>.

Appendix A. Descriptions of Selected Questions

This section provides a more complete description of several assessment test questions referenced in the body of the paper. Note that some of the details have been modified because the test is still in use.

A.1. Question 11

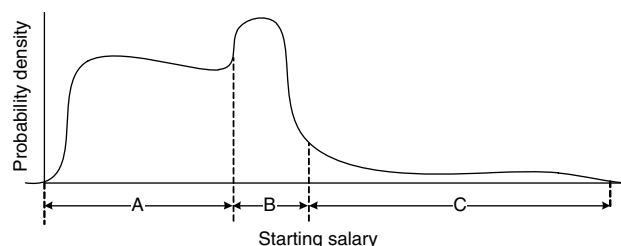
Choose which one of these five graph types would be most appropriate to investigate the relationship between *the time a student needs to complete a midterm exam* and *the time he or she requires for the final exam*: ogive, histogram, scatter plot, pie chart, box-and-whisker plot.

A.2. Questions 22–24

Scott and Susan take the same five classes: Math, Econ, French, History, and English. After school today, each of them will study one and only one of these five subjects. Each student randomly selects which one of the five subjects to study, with each subject having a 20% chance of being selected. Each student keeps his or her selection secret from the other student. What is $P(\text{Susan studies Math today} \mid \text{Scott studies Econ today})$? What is $P(\text{Scott studies Math today and Susan studies Math today})$? What is $P(\text{Scott studies Math today} \mid \text{Scott studies Econ today})$?

A.3. Questions 32–33

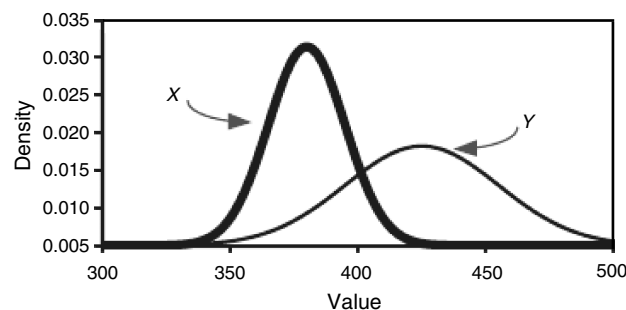
Consider this graph showing the continuous pdf for the salary of a newly hired employee:



Which range (A, B, or C) is most likely to contain the salary of a randomly selected newly hired employee? Which range corresponds to the highest starting salaries?

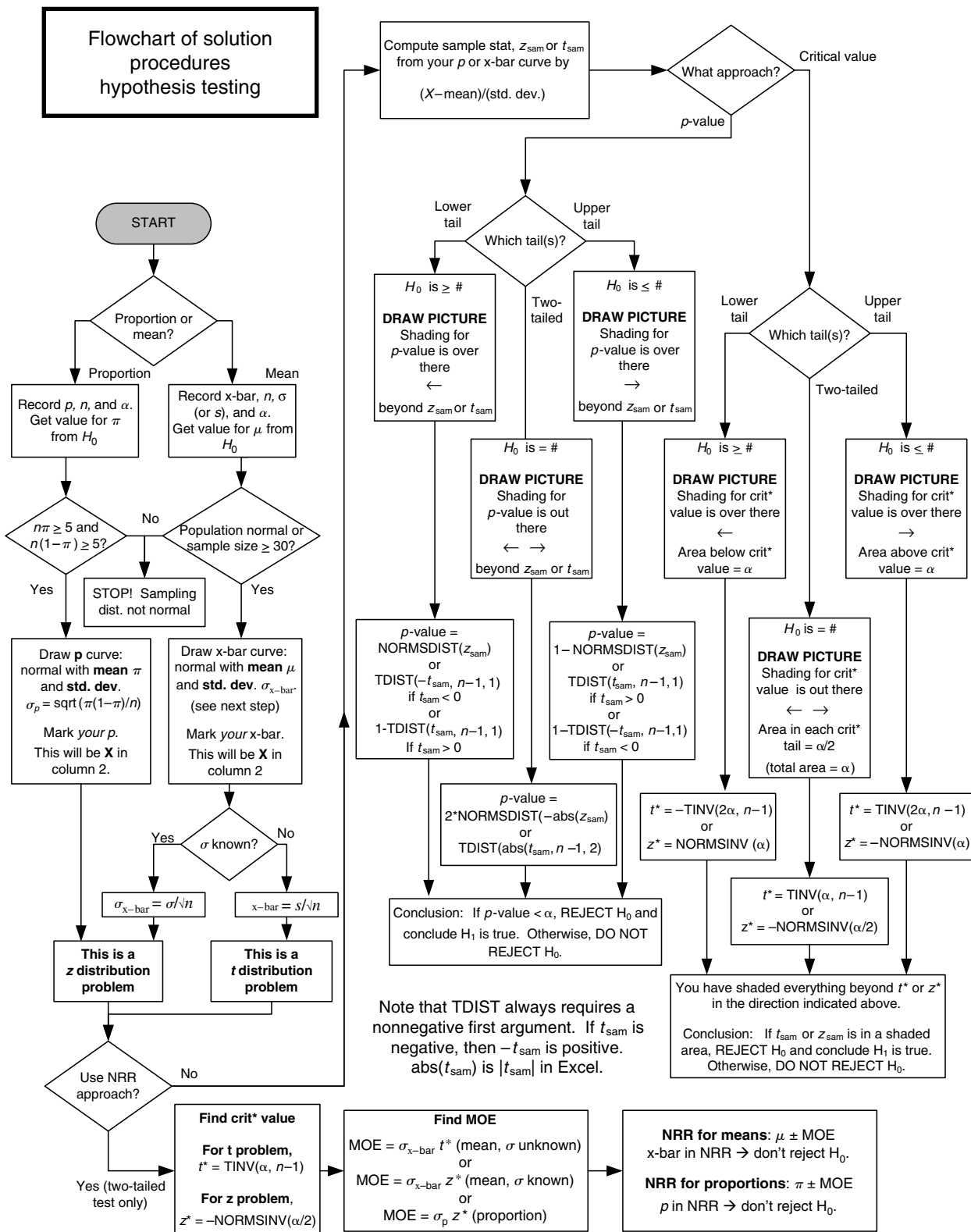
A.4. Question 34

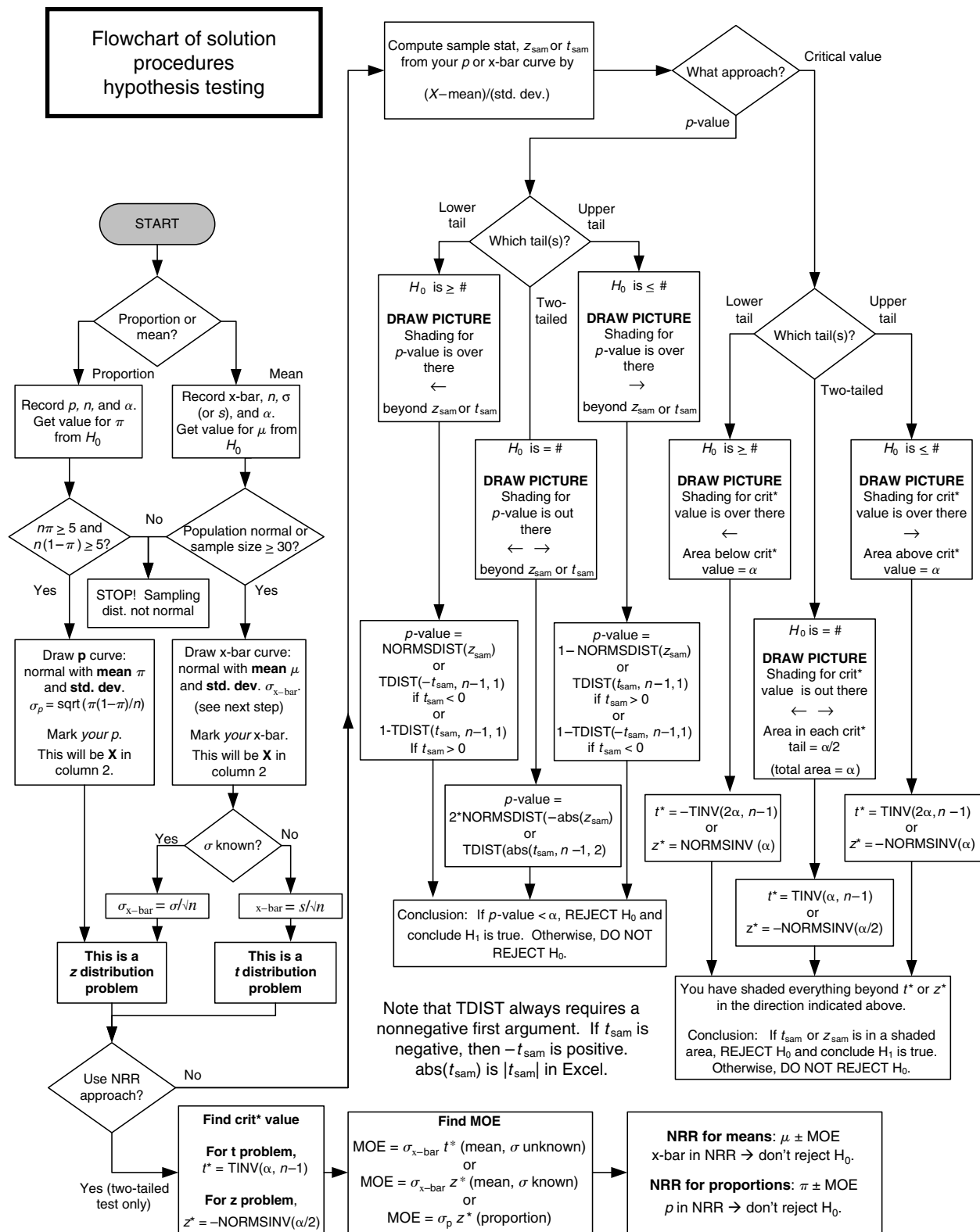
Consider the distributions of two normally distributed random variables X and Y shown below:



Compare the means and standard deviations of X and Y .

Appendix B. Flowcharts





References

- Albert, J. H. 2003. College students' conceptions of probability. *Amer. Statistician* 57(1) 37–45.
- Bain, K. 2004. *What the Best College Teachers Do*. Harvard University Press, Cambridge, MA.
- Bell, P. C. 2001. Teaching business statistics with Microsoft Excel. *INFORMS Trans. Ed.* 1(1) 18–26. <http://ite.pubs.informs.org/>.
- Ben-Zvi, D., J. Garfield. 2004. Research on reasoning about variability: A forward. *Statist. Ed. Res. J.* 3(2) 4–6.
- Brightman, H. 1977. An evaluation of two models of undergraduate statistics. *Decision Sci.* 8(1) 287–299.
- Brightman, H., M. Broida. 1975. On problem solving, motivation, and statistics. *Amer. Statistician* 29(4) 164–166.
- Chance, B., R. delMas, J. Garfield. 2004. Reasoning about sampling distributions. D. Ben-Zvi, J. Garfield, eds. *The Challenge of Developing Statist. Literacy, Reasoning, and Thinking*. Kluwer, Dordrecht, The Netherlands, 295–323.
- Clayton, H. R., C. S. Sankar. 2009. Using spreadsheets to enhance learning in the affective domain for undergraduate statistics students. *INFORMS Trans. Ed.* 10(1) 10–17. <http://ite.pubs.informs.org/>.
- Cochran, J. J. 2010. All of Britain must be stoned! An effective introductory probability case. *INFORMS Trans. Ed.* 10(2) 62–64. <http://ite.pubs.informs.org/>.
- Dambolena, I. G., S. E. Eriksen, D. P. Kocso. 2009. An intuitive introduction to hypothesis testing. *INFORMS Trans. Ed.* 9(2) 53–62. <http://ite.pubs.informs.org/>.
- delMas, R., J. Garfield, A. Ooms, B. Chance. 2007. Assessing students' conceptual understanding after a first course in statistics. *Statist. Ed. Res. J.* 6(2) 28–58. <http://www.stat.auckland.ac.nz/serj>.
- Dinov, I. D., N. Chistou, J. Sanchez. 2008. Central Limit Theorem: New SOCR applet and demonstration activity. *J. Statist. Ed.* 16(2). <http://www.amstat.org/publications/jse/v14n1/dinovr.html>.
- Evans, D. L., G. L. Gray, S. Krause, J. Martin, C. Midkiff, B. M. Notaros, M. Pavelich, D. Rancour, T. Reed-Rhoads, P. Steif, R. Streveler, K. Wage. 2003. Progress on concept inventory assessment tools. 33rd ASEE/IEEE Frontiers Ed. Conf. Boulder, CO.
- Garfield, J. B. 1991. Evaluating students' understanding of statistics: Development of the statistical reasoning assessment. *Proc. Thirteenth Annual Meeting of the North Amer. Chapter Internat. Group Psych. Math. Ed.* Vol. 2, Blacksburg, VA, 1–7.
- Garfield, J. B. 2000. Evaluating the impact of educational reform in statistics: A survey of introductory statistics courses. Final Report for NSF Grant REC-9732404. <http://www.cehd.umn.edu/edpsych/Projects/Impact.html>.
- Garfield, J. B. 2002. The challenge of developing statistical reasoning. *J. Statist. Ed.* 10(3). <http://www.amstat.org/publications/jse/v10n3/garfield.html>.
- Garfield, J. B., A. Ahlgren. 1988. Difficulties in learning basic concepts in probability and statistics: Implications for research. *J. Res. Math. Ed.* 19(1) 44–63.
- Garfield, J. B., R. delMas, B. Chance. 2002a. The assessment resource tools for improving statistical thinking (ARTIST) project. NSF CCLI grant ASA-0206571. <https://app.gen.umn.edu/artist/>.
- Garfield, J. B., B. Hogg, C. Schau, D. Whittinghill. 2002b. First courses in statistical science: The status of educational reform efforts. *J. Statist. Ed.* 10(2). <http://www.amstat.org/publications/jse/v10n2/garfield.html>.
- Grant, T. S., M. I. Nathan. 2008. Students' conceptual metaphors influence their statistical reasoning about confidence intervals. Working Paper 2008-5, Wisconsin Center for Education Research. <http://www.wcer.wisc.edu/publications/workingPapers/papers.php>.
- Hall, M. R., G. H. Rowell. 2008. Introductory statistics education and the National Science Foundation. *J. Statist. Ed.* 16(2). <http://www.amstat.org/publications/jse/v16n2/rowell1.html>.
- Haller, H., S. Krauss. 2002. Misinterpretations of significance: A problem students share with their teachers? *Methods Psych. Res. Online* 7(1). <http://www.mpr-online.de>.
- Harrington, C. F., T. Schibik. 2004. Methods for maximizing student engagement in the introductory business statistics course: A review. *J. Amer. Acad. Bus. Cambridge.* 4(1/2) 360–364.
- Hirsch, L. S., A. M. O'Donnell. 2001. Representativeness in statistical reasoning: Identifying and assessing misconceptions. *J. Statist. Ed.* 9(2). <http://www.amstat.org/publications/jse/v9n2/hirsch.html>.
- Hogg, R. V. 1972. On statistical education. *Amer. Statist.* 26(3) 8–11.
- Holmes, P. 2003. 50 years of statistics teaching in English schools: Some milestones. *The Statistician* 52(4) 439–474.
- Kaufman, L., P. Rousseeuw. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, New York.
- Konold, C. 1990. ChancePlus: A computer-based curriculum for probability and statistics. Final Report to the National Science Foundation Scientific Reasoning Research Institute, University of Massachusetts, Amherst.
- Love, T. E., D. K. Hildebrand. 2002. Statistics education and the Making Statistics More Effective in Schools of Business conferences. *Amer. Statist.* 56(2) 107–112.
- Mills, J. D. 2002. Using computer simulation methods to teach statistics: A review of the literature. *J. Statist. Ed.* 10(1). <http://www.amstat.org/publications/jse/v10n1/mills.html>.
- Ord, J. K. 2010. Statistics in B-schools: Millstone or cornerstone? *Decision Line* 41(4) 4–13.
- Sotos, A., S. Vanhoof, W. Van den Noortgate, P. Onghena. 2007. Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Res. Rev.* 2(2) 98–113.
- Tsai, W., D. G. Wardell. 2006. An interactive Excel VBA example for teaching statistics concepts. *INFORMS Trans. Ed.* 7(1) 125–135. <http://ite.pubs.informs.org/>.
- Vallecillos, A. 2002. Empirical evidence about understanding of the level of significance concept in hypotheses testing. *Themes Ed.* 3(2) 183–198.
- Ward, J. H., Jr. 1963. Hierarchical grouping to optimize an objective function. *J. Amer. Statist. Assoc.* 48(301) 236–244.
- Watson, J. M. 2007. The role of cognitive conflict in developing students' understanding of average. *Ed. Stud. Math.* 65(1) 21–47.
- Watson, J. M., B. A. Kelly. 2007. Assessment of students' understanding of variation. *Teaching Statist.* 29(3) 80–88.
- Willingham, D. T. 2009. *Why Don't Students Like School?* Jossey-Bass, San Francisco, CA.
- Zieffler, A., J. A. S. Garfield, D. H. K. Dupuis, B. Change. 2008. What does research suggest about the teaching and learning of introductory statistics at the college level? A review of the literature. *J. Statist. Ed.* 16(2). <http://www.amstat.org/publications/jse/v16n2/zieffler.html>.

CORRECTION

In this article, “Identifying Addressable Impediments to Student Learning in an Introductory Statistics Course” by Scott P. Stevens and Susan W. Palocsay (*INFORMS Transaction on Education*, DOI:10.1287/ited.1120.0085), the acceptance date of the paper was corrected to January 2012.