A Unified Introduction to Inference

Daniel T. Kaplan

July 8, 2019

Abstract

The paper introduces a new approach, called "RF," to teaching inference in college-level introductory statistics. The RF approach can be used to streamline the inference curriculum and unify the various settings distinguished in the traditional curriculum. The RF approach may be especially valuable to instructors who want to include appropriate methods for incorporating covariates into statistical inference. The RF approach is appropriate even for instructors who avoid using computing in their courses, but can also be used by instructors who embrace computing for multiple regression or for a bootstrapping/simulation pedagogy.

1 Introduction

A familiar and widespread approach to teaching statistical inference in a first university-level course refers to a series of statistics and their standard errors in several specific contexts: the difference between two proportions, the difference between two means, and the slope of a regression line. Prior to engaging these settings, the concept of a standard error is introduced in the context of a single sample proportion or sample mean.

I'll refer to the familiar curriculum as the "SE curriculum," as it is so strongly oriented to Standard Errors. This paper presents a new and different perspective on introducing inference that unifies the various settings of the SE curriculum, potentially allowing instructors to streamline their courses and thereby extend the range of topics to include multivariable settings, such as recommended by the 2016 GAISE College report. [Cite: GAISE]

The challenges faced by students in the SE curriculum are well known to instructors: the formulas for the standard errors are complicated; the formulas are different for the various settings yet similar enough to be mistaken for one another; the connection of the formulas to the underlying idea of sampling variation is not obvious; the underlying concept of a sampling distribution is nuanced and difficult to assimilate; and cognitive load is imposed by the repeated use of the words sample, standard, error, statistic in the vocabulary, as in the phrases "standard deviation," "standard error," "margin of error," "sample statistic," "sampling variation," and "test statistic."

Inference topics that follow the SE settings, chi-squared test and ANOVA, drop the use of standard errors entirely and have little connection to the earlier settings. This creates additional potential for confusion and draws undue attention to the p-value, which is the one quantity that appears in all of the SE inferential settings.

A recent trend is to unify the procedures of inference across the various settings by basing them on a couple of conceptually simple operations: resampling and shuffling. (See Lock5, Tintle.) This simulation-based approach is intrinsically rooted in computing that cannot practicably be done on a calculator or in a spreadsheet. Instead, the computing is done using professional-level software packages (such as R) or custom-built, interactive web apps. (See software sites for Lock 5, Tintle.)

This note describes another way of unifying inference procedures – which I call the RF approach – by adopting a standard graphical presentation and making calculations of confidence intervals and p-values without explicit reference to the standard error. Section 2 describes the graphical presentation and how it accomodates each of the inference settings in the conventional approach. Section 3 shows how confidence intervals and p-values can be calculated in a straightforward way without computing a standard error. Section 4 demonstrates how the RF approach can be used by instructors who eschew computing in their classes. Section 5 anticipates and addresses several possible criticisms of the new approach.

2 *A unifying graphical presentation*

To start, it helps to make a small change in nomenclature. In the conventional approach to inference, the phrases "response variable" and "explanatory variable" or their equivalent are used with simple regression, while the difference-between-means and difference-between-proportions settings refer to "two samples," even though there is really one sample with two variables: a response variable and a dicotomous explanatory variable that defines the two groups whose means or proportions are being compared. I use "response" and "explanatory" for all settings. (Later, I will use "covariate" to stand for a second explanatory variable.)

The unifying graphic is a point plot of the response variable versus the response variable. The various conventional inference settings differ in whether the response and explanatory variables are numerical or dicotomous categorical variables:

a. Difference between means: response is numeric, explanatory is dicotomous categorical.

b. Difference between proportions: response is dicotomous categorical as is the explanatory variable.
c. Simple regression: response is numeric as is the explanatory variable.

In principle, there is a fourth possibility:

d. Binary regression: response is dicotomous categorical, response is numeric.

Setting (d) is not found in the SE curriculum, but appears naturally in the RF approach.

In contemporary statistical graphics, (e.g. [cite grammar of graphics, ggplot2 book]) a categorical variable in a point plot is represented by "dodging": each level is assigned a discrete position on the coordinate axis. Other useful techniques in forming the point plot are "jittering" and transparency. Jittering displaces each point by a small random perturbation from its assigned discrete position. Jittering and transparency can be used together to avoid overplotting one data point on another, thereby making it clear to the eye the density of points at each location.

Using dodging, jittering, and transparency as needed, the four inference settings above can be effectively shown in a point plot, as shown in Figures 1 a-d, which displays some numerical and categorical variables from the `mosaicData::CPS85` data frame. [CITE mosaic Data and give link to direct URL of CSV]

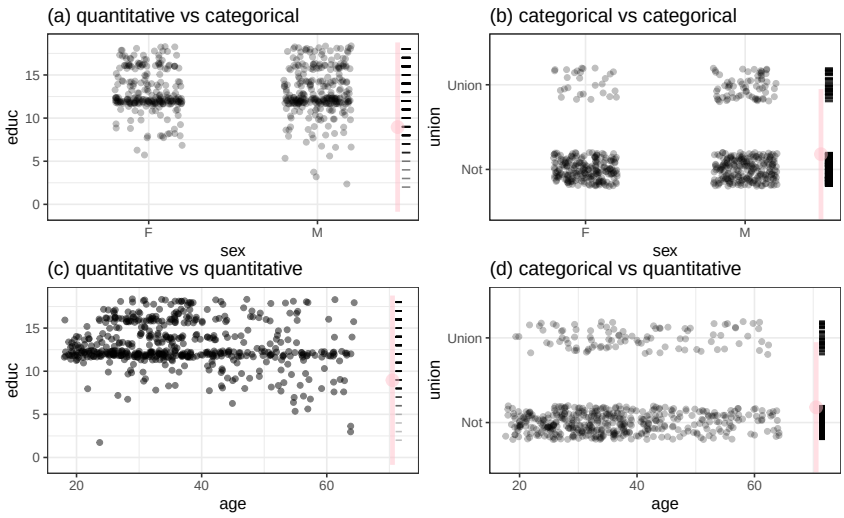

Figure 1: Figure 1. Point plots in various settings for inference.

In addition to the point plot itself, each graphic includes a rug plot of the values of the response variable along with a point-range bar

showing the mean ± 2 standard deviation of those values. This summary of the response will be used later in the inference calculations.

Another change in nomenclature helps to unify the sample statistics in the various SE curriculum inference settings. Rather than referring to groupwise means, groupwise proportions, and slopes, we'll display "model values" of the response variable as a function of the explanatory variable. The model value for a particular data point corresponds to the mean response for that point's group or the value of a regression line at that point's explanatory value. The corresponding plot layer (Figure 2) is similar in format to the data plot.

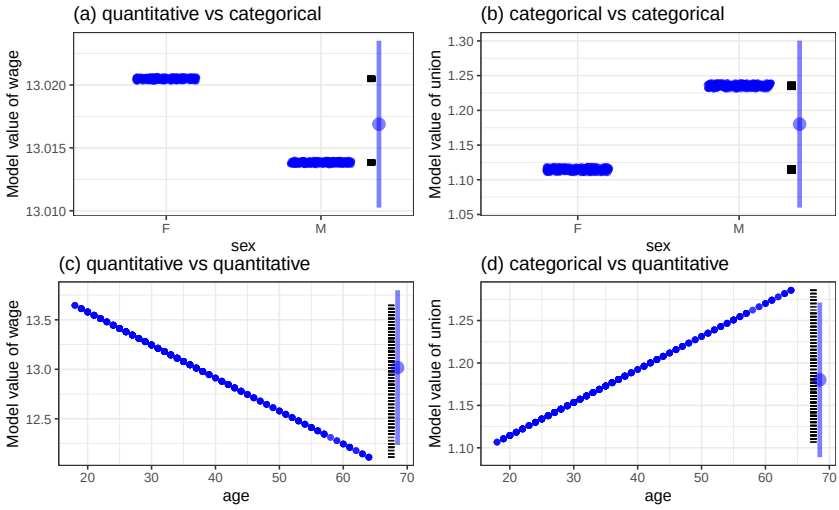

Figure 2: Figure 2. Model values in four settings for inference.

For a dicotomous response value, the model values are calculated using a 0-1 coding. In this coding, means are equivalent to proportions.

A striking feature of Figure 2 is that there are only two "shapes" for the plot, even though there are four settings for the models. The reason is that the vertical axis is being automatically scaled to the range of the model values. Every plot of model values will have one of these two shapes – two horizontal bars for a categorical explanatory variable and a line for a numeric explanatory variable, independent of the response variable. The only variations are whether the slope of the line is positive or negative (or zero), or whether the left bar is higher or lower (or the same level) as the right bar.

It is only when the model-value plot is overlaid on the same scale as the data plot that the regression contexts become apparent, as in Figure 3, which I will call the "standard presentation."

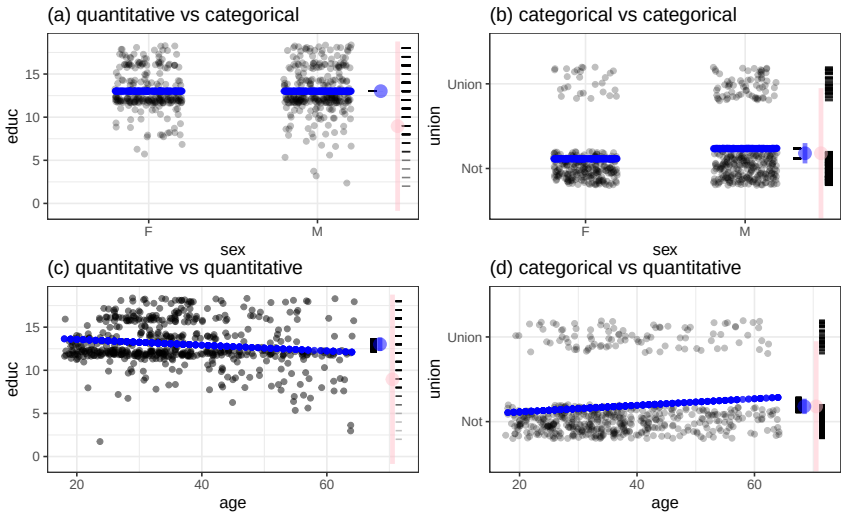

Figure 3: Figure 3. Showing both data and model values.

3 *F is for inference*

Several descriptive statistics can be visualized from the standard presentation: groupwise means and proportions and their differences, the slope of a regression line, the standard deviation of the response variable and of the model values. These, together with the sample size n are the basic inputs to the calculations for formal inference.

For notation, I'll use B to stand for the (unstandardized) effect size. B corresponds to the difference in means, the difference in proportions, or the slope of the regression line, depending on the context of the problem. The sample standard deviation of the response variable will be denoted as s_{raw} while the standard deviation of the model values will be s_{model}.

A basic inferential statistic applicable to all settings is the coefficient of determination, R^2, which is equal to the square of the ratio of the model and raw standard deviations:

$$\text{Eq. 1} \qquad R^2 = s^2_{model}/s^2_{raw}.$$

An important inferential quantity based on R^2 and sample size n is the F-statistic. For the settings (a) through (d), which all have one degree of freedom, the F statistic given in Eq. 2. (When there is more than one degree of freedom, see Eq. 4.)

$$\text{Eq. 2} \qquad F = (n-1)\frac{R^2}{1-R^2}.$$

The 95% confidence interval on the effect size B is simply expressed by Eq. 3.

Eq. 3 $$95\% \text{ confidence interval} = B(1 \pm 2/\sqrt{F}).$$

The 2 in Eq. 3 corresponds to the traditional 1.96 for 95% confidence from the normal distribution.

In the SE curriculum, a variety of test statistics is used: z for the difference in proportions, t for the difference in mean or regression slope, F for ANOVA and multiple regression. Given these different scales, it's sensible to summarize the results of a null hypothesis test with a common scale: the p-value which puts all results on a scale of 0 to 1.

In the RF approach, F is used for all inference settings. Following the traditional practice, F can be translated to a p-value can be determined by reference to a table, software, or simply the graph in Figure 6. There are $n - 1$ degrees of freedom. On the other hand, since F is used for all of the inference tests, it's possible to use F directly as the measure of implausibility of the null hypothesis. F > 4 corresponds to $p < 0.05$ while F > 7 corresponds to $p < 0.01$.

The graphical format of situations (a) through (d) is readily generalized to include multiple levels for the categorical explanatory variable or for multiple explanatory variables. As an example, Figure 4 shows wage modeled by sector of the workforce and sex. The lengths of the s_{raw} and s_{model} bars give R ≈ 0.5.

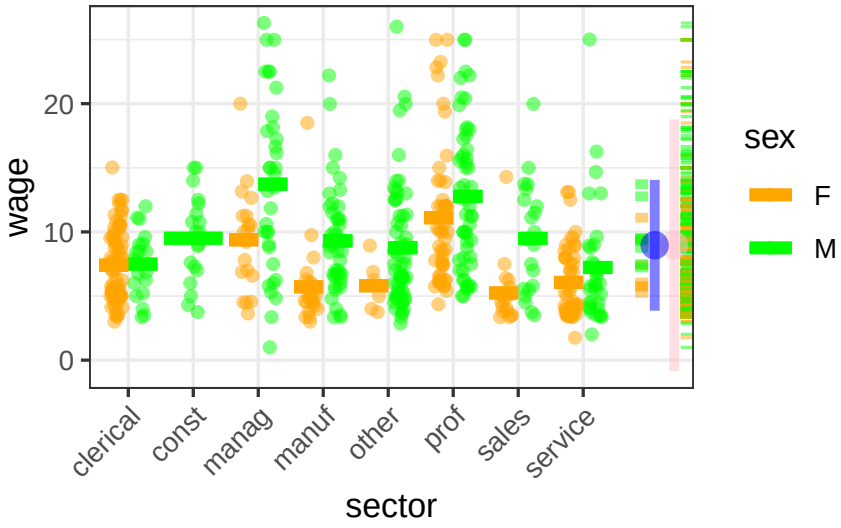

Figure 4: Figure 4: The standard graphic with two explanatory variables.

The form of F changes when there the multiple degrees of freedom in the model. For example, when there are k groups defined by the

categorical variable, F is given by Eq. 4.

$$\text{Eq. 4} \qquad F = \frac{n-(k-1)}{k-1}\frac{R^2}{1-R^2}.$$

For multiple degrees of freedom, whole-model p-values are calculated directly from F. However, there is no simple F-based formula such as Eq. 3 for confidence intervals on the multiple-regression model coefficients or individual group means.

4 *A graphical pedagogy*

The primary goals of the RF approach are:

1. to simplify the topic of inference while keeping it authentic by reducing the several settings of the traditional approach to a single setting that can be handled by a single graphical format.
2. to smooth the path for students to work with two or more explanatory variables.

The graphic at the heart of the RF approach can be constructed with statistical software (such as R) or by a simple web application that requires no installation of software and no experience with code. (A draft of such a web app is available at `https://dtkaplan.shinyapps.io/LA_explain/`.)

A distinctive feature of lower-level university mathematics and statistics is that many instructors largely eschew modern computing in favor of a calculator. Among the justifications given for this practice are: 1. the lack of availability of computing infrastructure beyond calculators; 2. the belief – right or wrong – that drill with hand computation of quantities such as the mean and standard deviation informs a student's understanding of these quantities; 3. using calculators helps to avoid students cheating on exams, since calculators generally lack the communications capabilities of computers.

One consequence of the exclusive use of calculators for teaching inference is that the formulas of the SE curriculum are seen as essential components of statistical procedure, since the value of those formulas can be worked out by plugging in numerical estimates of simple, "sufficient" statistics such as the mean and standard deviation.

The RF approach to inference should be particularly attractive to those who avoid computing in their statistics classes. The standard graphic of the RF approach (see Figure 3) can be generated by a web app or can be presented to students in a printed format.

With a little practice, the value of $R \equiv \sqrt{R^2}$ can be read by eye directly from the graphic by comparing the length of the point-line

marker for the model values to that of the raw values of the response variable.

The remaining calculations involved in statistical inference can also be accomplished by eye, using graphs that encapsulates the formulas and probability tables of the SE curriculum. Three such graphs may be particularly useful:

1. Calculating the value of F from R and n. (Figure 5)
2. Computing from F and n confidence intervals of the effect size B. (Figure 6)
3. Calculating p-values from F and n. (Figure 7)

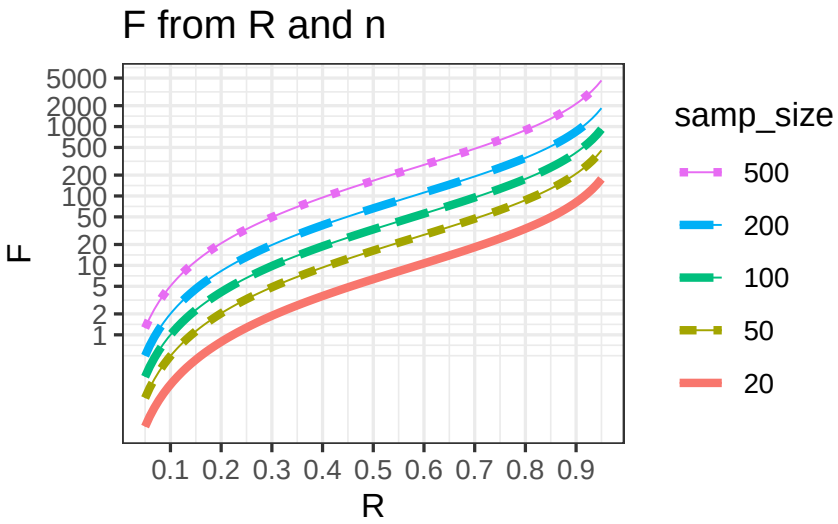

Figure 5: Figure 5. Computing F from R and n

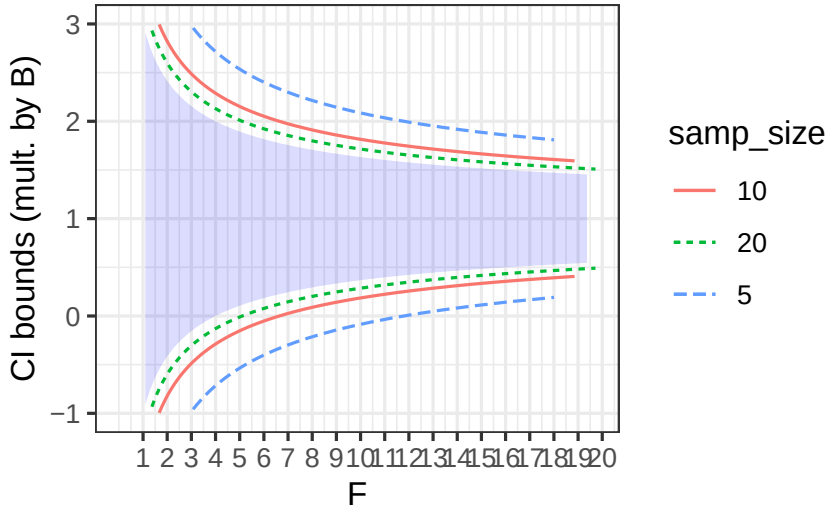

Figure 6: Figure 6: Confidence intervals from F. Multiply the effect size B by the upper and lower bounds at the appropriate F value. The central blue band is for large n. Bounds are also shown for several small n.

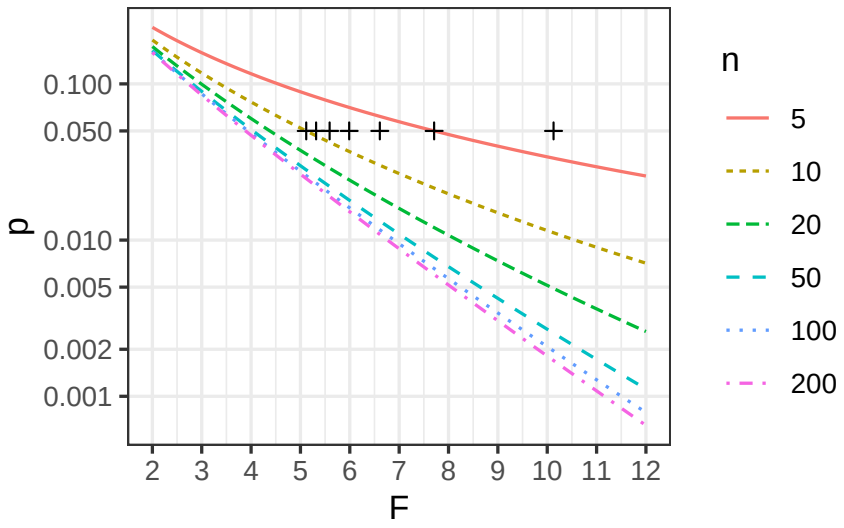

Figure 7: Figure 7. p-values from F and n. 0.05 critical values are shown for n of 10 and fewer.

These three graphs and an example from Figure 3 can be placed on the two sides of an ordinary index card and provided to students as the computing infrastructure needed for inference.

5 *Pros and cons*

The traditional approach to introductory inference was developed over many decades. Scores of textbook authors have had the opportunity to contribute their improvements and perspectives. The most recent major innovation is the use of randomization procedures to introduce inference. [Cite Lock, Tintle, Open Intro]. It's a truism that almost everyone teaching statistics who has significant formal training in that field has seen and mastered, to the extent possible, the SE curriculum. A very substantial fraction of instructors teaching introductory statistics, including this author, were not trained in statistics and learned the SE curriculum from the textbook used in class.

It's natural that people who are familiar with the SE curriculum think about the underlying statistical problem in those terms. For these people, any substantial deviation from the tried and true will seem initially more difficult. In addition, there are components of the RF approach that will be unfamiliar to the many instructors whose training consists of teaching an introductory course. For example, R^2 is not encountered in a many textbooks' chapters up through simple regression. Instead the emphasis is on the Pearson product-moment correlation coefficient, r. Many textbooks have a summation formula process for calculating F; the calculation via R^2 is not seen in many introductory textbooks. Of dozens of statistics educators to

whom I've demonstrated the RF approach of calculating confidence intervals from F, none had any initial idea that this was possible. (It helps to remind instructors that, with one degree of freedom in the numerator, F = t^2. And to point out that t is the effect size B divided by the standard error. Thus, F is linked to the standard error and the confidence interval.)

Our students initially have no such mastery of the SE curriculum. Properly judging the RF approach's difficulty for students must necessarily involve trying it in the classroom. Encouraging instructors who see face validity to the RF approach to try it in the classroom is the major purpose of this article.

As a mathematical object, there's nothing about the SE curriculum that demands fixing. But the mathematical object was developed to help researchers deal with contemporary problems, and the nature of contemporary problems has changed substantially in the many decades since the SE curriculum was developed. In this regard, it's worth noting that approaches to inference outside of the tradition, e.g. statistical/machine learning, are widely taught with an entirely non-traditional infrastructure, for instance, cross-validation. [cite: Machine era statistical inference.]

Putting aside the objections to the RF approach that will necessarily be fostered of lack of experience with it, I now consider some potential statistical-theory related criticisms.

First, it's not unreasonable to see the standard error as the center of statistical inference, at least for the methods likely to be encountered in intro stats. With this view, isn't by-passing the standard error a disadvantage of the RF approach? Yes, in the sense that it's generally better to know more than less. But teaching the standard error comes with its own costs. Students can be confused by the similar sounding terms "standard deviation" and "standard error." For most purposes, the standard error is only an intermediate result for calculating a confidence interval or a t statistic, and adds a bit of complexity to the overall process. In any event, the standard error is easily introduced in the RF context via the simple formula $\text{se} = B/\sqrt{F}$.

Second, using F rather than t takes the rug out from under the traditional topic of one-tailed versus two-tailed tests. But one-tailed tests are controversial and easily mis-used. Recently, the American Statistical Association has stated an interest in de-emphasizing p-values as an instrument of inference. [CITE TAS] A lead editorial in this journal (TAS) stated:

> *We conclude, based on our review of the articles in this special issue and the broader literature, that it is time to stop using the term "statistically significant" entirely. Nor should variants such as "significantly different," "$p < 0.05$," and "nonsignificant" survive, whether expressed in words, by*

asterisks in a table, or in some other way.

In this context, it hardly seems worthwhile to encourage students to distinguish between 0.10 and 0.05 as particular values of an arbitrary threshold. Besides, there are much more important factors at work in constructing a meaningful p-value that are not covered quantitatively in introductory statistics, e.g. the problems of "researcher degrees of freedom," multiple testing, and covariates.

Third, no room is given in the RF approach for calculating confidence intervals and p-values in the so-called "one-sample" setting. ($R^2 = 0$ in the one-sample settings) But these settings can be taught in other non-traditional ways, for instance by bootstrapping. And, insofar as one-sample settings are introduced to lead students to "two-sample" and other inference procedures involving explanatory variables, this is not a central loss.

Fourth, using graphs for calculating p-values (Figure 7) and bounds of confidence intervals (Figure 2) does not result in sufficient precision. The shorthand of using 2 for what should be z* or t* dramatically understates the width of the confidence interval and overstates the p-value for $n \lessapprox 10$. In response, I offer two suggestions. 1. It's misleading to present a p-value as a precise value, since we know there is a huge amount of sampling variation in it. [Cite: Statistical significance is not statistically significant] 2. Arguably the dominant interest in data in today's world is in large n. 3. The presentation in Figures 6 and 7 actually does a good job in showing the situation for small n. Figure 6 can easily be augmented to show individual curves for whatever values of n are desired. Figure 6 already does this for the 0.05 critical values.

Fifth, for the "difference of two probabilities" (the setting of Figure 2b) it's conventional to use a z distribution for inference. In contrast, the RF approach effectively uses a t distribution. Still, for the range of n commonly encountered in examining the difference between two probabilities, the t distribution closely approximates z. Indeed, using a critical value of 2 for all the settings suffices even when $n \gtrapprox 10$.

Sixth, the use of linear regression in a setting with a dicotomous response variable and continuous explanatory variable (setting (d) in Figure 2) fails to impose the natural constraint that probabilities must be between zero and one. Proper alternatives are readily available, such as logistic regression, in which log odds rather than raw probabilities are used. Similarly, there are appropriate statistical/machine-learning classifier techniques such as linear and quadratic discriminant analysis and support vector machines. Still, I think that linear regression provides a valuable introduction to students and helps point out the need for more advanced techniques. The pedagogy of

starting with linear models in order to progress to logistic regression has been used for at least a decade. [Cite: Statistical Modeling]

Seventh, the chi-squared test is not incorporated in the RF approach. True, but there's nothing to prevent an instructor from continuing to teach chi-squared in the traditional way alongside the RF approach. And, insofar as many traditionally chi-square examples involve two categorical levels in one of the variable, a modeling approach in the style of Figure 3 (b) & (d) may be more appropriate, since it produces an effect size in addition to a p-value.

6 Conclusion

Whether it makes sense to use the traditional SE curriculum or the RF approach, depends on the instructor's priorities. The SE curriculum was developed in the context of small n and costly computation. But it does not generalize to the use of multiple explanatory variables or even to multiple levels of a single categorical explanatory variable.

The RF procedure unifies and generalizes well to inferential settings with covariates. RF imposes a somewhat heavier computational load (to calculate model values), although a graphical approach to pedagogy allows closely approximate results to be constructed with calculations by eye.

Both the SE and the RF approaches can be streamlined considerably by putting aside situations with small n and the consequent need for tables of critical values. In addition, the RF approach obviates any consideration of dubious aspects of traditional inference such as one-tailed p-values or the spurious precision of the "unequal variance" t-test. RF, being so closely related to ANOVA, also provides a meaningful basis for introducing the comparison of multiple models, something absent from the SE curriculum.