Original Article

Statistics using just one formula

Jeffrey S. Rosenthal
Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada
e-mail: jeff@math.toronto.edu

Summary This article advocates that introductory statistics be taught by basing all calculations on a single simple margin-of-error formula and deriving all of the standard introductory statistical concepts (confidence intervals, significance tests, comparisons of means and proportions, etc) from that one formula. It is argued that this approach will better emphasize the core concepts and commonality of statistics and range of statistical applications, without forcing the students to memorize lots of different equations.

Keywords: Teaching statistics; Introductory statistics; Margin of error; Confidence interval; Significance test; Comparison of means.

INTRODUCTION

In our data-rich world, statistical analysis and education are more important than ever. I recently developed (http://probability.ca/sta130report) a new introductory statistics course, in which I tried to give the students a broad overview of the subject: a bit of probability theory, some discussion of *p*-values, calculation of confidence intervals, use of statistical software, applications to real data problems, statistical writing and communication, etc. The course was reasonably successful (mean student rating 4.3 out of 5, with many positive comments), but some students found the derivations of formulae *boring* and the calculations *tedious*. Then, at a recent social event, I met a woman who complained (as many do) about her own statistics course from her student days, lamenting all the equations she had to memorize – and I feared that some of my students might feel similarly. This led me to wonder, can the basic ideas of statistics be communicated reasonably, in a way that can actually be understood and used for applications, but with fewer equations and formulae and calculations?

Such concerns are not new. Traditionally, statistics was introduced with fairly mathematical courses involving lots of formulae, but even in the simplest cases, the individual formulae can look very different especially to less quantitatively inclined students and thus cause confusion and negativity and despair. In recent years, statistics has sometimes instead been taught in ways that do not use equations at all, concentrating instead on data analysis and the use of statistical computing software, but this risks giving the impression that statistics is a strange mystery that only computers can understand.

My opinion now is that we shouldn't totally eliminate formulae. In the end, I decided that most simple statistical inference problems can be solved effectively using essentially just a single formula, as I now explain. Introducing a single common approximate equation for confidence intervals for the simplest cases of one and two means and proportions, we can provide a solid foundation of basic statistical concepts with less entanglement in formula details.

SET-UP

Much of statistics involves taking a *sample* of measurements of some quantity and attempting to draw inferences about its true underlying *average* (or *mean*) in the entire population or general situation. For example, perhaps we sample some men's heights and wish to infer the average height of all men. Or perhaps, we measure the effect of a new medication on the blood pressure of a sample of patients and wish to draw conclusions about its average effect on everyone.

To that end, suppose we have a random sample of n different measurements of some quantity. Then, we can *estimate* the true underlying average by computing the average of our sample, but this estimate probably won't be *exactly* correct, due to the randomness of the sample. (Just like if you flip 100 coins, you probably won't get *exactly* half heads.) So, the question becomes, how close will our sample average probably be to the true underlying average? Or, to put it differently, how accurately can we estimate the true underlying average based only on our sample?

THE ONE FORMULA

The one formula that we shall use is as follows. Suppose $x_1, x_2, \ldots, x_n$ is a sample of n measurements of some quantity, and we estimate the true underlying average by our sample average (or *point estimate*) $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Then usually, i.e. about 95% of the time, i.e. about 19 times out of 20, $\overline{x}$ will be within the *error* of about $2\sqrt{v/n}$ of the true underlying average, where $v = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$ is the sample's average squared deviation from its average. Equivalently, about 95% of the time, the true average will be somewhere within the *95% confidence interval* given by $\left[\overline{x} - 2\sqrt{v/n}, \overline{x} + 2\sqrt{v/n}\right]$.

We shall see that this one approximate formula is all that we need to make lots of fairly accurate statistical inferences, for four basic one-sample and two-sample situations for both means and proportions, quickly and all at once.

APPROXIMATION?

Like most of the statistics, the above formula involves some *approximations*, as I now discuss. (It is useful for the instructor to keep these issues in mind – although I recommend *against* sharing them with the students right away as I will explain.)

First of all, v is an approximate estimate of the *variance* of the distribution of x_i. In fact, statisticians usually divide by $n - 1$ instead of n, which makes v an *unbiased* estimator and the pivotal statistics a t-statistic, although I find that this tends to confuse students and can actually make the estimate worse in some ways (see e.g. my article (http://probability.ca/varmse)), and in any case, it makes little difference if n is large.

Once we accept v as the variance, it then follows (with no further approximation) that the sample average $\overline{x}$ has variance v/n. Also, the *expected* value of $\overline{x}$ always exactly equals the true mean, say μ. It then follows that $(\overline{x} - \mu)/\sqrt{v/n}$ has mean 0 and variance 1.

Furthermore, for sufficiently large n, it follows from the Central Limit Theorem that $(\overline{x} - \mu)/\sqrt{v/n}$ has approximately a *standard Normal distribution*. Once we accept that, then that quantity is 95% likely to be between about -1.96 and 1.96, and for simplicity, we further approximate this *multiplier* 1.96 by 2. Our above formula then follows from this.

As a further refinement, the $(\overline{x} - \mu)/\sqrt{v/n}$ actually has (assuming the normal approximation) a t distribution, not a Normal distribution, which means that the multiplier 1.96 should actually be slightly larger in a way that depends on the sample size n, e.g. if $n = 10$, it's 2.26; if $n = 20$, it's 2.09; and if $n = 100$, it's 1.98. Also, these multipliers all correspond to the 95% confidence level and should be adjusted for other levels, e.g. for a 99% confidence level, the multiplier 1.96 should instead be 2.58.

However, in my opinion, it is not necessary to discuss any of these approximation issues with the students until later on (see 10). For an initial introduction, I think it is sufficient to stick to the simple formula given above with multiplier 2, to emphasize the core concepts and commonality of approach.

EXAMPLE: BABY WEIGHTS

Ten babies born in a hospital in North Carolina were measured (http://www.math.hope.edu/swanson/data/nc200.txt) to have the following weights, in pounds: $x_1 = 9.88$, $x_2 = 9.12$, $x_3 = 8.00$, $x_4 = 9.38$, $x_5 = 7.44$, $x_6 = 8.25$, $x_7 = 8.25$, $x_8 = 6.88$, $x_9 = 7.94$ and $x_{10} = 6.00$.

For these data, $n = 10$, and we compute that $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \doteq 8.11$ lb, and then $v = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \doteq 1.362$. On the basis of these data, we can be 95% confident that the true average baby weight in North Carolina is within $2\sqrt{1.362/10} \doteq 0.74$ of 8.11, i.e. that it is between $8.11 - 0.74 = 7.37$ and $8.11 + 0.74 = 8.85$ lb.

That is, our 95% confidence interval for the true average weight (in pounds) of babies in North Carolina is $7.37, 8.85$.

STATISTICAL SIGNIFICANCE

A claim based on a sample is traditionally called *statistically significant* at the 5% level if we are 95% confident that it holds for the whole population, i.e. if it is probably a genuine result rather than just an artefact due to the random luck of the sample. There are many specialized ways of computing significance levels in various settings. But a simple rule in our case is that a claim is statistically significant if it holds for all values in the confidence interval.

In the baby weight example, our sample average was 8.11 lb, which is certainly more than 8 lb. However, the claim that the true average baby weight is over 8 lb is *not* statistically significant, since the confidence interval includes weights lower than that. That is, our sample average being above 8 lb could have been just due to luck. On the other hand, the claim that the true average baby weight is over 7 lb is a statistically significant conclusion, since it holds for all values in the confidence interval.

Now, the usual assessment of statistical significance involves *hypothesis tests* and *p-values* and *probability tables* and so on. However, for the simple inference problems considered here, the usual statistical significance is exactly *equivalent* to the simple notion of 'holds for all values in the confidence interval' presented here. That is, these significance tests are still subject to the various approximations as in 4, but no more: they give exactly the same answers as the more traditional significance tests when using the same approximations. And, just as in 4, different confidence levels (besides 95%) can be achieved by adjusting the multiplier value – although again, I recommend against raising these issues with the students right away.

PROPORTIONS

An important special case is when each data value x_i is either 1 or 0, corresponding to a Yes/No outcome like winning/losing a game, or agreeing/disagreeing in a public opinion poll. In that case, $\overline{x}$ is simply the *fraction* (or *proportion*) of Yes outcomes. Also, since the sample has $n\overline{x}$ values which equal 1, and $n - n\overline{x}$ values which equal 0, $v = \frac{1}{n}\left[n\overline{x}(1-\overline{x})^2 + (n - n\overline{x})(0-\overline{x})^2\right]$, which reduces to simply $v = \overline{x}(1-\overline{x})$. Hence, for proportions, the true fraction is probably within about $2\sqrt{\overline{x}(1-\overline{x})/n}$ of the sample fraction $\overline{x}$. In this context, the quantity $2\sqrt{\overline{x}(1-\overline{x})/n}$ is often called the *margin of error*. (Also, it takes its maximum when $\overline{x} = 1/2$, so it is always $\leq 1/\sqrt{n}$, a useful upper bound.)

For example, a recent poll (http://www.theglobeandmail.com/news/politics/more-than-half-of-canadians-approve-of-trudeau-poll/article28210076/) claimed that *more than half* of Canadians approved of the government. The poll actually sampled 1,500 Canadians, of whom 53% replied Yes when asked if they approve. Since the sample was random, this doesn't imply that the true fraction is *exactly* 53%. But does this imply that is it more than 50%? Well, here $\overline{x} = 0.53$. So, as above, the margin of error is $2\sqrt{0.53(1-0.53)/1500} \doteq 0.026$, so our 95% confidence interval is $[0.53 - 0.026, 0.53 + 0.026] = [0.504, 0.556]$. Since all of the values in this interval are above 0.5, we can indeed conclude that the true fraction is (probably) more than half. (On the other hand, we could not conclude that it is more than 51%, since not all values in the interval are above 0.51.)

Do polling companies really use this formula? Yes indeed. The above poll claimed that 'The margin of error …is ±2.6%, 19 times out of 20'. In fact, one leading pollster provides (http://www.forumresearch.com/tools-margin-of-error.asp) a whole table of margins of error for various different sample sizes and observed proportions, each of which amounts to just plugging values into the above formula $2\sqrt{\overline{x}(1-\overline{x})/n}$. (Of course, these margins of error all assume that the sample was truly random and was not biased due to non-responses, misleading questions, dishonest answers, etc – all complicated issues that we do not address here.)

It is instructive to apply this formula to familiar situations. For example, suppose you flip n coins. How close to 0.5 will your proportion of heads be? Well, here $\overline{x}$ is near to 0.5, so the margin of error is about $1/\sqrt{n}$. If $n = 10$, this is about 0.3, so your fraction of heads will probably be somewhere in the interval $[0.2, 0.8]$. If $n = 100$, the interval becomes about $[0.4, 0.6]$. (Try it and see!) If $n = 400$, it's about $[0.45, 0.55]$, whereas if $n = 1000$, it's about $[0.47, 0.53]$, and if $n = 10{,}000$, it's about $[0.49, 0.51]$. So, the interval is indeed narrowing around 0.5, but rather slowly, due to the $\sqrt{n}$ factor in the denominator.

COMPARISON OF MEANS

Some of the most interesting statistical questions involve *comparing* two different average values, especially of the same quantity for different groups or at different times. Suppose our first sample is $x_1, \ldots, x_n$, with sample mean $\overline{x}$ and squared deviation v, and our second sample is $y_1, \ldots, y_m$, with sample mean $\overline{y}$ and squared deviation w. We are interested in the difference of the second true mean minus the first true mean. We can estimate this by the sample difference $\overline{y} - \overline{x}$. But how much uncertainty do we have?

We can answer this again using our one formula but with a slight modification. Our formula says that the first true mean is (probably) within $2\sqrt{v/n}$ of $\overline{x}$, and the second true mean is (probably) within $2\sqrt{w/m}$ of $\overline{y}$. For the difference of means, we *add* the two uncertainty quantities v/n and w/m together. That is, the difference of the true means is (probability) within $2\sqrt{v/n + w/m}$ of the sample difference $\overline{y} - \overline{x}$. With this one slight modification, our one formula applies to differences of means as well.

For example, I had my students measure (http://probability.ca/sta130/studentdata.txt) the circumference of their wrists. The $n = 39$ female students had sample mean $\overline{x} = 14.49$ (in cm) with squared deviation $v = 0.622$, whereas the $m = 41$ male students had sample mean $\overline{y} = 16.74$ with squared deviation $w = 0.947$. So, on average the male students were larger, but was this significant? Here the sample mean difference is $16.74 - 14.49 = 2.25$, with uncertainty $2\sqrt{v/n + w/m} = 2\sqrt{0.622/39 + 0.947/41} \doteq 0.40$. So, the 95% confidence interval for the true mean difference is $[2.25 - 0.40, 2.25 + 0.40] = [1.85, 2.65]$ cm. These values are all positive, so yes, the data do indicate that male students have statistically significantly larger wrists than female students on average.

COMPARISON OF PROPORTIONS

In the special case where each x_i and y_i is either 1 or 0, the above uncertainty value becomes $2\sqrt{\overline{x}(1-\overline{x})/n + \overline{y}(1-\overline{y})/m}$, and similar considerations apply.

For example, CBS News conducted a series of polls asking Americans if they supported the legalization of marijuana. In 2012, they sampled 1,100 adults and found (http://www.cbsnews.com/news/poll-nearly-half-support-legalization-of-marijuana/) that 47% said yes. In 2014, they sampled 1,018 adults and found (http://www.cbsnews.com/news/majority-of-americans-now-support-legal-pot-poll-says/) that 51% said yes. In 2015, they sampled 1,012 adults and found (http://www.cbsnews.com/news/poll-support-for-legal-marijuana-use-reaches-all-time-high/) that 53% said yes. So, does this indicate that support for legalizing marijuana was growing? That is, are these increases statistically significant?

Let's first compare 2012 and 2014. There, $n = 1{,}100$ and $\overline{x} = 0.47$, whereas $m = 1{,}018$ and $\overline{y} = 0.51$. Hence, the true fraction of Americans who support legalization in 2014, minus the true fraction in 2012, is probably within $2\sqrt{\overline{x}(1-\overline{x})/n + \overline{y}(1-\overline{y})/m}$ of $\overline{y} - \overline{x}$, i.e. within $2\sqrt{0.47(1-0.47)/1100 + 0.51(1-0.51)/1018} \doteq 0.043$ of $\overline{y} - \overline{x} = 0.51 - 0.47 = 0.04$. Thus, the 95% confidence interval for this difference is $-0.003, 0.083$. So, the difference could be as high as 8.3%, but it could also be (barely) negative. Thus, we cannot (quite) conclude a statistically significant difference between the years 2012 and 2014.

So, let's instead compare 2012 and 2015. So, still $n = 1{,}100$ and $\overline{x} = 0.47$, but now $m = 1{,}012$ and $\overline{y} = 0.53$. Hence, the true fraction of Americans who support legalization in 2015, minus the true fraction in 2012, is probably within $2\sqrt{0.47(1-0.47)/1100 + 0.53(1-0.53)/1012} \doteq 0.043$ of $\overline{y} - \overline{x} = 0.53 - 0.47 = 0.06$. Thus, the 95% confidence interval for this difference is $0.017, 0.103$. So, the difference could be as high as 10.3%, or as low as 1.7%, but all of these values are positive. Thus, this time, we can conclude a statistically significant increase in support for legalizing marijuana between the years 2012 and 2015.

WHAT NEXT?

The above approach provides useful tools for many elementary statistical analyses, including confidence intervals and significance tests for one and two means and proportions, all essentially using just one formula. Introductory statistics classes can begin with this material, and then apply it to lots and lots of interesting examples and applications and real world data, all without bogging down the students with a multitude of confusing equations.

Once this basic material is mastered and widely applied, and then there are many options for follow-up material.

One possible follow-up is to study the approximations summarized in 4 above, such as the Central Limit Theorem and t distribution and multiplier adjustments, to gain a more mathematical understanding of their derivations and assumptions and limitations, and to see how the approximations can be improved and refined in a variety of different settings as in traditional mathematical statistics classes.

Another is to reconsider statistical significance from a probability point of view, by computing actual p-values as the probability of observing an equally or greater deviation through pure chance alone under the null hypothesis of no actual effect. This can be done first for simple discrete examples, like the probability of winning five games in a row if you *really* have equal chance of winning or losing each time. It can then be done for continuous examples using normal approximations as in 4 This can then be used to demonstrate, as discussed in 6, that a result is significant at the 95% level from a probability point of view (i.e. has p-value less than 0.05), if and only if, the statement holds for all values in the corresponding 95% confidence interval – i.e. traditional statistical significance is *exactly equivalent* to the 'holds for all values in the confidence interval' notion presented herein.

Another recommended topic is *correlation*, which measures the relation between two continuous quantities: when one quantity increases, does the other quantity tend to also increase, or to decrease, or is it unaffected? Unfortunately, correlation does not quite fit into our one-formula approach, since it requires a separate formula $r = \frac{1}{(n-1)\sqrt{vw}} \Sigma_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$ for samples $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$, although that quantity can (and usually is) computed automatically by statistical software such as the free (https://cran.r-project.org) package *R*. Correlations are always between -1 and 1, with 0 indicating no (linear) relationship, and ± 1 the strongest relationships. For example, the percentage of adults who smoke (https://www.tobaccofreekids.org/research/factsheets/pdf/0176.pdf) and the average income per capita (http://www.infoplease.com/ipa/A0104652.html) for each of the 50 US states have correlation about -0.42, which is negative and suggests states with higher smoking rates tend to have smaller incomes and vice versa. Our one formula then *approximately* applies with $v = 1$ to give a 95% confidence interval $\left[-0.42 - 2/\sqrt{50}, -0.42 + 2/\sqrt{50}\right] \doteq [-0.70, -0.14]$. Since the values in this interval are all negative, we conclude that there is indeed a statistically significant negative correlation between smoking rates and average income. (One might wonder *why* this is so. That question is subtle, since *correlation does not imply causation*. In this case, it appears that the negative correlation is explained by the fact that people with less *education* tend to both smoke more and earn less.)

In a different direction, due to the very large data sets involved, much of modern statistical analysis is performed using a statistical software package such as the free (https://cran.r-project.org) package *R*. So, statistical software and applications should surely be included in any modern introductory statistics course, as well, together with (as noted) lots of interesting real world examples.

Of course, many other refinements and improvements, together with theoretical justifications and applications to many different areas, can be found in more advanced statistics courses. The simple approach herein will in turn inspire students to later study this important subject more deeply from more subtle perspectives.

In any case, I feel that the approach described herein provides a reasonable solution to most simple statistical inference problems, using essentially just a single formula as opposed to the multitude of formulae that arise in traditional statistics classes. Indeed, if I were to design an introductory statistics class again, I might well use this *one-formula* approach, and other statistics instructors might want to consider it too. Then that woman at the social event might finally stop complaining!