Statistics Using Just One Formula

Jeffrey S. Rosenthal[1], March 19, 2017

1. Introduction. In our data-rich world, statistical analysis and education are more important than ever. I recently developed[2] 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, but some students found the derivations "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 formulas she had to memorise – 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 used and understood, but with fewer equations and formulas and calculations?

In the end, I decided that most simple statistical inference problems can be solved effectively using just a single formula (and a few slight generalisations), as follows.

2. 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. 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 population average by computing the average of our sample. But this estimate probably won't be *exactly* correct, due to the randomness of the sample. So, the question becomes, how close will our sample average probably be to the true population average?

3. The One Formula. Suppose $x_1, x_2, \ldots, x_n$ is a sample of n measurements of some quantity, and we estimate the population average by our sample average $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Then it turns out that usually, i.e. about 95% of the time, i.e. about 19 times out of 20, $\overline{x}$ will be within about $2\sqrt{v/n}$ of the true population 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 $[\overline{x} - 2\sqrt{v/n},\ \overline{x} + 2\sqrt{v/n}]$. And this one formula is all that we need to make lots of fairly accurate statistical inferences.

(As an aside, you may wonder *why* this formula holds. Roughly speaking, v is a good estimate of the *variance* of the x_i values, so the sample average $\overline{x}$ has variance about v/n.

[1]Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 3G3. Email: jeff@math.toronto.edu. Web: http://probability.ca/jeff/ Supported in part by NSERC of Canada.

[2]see http://probability.ca/sta130report

Also, the *expected* value of $\overline{x}$ is the true mean, say μ. It follows that $(\overline{x} - \mu)/\sqrt{v/n}$ has mean 0 and variance 1. Furthermore, by the Central Limit Theorem, it has approximately a normal distribution. It is thus 95% likely to be between about -2 and 2.)

4. Example: Baby Weights. Ten babies born in a hospital in North Carolina were measured[3] to have the following weights, in pounds: $x_1 = 9.88$, $x_2 = 9.12$, $x_3 = 8.00$, $x_4 = 9.38$, $x_5 = 7.44$, $x_6 = 8.25$, $x_7 = 8.25$, $x_8 = 6.88$, $x_9 = 7.94$, $x_{10} = 6.00$.

For this data, $n = 10$, and we compute that $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \doteq 8.11$ pounds, and then $v = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \doteq 1.362$. Based on this 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 $8.11 + 0.74 = 8.85$ pounds. That is, our 95% confidence interval for the weight (in pounds) of babies in North Carolina is [7.37, 8.85].

5. Statistical Significance. A claim based on a sample is called *statistically significant* 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 artifact due to the random luck of the sample. There are many specialised ways of computing significance levels in various settings. But a simple rule in our case is: a claim is statistically significant if holds for all values in the confidence interval.

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

6. 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, $v = \frac{1}{n}[n\overline{x}(1-\overline{x})^2 + (n - n\overline{x})(0 - \overline{x})^2]$, 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[4] claimed that "more than half" of Canadians approved of the goverment. 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

[3] see `http://www.math.hope.edu/swanson/data/nc200.txt`

[4] see `http://www.theglobeandmail.com/news/politics/more-than-half-of-canadians-approve-of-trudeau-poll/article28210076/`

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[5] a whole table of margins of error for various different sample sizes and observed proprtions, 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], while 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.

7. Comparison of Means. 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[6] the circumference of their wrists. The $n = 39$ female students had sample mean $\overline{x} = 14.49$ (in cm) with squared deviation $v = 0.622$, while the $m = 41$ male students had sample mean $\overline{y} = 16.74$ with squared deviation $w = 0.947$. So, on average the males 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.

8. 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 consider-

[5] see `http://www.forumresearch.com/tools-margin-of-error.asp`

[6] see `http://probability.ca/sta130/studentdata.txt`

ations apply.

For example, CBS News conducted a series of polls asking Americans if they supported the legalization of marijuana. In 2012 they sampled[7] 1100 adults and found that 47% said yes. In 2014 they sampled[8] 1018 adults and found that 51% said yes. In 2015 they sampled[9] 1012 adults and found 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, $n = 1100$ and $\overline{x} = 0.47$, while $m = 1018$ 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 $n = 1100$ and $\overline{x} = 0.47$, but now $m = 1012$ 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.

9. Correlation. The above provides fairly useful tools for many elementary statistical analyses, all essentially using one formula. But there is one essential statistical concept which does not quite fit neatly into this approach, namely *correlation*. Suppose we have samples of two *different* quantities on the *same* n objects, say $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$, and we are interested in the *relation* between them. Specifically, when one quantity increases, does the other quantity tend to also increase, or to decrease, or is it unaffected?

To answer this question, we must compute the sample correlation. Unfortunately that requires another formula (contrary to this paper's title!). Specifically, the sample correlation is the average of the products of the two samples, after they have been *normalised* by subtracting off their means and dividing by the squareroots of their squared deviations. That is, the sample correlation is given by $r = \frac{1}{n\sqrt{vw}} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$. (Fortunately, we don't really need to *know* this formula, since it can be computed automatically by any statistical software package, such as the free[10] package "R".) Correlations are always between -1 and 1, with 0 indicating no (linear) relationship, and ± 1 the strongest relationships.

For example, consider the percentage of adults who smoke[11] and the average income per capita[12] for each of the 50 U.S. states. Using the above formula, we compute the correlation

[7] see `http://www.cbsnews.com/news/poll-nearly-half-support-legalization-of-marijuana/`

[8] see `http://www.cbsnews.com/news/majority-of-americans-now-support-legal-pot-poll-says/`

[9] see `http://www.cbsnews.com/news/poll-support-for-legal-marijuana-use-reaches-all-time-high/`

[10] see `https://cran.r-project.org/`

[11] see `https://www.tobaccofreekids.org/research/factsheets/pdf/0176.pdf`

[12] see `http://www.infoplease.com/ipa/A0104652.html`

to be about -0.42. Since this value is rather negative, it means that on average, states with higher smoking rates have smaller incomes, and vice versa.

But is this relationship statistically significant? For this we go back to our one formula! The only issue is what value of v should be used. That isn't so straightforward in general, but a reasonable approximation (which is exact in the independent case) is to simply take $v = 1$, leading to a 95% confidence interval of about $[r - 2/\sqrt{n},\ r + 2/\sqrt{n}]$. For our U.S. states example, this equals $[-0.42 - 2/\sqrt{50},\ -0.42 + 2/\sqrt{50}] \doteq [-0.70,\ -0.14]$. 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.)

10. Discussion. 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 formulas which arise in typical 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!

Of course, the above discussion is just an *approximate* summary of certain basic statistical inference procedures. For example, in the formula for v (and also for r), statisticians usually divide by $n - 1$ instead of n for reasons which are somewhat subtle and controversial[13], though this makes little difference if n is large.

Also, the constant "2" in our margin of error formula should actually equal 1.96 for the normal distribution case, or various larger values for the t distribution case (depending on the sample size n, e.g. if $n = 10$ it's 2.26, if $n = 20$ it's 2.09, if $n = 100$ it's 1.98, etc.). But 2 is a reasonable approximation and is close enough for most practical purposes. Related to this, we can be (say) 99% confident instead of just 95% confident by increasing the constant a little bit: from 1.96 to 2.58 in the normal distribution case.

In a different direction, the usual assessment of statistical significance involves "hypothesis tests" and "P-values" and "probability tables", not discussed here. However, for the simple inference problems considered here, the usual statistical significance turns out to be *equivalent* to our simple notion of "holds for all values in the confidence interval".

Many other refinements and improvements, together with theoretical justifications and applications to many different areas, can be found in more advanced statistics courses. Hopefully the brief look herein will inspire you to study this important subject more deeply!

[13] see e.g. my article at `http://probability.ca/varmse`