Simple Methods for Estimating Confidence Levels, or Tentative Probabilities, for Hypotheses Instead of *P* Values

23 April 2018

Michael Wood University of Portsmouth, Portsmouth, UK michaelwoodslg@gmail.com or michael.wood@port.ac.uk http://woodm.myweb.port.ac.uk/

Abstract

In many fields of research null hypothesis significance tests and *p* values are the accepted way of assessing the degree of certainty with which research results can be extrapolated beyond the sample studied. However, there are very serious concerns about the suitability of *p* values for this purpose. An alternative approach is to cite confidence intervals for a statistic of interest, but this does not directly tell readers how certain a hypothesis is. Here, I suggest how the framework used for confidence intervals could easily be extended to derive confidence levels, or "tentative probabilities", for hypotheses. I also outline four quick methods for estimating these. This allows researchers to state their confidence in a hypothesis as a direct probability, instead of circuitously by *p* values referring to an unstated, hypothetical null hypothesis. The inevitable difficulties of statistical inference mean that these probabilities can only be tentative, but probabilities are the natural way to express uncertainties, so, arguably, researchers using statistical methods have an obligation to estimate how probable their hypotheses are by the best available method. Otherwise misinterpretations will fill the void.

Key words: Confidence, Null hypothesis significance test, p value, Statistical inference

Contents

Abstract1
Introduction
An example3
Table 1.Data and conventional analysis for small and large samples4
Fig. 1. Confidence / tentative probability distributions for the difference between the mean scores in two populations based on small and large samples4
Confidence levels for hypotheses
Table 2: Confidence levels / tentative probabilities for hypotheses about how the mean score in Population A compares with Population B
Tentative probabilities or confidence levels?
Methods for quick estimates of tentative probabilities, or confidence levels, for hypotheses
Method 1: Using the confidence interval routine in statistical packages7
Method 2: Using p values8
Method 3: Using the normal or t distribution to extrapolate from p values or confidence intervals 8
Method 4: Bootstrapping9
Fig. 2. Bootstrapped confidence / tentative probability distribution
How should we describe the strength of the evidence based on a sample: confidence levels or p values?
Conclusions11
References

Spreadsheets to which this article refers:

<u>http://woodm.myweb.port.ac.uk/SL/popsab.xlsx</u>: the data used in this article <u>http://woodm.myweb.port.ac.uk/CLIP.xls</u>: estimating confidence levels from ci's or *p* values <u>http://woodm.myweb.port.ac.uk/SL/resamplepopsab.xlsx</u>: bootstrapping

Introduction

There have been extensive criticisms of null hypothesis significance tests and p values in the literature for more than 50 years (e.g. Morrison & Henkel 1970; Nickerson 2000; Nuzzo 2014; Wasserstein & Lazar 2016): they are widely misinterpreted, and do not, as is often assumed, give the probability of a hypothesis being true, nor a measure of the size or importance of the effect. My aim here is not to add to this critical, but often ineffective, literature, but to suggest a simple alternative to p values. Sometimes, of course, p values are a sensible approach (see <u>Wood, 2016</u>), but often, perhaps usually, they are not.

My proposal is simply to use confidence distributions (Xie & Singh, 2013), which are the basis of confidence intervals, to estimate confidence levels for whatever hypotheses are of interest. Instead of giving a p value (e.g. p=4%) to indicate our degree of confidence in a statistical conclusion, we could give a confidence level for the conclusion (e.g. CL= 98%). This is far more straightforward and informative than the p value, but it is important to be aware that these confidence levels can only be tentative. This tentativeness applies equally to confidence intervals, although this is rarely acknowledged.

There are a variety of methods of deriving confidence distributions: to review these would take me too far from my main concern, but I will explain how estimates of confidence levels can be obtained from the p values and confidence intervals produced by computer packages, and from bootstrapping. I will also suggest that the distinction between confidence and probability is an unnecessary complication, and that confidence levels should be viewed as "tentative probabilities".

Given a confidence distribution, the mathematics of my proposal is trivial and best explained by means of an example, although the scope of the approach is far wider than this example. My aim here is to suggest how a well-established statistical framework could easily be extended to make it far more useful for research by providing a straightforward answer to the obvious question: "How likely is it that this hypothesis is true?" (P values, of course, do not do this.)

An example

I will use the fictional "data" in Table 1 to illustrate the approach. These are scores from random samples from two populations, A and B. (The scores might be from a psychological test, or measures of the effectiveness of two treatments; the samples might be from two countries or different treatment groups in a randomized trial.) One of the researcher's hypotheses was that the mean score in one of the populations (A) would be higher than in the other (B). The difference between the means is small (0.3), and with the small sample of 10 in each population, the p value is high (0.673) and the 95% confidence interval for the difference (-1.2 to +1.8) encompasses both positive and negative values indicating that we cannot be sure which population has the higher mean score.

Ten is obviously an unrealistically small sample, but it is helpful to show the contrast with a more realistic sample of 400 in each population. With the larger sample (see Table 1), the difference is significant at the 1% level, and the confidence interval is entirely in the positive range indicating that the mean score in Population A is likely to be more than in B. However, the apparent strength of these conclusions makes it easy to forget that the estimated difference between the population means is only 0.3.

Data and analysis for small samples (n=10 from each population)				
	Sample from Population A	Sample from Population B		
	8	8		
	5	7		
	6	5		
	6	6		
	6	3		
	5	8		
	8	6		
	7	6		
	7	3		
	5	8		
Means	6.3	6		
<i>p</i> value (two tailed) for the difference of means: 0.673 (67.3%)				
95% Confidence interval for the difference of means: -1.2 to +1.8				
Large samples comprising 40 copies of small samples (n=400 from each pop)				
Obviously the means are the same as for the small sample				
p value (two tailed) for the difference of means: 0.004 (0.4%)				
95% Confidence interval for the difference of means: +0.1 to +0.5				
The <i>p</i> val	ues and confidence intervals are base	d on the standard method with the t		
distrik	oution, using the Independent samples	s t test in SPSS, or the formulae in		

Table 1.Data and conventional analysis for small and large samples

http://woodm.myweb.port.ac.uk/SL/popsab.xlsx.

The confidence interval for the difference of the means based on the small samples is derived from the confidence distribution in Figure 1 (bold curve). It should be roughly obvious that 95% of the "confidence" lies between -1.2 and +1.8, with 2.5% in each of the tails since the distribution is symmetrical. (The vertical lines, and the phrase "tentative probability", are explained below.)

Fig. 1. Confidence / tentative probability distributions for the difference between the mean scores in two populations based on small and large samples



Confidence levels for hypotheses

Conventionally, the word "confidence" is only used in relation to intervals, usually 95% ones. In practical terms this is odd because the 95% is based on arbitrary statistical convention, rather than the requirements of the problem. My suggestion here is that confidence distributions could be used to derive confidence levels for more general hypotheses.

Figure 1 can be used to derive a confidence level for the mean score in Population A being more than in B - i.e. the difference of the means being positive. This is the proportion of the area to the right of the line representing a difference of zero. To derive this from a confidence interval, we need one of the limits for the interval to be zero. This can be estimated by trial and error with a package such as SPSS: the required confidence interval for the small samples is the 32.7% interval which leaves 33.65% in each of the tails, and the "confidence" to the right of the solid vertical line is 66.35% (32.7%+33.65%) - which is our confidence that the mean score of Population A is greater than that of Population B. Alternatively, the same result could be obtained directly from the formulae used to draw Fig. 1 (see http://woodm.myweb.port.ac.uk/SL/popsab.xlsx).

This principle can easily be extended. For example, we might decide that small differences between the populations are not of interest so we might have three hypotheses: the mean scores in the two populations are within one unit of each other, A is substantially more than B (more than one unit more), and vice versa. Table 2 brings all these results together for both the small and big samples in Table 1.

-	<u> </u>			
Hypothesis about mean	Difference	Confidence level or	Size of sample from	
score	(A - B)	tentative probability	each population	
A > B	> 0	66%	10	
A < B	< 0	34%	10	
A >> B	> 1	17%	10	
$A \approx B$	Between -1 and 1	79%	10	
A << B	< -1	4%	10	
A > B	> 0	99.8%	400	
A < B	< 0	0.2%	400	
A >> B	> 1	0.0%	400	
$A \approx B$	Between -1 and 1	100.0%	400	
A << B	< -1	0.0%	400	
The confidence levels / tentative probabilities are based on the distributions in Figure 1.				

Table 2: Confidence levels / tentative probabilities for hypotheses about how the mean score in Population A compares with Population B

The confidence levels / tentative probabilities are based on the distributions in Figure 1. ≈ means approximately equal which is defined as being within one unit. >> means substantially more which is defined as more than a unit.

Table 2 shows the obvious contrast between the big and small samples. For the big sample the evidence is almost conclusive (100.0%) that A has *approximately* the same mean score as B, whereas for the small sample this hypothesis only has a 79% confidence.

The big sample result illustrates one of the problems with p values. The low p value (0.004) suggests that the null hypothesis of *exactly* the same mean scores in both populations should be rejected. Population A does have a higher mean that Population B. However the tentative probability of

100.0% of the mean score in the two populations being *approximately* equal (within one unit) indicates that although differences do exist between the two populations, they may be too small to matter in practice. Over-reliance on *p* values may obscure this conclusion.

This approach would obviously work for any hypothesis based on a numerical statistic for which a confidence distribution like Figure 1 can be derived. Other standard examples include regression and correlation coefficients, and the difference of two proportions. Figure 1 is symmetrical, so the confidence level in each tail is obviously the same: where this is not the case the method would need to be adapted in an obvious way. The method can even be extended to hypotheses that are not defined by a single numerical statistic - see under <u>Bootstrapping</u> below for an example.

Tentative probabilities or confidence levels?

It is possible to interpret confidence levels within the Bayesian statistical paradigm. This uses Bayes' theorem and a prior probability distribution - reflecting our prior knowledge of the situation - to derive a posterior probability distribution on which "credible intervals" are based. If we make the neutral assumption that all values of the horizontal axis in Fig 1 are equally likely (a flat prior distribution), then confidence intervals are often identical to Bayesian credible intervals, which means that confidence levels for hypotheses can be interpreted as Bayesian posterior probabilities. This is not true in general, but it is true for many distributions provided that the prior distribution is flat (Bolstad 2007, Xie and Singh, 2013). Xie and Singh (2013) say this "appears to be a mathematical coincidence", which ignores the fact that such coincidences can be analyzed mathematically to uncover the circumstances in which they occur (see the Appendix of Wood, 2014).

However, according to the dominant, frequentist, version of statistical theory, confidence levels are *not* probabilities because there is just one true value of the statistic so probabilities are irrelevant. Probabilities apply to uncertain *events*, like whether a coin lands on heads or tails, *not* to beliefs or hypotheses which are either true or false. However, this depends on the meaning we choose to attach to the word "probability": there seems little reason why statistics should not follow everyday language and extend the idea of probability to cover beliefs and hypotheses.

We have seen that, with a flat prior distribution, confidence intervals are often numerically identical to their Bayesian equivalent, and Xie & Singh 2013, in a strictly frequentist review of confidence distributions, state that they can be used "in the style of a Bayesian posterior" probability distribution. Bayesian posterior probability distributions *are* probability distributions: the main difference between the (standard) frequentist and Bayesian schools of probability theory being that that the former defines probability in terms of frequencies of events, whereas the latter has a more flexible definition involving some kind of degree of certainty. However, in practice, they both obey the same rules, and in everyday discourse the idea of probability is used in the broader, Bayesian sense. For example, if someone gives a probability for life existing on one of Jupiter's moons, this probability can only be interpreted as a degree of belief. And on the website of the medical journal BMI (2017), confidence intervals are explained in these terms: "this is called the 95% confidence interval, and we can say that there is only a 5% chance that the range 86.96 to 89.04 mmHg excludes the mean of the population", which is a probabilistic explanation if we assume that "chance" means probability. For all these reasons, there seems little reason to distinguish between

confidence and probability: using the term probability in both contexts would avoid the confusion of the extra term "confidence".

Despite this I will continue to use the term confidence in this article simply because of its familiarity: people know what confidence intervals are and using the term "confidence" taps into this knowledge. But I would argue that confidence levels should be regarded as probabilities.

However, the probabilities derived from confidence distributions must be regarded as tentative. There is a strong argument that there cannot be a rigorous, general method of calculating probabilities for hypotheses, so any such probability should be regarded as provisional. Imagine, with the data in Table 1, that we subsequently found that the data was a hoax and all samples came from the same population. This would mean that the mean population scores were identical and the observed difference was just sampling error for both big and small samples. This illustrates the principle that prior beliefs *must* have an impact on sensible conclusions: if we have a hypothesis which we are sure is false, no evidence will suffice to overturn it. This is where Bayesian statistics is helpful. A suitable prior distribution incorporating our prior beliefs will ensure that we get a sensible answer. On the other hand, if we have no definite prior information, a Bayesian Interpretation of confidence levels has the advantage of yielding a probability, and of clarifying the main condition for the validity of the probability - namely that the prior distribution should be flat.

There is a plethora of other concepts in this area - fiducial probabilities, Bayes' factors, etc. However, none of them are widely used, probably because they don't produce useful and intuitive measures of the certainty of a hypothesis. My contention here is that Bayes' theorem gives an answer in principle, but in practice we have to make simplifying assumptions about prior probabilities, and that extending the confidence interval framework is a good compromise.

Methods for quick estimates of tentative probabilities, or confidence levels, for hypotheses

The general method is simply to use the confidence distribution used for confidence intervals to make the appropriate estimates, or to use Bayes theorem as discussed <u>above</u>. However, this may not always be practicable. Even if we only have the confidence intervals or p values cited in a conventional analysis, it may still be possible to estimate tentative probabilities for the hypotheses of interest (Methods 2 and 3 below). And if we have the data, but no mathematical model of the confidence distribution, it may be possible to use the confidence interval routine built into statistical packages to reverse engineer the required confidence levels (Method 1), or to use bootstrapping to simulate a confidence distribution (Method 4). Bootstrapping is a very flexible simulation method: it can, for example, be used to estimate a tentative probability for the relationship between two variables being an inverted U shape (see below and <u>Wood, 2013</u>, p. 7).

Method 1: Using the confidence interval routine in statistical packages

If the confidence intervals have equal confidence levels in each tail (this is usually the case), trial and error trying different confidence levels can be used to estimate tentative probabilities as explained <u>above</u> and illustrated by <u>Figure 1</u>.

Method 2: Using p values

Table 1 shows that the p value based on the big sample is 0.4%, and Table 2 shows that the tentative probability of the mean of Population A being greater than B is 99.8%. The obvious relation between these figures is that 99.8% = 100% - 0.4%/2. For the small samples, to one decimal place, the p value is 67.3% and the tentative probability is 66.3%, and the same relationship holds (66.3%=100% - 67.3%/2 except for the rounding error). The tentative probabilities that the mean of A will be less than B is simply p/2: 0.2% for the large samples and 67.3%/2 = 33.65 for the small samples.

This is not a coincidence but is a consequence of the way confidence intervals are derived: see, for example, Smithson (2003: 3-9)¹. When the confidence distribution is a symmetrical curve like Figure 1, this leads to the following conclusions:

Suppose we have a two tailed p value for a statistic based on a null hypothesis value of zero. Suppose further that the sample value of the statistic is positive but negative values would be possible. The difference of means in Table 1 is an example; other examples include correlation and regression coefficients, but not chi squared which is always positive. Then *Tentative probability that population value of statistic is positive = 1 - p/2 Tentative probability that population value of statistic is negative = p/2* If the sample value is negative these probabilities would be reversed.

These formulae can easily be adapted if the information we have about p is an inequality. For example if p < 0.1% then the first equation above becomes *Tentative probability that population value of statistic is positive* > 99.95%

This method only works if the cutoff value for the hypotheses is zero. Method 3 is more general.

Method 3: Using the normal or *t* distribution to extrapolate from *p* values or confidence intervals

If we have the data we can either work out the mathematics of the confidence distribution ourselves, or reverse engineer confidence levels for hypotheses from the confidence interval routines in standard packages. However, even in the absence of the data it is still possible to make a reasonable estimate if we have either a confidence interval or a p value. If we assume that the confidence distribution is a normal distribution (or a *t* distribution, but this usually makes little difference), it is possible to use the information we have to reverse engineer the confidence distribution of normality might seem suspect, but many distributions are roughly normal, and normal approximations are widely used in statistics even when they are not very close (e.g. the for binomial or Poisson distributions).

¹ Very briefly, using the large sample result as an example, the 99.6% confidence interval can be defined as the set of possible null hypothesis values for the unknown population statistic which would yield a *non-significant* result with a *p* value of 0.4%. Now imagine that the population statistic were actually 0.6 in Figure 1. The sampling distribution is simply the confidence distribution shifted 0.3 units to the right, and probability of the sampling statistic being less than the observed value of 0.3 is obviously 0.2% and the two tailed *p* value will the twice this or 0.4%. If the population statistic were more than 0.6, the result would be significant so these points are not in the confidence interval; if the population statistic were between 0 and 0.6 the result would not be significant so these points are in the confidence interval.

As an example suppose we wanted to estimate the confidence level for the mean of Population A being at least one unit more than B: Table 2, based on the full data set, shows this is 17% based on the small samples. However, if we *only* knew that the sample value of the difference is 0.3, and the *p* value is 67.3%, the spreadsheet at <u>http://woodm.myweb.port.ac.uk/CLIP.xls</u> is designed to estimate confidence levels. The confidence level in question comes to 16% using the normal distribution - which is close enough to be useful in practice. Alternatively we could use the 95% confidence interval (-1.2 to 1.8) to derive a similar result from the same spreadsheet.

Method 4: Bootstrapping

Bootstrapping is a widely used method of deriving confidence distributions (Xie and Singh, 2013). It relies on computer simulation which can estimate approximate probabilities without detailed mathematical models. The spreadsheet at <u>http://woodm.myweb.port.ac.uk/SL/resamplepopsab.xlsx</u> is set up to estimate confidence levels from the data in the small sample in Table 1 by bootstrapping, and provide a rough explanation of how and why the method works. Because bootstrapping is a simulation method, it will produce a slightly different distribution each time it is run. Figure 2 shows one such bootstrapped confidence distribution.





When I used the spreadsheet, which analyzes 1000 simulations each time, the results for the confidence level of the mean of Population A being at least one unit more than B ranged from 14% to 17%. This is close² to the answer of 17% from the confidence distribution in Figure 1

One of the strengths of bootstrapping is its flexibility in terms of the problems it can model. The spreadsheet used for Figure 2 can easily be adapted to analyze other statistics such as regression and correlation coefficients. Wood (2013) describes how bootstrapping can be used to estimate a confidence level (65%) for a hypothesis of an inverted U shape between two variables. The original paper on whose data this analysis is based (Glebbeek and Bax, 2004) merely produced two non-significant p values for the two parameters of a quadratic function.

² But not identical: the bootstrap estimate is slightly less, possibly because small samples are less likely to include extremes and so may underestimate the variability in the population. However, worrying unduly about the precision of the confidence levels seems unreasonable because the sample is small so everything is just a rough guess.

How should we describe the strength of the evidence based on a sample: confidence levels or p values?

Sometimes p values are a sensible approach. In an experiment to detect telepathy described in Wood (2016, p. 4), a volunteer chooses a card at random from a pack of 50 cards, concentrates hard on it, and then another volunteer tries to work out what the chosen card is without being able to see the card or communicate with the first volunteer in any way (except by telepathy). If the second volunteer gets the right answer, this represents a p value of 1/50 or 2% based on the null hypothesis that the second volunteer is guessing. There is no confidence distribution here, and no obvious easier way of coming up with a probability for telepathy than the use of Bayes' theorem, which requires, of course, a prior probability for telepathy. The p value(2%) is a simple and natural way to summarize the conclusion about the strength of the evidence. But this example is unusual: in many cases there is a confidence distribution, and p values are neither simple nor natural.

Redelmeier and Singh (2001, p. 955) concluded that "life expectancy was 3.9 years longer for Academy Award [Oscar] winners that for other, less recognized performers (79.7 vs. 75.8 years, p = 0.003)." No confidence interval is given, so readers get no feeling for how accurate the result is likely to be. The p value gives an indication of how susceptible the result may be to sampling error, but only for readers who understand the role of the unstated null hypothesis. (The conclusions of this study were challenged by Sylvestre et al, 2006, in a later article in the same journal. The challenge, however, has nothing to do with the p values.)

More recently, confidence intervals are widely cited in medical journals: for example , Wallis et al (2017), concluded that "patients treated by female surgeons were less likely to die within 30 days (adjusted odds ratio 0.88; [95% confidence interval] 0.79 to 0.99, P=0.04), but there was no significant difference in readmissions or complications." The confidence interval here gives an indication of the likely error but the p value is still the statistic given to quantify the strength of the evidence for the conclusion about the difference between male and female surgeons. And the phrase "no significant difference" implies that it is the difference which is "significant", whereas it is actually the *evidence for* the difference which is significant. Phrases like "significant difference" can only reinforce the common misunderstanding that a significant difference is a big or important difference.

In many other disciplines p values remain the sole statistic used to quantify uncertainty due to sampling error; confidence intervals are not given. We have seen <u>above</u> how Glebbeek and Bax (2004) could only give two non-significant p values in support of their hypothesis of an inverted U shape between two variables. They also looked at a linear regression model between the same two variables and concluded that "a 1% increase in turnover equals a loss of 1780 Dutch guilders [this was before the introduction of the Euro] From a management point of view, this is quite substantial" (p. 283). However the only statistic quoted to quantify the strength of the evidence for this is p<0.01. The confidence interval (-3060 to -500: see Wood, 2013) is not given.

I think there is a very strong case for using confidence levels for the hypotheses of interest to the research in all these papers. So instead of the p value quoted above for the comparison between male and female surgeons we could (using Method 3 above) write:

Patients treated by female surgeons were slightly less likely to die within 30 days (confidence level: 98%).

And instead of the potentially misleading statement that "there was no significant difference in readmissions or complications", the fact that there was a difference but that the evidence that this would apply to a wider population was weaker could be acknowledged by stating a confidence level (about 90% for the readmission rate).

Or, taking a different perspective, we might feel that, given the likely inaccuracies of the matching and adjustment processes, a difference in the readmission rates of less than 10% is not meaningful. Method 3 can be used to estimate the confidence level for the difference being less than 10%: this comes to about 98% (by estimating the probability of the risk ratio - which is identical to the odds ratio in this case - being <0.9 and >1.1 and subtracting these two probabilities from 100%). This represents strong evidence that there is not a meaningful difference.

The conclusions of the other two studies could be rephrased in a similar way. The confidence level for Oscar winners living longer is 99.85% (Wood, 2014), and for increasing turnover leading to a loss in the Glebbeek and Bax's (2004) study the corresponding confidence level is 99.7% (Wood, 2013). The evidence for the inverted U shape cannot be adequately measured by p values: the confidence level of 65% makes the conclusion far clearer than two non-significant p values.

Conclusions

Despite being an obvious extension to the idea of confidence intervals, confidence levels such as those above are never cited. As far as I can see, the idea is never even considered, despite the fact that it gives clear answers to questions about the strength of evidence, whereas the ubiquitous p values are widely misinterpreted and often fail to provide much useful information. Instead of qualifying the conclusion that patients treated by female surgeons were slightly less likely to die within 30 days with the statement p=0.04 (Wallis et al, 2017), we could qualify it with by writing confidence level = 98%.

The reason why these confidence levels are not used may be that a p value is a definite probability based on a clearly defined (although usually unstated) null hypothesis, whereas confidence levels and confidence intervals, however we choose to define them, depend on assumptions whose validity is always tentative. Some of the confidence levels above depend on an additional assumption that a normal approximation is reasonable, but even if we go back to the raw data, the estimation of confidence levels can only be tentative.

However, confidence intervals are used, despite their tentative status, and they are inevitably likely to be the basis of informal calculations of confidence levels such as those I am suggesting here ("The confidence interval does not include zero so ..."). Why not make these informal calculations explicit?

The answer, I suspect, is that doing so would make the conclusions clearer, which would expose them to more scrutiny. If, for example, only 50% of hypotheses which achieve a confidence level of 95% or more turn out to be true, this would suggest something is wrong with the assessment of confidence levels.

I have also argued that the definition of probability should be expanded so that confidence levels could be regarded as tentative probabilities. Confidence is an unnecessary concept in an already overcrowded conceptual framework. It is also a bad name for a concept in which we should not have too much confidence. However, in the short term, keeping the word confidence may be helpful because it gives the idea of tentative probabilities for hypotheses the support of an established, and so relatively unquestioned, theoretical framework.

References

BMJ (2017). Website at <u>http://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/4-statements-probability-and-confiden</u> accessed on 26 November, 2017.

Bolstad, W. M. (2007). Introduction to Bayesian statistics (2nd edition). Hoboken, New Jersey: Wiley.

Glebbeek, A. C., & Bax, E. H. (2004). Is high employee turnover really harmful? An empirical test using company records. *Academy of Management Journal*, *47*(2), 277-286.

Morrison, D. E. & Henkel, R. E. (eds). (1970). *The significance test controversy: a reader*. Chicago: Aldine Pub. Co.

Nickerson, R. S. (2000). Null Hypothesis Significance Testing: A Review of an Old and Continuing Controversy. *Psychological Methods*, *5*, 241-301.

Nuzzo, R. (2014, 12 February). Scientific method: statistical errors. Nature, 506, 150-2.

Redelmeier, D. A., & Singh, S. M. (2001). Survival in Academy Award-winning actors and actresses. *Annals of Internal Medicine*, *134*, 955-962.

Smithson, Michael. (2003). Confidence intervals. Thousand Oaks: Sage.

Sylvestre, M.-P., Huszti, E., & Hanley, J. A. (2006). Do Oscar winners live longer than less successful peers? A reanalysis of the evidence. *Annals of Internal Medicine*, *145*, 361-363.

Wallis, C. J. D., Ravi, B., Coburn, N., Nam, R.K., Detsky, A. S. & Satkunasivam, R. (2017). Comparison of postoperative outcomes among patients treated by male and female surgeons: a population based matched cohort study. *BMJ* 2017;359:j4366 doi: 10.1136/bmj.j4366.

Wasserstein, R. L. & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose, *The American Statistician*, DOI: 10.1080/00031305.2016.1154108

Wood, M. (2013). Making statistical methods more useful: some suggestions from a case study. *Sage Open, 3* (1) <u>https://doi.org/10.1177/2158244013476873</u>

Wood, M. (2014). P values, confidence intervals, or confidence levels for hypotheses? <u>https://arxiv.org/abs/0912.3878</u>.

Wood, M. (2016). Beyond p values: practical methods for analyzing uncertainty in research. <u>https://arxiv.org/abs/1609.09803</u>.

Xie, M. & Singh, K. (2013). Confidence Distribution, the Frequentist Distribution Estimator of a Parameter: A Review. *International Statistical Review*, 81, 1, 3–39, doi:10.1111/insr.12000, (p.3).