Data as evidence

Peter Cahusac^{1,2}

¹ Department of Pharmacology & Biostatistics, College of Medicine, Alfaisal University, Riyadh, Saudi Arabia

² Department of Comparative Medicine, King Faisal Specialist Hospital and Research Centre, Riyadh, Saudi Arabia

Correspondence:

Peter Cahusac, College of Medicine, Riyadh 11533, Saudi Arabia.

Email: pcahusac@alfaisal.edu

Edited by: Jeremy Ward

Abstract

In an earlier article, Gordon Drummond summarized ongoing changes in how statistics are being used in experimental physiology. He described the near-ubiquitous use of the *P* value, cautioning against declaring an analysis 'statistically significant'. He mentioned alternative approaches, including Bayesian and likelihood approaches. This article focuses on the latter approach, although I initially take another look at the *P* value. Then the likelihood approach is introduced with a very artificial example to enable the concept to be grasped easily. Next, a more realistic example is described, with associated calculations. A further example using real categorical data is explained, showing how it relates to and is superior to the oft-used χ^2 test. A final discussion reveals that the likelihood approach, although mathematically and statistically accurate, is poorly supported by literature and training.

KEYWORDS

evidential approach, likelihood, *P* values, support

1 | WE HAVE DATA

We have some data, carry out a statistical test, obtain a *P* value and decide that there is an effect. Apparently, this is all easy and comes naturally to all of us. This is how we do statistics. But then one asks, 'What is the *P* value, and what does it mean?'. We would rather avoid that question; after all, we know that the smaller it is the more statistically significant it is, right? And if we get a really small *P* value it's a licence to publish. A previous Myths and Methodologies article has discussed the *P* value and going beyond it (Drummond, 2020).

By historical accident, and without the mutual consent of the two statistical inference approaches of Fisher and Neyman–Pearson, the *P* value plays a double duty. First, it tries to tell us the probability of obtaining our particular result, or a more extreme result, if the null hypothesis (H_0) were true (Fisher, 1956). Second, it provides control in the long run for errors of claiming a result significant when the null is true (type I errors; Neyman & Pearson, 1928). The latter is the reason why once we have collected data and done a statistical test we cannot collect more data (you mean, you didn't know?). It is the reason we need to make corrections à la Bonferroni. The

P value cannot be expected to provide evidence against a hypothesis at the same time as providing a probability measure for obtaining that evidence.

A popular fallacy is to think that if a *P* value fails to reach statistical significance then the treatment is ineffective, and H_0 is true (Dixon, 2003). However, absence of evidence does not mean evidence of absence (Altman & Bland, 1995). It was appreciation of this issue that caused the US Supreme Court in 2011 to reject the 'bright-line rule' of statistical significance (retrieved from <https://www.supremecourt.gov/opinions/10pdf/09-1156.pdf>). For the case in question, the pharmaceutical company claimed that their medication was safe because the link between its use and anosmia failed to reach statistical significance. Misuse of the *P* value probably also led British law courts to rely exclusively on the likelihood ratio (LR) for statistical evidence (see p. 17 of http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf). Criminal cases must meet a higher standard of evidence compared with scientists, requiring the prosecution to prove guilt 'beyond reasonable doubt' (for more details, see the footnote to Table 1).

Some readers may find it helpful to consult statistical definitions given in Box 1 (*Common statistical terminology*).

BOX 1**Common statistical terminology**

H_0	The null hypothesis, which typically represents no effect of a treatment on our measurements.
H_1	The alternative or experimental hypothesis, which represents some/any effect produced by a treatment on our measurements. H_2 can be used to represent any other hypothesis.
Standard deviation	Roughly speaking, it is the average variability of individual data points. It tells us how much data varies. It is the square root of the variance.
Standard error	Like the standard deviation but for a statistic, such as a mean. It tells us how much a statistic, such as a mean, varies. It is the standard deviation for the sampling distribution.
Sampling distribution	The distribution of a statistic, such as a mean, when repeated samples of specified size are taken from a large population. This is best demonstrated by computer simulation. For a sampling distribution of the mean (which we are most often interested in), the sampling distribution becomes more normally distributed the larger the sample size. The standard deviation of the sampling distribution gives us the standard error.
z	Obtained from the standard normal distribution, this statistic represents the number of standard deviations a value is from a specified mean. If this concerns the sampling distribution, it represents the number of standard errors from the specified mean. It requires the population standard deviation to be known.
t	The commonest statistical test produces t values, which represent the number of standard errors difference of a mean from a null value or specified value. We use this instead of z when the population standard deviation is not known and has to be estimated from the sample.
F	This is obtained in an analysis of variance (ANOVA) and represents the variance ratio from two estimates of population variance: the between-groups to the within-groups variance. If there are only two samples then the square root of the F value will give the t value (and identical P values). More generally, F is used to assess the fit of data to a model, such as in regression.
χ^2	This is the sum of independent squared standard normal (z) values whose degrees of freedom depend upon how many terms there are. It pops up everywhere and is related to other statistics here (z , t and F) but most commonly associated with categorical analyses.
R^2	In a bivariate correlation, it is the correlation coefficient r squared. More generally, it represents the proportion of variability (variance) explained by a model.
MLE	Maximum likelihood estimate is a statistic which, for a given model, has the highest probability of being predicted from the data. It is therefore the value with the highest likelihood given the data. As the data sample increases to infinity it has desirable properties, such as statistical consistency (converges to the true value) and statistical efficiency (no other estimator is more efficient).
P value	What we obtain from a statistical test which tells us the probability of obtaining our data or more extreme data assuming the null hypothesis is true. Often misunderstood.

TABLE 1 Interpreting support values

LR (1/LR)	S	Interpretation H_1 versus H_2
1 (1.00)	0	No evidence either way
2.7 (0.37)	1	Weak evidence
7.4 (0.14)	2	Moderate evidence
20 (0.05)	3	Strong evidence
55 (0.02)	4	Extremely strong evidence

The first column gives the likelihood ratio (LR) and inverse (in parenthesis). The second column gives the natural logarithm of LR (log LR), known as the support (S). The third column gives the verbal description for the level of support for hypothesis H_1 versus H_2 for each value of S . Negative values for S represent support for H_2 versus H_1 . Support values can be given to one decimal place. This scale provides graded evidence without a threshold, unlike P values. The scale runs from $-\infty$ to $+\infty$; the midpoint of zero represents no evidence either way. This table has been adapted from that suggested by Goodman & Royall (1988). British courts use the same LR scale, but ramped up, meaning that the court considers $S = 4$ as only 'moderate evidence', and $S = 8.6$ as 'strong evidence'. Therefore, compared with scientists, courts require more than twice the evidence to influence their judgements.

2 | LIKELIHOOD

Likelihood is at the very heart of statistical inference (Edwards, 2015). It uses the standard error, exactly like statistical testing. In fact, all the statistics with which we are familiar, such as z , t , F and R^2 , are used to calculate likelihoods and their ratios. The likelihood approach exploits maximum likelihood estimation to obtain a maximum likelihood estimate (MLE), and this has desirable statistical properties. An MLE is basically a statistic, such as the observed mean ($\bar{X}$) that we obtain from our sample of data. As its name suggests, it is the most likely estimate for the parameter that we might be interested in, such as the population mean (μ).

The likelihood approach is one of the main statistical approaches, along with frequentist hypothesis testing (P values and 95% confidence intervals), Bayesian and information criteria approaches. Despite providing the mathematical foundations for these other approaches it is, surprisingly perhaps, much less known. It is the Cinderella of inferential statistics.

FIGURE 1 The distribution of heights for men (m) and women (w) from 20 countries (aggregate of Europe, North America, Australia and East Asia) in a cohort born between 1980 and 1994. The curves are likelihood functions calculated using the obtained means and standard deviations. Likelihood functions are scaled so that their maximum occurs at one (see vertical axis). The mean for each distribution occurs at the peak likelihood value. The curve for men (dashed) is slightly wider because the distribution has a larger standard deviation. A randomly selected person has a height of 180 cm and is indicated by the thin vertical line. The likelihoods, shown as horizontal dashed lines, represent the heights of the two curves where the thin vertical line hits them. The brackets are labelled with the likelihood values, and at the top right is given the likelihood ratio (LR) calculation. At the top left of the figure, the log LR, the support (S), is shown

3 | WHERE TO START?

Let us start with a simple, and very artificial, example to explain the concept. It is well known that height has an almost normal distribution. The heights for a large cohort of men and women across 20 countries were measured (see this fascinating webpage: <https://ourworldindata.org/human-height#height-is-normally-distributed>). From these data, we have the means and standard deviations, which allows us to plot the normal distributions for men and women (Figure 1).

If we had the raw data, we could have plotted frequency histograms for men and women separately, where the vertical axis would be number of men and women in each respective histogram bin. Adding smoothed curves over the histogram bars would give us something very similar to Figure 1. Instead, we could make plots knowing the means and standard deviations for men and women and assuming the distributions were normal. This would give us a bimodal plot, with a peak for the mean for each sex. The vertical axis would now represent the probability density for height as we move along the horizontal axis from short to tall people. If we scaled this plot so that each of the peaks was one, then these distributions would be the likelihood functions for each sex. A short-cut to doing this is to use the following simple equation for each sex:

$$\text{Likelihood} = e^{-\frac{z^2}{2}} \quad (1)$$

This would be applied to values for height starting at, say, 140 cm, and incrementing by 0.1 cm until 210 cm. This could be done for women first (using their mean and standard deviation), producing the left-hand curve, and repeated for men. This is easy in Excel; try it!

There are clearly two separate curves, but there is a fair amount of overlap between them. The maximum height of the curves, at the respective means, is one. Now imagine that we pick a person at random from the population of the 20 countries used for the data. We measure the height of this person as 180 cm. From its position in Figure 1, the vertical line, we can see that this is near the mean for men and far from the mean for women. We can calculate the likelihoods, which are the same values as where the vertical line cuts the women's and men's curves. Callout values are given for these as 0.096 and 0.978, respectively. These are merely the heights at those points of the curves, remembering that their maximum heights are one. We can calculate each of these manually from the z value and Equation 1 given above. The z value for women is:

$$z_w = \frac{\bar{X} - 180}{SD} = \frac{164.7 - 180}{7.07} = -2.1641$$

The likelihood for women is then:

$$L_w = e^{-\frac{z_w^2}{2}} = e^{-\left(\frac{-2.1641^2}{2}\right)} = 0.09617$$

The same for men, using their mean and standard deviation:

$$z_m = \frac{178.4 - 180}{7.59} = -0.2108$$

And likelihood:

$$L_m = e^{-\left(\frac{-0.2108^2}{2}\right)} = 0.97803$$

FIGURE 2 Graphical illustration of the different scales for likelihood ratio (LR) and support (S). The LR scale on the left shows that all the LRs in favour of hypothesis H_2 are 'compressed' between zero and one. Above one, the LR favours H_1 , proceeding up to infinity. On the right is the linear scale provided by the logarithmic transform of the LR, S . This has as its midpoint zero, representing no evidence either way. Above this are values in favour of H_1 that proceed to $+\infty$. Below this are negative values in favour of H_2 that proceed to $-\infty$. If the null hypothesis, H_0 , is used, then this would replace H_2

The likelihood ratio (LR) for these, say the likelihood of the person being a man rather than a woman, is:

$$LR_{mw} = \frac{L_m}{L_w} = \frac{0.97803}{0.09617} = 10.17$$

This means that, given the data (180 cm), this person is >10 times more likely to be a man than a woman. If we invert the LR then we would get $LR_{wm} = 0.098$, meaning that the person is less than one tenth as likely to be a woman as a man. Taking the natural logarithm of 10.17 gives us the support (S) of 2.3. For the inverse, with perfect symmetry, our $S_{wm} = -2.3$. Consulting Table 1, we see that this would be regarded as moderately strong evidence in favour of a man rather than a woman. This calculation flows naturally from the visualization of the problem. We have two hypotheses, and we have calculated which of these is most likely given our data.

Figure 2 shows a graphical representation of Table 1, comparing the LR scale with the S scale. The LR scale is compressed at the lower end, whereas the S scale is linear from $-\infty$ to $+\infty$. For the LR scale, one represents no evidence either way, whereas the corresponding value on the S scale is zero.

An analysis using P values is less obvious, but let's try. We could start by specifying that the null hypothesis is that the 180 cm person is a woman. Using the women's z value of -2.164 gives us a tail probability of 0.015. Convention is to use a two-tailed test, so we need to double our P value. We can declare the result significant with $P = 0.03$, rejecting our null hypothesis that the person was a woman (i.e. decide the person is a man). There are a couple of puzzling things. First, it is unclear why we need to consider a tail region containing data we have not collected, such as women taller than 180 cm. Second, what happens if we select a short person, with height of 150 cm? The same analysis would lead us again to reject the null hypothesis ($P = 0.038$), but for what? The way to deal with this is to use a one-tailed test, meaning that only values greater than the mean for women can be considered as evidence against it being the height of a woman. Given that the likelihood analysis includes both hypotheses (woman versus man), the same 150 cm person would be clear and unambiguous; it would provide extremely strong evidence ($S = 4.8$) that the person was a woman rather than a man.

One interesting solution using P values is to convert them to so-called 'surprisal' values (Greenland, 2019). This means taking the logarithm base 2 of the inverse of the one-tailed P value. For our 150 cm person, this would provide positive evidence for being a woman of 5.7 bits of information, whereas for the 180 cm person this would be 6.0 bits of information in favour of a man.

For completeness, we could do a Bayesian analysis for the 180 cm person. We need to assume that there are equal numbers of men and women in the population, giving us even prior odds. From the LR of 10.17, we convert these odds to a probability. This gives us 0.91, which means that there is a 91% posterior probability of the person being a man. This is correct and perfectly legitimate, given, of course, that we know the prevalence beforehand.

From this admittedly artificial example, we can move to something more realistic. In the above example, we assumed equal numbers of men and women in the population, meaning that we were justified in calculating a posterior probability using the Bayesian approach. But what happens if we do not know the prevalence? Generally, we do not know the prior probability of hypotheses we are interested in. For this reason, it is problematic using a simple Bayesian approach when analysing data from experiments, although there are more sophisticated semi-Bayes and empirical Bayes approaches that can be used (Greenland, 2006)

4 | LET'S GET REAL

Consider a more realistic example. We are measuring blood levels of a protein in a group of volunteers. Our procedure is to take a baseline measurement and then another measurement after an intervention. The intervention is supposed to increase the blood level of the protein. A previously published study by another research group showed that the intervention increases the level of the protein by 3 units. These are our previous data (hypothesis H_2). For the intervention to be regarded as effective, an expert clinician says it should cause an increase of ≥ 2 units. This is our minimal effect size (hypothesis H_1).

We carry out our study in 11 women, where the intervention causes a mean increase of 2.2 units with standard deviation of 1.75 units. We use these data to produce a likelihood function for the mean difference

FIGURE 3 The likelihood function for the observed data. The function is centred on the mean value of 2.2, which is the maximum likelihood estimate, indicated by the dashed vertical line. The fine lines hitting the curve represent the different hypothesis means. These are labelled with callouts for the null, the minimal and previous values. The lines for the null value are so low that they are not visible. Each has a horizontal line from the curve to the nearest vertical axis, and the likelihood value given. These represent their height as a proportion of one (the maximum). The thicker rectangular lines represent the likelihood interval for $S = -2$, which reaches $e^{-2} = 0.1353$ on the vertical axis

from baseline, as shown in Figure 3. We already plotted likelihood functions for the heights of men and women. We used Equation 1 to do this, indicating this can be done easily in Excel. To see how to obtain Figure 3, go to Box 2 (*How to create a likelihood function*).

The shape of the likelihood function is identical to the sampling distribution of the mean around the mean of 2.2, except for its vertical axis maximum which, for convenience, is one. If we were doing a null hypothesis test then this distribution would be situated over the null value, here zero, and we would be calculating the tail area from 2.2 and up. In the likelihood approach, we situate the distribution over the observed mean of 2.2 units. The height of the curve at different hypothesis values is indicated by the thin lines in Figure 3. The minimal effect size is shown at 2 units, and 0.924 shown on the left side. This is the likelihood for the minimal effect size. We also see a line for the published study of 3 units, and the likelihood of 0.320 is shown to the right. There is a third likelihood shown at the bottom left of the graph. This likelihood of 0.0039 is for the null of zero units. The line hitting the curve is so low that it is difficult to see. The values illustrated in Figure 3 are presented more fully in Table 2.

The t value is calculated using the usual equation:

$$t = \frac{\mu - \bar{X}}{s/\sqrt{n}} \quad (2)$$

The hypothetical mean is represented by μ , from which the observed mean is subtracted. The observant reader will notice a minor difference here from the usual formula, which has the observed mean minus the hypothetical mean. This is because in the likelihood approach the observed data are fixed and the hypotheses vary. The denominator in the equation is the usual standard error. The likelihoods in the fifth column in Table 2 are calculated using t with this general equation:

$$\text{Likelihood} = \left(1 + \frac{t^2}{\text{d.f.}}\right)^{-N/2} \quad (3)$$

Where d.f. is the degrees of freedom for t , and N is the total sample size. For related one-sample and paired designs, $N = n$. If there are two independent groups, $N = n_1 + n_2$.

The sixth column of Table 2 gives the reciprocal of the likelihood. For example, this tells us that the mean value of 2.2 is 254.8 times more likely than the null value of zero. This is also known as the maximum likelihood ratio and resembles the P value insofar as it simply represents the evidence against the null hypothesis. Here it translates to a support of 5.5 in the last column, indicating more than extremely strong evidence (see Table 1). The frequentist analysis gets $P = 0.002$. The other $S_{1/L}$ values tell us the strength of evidence for the mean versus the other hypotheses. Considering the minimal value (H_1), there is close to zero evidence (a support value of 0.1), which might be expected because the value is so close to the mean of the sample, which is 2.2. Finally, considering the previous finding of 3.0 (H_2), the evidence is weak (1.1). These support values for reciprocal likelihoods will always be positive because the likelihoods are always ≤ 1 .

We can now compare these likelihoods by forming ratios of them. We can do this for any hypotheses of interest. Table 3 shows the calculations for different hypotheses that we wish to compare. The top row gives the LR and S for the minimal effect size versus the null (which we can denote as LR_{10} and S_{10} , respectively) and the second row gives the LR and S for the previous value of 3.0 versus the null (LR_{20} and S_{20}). Given that these support values both comfortably exceed four, we can see that, given the data, there is extremely strong evidence for both the minimal effect size and the previous value versus the null. The bottom row gives the previous versus the minimal effect size (LR_{21} and S_{21}). The negative support value represents evidence in favour of the minimal effect size versus the previous value, but the evidence is weak because it is close to unity (see Table 1).

This approach allows us to simply compare any two heights on the likelihood function using a ratio. This gives us an immediate sense of how likely one hypothesis is compared with another. We see from our

BOX 2**How to create a likelihood function**

In Excel, label the first row of three columns with **Mean**, **Likelihood** and **t**. In the first column, enter -1 and -0.95 , representing the mean increase in protein, as shown here:

	A	B	C
1	Mean	likelihood	t
2	-1	$=(1+C2^2/(10))^{-(11/2)}$	$=(A2-2.2)/0.52764485$
3	-0.95		
4			

In the second and third columns, type in the formulae exactly as shown above, pressing Enter after each one. Column C calculates t values, using Equation 2, referencing the column A value in the same row (cell A2). The respective values for the mean and standard error are included. Column B calculates the likelihood, using Equation 3, referencing the t value calculated in column C (C2). The degrees of freedom and n are within the formula.

Next, select the top two values in column A and use the **AutoFill** feature to complete the sequence of values (each incrementing by 0.05) until you reach 5.0. Return to the top of the sheet and select the two cells containing formulae (B2 and C2) and use **AutoFill** to replicate the formulae to the end of the sheet (row 122). Columns B and C will now contain the calculated likelihood and t values, and the first few, the middle few and last few rows should look like this:

Mean	likelihood	t
-1	0.00021	-6.06469
-0.95	0.00024	-5.96992
-0.9	0.00027	-5.87516
.	.	.
2.15	0.99508	-0.09476
2.2	1.00000	0.00000
2.25	0.99508	0.09476
.	.	.
4.9	0.00085	5.11708
4.95	0.00073	5.21184
5	0.00063	5.30660

Select the first two columns and click on **Scatter** plot from the **INSERT** tab. This will give you the basic likelihood plot, shown below left, that is used in Figure 3. The same can be produced in the R statistics package using a single line of code, to give the plot shown below right:

```
curve((1 + ((2.2-x)/0.5276448530)^2/10)^{-(11/2)}, xlim = c(-1,5), ylab = "Likelihood")
```

TABLE 2 Likelihood and support values calculated from t values for hypothesized population mean values

H	Represents	Value	t	L	$1/L$	$S_{1/L}$
0	Null	0	-4.1695	0.0039	254.8447	5.541
1	Minimal	2	-0.3790	0.9245	1.0816	0.078
2	Previous	3	1.5162	0.3204	3.1206	1.138

The three different hypotheses (H) are numbered in the first column, with the second column giving what they represent and the third column being the hypothesized mean value for each of them. The likelihood (L) value is calculated from the t value. The next column gives the reciprocal of the likelihood, and the last is the support (S) for the reciprocal.

TABLE 3 Likelihood ratios (LR) and support (S) values given for three pairs of hypothesis comparisons

Hypotheses	LR	S
H_1 versus H_0	235.6143	5.462
H_2 versus H_0	81.6649	4.403
H_2 versus H_1	0.3466	-1.060

data, for example, that the minimal effect value is 236 times more likely than the null value. The natural logarithm of the LR gives us the support, S , which represents the strength of evidence (consulting Table 1 for its interpretation).

It is always useful to construct likelihood intervals (also known as support intervals) for the same reasons that confidence intervals are useful in frequentist statistics (Cumming & Calin-Jageman, 2017). A likelihood interval is also useful when we are not sure which particular hypotheses to compare. An $S-2$ likelihood interval gives us a range of values around the MLE that would produce weaker evidence than $S = -2$ (i.e. $e^{-2} = 0.1353$) when compared against our observed mean of 2.2. The $S-2$ likelihood interval is shown in Figure 3 between 1.10 and 3.30 and demonstrates a close numerical correspondence to the 95% confidence interval from 1.02 to 3.38.

TABLE 4 Likelihood ratios (LR) and support (S) values when $N = 33$

Hypotheses	LR	S
H_1 versus H_0	6805159.2786	15.73
H_2 versus H_0	338997.6799	12.73
H_2 versus H_1	0.0498	-3.00

5 | LET'S ADD DATA

Using the evidential approach, we are free to add more data to existing data. Let's say we added another 22 participants. Assume that the observed mean and standard deviation remain the same at 2.2 and 1.75 units. The likelihood function, like the sampling distribution for the mean, narrows as the sample size increases. This strengthens the existing extremely strong evidence for the minimal effect and the previous data versus the null, meaning that the support values are now in double figures (Table 4). We also notice that the evidence in support of the previous data versus the minimal effect size is weakened. This suggests an exaggerated earlier finding (as is often the case; Alahdab et al., 2018). In fact, there is a linear increase in the strength of evidence as the sample size increases (three times the data gives us approximately three times the S value). We can see this in Figure 4, which demonstrates the last of the hypothesis comparisons (H_2 versus H_1), starting at $N = 5$ and incrementing by five until 100.

As indicated earlier, we would not be justified in adding further data to our initial sample if we used conventional frequentist statistics. This is tempting to do when the initial sample fails to produce $P < 0.05$ and the sample mean looks promising. If further data are added to existing data then eventually the investigator will always obtain $P < 0.05$, even if the null hypothesis is true. Indeed, continued long enough, any minuscule P value ($P < 0.01$ or 0.001 etc.) will be obtained 100% of the time. Such a process would produce an inferential error, specifically a type I error.

In frequentist statistics a second error can be made, a type II error, which occurs if insufficient data are collected and the null hypothesis is false. This contrasts with the evidential approach where no such errors

FIGURE 4 Sample size (N) increases the evidence in a linear fashion. This plot represents the support (S) values calculated when comparing hypothesis H_2 versus H_1 . The negative values for S are interpreted as evidence against H_2 relative to H_1 , i.e. weakening of evidence. Alternatively, the evidence for H_1 relative to H_2 is increasing. Using z rather than t would have produced a line that went through zero: no data = no evidence

are made. This is difficult to believe and will need a little explanation. The evidential approach provides evidence, and this evidence can either be misleading or it can be weak (Royall, 1997, 2000). Misleading evidence is evidence that exceeds some specified strength of evidence (e.g. $-2 > S > 2$) but in support of the wrong hypothesis. This can happen, but the probability of it happening is related to the S value obtained and is equal to $P(\text{misleading evidence}) \leq e^{-\text{abs}(S)}$. That is, the probability is exponentially related to the absolute value of S . Let us say that we obtained $S = 3$, $e^{-3} = 0.05$. This means that the probability of misleading evidence would be $\leq 5\%$. The same probability of misleading evidence would be obtained for $S = -3$. Remember now that as more data are collected the S value changes linearly in proportion to the amount of that data; hence, S would continue to increase beyond two or decrease below minus two. The only exception would be the unlikely situation where the observed mean was exactly between the two hypothesis values used for the LR . If there was any doubt here, then a likelihood interval would settle the matter. Weak evidence means that it is not strong enough, i.e. $-2 < S < 2$. What do we do? Well, we simply need to collect more data!

Thinking that this evidential approach can easily be abused, consider the following. I have a preferred hypothesis, H_1 . I am interested in obtaining support for it against H_0 . I start collecting data. After each additional data point, I calculate S_{10} . Because of my bias, I will ignore any evidence in favour of H_0 , however strong it might be. I will stop only when I obtain sufficient evidence in favour of my preferred hypothesis, H_1 . If H_0 is true then the probability is $> 1 - e^{-S}$ that my research would never end. If we were to use $S_{10} = 3$, then the probability of carrying on forever would exceed $1 - e^{-3} = 0.95$. The stronger the evidence I would require for my false pet theory, the less chance I have of ever obtaining it.

This all seems better than using P values. Even if I follow the rules of not adding to my initial sample, the probability of making a type I error is constant, regardless of the sample size I choose. With the evidential approach, the probability of misleading evidence decreases with sample size.

6 | MORE STATISTICS

Likelihood is a complete approach to statistics. It can obtain evidence S for regression, correlation, ANOVA and non-parametric analyses, and any others you can think of. Categorical data are easily analysed. Consider the following coronavirus disease 2019 (COVID-19) data from South Korea (retrieved from <https://twitter.com/DrEricDing/status/1239226811185344517>):

Sex	Health status		Total
	Dead	Alive	
Male	41	3095	3136
Female	34	4992	5026
Total	75	8087	8162

The grand total of 8162 represents those who had tested positive for COVID-19. Fewer men than women tested positive (3136 versus 5026). Despite this imbalance, fewer women died. The usual frequentist test of association would give $\chi^2(1) = 8.44$, $P = 0.004$.

The evidential approach calculates the support using the general equation with natural logarithms:

$$S = \sum_i \left\{ O_i \times \log \left(\frac{O_i}{E_i} \right) \right\} \quad (4)$$

Like the χ^2 analysis, observed (O) and expected (E) values are used in the equation. The expected values for an association analysis are worked out in the same way, using the marginal totals (these are the vertical and horizontal totals in the margins of the contingency table). The four terms that we sum would be:

$$S = \sum_i \{ 14.4576 + (-) 12.1596 + (-) 10.4129 + 12.1984 \}$$

$$S = 4.084$$

FIGURE 5 The support function for differences in the dead male cell counts. The support intervals for two and four are shown by dashed lines. The changes assume fixed marginal totals. The term 'cell' here refers to one of the four values located within the body of the coronavirus disease 2019 (COVID-19) data shown in the main text

FIGURE 6 The support function for the odds ratio, which is maximal at the observed odds ratio of 1.945. The support intervals for two and four are shown by dashed lines. The function assumes fixed marginal totals

Referring to Table 1, we see that this represents extremely strong evidence for the interaction between sex and health status *versus* no association. We can obtain an exact support function, which informs us of the degree of departure from the observed values that could be expected given the data.

Figure 5 shows the support function (log LR) for departures of cell counts assuming that the marginal totals are fixed. For example, if there were nine fewer male deaths, giving the cell count of 32, then this would result in moderate evidence for a difference from the observed data (the point falls below dashed line for support minus two). The expected number of male deaths, given the marginal totals (for the null model), is 28.8165. This is 12.1835 difference from the observed 41, falling slightly below the dashed support interval line for minus four. More precise calculation with the function gives us support -4.084 . This is made positive to $S = 4.084$ because inversion of the likelihood, as we saw in Table 2, changes the sign of the support. This value agrees with our earlier calculation for the interaction. The same support function can be calculated for each of the four cells. They look similar because the marginal totals are fixed.

We can produce other support functions for these data, such as relative risk and gamma. A popular statistic is the odds ratio. For these data, we obtain an odds ratio of 1.945. The support curve is shown in Figure 6. Consistent with the analysis of the interaction for these data, the observed odds ratio compared with the null value of one gives the same $S = 4.084$.

The other components that can be analysed give main effects of $S = 220.8238$ for sex and a massive (as expected) $S = 5231.0812$ for health status. These total to 5455.9885, which is exactly the value obtained by an analysis where $8162/4 = 2040.5$ is the expectation in each of the four cells. This total is constrained only by the grand total and therefore has three degrees of freedom. These analyses can be extended to tables of any size and dimension, as explored by log-linear analysis.

7 | CONCLUSION

The evidential approach requires a different way of thinking. We need to consider the weight of evidence for one hypothesis *versus* another, given our data. It does not involve probabilities or the probability tail areas that provide evidence only against a single hypothesis. It does not involve using a single threshold value, such as 0.05, but a graded measure of evidence for two competing hypotheses using values from $-\infty$ to $+\infty$.

Using likelihoods for statistical inference is not without its detractors. Perhaps the most vocal is Deborah Mayo, who has posted several blog pieces, with one appearing to show that the likelihood approach exaggerates evidence (retrieved from <https://errorstatistics.com/2014/11/25/how-likelihoodists-exaggerate-evidence-from-statistical-tests/>). However, this is also a charge previously levelled against P values (Goodman, 1993; Goodman & Royall, 1988).

For many, the evidential approach is so very different that it is hard to think clearly about it. It's as if we've been brainwashed into the statistical testing P value approach. This is even though the evidential approach is conceptually easier to grasp and the calculations are more straightforward to perform. Moreover, in our calculations we do not need to worry about look-up tables for critical values or exact integrals for P values.

To mix metaphors, likelihood is the real deal; it does exactly what it says on the tin, nothing more, nothing less. You're thinking, 'This evidential approach is grand, sign me up. When can I start?'. Well, that's the problem. There is virtually no support for the approach in any of the standard software packages, including R, Minitab, SPSS, SAS, GraphPad Prism and Stata. This means that calculations, although simple, have to be done by hand, in an Excel spreadsheet, or using a series of programming commands in a package such as R. Moreover, the evidential approach hardly gets a mention in most statistical courses. Finally, there are relatively few books that explain how to use it. The

best books are those by Edwards (1992) and Royall (1997), although neither is very accessible to the average researcher. Then there are the books by Clayton & Hills (2013) for epidemiology and Aitkin & Taroni (2004) for forensic evidence, neither of which is an appropriate text for those carrying out laboratory experiments. An attempt has been made to provide a practical step-by-step guide for researchers (Cahusac, *in press*).

COMPETING INTERESTS

None declared.

ORCID

Peter Cahusac <https://orcid.org/0000-0003-4976-2834>

REFERENCES

- Aitkin, C. G. C., & Taroni, F. (2004). *Statistics and the evaluation of evidence for forensic scientists* (2nd ed.) Chichester: John Wiley & Sons.
- Alahdab, F., Farah, W., Almasri, J., Barrionuevo, P., Zaiem, F., Benkhadra, R., ... Wang, Z. (2018). Treatment effect in earlier trials of patients with chronic medical conditions: A meta-epidemiologic study. *Mayo Clinic Proceedings*, 93, 278–283.
- Altman, D. G., & Bland, J. M. (1995). Absence of evidence is not evidence of absence. *British Medical Journal*, 311, 485–485.
- Cahusac, P. M. B. (In Press). *Evidence-based statistics: An introduction to the evidential approach – from likelihood principle to statistical practice*. Chichester: Wiley.
- Clayton, D., & Hills, M. (2013). *Statistical models in epidemiology*. Oxford: Oxford University Press.
- Cumming, G., & Calin-Jageman, R. (2017). *Introduction to the new statistics*. New York: Routledge.
- Dixon, P. (2003). The p-value fallacy and how to avoid it. *Canadian Journal of Experimental Psychology/Revue canadienne de Psychologie Expérimentale*, 57, 189–202.
- Drummond, G. (2020). A world beyond P: Policies, strategies, tactics and advice. *Experimental Physiology*, 105, 13–16.
- Edwards, A. W. F. (1992). *Likelihood* (2nd ed., Vol. 2). Baltimore: John Hopkins University Press.
- Edwards, A. W. F. (2015). Likelihood in statistics. In J. D. Wright (Ed.), *International encyclopedia of the social & behavioral sciences* (2nd ed., pp. 116–119). Oxford: Elsevier.
- Fisher, R. A. (1956). *Statistical methods and scientific inference*. Edinburgh: Oliver & Boyd.
- Goodman, S. N. (1993). *p* Values, hypothesis tests, and likelihood: Implications for epidemiology of a neglected historical debate. *American Journal of Epidemiology*, 137, 485–496.
- Goodman, S. N., & Royall, R. M. (1988). Evidence and scientific research. *American Journal of Public Health*, 78, 1568–1574.
- Greenland, S. (2006). Bayesian perspectives for epidemiological research: I. Foundations and basic methods. *International Journal of Epidemiology*, 35, 765–775.
- Greenland, S. (2019). Valid *P*-values behave exactly as they should: Some misleading criticisms of *P*-values and their resolution with *S*-values. *The American Statistician*, 73(Suppl 1), 106–114.
- Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika*, 20A, 175–240.
- Royall, R. M. (1997). *Statistical evidence: A likelihood paradigm*. London: Chapman & Hall.
- Royall, R. M. (2000). On the probability of observing misleading statistical evidence. *Journal of the American Statistical Association*, 95, 760–768.

How to cite this article: Cahusac P. Data as evidence. *Experimental Physiology*. 2020;105:1071–1080.
<https://doi.org/10.1113/EP088664>