Chemometrics and Intelligent Laboratory Systems
Available online 29 October 2019, 103884
In Press, Journal Pre-proof

The use and misuse of p values and related concepts

Richard G. Brereton ✉
Show more

https://doi.org/10.1016/j.chemolab.2019.103884

Highlights

- Historical review of the concept of p values.
- Relationship between common concepts such as False Positive Rates and p values and type 1 and type 2 errors.
- Reproducibility of p values.
- Current controversy over use of p values.

Abstract

The paper describes historic origins of p values via the work of Fisher, and the competing approach by Neyman and Pearson. Concepts of type 1 and type 2 errors, false positive rates, power, and prevalence are also defined, and the merger of the two approaches via the Null Hypothesis Significance Test. The relationship between p values and false detection rate is discussed. The reproducibility of p values is described. The current controversy over the use of p values and significance tests is introduced.

Keywords

p value; Type 1 error; Type 2 error; False positive rate; Hypothesis test; Significance test; Null hypothesis

1. Introduction

Historically, chemometrics was firmly based within analytical and physical chemistry. Merriam-Webster defines chemometrics as "*The application of statistics to the field of chemical analysis.*

The system doesn't directly measure boiling points and other physical properties; rather, it used chemometrics to infer them from the process stream's chemical composition." [1]. The International Chemometrics Society defines chemometrics as "*Chemometrics is the science of relating measurements made on a chemical system or process to the state of the system via application of mathematical or statistical methods.*"

B

chromatogram, or NIR spectrum. A major aim of classical chemometrics, for many decades, would be to take chemical measurements, primarily from spectra or chromatograms, and use them to estimate something such as the concentration of an analyte, the rate of a reaction or whether a drink is adulterated. In almost all cases, as in classical analytical chemistry, the answer (usually quantitative) is known in advance for the purpose of method development. For example, can we use uv/vis spectroscopy to estimate the concentration of an additive? We prepare accurately known standards and then use the spectra to assess the concentration as accurately as we can: different methods can be compared by how well they estimate the concentration. Sometimes confidence limits can be obtained. We can produce a series of solutions, for example of orange juices from different sources, and use NIR to determine their origin – we know in advance which extract comes from which source and so can compare methods to see which result in the lowest error in grouping the orange juices. Exploratory methods such as PCA (Principal Components Analysis) continue to complement these predictive approaches sometimes as a first step in data analysis and also have a long vintage in chemometrics [2], but this paper focuses primarily on hypothesis testing.

Statistical experimental design (DoE) is an exception in chemometrics, where p values are often used, sometimes inappropriately, to determine the probability a factor (either a variable or interaction) is significant. However there is commonly a lack of understanding in chemometrics, for example p values do not actually provide information about the significance of a factor, but about the significance of a model without this factor [3]. In addition, many experimenters use sharp cut-offs such as $p=0.05$ to determine whether they feel a factor is significant or not. A better understanding of the history and usage of p values can help the user of DoE packages to better appreciate the problems of basing decisions solely on p values.

Almost all classical texts on chemometrics take the approach that methods are described and compared according to how accurately or precisely they can predict either the provenance of a sample or a quantitative parameter usually using multivariate measurements. However over the past decades an important new focus of chemometrics has been in sciences such as metabolomics and heritage studies. In both cases it is often not possible to obtain standards of perfectly known provenance, and the aim is not so much to predict the known parameters or provenance of samples to a high degree of accuracy, but to test hypotheses. For example, if we take samples of pottery from three geographic regions, does their method of manufacture or their origins come from the same source? We are not sure, and as the pottery may be thousands of years old, we clearly cannot do a modern day experiment. What we are interested in is how certain are we of one or more underlying hypotheses, which can be done using multivariate hypothesis testing, resulting in a probability that the three groups of pottery are from the same source.

Where the data are multivariate, chemometrics techniques often come into play. But many chemometricians, schooled in traditional analytical or physical chemistry, find hypothesis based science hard to appreciate. For example a chemometrician may want to compare classification techniques, if one method is preferred to another. The traditional approach is to take one or more datasets with known groupings and see which method predicts these groups with lowest error rates. However, in many modern applications we are not using these classification techniques for this purpose. We are asking whether the measurements support a hypothesis, for example that extracts from LCMS of serum can predict whether a subject has the early stage of a disease. A method that “perfectly separates” two groups may not be the most appropriate: the samples may not be perfectly separable, for example there may be other confounding factors such as genetics, age, diet, personal habits etc. that interact with the disease progression; there may be outliers, misdiagnoses etc. What we are trying to do is attempt to study a hypothesis.

Outside the core physical sciences, most experiments involve studying one or more hypotheses. The majority involve obtaining a probability that some idea is significant. Can we hypothesize that historic textiles from a region in Asia were coloured with a specific dye extracted from a plant grown in another region, and so deduce about trade routes especially when the colours have degraded with time? Hence many chemometricians are no longer primarily aiming to reach a known answer with a given level of accuracy, but to test a hypothesis. This change in core aims over the past two decades has been slow to be appreciated in many chemometrics circles, as its applications move from the core physical sciences to areas such as biology, heritage science or forensics.

Classical chemometrics has primarily been about algorithm development and improvement. Most classical chemometricians have been schooled in the physical sciences and hence the majority of the classical literature has been about announcing new and improved methods and showing they can reach a known answer with greater accuracy or precision – sometimes involving defining different criteria of success. PhDs and conference presentations often involve demonstrating the so-called superior quality of a new computational method. Most classical chemometrics does touch on

tl

Martens and Næs [4] that represented an important area of chemometric thinking in the 1980s and 1990s, whereas discussing in depth statistical concepts of estimation, never mentions p values and hypothesis testing.

Hence a whole literature has developed with very little discussion about hypothesis testing. Ideas such as p values, and Bayesian statistics, have very limited role in classical chemometrics, probably because most classical workers came from the fields of analytical and physical chemistry. Classical experimental design is an exception, but until recently where methods such as ASCA (ANOVA Simultaneous Components Approach), ANOVA-PCA (Analysis of variance PCA) and multilevel approaches have been introduced, were originally treated by many as a separate area of chemometrics largely involving analysis of univariate responses.

However, as the applications of chemometrics change, the concepts of p values, and related hypothesis and significance testing, will assume greater importance than in the past, and the aim of this article is to introduce these in a chemometric context. Within mainstream statistics, the use of p values have a long vintage and are currently subject to major controversy. Areas such as biology, medicine and psychology often require the calculation of p values for publication, however many practitioners have limited understanding. This article will discuss the history of p values, how they relate to concepts such as α and β, and false positive rates, the null hypothesis, and the some of the current controversies.

2. History

2.1. Early concepts

As often is the case, credit to ideas usually is given to the person that first reports them in a formal manner. Yet key ideas sometimes are reported in different forms over many decades or even centuries before being recognised, especially in mathematics and statistics.

Early credit to a concept we might now recognise as the p value is given to John Arbuthnot in the early 18th century [5]. He analysed the number of males and females born over 82 years, as registered in London, and found that each year there were more males than females. His hypothesis was that in the absence of any other explanation (we would call this the null hypothesis), each year there would be an equal chance of males than females, so, if this hypothesis were true, the chances males exceed females each year would be 0.5^{82} or 2×10^{-25}. This very small number we would now call a p value. His argument was that this value was so small that we can discount the hypothesis that a child was equally likely to be male than female, and have to seek an alternative explanation. He used this as evidence of divine intervention: nowadays we may look for other explanations for example as females were economically less valuable to males, some were not registered and given away or even left to die.

There are a series of descriptions of applications of similar concepts over the next two centuries, but it waited until Karl Pearson in 1900 to first formulate this, in the context of the χ^2 test [6].

2.2. RA Fisher, p values, the null hypothesis and significance test

Many methods we use in multivariate analysis were formally defined by R A Fisher in the 1920s and 1930s, from which we take much of modern terminology, and he carried on the work and popularised the idea of p values and the null hypothesis.

He introduced the famous example of the lady tasting tea [7,8]. A lady, Muriel Bristol, who was in Oxford in the 1920s, claimed she could determine by taste whether milk was added before or after tea: in England at that period it was traditional to drink tea with milk. He defined the “null hypothesis” as the hypothesis she could not taste the difference. She was presented with 8 cups of tea, 4 had milk in before and 4 after. She was asked to taste each cup and decide whether the milk had been added before or after.

She managed to correctly assign all 8 cups. His reasoning was that if she had been unable to distinguish (the null hypothesis), she would get this result 1 time in 70 (=4! 4!/8! = 0.014 using the binomial theorem). This he called the p value, and represents the proportion of times this result would occur if the null hypothesis were true. It was then usual to put a threshold often of 0.05. If the p value is less than this (ie the result would happen less than one time in twenty if the null hypothesis were true) there was sufficient evidence to reject the null hypothesis, and something was probably happening in the background. Fisher called this a significance test.

Fisher only considered the null hypothesis. He was not interested in the alternative. In his terminology the null hypothesis is an exact hypothesis. Consider another example. We are interested in whether the population of mice in field A have larger

minus that from field B equals 0. This is an exact hypothesis we are trying to disprove. The alternative hypothesis may be that the difference of means is 1 cm, or that those from field A have shorter tails because they fight a lot and bite each others' tails. We are not so interested in this and just want to see how frequently the difference in mean tail lengths we observe from a sample of mice would be observed if there were no such difference (ie a consequence of sampling rather than any underlying difference).

Arbuthnot's original observations about relative male and female births in London over the years, is a similar case. All his calculations were doing is saying how unlikely it was that there were equal numbers of males and females and variations over the years were just by chance. The reason behind this difference, whether divine providence or just that some females were left to die at birth as they were unwanted, is neither proven nor tested.

2.3. Neyman and Pearson, α and the alternative hypothesis

A different approach was pioneered by Ergon Pearson and Jerzy Neyman at around the same time as Fisher [9]. Although their approach and Fisher's were quite different, the concepts have often been combined in modern thinking.

Neyman and Pearson introduced the idea of an alternative hypothesis. The idea of hypothesis testing was to choose between the null hypothesis and the alternative. To do this they introduced further concepts.

The type 1 error rate of α was the probability of falsely rejecting the null hypothesis. This has some analogies to the p value, but as we will see, is different philosophically. The type 2 error rate of β is the converse, falsely rejecting the alternative hypothesis. The power of a test is defined to be 1 - β, ie how effective it is on the samples that represent the alternative population.

There were many further differences between these approaches. Neyman and Pearson introduced the terminology hypothesis test to distinguish from Fisher's significance test. They used a cut-off eg $\alpha = 0.05$ rather than a continuum that Fisher preferred.

Fisher's significance test was empirical, ie followed on from the observed data, whereas the Nyman-Pearson test was computed prior to measurements. Fisher might calculate from a series of experiments, for example that $p = 0.032$ that the null hypothesis is correct, and then use a significance level, often 0.05, that the null hypothesis is rejected. Neyman-Pearson would calculate a significance level of, for example $\alpha = 0.05$, in advance of experimentation, and then if the observed data was below this level after experimentation the null hypothesis could be rejected and the alternative hypothesis accepted. However the Neyman-Pearson approach also introduced β, so alternatively a value of β could also be defined as a criterion. The result of this approach was always to accept one of the two hypotheses, whereas Fisher would simply decide whether to accept or reject the null hypothesis, and involved no alternative hypothesis.

2.4. Combining and contrasting the two approaches

The two approaches by Fisher and Neyman-Pearson were considered by the original advocates as quite distinct. The NHST (Null Hypothesis Significance Test) was developed in the 1940s. Most modern texts and courses in statistical hypothesis formulation do not differentiate between the methods. Often a p value is calculated in order to test against α. Concepts such as specificity and sensitivity, requiring two hypotheses, are often mixed up with p values. Outside specialist statistics, most chemometrics experts would not distinguish the approaches.

Over the past few years, with increasing use of p values in the literature, there have been several important articles that try to disentangle the approaches. We will reference two. Jose Perezgonzalez discusses in detail the different philosophies and their merger [10], and states " *Both theories had enough similarities to be easily confused ……, especially by those less epistemologically inclined; a confusion fiercely opposed by the original authors …. —and ever since …. —but something that irreversibly happened under the label of null hypothesis significance testing. NHST is an incompatible amalgamation of the theories of Fisher and of Neyman and Pearson …* ". An excellent on-line paper by Raymond Hubbard and Susie Bayarri [11] is worth reading also, although somewhat mathematical. They state " *… there is a widespread failure to appreciate the incompatibility of Fisher's evidential p value with the Type I error rate, α, of Neyman–Pearson statistical orthodoxy. The distinction between evidence (p's) and error (α's) is not trivial. Instead, it reflects the fundamental differences between Fisher's ideas on significance testing and inductive inference, and Neyman–Pearson views of hypothesis testing and inductive behavior. Unfortunately, statistics textbooks tend to inadvertently cobble together elements from both of these schools of thought …* ".

C

to contrast the two approaches in depth, except to point out that by origins they are quite distinct, but in modern practice except in the more rigorous papers, are in practice merged.

3. Definitions

In this section we will look at various definitions arising from the two approaches for hypothesis and significance testing.

In Fig. 1 we illustrate the p value. The distributions are illustrated using normal probability distribution functions (pdfs) for simplicity, although the principles are relevant to any type of pdf. The dark blue area represents that proportion of an underlying null distribution exceeding a limit of p, the light blue representing the remainder of the null distribution. By definition, if a series of observations from the null distribution is repeated 20 times, on average only 1 will have a value equal to or exceeding a value corresponding to p = 0.05 (technically this is the one tailed p value and there is also an equivalent two tailed definition). So if the sample size is 100, 5 of the values will exceed the value corresponding to p = 0.05. Therefore if a sample size is quite large, we would still expect occasionally to exceed this value at p = 0.05. Formally, in the case illustrated, p can be defined by $\Pr(X \geq l \,|H)$ where H refers to a hypothetic distribution (eg a Gaussian of defined mean and standard deviation in the case illustrated), X is the observed data, l is the limit corresponding to the p value, and the | symbol means “given that".

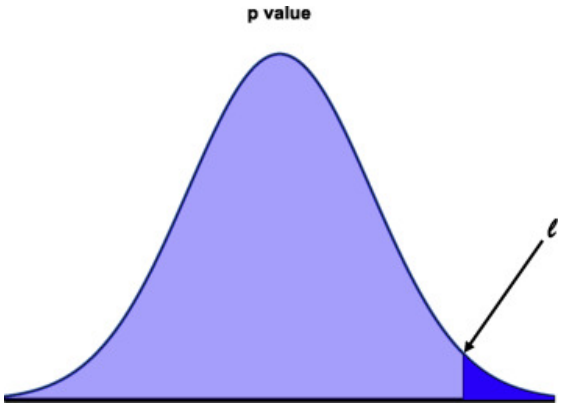


Fig. 1. Representation of p value; the dark blue area is the proportion of the distribution equal to p and l represents the limit corresponding to the p value. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Fig. 2 represents α or the type 1 error. The vertical line represents the decision threshold of a test: to the left the test is assumed negative (or null) and to the right it is positive (or representing the alternative hypothesis). In this case there is now an alternative distribution, although the type 1 error only requires the null distribution for its computation and appears superficially similar to p. In modern practice α is often used as a limit, and p as the experimentally determined value, but this was not the origin of this terminology. The concept of specificity is also sometimes used especially in clinical tests and equals 1-α. Definitions in Fig. 2 and later figures can be understood using a contingency table as in Table 1, where H_0 is considered to represent the negative hypothesis. TN represents the number of measurements or area under the curve of the null (blue) distribution that is to the left of the decision threshold, and α = FP/(TN + FP).

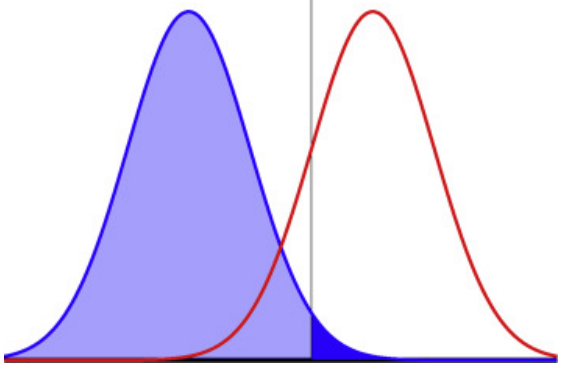


Fig. 2. Representation of α or type 1 error: the blue distribution is the null distribution and the red the alternate distribution; the dark blue area is the proportion of the blue distribution equal to α and the decision threshold is represented by a vertical line. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 1. Simple contingency table, P = positive and N = negative; T = true and F = false.

	Fail to reject H_0	**Accept H_A**
H_0 true	TN	FP
H_A true	FN	TP

Fig. 3 represents β or the type 2 error. Now only the alternative (sometimes called positive) hypothesis is required for computation. Dependent on application either the type 1 or type 2 error is more important to minimise. In forensics, if we define the null hypothesis that the defendant is innocent, it is important to reduce false positives (type 1) because in court it is essential to be sure of convictions and only convict a defendant if there really is a very small chance of being wrong. However in clinical diagnosis, it is often important to look into all possible issues even if some are false negatives, and so minimise the type 2 error with the risk of treating some patients who are actually healthy. Numerically we can define $\beta = FN/(TP + FN)$.

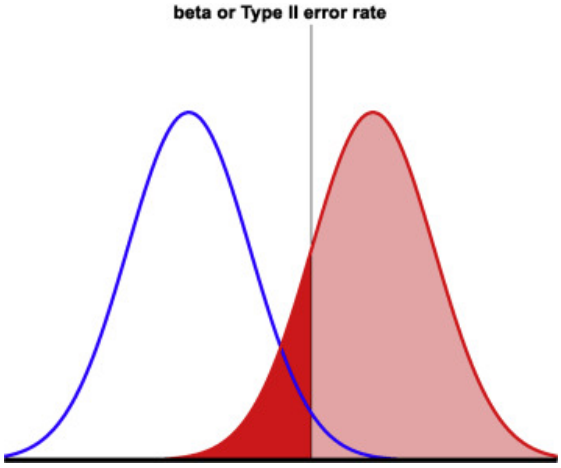


Fig. 3. Representation of β or type 2 error: the blue distribution is the null distribution and the red the alternate distribution; the dark red area is the proportion of the red distribution equal to β and the decision threshold is represented by a vertical line. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

The Neyman-Pearson approach introduces an important concept, the power of test (or sensitivity) as illustrated in Fig. 4. The greater the power, the better the test is (for positive cases). So a power of 0.9 means that if a case is positive (or belongs to the alternative distribution) in 90% of the cases, it will also test positive. Such a situation is not part of Fisher's original approach as he only considered the null distribution, his view being that the alternative hypothesis was not an exact hypothesis and so harder to test. Returning to Arbuthnot's original work, he only could test the null hypothesis easily. However in many applications of chemometrics eg in metabolomics to see whether an extract from a person's plasma suggests they have an early stage of a disease, we may be interested in the alternative hypothesis: the null hypothesis might suggest they are not diseased, but will not give us an idea of what disease they could be suffering from. Numerically we define 1-β = TP/(TP + FN).

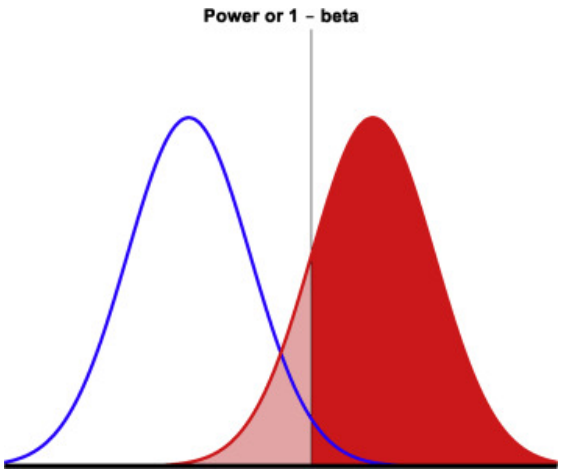


Fig. 4. Representation of 1-β or power: the blue distribution is the null distribution and the red the alternate distribution; the dark red area is the proportion of the red distribution equal to 1-β and the decision threshold is represented by a vertical line. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

In chemometrics, we are often interested in other aspects, and a very important one is the false positive rate (FPR). For example, if we are screening several metabolites for their biological activity or as diagnostic for a particular condition or genetic trait, we want to know the chance a given potential marker is a true one. The FPR is illustrated in Fig. 5. Both the null and alternative hypotheses (negative and positive) are required. All observations to the right of the vertical line are deemed positive. The blue area represents the portion of observations to the right of this line that are false. In the minds of some, the FPR and the p value of a potential marker are often confused, and we will look at this confusion in the next section. Numerically FPR = FP/(FP + TP) or can be defined in terms of α, β and $\mathcal{T}$ (prevalence) as defined in Section 4.1.

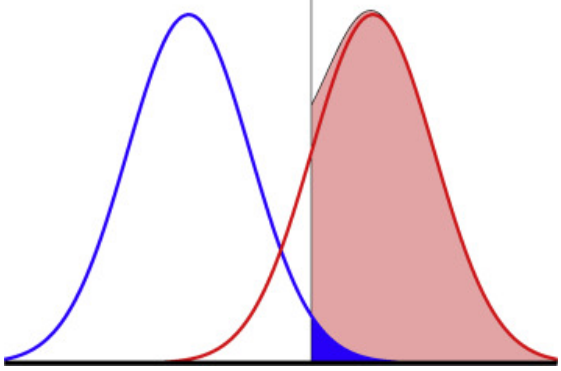


Fig. 5. Representation of false positive rate; the dark blue area is the proportion of the joint distributions to the right of the decision threshold (represented by a vertical line) equal to the FPR. The line red area above the curve equals the dark blue area. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Finally we introduce the concept of prevalence in Fig. 6 for two distributions. The red area represents the true positives whereas the grey area everything else. In many situations encountered in chemometrics, the prevalence is often quite low as in the bottom diagram. If we screen for markers, we usually expect only a very small portion of the peaks tested to be true markers. Numerically we can define prevalence as $T = (TP + FN)/(TP + TN + FP + FN)$, ie the proportion of all observed peaks that represent true + markers in the case discussed above.

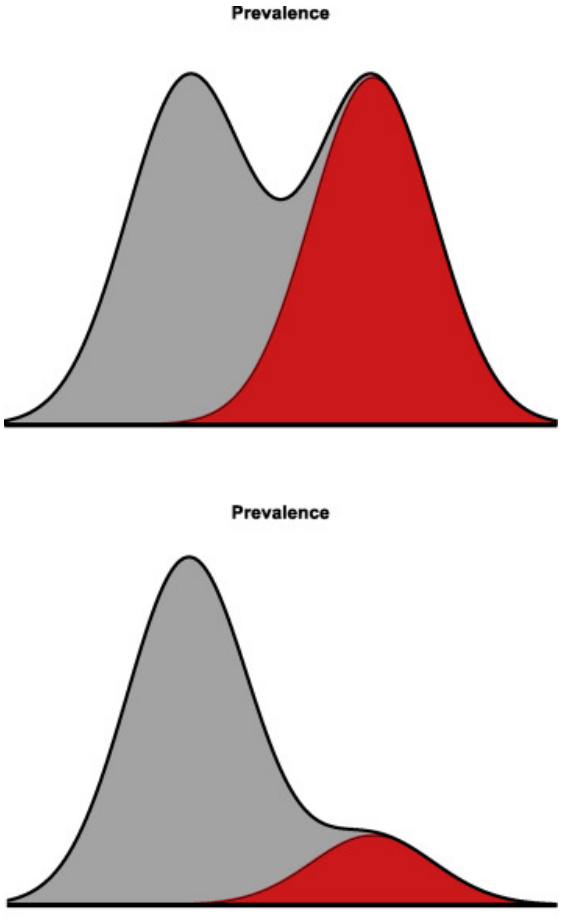


Fig. 6. Prevalence of samples in the alternate distribution, for two cases, top where the two distributions are equal in area, and bottom where they are different; the dark red area is the proportion of the joint distributions equal to the prevalence. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

As we will see, these concepts are related, but often confused in the literature.

4. Confusions and controversies

4.1. False positive rates and p values

Of particular importance in chemometrics is the FPR (False Positive Rate), for example when screening for active metabolites, which peaks or compounds thought to be markers are actually not. For example, we may screening 1000 LCMS peaks, decide 100 are above a certain statistical threshold eg as assessed by an F statistic, but in fact only 75 are real markers, so the FPR (sometimes called false detection rate) is 0.25.

The FPR is sometimes confused with p values, but as we will see is very different. Although p and α are theoretically distinct, their method of calculation is the same, so we will assume a value of α has been defined, eg for a specific marker, the value of α strictly speaking is a threshold, whereas p is an experimentally determined value – however the Neyman-Pearson approach introduces the "alternative hypothesis" which for the purpose in this section is essential when defining FPRs, so we cannot avoid confusing the two approaches. The prevalence, ie what proportion of potential markers are true positives we will denote by T, and define the power by (1-β). Note that in the literature there is no real distinction between p and α in this situation: strictly speaking p values should only be used when there is only a null hypothesis, so we should use α in this case – also p values are experimentally determined, but we cannot be certain of T and only guess or estimate it. However many papers still use p for example when employing ANOVA to determine the significance of a specific variable in a model. We will not attempt to distinguish these concepts in this section, so as to be compatible with the literature and for brevity, although an entire series of papers could be developed to argue to case.

We can define

$$FPR = \alpha (1-T) / [(1-\beta)T + \alpha (1-T)]$$

This can be demonstrated by inspection of Fig. 5, the solid blue region representing α (1-T) of the entire dataset and is discussed in detail by Colquhoun [12]: the distinction between p and α is not made in his paper.

In many cases T is quite low as we do not expect many of the potential markers we screen to be true positives, analyse 1000 LCMS peaks and only a few would show activity. If we set α=0.05, then if T=0.1 (10% of the peaks screened are actually true markers), and if the test has quite a high power (or selectivity=1-β) of 0.8 then

$$FPR = 0.05 \times 0.9 / [0.8 \times 0.1 + 0.05 \times 0.9] = 0.36$$

This is far higher than the original value of α of 0.05. (Using α we are setting a limit, so if peaks have a p value at 0.05 their FPR is 0.36, many papers do not distinguish p and α in this context, p value would convert the observed value eg of an F or t statistic for a specific maker to a probability that the marker is a false positive but we will not dwell in detail on this distinction for brevity). This means that although a peak may appear to have a p value (or be below α) of 0.05 and so appear at first to be a good candidate for pursuing as a possible marker, in the example above, but in fact there is a probability of over one third (rather than one twentieth) it is a false positive. Hence when screening potential markers we should estimate FPRs rather than calculate p values or Type 1 errors.

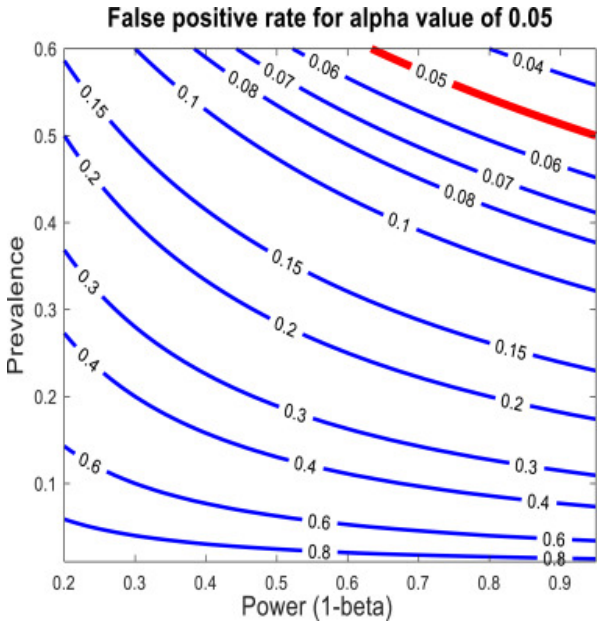
Of course, we do not now the prevalence in advance, so can only estimate the FPR. Smellke et al. [13] calculate typical FPRs for different p values. Using intermediate assumptions, these are presented in Table 2. As can be shown a p value of 0.05 may actually correspond to a FPR of 0.5, dependent on the power and sensitivity of an analysis. A contour plot of FPR against power and prevalence for a value of α (or p) of 0.05 is presented in Fig. 7. In many practical situations in chemometrics, the results will be in the bottom right corner, as prevalence tends to be quite low (only a fraction of candidate peaks or potential markers screened will be true markers), whereas the power of a good method is usually quite high. So for p=0.05, the FPR may in fact be around 0.6 in many typical cases. This means that really low p values such as 0.01 or less are required to have any confidence that a potential marker has a real effect. This means that the traditional

a

model; tells very little as to whether they really have important biological effects. Ioannidis has published a useful paper "*Why Most Published Research Findings Are False*" which expands on this topic in more detail [14].

Table 2. Typical FPRs for different p values using middle of the road assumptions – see Fig. 7 for p = 0.05

p value	Typical FPR
0.05	At least 0.23 (typically 0.5)
0.01	At least 0.07 (typically 0.15)



Download : Download high-res image (448KB) Download : Download full-size image

Fig. 7. Contour plot of FPR against power and prevalence at a value of p (or α) of 0.05.

4.2. Reproducibility of p values

Another problem with p values is that they can be irreproducible. Cumming has written several articles in which he performs repeated simulations, each time calculating p values using the same underlying model. He concludes that the range of p values is quite large. In Ref. [15] he claims "*… it is important to ask what p says about replication. The answer to this question is "Surprisingly little." In one simulation of 25 repetitions of a typical experiment, p varied from < 0.001 to 0.76, thus illustrating that p is a very unreliable measure. …. ., if an initial experiment results in two-tailed p = 0.05, there is an 80% chance the one-tailed p value from a replication will fall in the interval (0.00008, 0.44), a 10% chance that p < 0.00008, and fully a 10% chance that p > 0.44.*".

To test this, 20 observations were taken from two underlying distributions, as illustrated in Fig. 8: the two Gaussians represent two different populations, and the crosses represent 20 samples from each distribution. This sampling is repeated again 10 times, each time the significance of the difference in means is calculated using a one tailed *t*-test [16]. As can be seen in Table 3 there is a very wide range of observed p values, using the *t*-test. Cummings claims that this wide range is obtained however large the sample size is, although there is not enough research in this area yet. However he is correct that the reproducibility of p values has not been adequately studied and that alternative approaches, based on confidence intervals, are more robust.

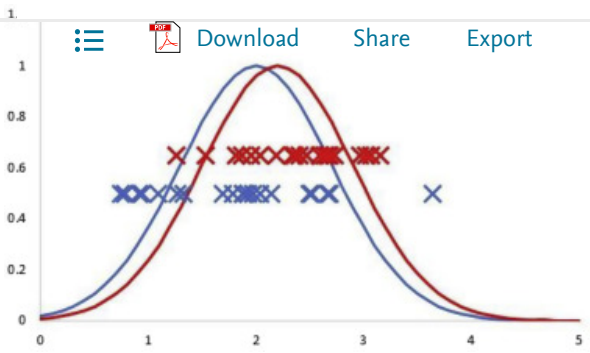


Fig. 8. A typical simulation: there are two underlying distributions, each of which are sampled 20 times to give the blue and red crosses representing one of the cases of Table 3. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 3. p values for one tailed *t*-test using the distributions of Fig. 8.

Simulation	p value
1	0.037203
2	0.110324
3	0.014832
4	0.088785
5	0.556651
6	0.210629
7	0.064671
8	0.138183
9	0.224906
10	0.008772

4.3. The current controversy

The sections above discuss two important issues about p values that are of special interest to readers of this article.

However within the statistics community, there have been controversies surrounding the use of p values for many years. In particular, journals in areas such as biology and medicine often routinely require investigators to quote p values in order to get work accepted. Often a p value of 0.05 or less is required for publication. In certain areas of science this criterion is so strongly engrained that PhDs, grants and even scientific jobs depend in this. In contrast, in mainstream physical sciences, such as chemistry, p values have a limited role. Early chemometricians, mainly analytical or physical chemistry, rarely saw the need to calculate p values. But as discussed in the introduction, chemometricians are now collaborating especially with biomedical scientists, and to get their work accepted and understood are often forced to use p values as their colleagues, or indeed members of their grant committee, are used to this, and so (often falsely) judge whether their hypothesis is reasonable or not on this criterion.

Over the past two decades there has been a substantial backlash against the use of p values. There are several hundred papers discussing the flaws.

perceived as a mathematically coherent approach to inference. There is little appreciation …….. that the methodology is an amalgam of incompatible elements, whose utility for scientific inference has been the subject of intense debate among statisticians for almost 70 years."

Rex Kline's book is one of the most comprehensive discussions [18]. In it he lists several fallacies about the use p values. Kline refers to over 400 papers, published by the time of his first edition in 2004, so we will only describe a few of the issues. For p values to have meaningful interpretation.

- It is necessary to specify a plausible null hypothesis
- It is necessary to study random samples
- The distributional assumptions should be checked
- The power of the method should be evaluated.

Very few studies in using chemometrics would comply with these requirements. For example, how many datasets satisfy multi-normality? And when taking samples, for example, from human population, how can we control the sampling and ensure it is random, which has a specific statistical meaning?

The American Statistical Association published an important paper in 2016 [19]. Their conclusions are summarised.

1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

A recent paper in Nature confirms this trend among statistically oriented researchers [20] comes to broadly similar conclusions.

This controversy has assumed considerable importance within mainstream statistical thinking, but is not well known in chemometrics. The definition of chemometrics includes statistical modelling, and so it is important that analytical chemists and others using chemometrics methods for hypothesis testing are aware.

5. Conclusion

There is still quite limited awareness about the concept of p values within the chemometric community. They are used, for example, when analysing designed experiments, so as to determine whether individual factors or variables are significant, but even there, most users do not fully understand the basis of the significance tests.

Within current practice, there is an increasing need for significance or hypothesis tests, especially to determine whether metabolites have a significant role, eg as a disease marker or to distinguish groups. Often p values are used instead of False Discovery Rates, or are employed without reference to the power of an analysis; of course the power does require an alternative hypothesis but that is essential for estimating FPRs. However, the literature is somewhat confusing, and although all definitions of chemometrics do include the statistical analysis of data, many practitioners are not aware of many basic statistical concepts. This paper does not delve into statistical details and is written for the general reader. Almost anyone that has come across p values, Type 1 errors and hypothesis tests should have some appreciation of the historic background and modern usage of these methods. As chemometrics moves more into hypothesis based science, for example in cultural heritage studies or biomedicine, an appreciation of the concepts discussed in this paper is important for safe interpretation of p values and related concepts.

Declaration of competing interest

I

References

[1] https://www.merriam-webster.com/dictionary/chemometrics, Accessed 24th Jun 2019

[2] O.M. Kvalheim
History, philosophy and mathematical basis of the latent variable approach - from a peculiarity in psychology to a general method for analysis of multivariate data
Chemometr. Intell. Lab. Syst., 26 (2012), pp. 210-217
CrossRef View Record in Scopus Google Scholar

[3] R.G. Brereton
Determining the significance of individual factors for orthogonal designs
J. Chemom., 32 (2019), Article e3124
Google Scholar

[4] H. Martens, T. Naes
Multivariate Calibration
Wiley, Chichester (1989)
Google Scholar

[5] J. Arbuthnot
An argument for Divine Providence, taken from the constant regularity observed in the births of both sexes
Philos. Trans. R. Soc. Lond., 27 (1720), pp. 186-190
Google Scholar

[6] K. Pearson
On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling
Lond. Edinb. Dublin Philos. Mag. J. Sci., 50 (1900), pp. 157-175
Series 5
CrossRef View Record in Scopus Google Scholar

[7] R.A. Fisher
The Design of Experiments
Oliver and Boyd, New York (1935)
Google Scholar

[8] D. Salsburg
The Lady Tasting Tea : How Statistics Revolutionised Science in the Twentieth Century
Henry Holt and Company, New York (2001)
Google Scholar

[9] J. Neyman, E.S. Pearson
On the use and interpretation of certain test criteria for purposes of statistical inference: part I
Biometrika, 20A (1928), pp. 175-240
CrossRef Google Scholar

[10] J.D. Perezgonzalez, Fisher
Neyman-Pearson or NHST? A tutorial for teaching data testing
Front. Psychol., 6 (2015), p. 223
Google Scholar

[11] R. Hubbard, N.J. Bayarri
P Values Are Not Error Probabilities
Duke University (2003)

[12] D. Colquhoun
An investigation of the false discovery rate and the misinterpretation of p-values
R. Soc. Open Sci., 1 (2014), Article 140216
CrossRef Google Scholar

[13] T. Smellke, M.J. Bayarri, J.O. Berger
Calibration of p values for testing precise null hypotheses
Am. Stat., 55 (2001), pp. 62-71
Google Scholar

[14] J.P.A. Ioannidis
Why most published research Findings are false
PLoS Med., 2 (2005), Article e124
CrossRef View Record in Scopus Google Scholar

[15] G. Cumming
Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better
Perspect. Psychol. Sci., 3 (2008), pp. 286-300
CrossRef View Record in Scopus Google Scholar

[16] R.G. Brereton
Chemometrics : Data Driven Extraction for Science
Wiley, Chichester (2018)
Google Scholar

[17] S.N. Goodman
Toward evidence-based medical statistics. 1: the P value fallacy
Ann. Intern. Med., 130 (1999), pp. 995-1004
CrossRef View Record in Scopus Google Scholar

[18] R.B. Kline
Beyond Significance Testing
(second ed.), American Psychological Association, Washington, DC (2013)
Google Scholar

[19] R.L. Wasserstein, N.A. Lazar
The ASA's statement on p-values: context, process, and purpose
Am. Stat., 70 (2016), pp. 129-133
CrossRef View Record in Scopus Google Scholar

[20] V. Amrhein, S. Greenland, B. McShane
Retire statistical significance
Nature, 567 (2019), pp. 305-307
CrossRef View Record in Scopus Google Scholar