JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY

Article

pubs.acs.org/JAFC

Biological Importance and Statistical Significance

David P. Lovell*

St. George's, University of London, Cranmer Terrace, London SW17 0RE, United Kingdom

ABSTRACT: Statistical ideas behind the analysis of experiments related to crop composition and the genetic factors underlying composition are discussed. The emphasis is on concepts rather than statistical formulations. Statistical analysis and biological considerations are shown to be complementary rather than contradictory, in that the statistical analysis of a data set depends on the experimental design, that no amount of statistical sophistication can rescue a badly designed study, and that good experimental design is crucial. The traditional null hypothesis significance testing approach has severe limitations, but *p* values and statistical significance still often seem to be the primary objective of an analysis. Emphasis instead should be on identifying the size of effects that are biologically important and, with the involvement of the "domain" scientist, using these to help design experiments with appropriate sample sizes and statistical power. The issues discussed here are also directly applicable to other areas of research.

KEYWORDS: *statistical significance, experimental design, power*

INTRODUCTION

The International Life Sciences Institute's (ILSI) International Food Biotechnology Committee (IFBiC) held its Plant Composition Workshop in September 2012. This date was close to the 50th anniversary of the death of R. A. Fisher in Adelaide, Australia, on July 29, 1962. Fisher was an outstanding geneticist and statistician. His wide-ranging contributions to the development of statistics, experimental design, and genetics underpin many of the topics discussed at the 2012 ILSI IFBiC workshop.

Through his work at Rothamsted and later at University College London and Cambridge University, Fisher laid the foundations for many developments, including the modern theory and basis for the design of experiments and the analysis of variance (ANOVA) methodology. Fisher was one of the founders of population genetics and initiated the concept of significance testing.[1,2] His contributions, such as Fisher's exact test, appear throughout any consideration of the statistical methods used in the interpretation of composition data.

The objective of this paper is to discuss statistical concepts, some originally developed by Fisher, used in the analysis of experiments related to the composition of crops and the genetic factors that underlie their composition. This paper concentrates on concepts rather than detailed statistical formulations and equations and is aimed at the reader who wants to know why something is done rather than how it is done. The concepts are also directly applicable to other areas of research. Two main arguments will be put forward: first, the over-riding importance of experimental design that precedes statistical analysis, and, second, the limitations of the concept of statistical significance and the mistaken equating of it with biological relevance or importance.

Investigations in the area of research on the composition of crops and the genetic factors that underlie their composition encompass a wide range of objectives from the assessment of agricultural field trials to the safety evaluation of foods derived from modern technologies. One of the objectives of this paper is to lay out some of the statistical issues related to these studies so that even if a statistical analysis is unable to provide a "perfect solution," an appreciation of the statistical issues can help in part of what could be termed a *weight of evidence approach*. Statistical considerations are a crucial aspect of an evidence-based approach.[3]

Some of the discussion here is based on work carried out by the Statistical Working Group of the European Food Safety Authority (EFSA) Scientific Committee to help EFSA scientific panels and committees in their assessment of biologically relevant effects.[4] An opinion produced by this group was primarily intended for EFSA experts and staff, with the objective of clarifying the main concepts and definitions associated with statistical significance and biological relevance. It was considered that this may also be useful for risk managers and risk communicators in general. EFSA has also produced a number of other documents in which experimental design and statistical analysis issues are discussed in areas ranging from field design to 90 day toxicology studies.[5−7]

EXPERIMENTAL DESIGN

The concepts of statistical significance and hypothesis testing dominate many scientists' thinking about the statistical analysis of experimental data almost to the exclusion and detriment of other aspects of statistical analysis. A primary question is "What is the purpose or objective of a statistical analysis?" In many discussions, a statistical analysis and a statistician's role are equated with the finding of statistical significance using, for example, one of the various statistical methods to test a null hypothesis of no difference between a treated group and a control group. However, the statistician has a far more important role in the design of studies.

Special Issue: Safety of GM Crops: Compositional Analysis

Received: March 12, 2013
Revised: August 2, 2013
Accepted: August 2, 2013
Published: August 2, 2013

dx.doi.org/10.1021/jf401124y | *J. Agric. Food Chem.* 2013, 61, 8340−8348

In practice, the design of the experiment can be considered the strategy and the statistical analyses used, the tactics. Statistical thinking is consequently important at the strategic experimental design stage, and the analysis is consequently secondary to and dependent on good experimental design. Statistical tests rely on the implicit assumption that the design is correct. No amount of statistical testing will rescue results obtained in a poorly designed or nondesigned study.

Researchers are frequently exhorted to consult a statistician before starting an experiment. The following quote from Fisher remains apposite: "To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of."[8]

The following quotation from Price and Underhill provides an example of a study design for composition research: "Studies are usually with the GM plant variety and its parent. Several field locations are selected, representing the area where the GM plant is expected to be grown. Within a location, growing plots of land are allocated to the experiment. Usually blocks of two or more plots are formed to ensure that the treated and control plants are grown under the same conditions."[9]

Experimental design in the agricultural sciences has been heavily influenced by Fisher, who first developed the factorial design: experiments where two or more factors are investigated in all possible combinations (based on Mendelian genetic concepts).[10,11] Fisher and co-workers developed a range of experimental designs, such as the randomized complete block, the split plot, the Latin square, and the factorial designs in a series of influential books and papers. Some of these designs are, to some extent, contrary to much of conventional teaching of varying just one factor and keeping all the others constant, called the one factor at a time (OFAT) approach.[12]

The factorial approach has developed into the field of design of experiment (DOE) methodology. Fisher demonstrated that these DOE approaches such as the factorial design (where there is systematic and simultaneous variation of experimental conditions such as the classical field trials involving nitrogen, phosphorus, and potassium fertilizers) are both economically and scientifically more efficient than the traditional OFAT approach. They can identify the multiple factors (and the interactions between them) that affect results appreciably. The DOE methodology further identifies the levels of these factors that optimize results as well as minimize the number of independent studies needed and the experimental units required. This approach is ideal for topics such as protocol development, in which there may be a number of factors that can affect the experimental results. The DOE approach is now widely used in industrial settings, such as the manufacturing and chemical industries, to identify optimal conditions for processes under which to operate and to reduce the number of runs needed to identify these. DOE methodology could, for instance, be an efficient approach to identifying optimum allocation or use of resources in studies in which there are a number of different factors for which conditions could be varied. The advantages of the factorial approach are, unfortunately, still not widely appreciated in many research areas. The agricultural basis for many experimental designs can be seen in the use of terms such as blocks and plots.

HYPOTHESIS TESTING AND STATISTICAL SIGNIFICANCE

The concept of statistical significance and hypothesis testing unfortunately dominates many scientists' thinking about the statistical analysis of their data, almost to the exclusion and detriment of other aspects of statistical analysis. The two concepts are different, and Cox[13] distinguished between statistical significance testing (the Fisherian approach) and hypothesis testing (the Neyman–Pearson approach).

One of Fisher's most well-known legacies is the significance test, which together with reports of the p value and asterisks adorn many tables of results. Fisher envisaged a test of significance that then provided him with evidence on which to base further work.[11] The test was based on a test of whether a single model fitted the data. There was no alternative hypothesis. The use of the statistical test for hypothesis testing derives from the work of Jerzy Neyman and Egon Pearson in the late 1920s and early 1930s.[14]

Statistical significance later became a concept associated with the use of a specific statistical test to test a null hypothesis of no difference between two groups such as a treated group or a control group. This has been referred to as the null hypothesis significance testing (NHST) approach.[15] The concepts are (or should be) familiar to many: Type I or Type II alpha and beta errors, power, and significance levels. The familiar 2×2 table can be used to illustrate the NHST approach. However, understanding the NHST approach is difficult in part because it is a mixture of two different concepts. It was (and remains) deeply controversial.

The standard NHST approach is a hybrid of the Fisher and Neyman–Pearson processes.[16] Fisher was initially interested in testing a single null hypothesis and not in the alternative hypothesis. This test of the null hypothesis was not particularly important and was to be used as a guide rather than a decision-making process. There was no alternative hypothesis. Neyman and Pearson wanted to compare two hypotheses that relate to the concept of the null and alternative hypotheses. This is particularly relevant for quality control type investigations.

Gigerenzer provides a detailed discussion of some of the issues around the null hypothesis and statistical significance testing.[17] He discusses the long-running (and often acrimonious) debate between Fisher and Neyman and Pearson over the context of significance testing, pointing to how Fisher had initiated the significance level but moved in later life to take a more nuanced view. On the other hand, Neyman and Pearson maintained their interest in the use of the decision rule approach.

In the Neyman–Pearson method, the only criterion is whether the hypothesis is rejected. It is a binary decision with the critical value for a test statistic equivalent to, say, $p = 0.05$ as the criterion. The result is then declared statistically significant at $p = 0.05$ and the alternative hypothesis is accepted in contrast to the null hypothesis. The Fisherian approach is to report the exact p value and use this value only to gauge where the null hypothesis has been rejected and not whether another hypothesis has been accepted.

Note that in the NHST approach, the test statistic and its associated p value depend on a number of factors such as sample size, statistical test used, and amount of variability. This means that the actual size of difference that is just significant (i.e., reached the critical value of the test statistic) will vary from study to study. In addition, each experiment is, in effect, one

from a population of possible experiments and is thus only an estimate (with a distribution) of the true difference. By "bad luck," the actual experiment performed can give an estimate in one of the tails and the study would be reported as "not significant" even when there is a real difference (a Type II error).

CRITICISM OF THE USE OF SIGNIFICANCE

Some think that producing statements about the presence or absence of statistical significance is the role of the professional statistician. Protocols often include language stating that a level of probability <0.05 will be accepted to indicate statistical significance for the comparisons. According to Dallal, "For better or worse, the term *statistically significant* has become synonymous with $p \leq 0.05$."[18]

As noted earlier, the p value and the significance test are two different things. Fisher was responsible for the first use of p values and the term *test of significance*: "Critical tests of this kind may be called tests of significance, and when such tests are available we may discover whether a second sample is or is not significantly different from the first."[11]

Importantly, the p value is the probability that a difference as large as or larger than that seen in the experiment would have occurred by chance alone if the treatment groups were in fact not different. It is *not* the probability that the null hypothesis is true, which is a frequent but serious misinterpretation.

Statistical comparisons are often reported as significance levels, with one asterisk to indicate $p \leq 0.05$, two asterisks to indicate $p \leq 0.01$, and three asterisks to indicate $p \leq 0.001$. Values >0.05 are often reported as nonsignificant or not significant, although Altman in particular does not recommend this.[19] The choice of the three levels relates to the tables for the three levels of 0.05, 0.01, and 0.001 in the statistical tables produced by Fisher and Yates.[20]

Many authors now argue against this convention. Yates accidentally introduced this "star" nomenclature but abhorred its subsequent use and vehemently opposed its unthinking use. In 1937, Yates produced a monograph on factorial experiments that became the standard work of reference on the ANOVA.[21] Asterisks were used to indicate successive footnotes, but were interpreted by adherents as a new standard notation. Yates regretted that this gave too much emphasis to the significance test and criticized some scientific research workers for paying undue attention to them.[22]

The term *significant* has a number of meanings, some of which are technical and others less so. In 2008, *Nature* published a list of disputed definitions that included the word "significance".[23] In regard to this list, Reese noted that "Too many scientists — and editors — take the line you reproach and use statistical significance as a criterion of importance."[24]

Salsburg[25] (p 98) provides an interesting discussion on how the word "significant" changed its meaning:

> *The word was used in its late-nineteenth-century meaning, which is simply that the computation signified or showed something. As the English language entered the twentieth century, the word significant began to take on other meanings, until it took on its current meaning, implying something very important. ... Unfortunately, those who use statistical analysis often treat a significant test statistic as implying something closer to the modern meaning of the word.*

Many statisticians abhor the use of the NHST and the "cult of the p value," yet is seems to have taken root as part of the basic requirements that the regulator and the referee/journal editors require. Attaining a p value <0.05 sometimes seems the sole objective of the experimenter. (It should be noted, though, that the R statistical package, developed by statisticians and with many enthusiastic adherents, attaches asterisks to the output of the ANOVA!)

There have now been many criticisms of the use of significance tests. In their book titled *The Cult of Statistical Significance*, Ziliack and McCloskey provide a detailed criticism of the concept of statistical significance and discuss some of the history and philosophy related to the development of the concept, as well as list many statisticians critical of the approach.[26] Nester maintains a large compendium of quotes (too numerous to reproduce here) from many leading statisticians.[27,28] Dallal also provides an interesting set of quotes criticizing significance testing.[18] One example is from Jacob Cohen, one of the developers of power calculations, who explained that he wanted to call the NHST process "statistical hypothesis inference testing" but according to anecdotes was (wisely) warned by his wife that his desire to reduce this to an acronym was not a good career move.[15] Bill Thompson also provides a list titled "402 Citations Questioning the Indiscriminate Use of Null Hypothesis Significance Tests in Observational Studies". This list was last updated in February 2001 but shows the long-standing and relevant criticisms of this method[29] that, as Ziliack and McCloskey suggest, date back at least to the correspondence of W. S. Gossett (the Student of Student's t test) with other statisticians in the 1920s.

These regular arguments against the use of and over-reliance on p values and significance testing are familiar to statisticians who have heard them many times before. Hubbard and Lindsay recently argued that p values are not useful in the field of psychology.[30] Many of their arguments are generalizable to other areas of research and link into the idea that p values exaggerate the evidence against the null hypothesis by, for instance, encouraging publication bias and by the lack of a relationship between the p value and the effect size obtained in a study. In a series of papers, especially his 2005 *PLoS* paper,[31] Ioannidis has highlighted the problems in a number of biological fields including genome-wide association studies, gene expression experiments, and the lack of reproducibility of highly cited papers that result from an over-reliance on statistical significance. Some statisticians have urged the replacement of the p value, suggesting in its place an estimation approach based on confidence limits or fundamentally different approaches built around Bayesian statistical methods (discussed below).

There have been very few supporters of the continued use of the p value. One example is Chow, who gave a lengthy and complex defense of the value of NHST in the area of theory corroboration,[32] which is summarized in a document by Fiona Fidler.[33] Partial support or at least an argument for the continued use of the p value is given by Moran and Solomon[16] and by Senn, who offered "a limited defense of P-values only" and stated that "P-values are a practical success, but a critical failure".[34]

In 2011, EFSA[4] made the following recommendations concerning statistical significance and p values:

Statistical significance, when expressed by a P-value, relates to the probability of having obtained results as (or more extreme than) those observed, given that the null hypothesis H0 is true... Statistical significance is considered as just one part of an appropriate statistical analysis of a well designed experiment or study... Identifying statistical significance should not be the primary objective of a statistical analysis.

EFSA made a further recommendation that in any discussion of statistical analysis, particularly when inference is to be carried out, the word "significance" is reserved for its specific/narrow interpretation.

BIOLOGICAL IMPORTANCE OVER STATISTICAL SIGNIFICANCE

EFSA defined a biologically relevant effect as "an effect considered by expert judgement as important and meaningful for human, animal, plant or environmental health. It therefore implies a change that may alter how decisions for a specific problem are taken."[4]

A biologically important or relevant effect can be related to the effect size and to the concept of power and sample size calculations. These can be carried out, for instance, using software packages where appropriate information is entered to obtain either the sample sizes associated with a particular power or, alternatively, the power for a given design. Power is the probability of rejecting the null hypothesis when this is false and thus not committing a Type II error. Defining what is a biologically relevant or important effect is not straightforward. It is not a statistical decision, although it has important consequences for the design, statistical analysis, and interpretation of an experiment. One approach at the design stage is to identify the size of effect that would be the "minimum difference that you can afford to miss".[35] This relates to concepts such as the minimum clinically important difference or the clinically relevant difference (CRD). The CRD is the smallest treatment effect that is clinically meaningful or clinically relevant.[36] This difference is also referred to in sample size calculations as the effect size and can be standardized by expressing it in standard deviation (SD) units. Cohen discusses the use of such standard effect sizes in calculating sample sizes and gives examples of small, medium, and large effect sizes,[37] whereas Lenth provides an alternative viewpoint that is critical of the unthinking use of such standardized effect sizes.[38]

The choice of an appropriate effect size requires an appreciation of the biological material and context being investigated, and there is likely to be an appreciable amount of expert, but nevertheless subjective, judgment in the choice. There may be no consensus as to what is the minimum difference that is considered tolerable if the effect might be considered a toxicological response. The choice of the effect size is, therefore, a decision for the domain scientist based on his or her expert knowledge.

An experiment is designed for what is called the primary end point in clinical trials, which is the measure that is most directly related to the main purpose of the study. The power associated with the design for the other secondary measures or end points will be based on the nature of the measure and the variability of the material. In practice, experiments often have multiple measures of interest, and estimating sample sizes to consider all measures simultaneously is more complex.[39]

Relating to a standardized effect size is one way to progress. A sample size of 20 has a power of approximately 85% (exact is 86%) to detect a 1 SD difference, if it really exists, as statistically significant at $p = 0.05$ in a two-sided test. The crucial issue is how biologically important a 1 SD difference is between two groups in, for instance, a hematological or clinical chemistry measure. Different measures will have different levels of detection or "resolution." A study designed to detect an effect in the primary end point based on a biologically important difference of, say, 0.5 SD will also show effects in other measures as significant at this level. In some cases, these differences, although significant, are not of biological importance. One example is in the standard 90 day toxicological studies. With relatively large animal studies (20 per group), significant effects can be detected in clinical chemistry or hematological measures that, although real treatment-related effects, are minor changes and not considered of biological importance by the study toxicologist. The ambiguous meaning of the term "significance" can be a source of controversy in such cases. Guidelines such as those from the Organisation for Economic Co-operation and Development (OECD) include phrases such as "Evaluation of data should include a discussion of both the biological and statistical significance"[40] or "Both biological and statistical significance should be considered together in the evaluation".[41]

Price and Underhill state the following for instructions for the statistical analysis of the composition data: "Compositional analysis is conducted using validated analytical methods, and the data are subjected to an appropriate statistical analysis, resulting in probability estimates."[9] The authors also state for summary statistics that "Data for each plant component should be assembled in tabular form. A mean, p or F value, and literature range are required."[9]

An experiment designed with 80% power for the primary end point may have higher or lower power for the biologically important difference for the other secondary end points. There are now large databases of information, such as the U.S. Environmental Protection Agency's Toxicity Reference Database (ToxRefDB),[42] that could provide some guidance to the size of effects likely to be detected with high power in some of the standard designs in areas such as composition analysis or toxicology. These specific experimental designs are to a large extent based on historical experience and are often based on having appropriate power to detect treatment-related increases in the incidence of, for instance, binary traits such as the presence or absence of a pathological condition. The power of standard studies to detect such increases as significant is low but is quite powerful for some quantitative measures.

Such databases could provide estimates of the relative variability of the measures used in these designs. These estimates of variability can then be used to provide a guide to the size of effects that would be detected in specific designs and provide the basis for discussion among domain scientists as to their biological importance.

The EFSA steering committee recommended that "... the nature and size of biological changes or differences seen in studies that would be considered relevant should be defined before studies are initiated. The size of such changes should be used to design studies with sufficient statistical power to be able to detect effects of such size if they truly occurred."[4]

ALTERNATIVES: CONFIDENCE INTERVALS AND ESTIMATION

Dissatisfaction with the NHST approach is partly behind the growing argument for more emphasis to be put on estimates of

the size of effects and the confidence associated with these estimates. This reduces the problem of a small and biologically unimportant effect being declared statistically significant and that of attempting to convert a result into a "binary" positive or negative conclusion on the basis of a p value of 0.049 or 0.051, for example.

The primary interest in food safety research and agriculture is to estimate the size of the interventions in an experiment. The question is often how big the effect is rather than is there a significant effect. Cochran and Cox point out that "In many experiments it seems obvious that the different treatments must have produced some difference, however small, in effect. Thus the hypothesis that there is no difference is unrealistic: the real problem is to obtain estimates of the sizes of the differences."[43]

It has been proposed for some time that confidence intervals should be reported in journals.[44,45] An increasing number of journals have now, in fact, expressed their preference for confidence intervals over p values (e.g., the *British Medical Journal*'s instruction for authors, http://bmjopen.bmj.com/site/about/guidelines.xhtml).

The information from a significance test is rather limited, especially when reduced to the statement that an effect is statistically significant or not by rejecting or accepting the null hypothesis. A confidence interval provides more information than in a significance test by giving an estimate of the size of the effect, such as the mean difference, together with information on the variability expressed by, say, the width of the 95% confidence interval. A two-sided 95% confidence interval also equates to a hypothesis test because if the interval does not straddle zero, this means that the null hypothesis of zero is rejected at $p = 0.05$. Confidence intervals can also be included on graphical presentations of the results. Weller provides a discussion of some of the history behind statistical guidelines.[46]

EQUIVALENCE TESTING: INFERIORITY AND SUPERIORITY

The primary function in food, agriculture, and ecology research is to estimate the magnitude of treatment effects. In an approach based on estimation, typical relevant questions might include the following: "By what proportion is mortality increased through obesity?" or "By how much does the mortality rate increase for each unit increase in cholesterol and at what level does this relationship cease to be linear?" rather than "Is there a significant difference in mortality between normal and obese males?" or "Is there a significant increase in mortality when cholesterol is increased?" As Cox and Snell commented, "In practice it is rarely necessary to find P at all precisely."[47] There has been a move away from the idea of hypothesis testing to one of testing for equivalence.

Two trends are evident: the concept of substantial equivalence, developed by the OECD in 1993[48] in the novel food area, and the concept of equivalence in testing pharmaceuticals. An important consideration is to distinguish between bioequivalence and substantial equivalence.

Substantial equivalence is a concept that maintains that a novel food (e.g., genetically modified foods) should be considered the same as and as safe as a conventional food if it demonstrates the same characteristics and composition as the conventional food.[49] Substantial equivalence is important from a regulatory perspective. Bioequivalence is a concept whereby two pharmaceutical products are judged to have sufficiently similar characteristics as to be considered to be essentially the same. Bioequivalence testing is an alternative approach built around the weakness of the NHST approach when the aim is to show an absence of an effect. The aim is to avoid the dangers of declaring no difference between the two groups (and wrongly claiming the null hypothesis accepted) through the use of small samples, poor experimental design, and inappropriate statistical methods.

The methods developed in the clinical trial field were developed to surmount the problems of using statistical tests to show, for instance, that two pharmaceutical formulations (e.g., a proprietary and a generic version) are similar rather than different. An example is the U.S. Food and Drug Administration (FDA) bioequivalence guidelines in which tests of point-null hypotheses have fallen from favor and equivalence procedures are mandated.[50]

Instead of trying to show that there is no statistically significant difference between two formulations, the objective of these methods is to show that although there may be a difference, this is sufficiently small to be considered as not biologically important or relevant. The approach was developed to show that equivalence had been attained if the 90% confidence intervals of the difference in the specified pharmacokinetic measures lay within a range between 80 and 125% of the ratio of the two formulations.

This concept is based on an estimation approach using a confidence interval rather than the NHST and has been extended in the pharmaceutical industry to develop the concept of inferiority and superiority trials. This trend away from formal hypothesis testing approaches to methods based on estimation (such as equivalence, noninferiority, and superiority tests) and to statistical modeling is likely to continue.

Critical to the design of equivalence studies is the need to predefine the acceptable intervals for sample size and power calculations. In bioequivalence, this is the term *delta* (Δ), which is the noninferiority margin or "... the largest difference that is clinically acceptable, so that a difference bigger than this would matter in practice".[51]

It is important to appreciate that the absence of a significant result is not a proof for the equivalence of the new formulation against the standard formulation. This is a clear example of the concept that "absence of evidence is not evidence of absence".[52,53]

There is an important contrast between the pharmaceutical and agriculture sectors. Bioequivalence tests are usually crossover designs in which the subject is his or her own comparator, unlike agricultural trials in which they are independent units. The choice of lower and upper limits or boundaries as 80−125% is not a statistical decision but rather one based on the expertise of the domain scientist. Considerable discussion within the sector was carried out by experts in the bioequivalence fields. Similarly, the choice of the 90% confidence intervals is based on a choice made by domain scientists as it corresponds to the 5% level used in statistical tests.

The conclusion of equivalence or noninferiority from a study depends on the choice of Δ, so it must be decided on at the design stage where the choice and justification should be predefined and specified in the study protocol. Similarly, the objectives of the trial (the comparator, measures or end points, populations, sample sizes, and statistical analysis plan) are all defined in the study protocol.

MULTIPLE COMPARISONS

Many studies have multiple measures, and there is the potential complication of the multiplicity of tests. For example, many toxicological studies have multiple measures. The 90 day chronic feeding studies, for instance, involve multiple measurements: body weights, organ weights, clinical chemistry, urine analysis, and hematology. As many as about 50 different measures may be made in all, together with repeated measures of body weight and food intake.

If a NHST approach is taken, there is a high probability of a number of comparisons being considered significant just by chance alone (Type I errors). With 200 independent comparisons in which there is no difference, 10 comparisons on average will be declared significant using the critical value associated with a p value <0.05. One approach to avoid this is the use of multiple comparison methods. In practice, however, many measures are often correlated with one another, and not taking these correlations into account makes many of the multiple comparison methods overly conservative.

Some of the most commonly used methods include the Bonferroni correction and Dunnett's test. However, others such as Duncan's or Student−Newman−Keuls are in widespread use depending on the particular research field. In the SPSS package, for instance, there are 18 different post hoc tests applicable for multiple comparisons in a one-way ANOVA. Each method asks different questions about the comparison and makes different assumptions. Multiple comparisons can be applied, for instance, to measures of the expression of many different genes in a microarray experiment or between values of a single variable across different treatment groups. Multiple comparisons may be a posteriori and differ, in general, from the a priori comparisons and contrasts that are built either explicitly or implicitly into the design (such as a dose−response relationship).

These methods effectively "dampen" the significance level. The Bonferroni correction can be very conservative, especially when there are many measures. Dunnett's test is designed specifically to adjust the error rate for multiple comparisons for comparing a number $(k - 1)$ of treatment groups with a control group and is widely used in toxicology after an initial ANOVA. Often the methods are used to make significant effects disappear because they are not considered biologically important. A trend test is possibly more relevant and more powerful when there are multiple dose levels.

The false discovery rate (FDR) developed in part to help analyze large-scale microarray data is a different approach in that it compares the actual number of differences identified as significant with those expected to be significant purely by chance if there were in fact no differences.[54] It produces a q value, which is the expected number of effects that are false positives if the p value associated with a particular comparison was chosen as the threshold for significance.

The FDR is less conservative than the Bonferroni method and has higher power to identify true positive results. An FDR adjusted p value (or q value) of 0.05 suggests that 5% of significant tests may be false positives, whereas a p value of 0.05 implies that 5% of all results will be false positives if there really is no difference.

The interpretation of the methods such as the Bonferroni and the FDR is complicated because the comparisons of the measures are, in general, not independent. For example, body weights on one day are likely to be correlated with those a little later in the study. Correlations between the measures mean that it is difficult to calculate the p values associated with the multiple comparisons accurately.

In confirmatory trials in the pharmaceutical industry, the decision is based very firmly on p values, and multiplicity issues associated with the Neyman−Pearson hypothesis-testing approach may be important. On the other hand, the Fisherian significance test approach may be a useful assessment device in exploratory/hypothesis-generating studies.

Some journals have been quite prescriptive in their requirement for specific multiple comparison approaches. This has resulted in some fairly strong responses from some statisticians. The *GenStat* manual[55] points out that many statisticians have reservation over this as they do not feel it is good statistical practice. Nelder[56] and Bryan-Jones and Finney[57] give an overview of this argument. These methods are also considered inappropriate for factorial-type designs or for data in which there are quantitative doses.[58,59] It is argued that multiple-comparison methods are considered unnecessary if there are only a few treatments or if the specific comparisons should have been identified beforehand (a priori) or the treatments have some sort of structure.[60]

MODELING

One of Fisher's other great legacies is ANOVA. Many people are familiar with it because of the complicated series of equations that were once needed to produce the ubiquitous ANOVA tables leading to the F test (named by George Snedecor in honor of Fisher). Less appreciated is that the ANOVA is an entry into the area of statistical model building and part of a "network" of models that link a wide range of statistical methods such as linear and multiple regression, ANOVA, and analysis of covariance (ANCOVA). ANOVAs are a special case of the general linear model, with the interconnectedness of statistical tests also illustrated by the pooled two-sample t test being a special case of a one-way ANOVA with only two groups.

The ANOVA is a convenient general method that, while making a number of assumptions that are difficult to check with small sample sizes (independent experimental units, normally distributed random errors, and homogeneity of within group variances), is generally robust to moderate violation of these assumptions.

One of the fruitful areas of research in statistics has been the development of relationships between different methods. The original ANOVA was a convenient way of doing the algebra, but matrix algebra opens up the general linear model (GLM) approach and unified ANOVA with regression modeling. Relaxing assumptions leads to more sophisticated modeling approaches by including, for instance, correlations between measures such as in repeated measures or time series approaches.

The GLM is a generalization of the ordinary least-squares approach (used in ANOVA, ANCOVA, and multivariate ANOVA) and is a special case of the generalized linear model (GLZ). The GLZ is a unified method used to extend the GLM approach to incorporate responses other than those based on the normal distribution. Nelder and Wedderburn developed the concept of the GLZ, which placed all of the commonly used models (binomial, logit, probit, and normal) in a unified framework.[61] The GLZ can be further generalized. Generalized linear mixed models (GLMM) are an extension of the GLZ with random effects and are also called generalized linear mixed-effects models. Generalized estimating equations

are another extension of GLZ involving algorithmic adjustments used to model longitudinal or clustered data and to estimate regression coefficients. It is to some extent a matter of choice whether to use simple methods rather than the more complex models.

Building empirical statistical models to try to make predictions is a common, but challenging, feature of many scientific areas. George Box famously said "All models are wrong, but some are useful".[62] Identifying useful models is critical to avoiding results that may look feasible but are, in fact, incorrect and in some cases seriously misleading.

Modeling leads into the area of data exploration, data mining, or pejoratively, data trawling. Considerable skill is needed to move from explanation to accurate prediction and to avoid overfitting (where the model describes the underlying "noise" in the system rather than the underlying relationship). A challenge is to build models in which there are complex relationships between the inputs while also needing to validate and test these models. Modeling approaches derive from both statistical modelers and from machine-learning practitioners and come with different methodologies and terminologies.

MULTIVARIATE AND GRAPHICAL METHODS

Food composition data are multivariate in that multiple variables are measured on the same sample. A wide range of methods are now available for the graphical presentation of data. Tukey developed the exploratory data analysis (EDA) approach with a concentration on graphical techniques as an approach for hypothesis generation rather than testing as providing an alternative approach to the analysis of large data sets.[63] Many multivariate techniques are available for the exploration of complex data sets in which there is some degree of correlations between the different varaibles.[64] Methods can be broken down into supervised learning approaches (e.g., canonical variate analysis), in which there are pre-existing groupings in the data set, as well as unsupervised learning approaches (e.g., principal component analysis and hierarchical cluster analysis), in which the aim is to identify possible groupings in a set of data. Such approaches are widely used in areas such as genomics and metabolomics and provide highly visual approaches to the representation of complex data sets. They can help illustrate the relationships between individual samples as well as provide insight into the levels of variability between and within groups. For example, the paper by Shewry in this issue illustrates how multivariate methods can be used to describe and portray the composition variability among 200 cereal lines.[65]

BAYESIAN APPROACHES

The debate about hypothesis testing and estimation can be deeply philosophical, bringing in the logic of science and the value system of scientists. Alternative statistical methods such as the differences between frequentists and Bayesians are based on fundamentally different philosophies and methods. Put very simply, frequentists will base their analyses on just the data available to them, whereas the Bayesians may introduce their degree of belief or prior probability into the analysis. In practice, the choice of a prior probability for inclusion in Bayesian analyses can be problematic.

It should now be clear that the Neyman–Pearson approach is not universally approved of and Bayesians reject the approach. Bayesians represent a statistical school of thought that argues that inferences about any unknown parameter or hypothesis should be incorporated into a probability distribution given the observed data. This is in contrast to frequentists, who base their support for a hypothesis or parameter value by an assessment of the probability of the observed data given the hypothesis or value.

Ideally, the probability that the scientist would like to know is, given a significant result, what the probability is that this is a true positive rather than a false positive rather than the frequentist's question of what is the probability of a hypothesis being false given the results

Bayesian statistics derive from a posthumous publication in the 18th century by the British nonconformist minister, Reverend Thomas Bayes. Although long appreciated by statisticians for its potential, it was the advent of increased computing power and the development and application in the 1990s of specialist algorithms such as the Gibbs sampler and the Metropolis–Hastings Markov chain Monte Carlo method and suitable software such as Bayesian inference using Gibbs sampler (BUGS) that have allowed its potential to be realized for real-life problems. Monte Carlo methods are computer simulations using random numbers and named after the town with a casino.

Although Bayesians begin from a different intellectual starting point, the results of analyses are often similar to those done by frequentists, and Bayesian analyses produce credibility intervals that are superficially similar to confidence intervals. Nevertheless, the approach is different and the conclusions can also be assumption dependent. It should also be recognized that when the results obtained by different methods provide different interpretations, this can provide important information about the study and the data and can help, rather than hinder, the true interpretation.

One recent example of the application of Bayesian methods is an analysis of food composition data from transgenic maize.[66] In this approach, the authors discuss the choice of an appropriate prior probability, the practicalities of applying the methodology, and the interpretation of the results. They contrast the results obtained with those obtained using the traditional significance testing approach. They base their methodology on guidelines produced by the FDA[67] for the use of Bayesian statistics in the assessment of clinical trials for medical devices. Harrison et al. suggest that the Bayesian approach has a number of advantages such as removing the need to correct for multiple comparisons or to make separate tests for significance or equivalence, easier presentation, and interpretation of probability statements and the use of credibility intervals.[66]

Bayesian methods now have many proponents and are an area of statistics whose influence is likely to grow in the future. Some proponents are particularly "bullish". For example, in an open letter, John Kruschke stated the following: "Null-hypothesis significance testing (NHST), with its reliance on p values, has many problems. There is little reason to persist with NHST now that Bayesian methods are accessible to everyone."[68]

In summary, this paper was aimed at providing an overview of some of the statistical concepts that relate to experimental work on the composition of crops and factors that influence the composition. First, it was stressed that experimental design precedes the statistical analysis and that without an appropriate design there is no valid analysis. Second, statistical significance should not be equated with biological importance and should

not be the primary objective of a statistical analysis. The *p* value that is often reported depends on the experimental design, the variability of material, and the statistical test used. It is not a measure of the size of an effect. Third, this paper aimed to point out that the NHST approach has been subject to criticism by many statisticians for some time, with alternative approaches such as estimation and confidence limits preferred to hypothesis testing being preferred by some, whereas others have suggested the development of Bayesian methods. Fourth, I aimed to highlight alternative approaches based around concepts such as equivalence testing and the use of multivariate methods for the exploration and visualization of complex data that can lend insight to the variability in samples. Finally, this paper suggests that it is important to relate the effect size of interest to concepts such as power and sample size so that appropriate experiments are designed and to encourage domain scientists to work toward a consensus on size of effects to be considered biologically important.

AUTHOR INFORMATION

Corresponding Author

*Phone: +44(0)20-8725-5363. Fax: +44(0)20-8725-2993. E-mail: dlovell@sgul.ac.uk.

Funding

The author received travel and accommodation support from the International Life Sciences Institute (ILSI) to present topics discussed in this paper at the 2012 ILSI International Food Biotechnology Committee Plant Composition Workshop.

Notes

The author was a member of the EFSA Scientific Committee from 2008 to 2012 and chair of an EFSA Working Group on Statistical Significance and Biological Relevance. All comments in this paper are his own opinions and do not represent the views of EFSA.

ABBREVIATIONS USED

ANCOVA, analysis of covariance; ANOVA, analysis of variance; BUGS, Bayesian inference using Gibbs sampler; CRD, clinically relevant difference; DOE, design of experiment; EDA, exploratory data analysis; EFSA, European Food Safety Authority; FDA, U.S. Food and Drug Administration; FDR, false discovery rate; GLM, general linear model; GLMM, generalized linear mixed models; GLZ, generalized linear model; IFBiC, International Food Biotechnology Committee; ILSI, International Life Sciences Institute; NHST, null hypothesis significance testing; OECD, Organisation for Economic Co-operation and Development; OFAT, one factor at a time

REFERENCES

(1) Box, J. F. *R. A. Fisher: The Life of a Scientist*; Wiley: New York, 1978.
(2) Edwards, A. W. R. A. Fisher. Twice professor of genetics: London and Cambridge or "a fairly well-known geneticist". *Biometrics* **1990**, *46*, 897−904.
(3) Smith, A. F. M. Mad cows and ecstasy: chance and choice in an evidence-based society. *J. R. Stat. Soc. Ser. A: Stat. Soc.* **1996**, *159*, 367−384.
(4) EFSA Scientific Committee. Statistical significance and biological relevance. *EFSA J.* **2011**, *9*, 2372 (17 pp).
(5) EFSA Panel on Genetically Modified Organisms (GMO). Guidance on the environmental risk assessment of genetically modified plants. *EFSA J.* **2010**, *8*, 1879 (111 pp).
(6) EFSA Panel on Genetically Modified Organisms (GMO). Statistical considerations for the safety evaluation of GMOs. *EFSA J.* **2010**, *8*, 1250 (59 pp).
(7) EFSA Scientific Committee. Guidance on conducting repeated-dose 90-day oral toxicity study in rodents on whole food/feed. *EFSA J.* **2011**, *9*, 2438 (21 pp).
(8) Fisher, R. Presidential address to the First Indian Statistical Congress. *Sankhya* **1938**, *4*, 14−17.
(9) Price, W. D.; Underhill, L. Regulatory perspectives on how composition data are interpreted: food and feed. *J. Agric. Food Chem.* **2013**, DOI: 10.1021/jf401178d.
(10) Fisher, R. A. Statistical methods in genetics. *Heredity* **1952**, *6*, 1−12.
(11) Fisher, R. A. *Statistical Methods for Research Workers*; Oliver and Boyd: Edinburgh, 1925.
(12) Box, G. E. P.; Hunter, W. G.; Hunter, J. S. *Statistics for Experimenters*; Wiley: New York, 1978.
(13) Cox, D. R. The role of statistical significance tests. *Scand. J. Stat.* **1977**, *4*, 49−70.
(14) Neyman, J.; Pearson, E. S. On the use and interpretation of certain test criteria for purposes of statistical inference: part I. *Biometrika* **1928**, 175−240.
(15) Cohen, J. The earth is round ($p < 0.05$). *Am. Psychol.* **1994**, *49*, 997−1003.
(16) Moran, J. L.; Solomon, P. J. A farewell to *P*-values. *Crit. Care Resuscitation: J. Australasian Acad. Crit. Care Med.e* **2004**, *6*, 130.
(17) Gigerenzer, G. Mindless statistics. *J. Socioeconomics* **2004**, *33*, 587−606.
(18) Dallal, J. *Why P=0.05?*; 2012; http://www.jerrydallal.com/LHSP/p05.htm.
(19) Altman, D. G. *Practical Statistics for Medical Research*; Chapman and Hall: London, UK, 1991.
(20) Fisher, R. A.; Yates, F. *Statistical Tables for Biological, Agricultural, and Medical Research*; Oliver and Boyd: London, UK, 1938.
(21) Yates, F. *The Design and Analysis of Factorial Experiments*; Technical Communication 35 of the Commonwealth Bureau of Soils; Commonwealth Agriculture Bureau: Farnham Royal, UK, 1937.
(22) Yates, F. The influence of statistical methods for research workers on the development of the science of statistics. *J. Am. Stat. Assoc.* **1951**, *46*, 19−34.
(23) Brumfield, G. Language: disputed definitions. *Nature* **2008**, *455*, 1023 (see comment)..
(24) Reese, R. A. Significant confusion in scientists' grasp of statistics. *Nature* **2008**, *456*, 315.
(25) Salsburg, D. *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*; W. H. Freeman: New York, 2001.
(26) Ziliak, S. T.; McCloskey, D. N. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*; University of Michigan Press: Ann Arbor, MI, 2008.
(27) Nester, M. Quotes criticizing significance testing, 1997; http://www.indiana.edu/~stigtsts/quotsagn.html.
(28) Nester, M.; Anderson, D. R. *A few notes regarding hypothesis testing*, 1997; http://warnercnr.colostate.edu/~anderson/nester.html.
(29) Thompson, B. 402 citations questioning the indiscriminate use of null hypothesis significance tests in observational studies, 2001; http://warnercnr.colostate.edu/~anderson/thompson1.html.
(30) Hubbard, R.; Lindsay, R. M. Why P values are not a useful measure of evidence in statistical significance testing. *Theory Psychol.* **2008**, *18*, 69−88.
(31) Ioannidis, J. P. Why most published research findings are false. *PLoS Med.* **2005**, *2*, e124.
(32) Chow, S. L. Précis of statistical significance: rationale, validity, and utility. *Behav. Brain Sci.* **1998**, *21*, 169−194.
(33) Fidler, F. From Statistical Significance to Effect Estimation: Statistical Reform in Psychology, Medicine, and Ecology. University of Melbourne, Melbourne, Australia, 2005.
(34) Senn, S. Two cheers for P-values? *J. Epidemiol. Biostat.* **2001**, *6*, 193−204.

(35) Carlin, J. B.; Doyle, L. W. Sample size. Continuing education: statistics for clinicians. *J. Paediatr. Child Health* **2002**, *38*, 300−304.
(36) Turner, J. R. *New Drug Development: Design, Methodology, and Analysis*; Wiley: Hoboken, NJ, 2007.
(37) Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed.; Lawrence Erlbaum Associates: Hillsdale, NJ, 1988.
(38) Lenth, R. V. Some practical guidelines for effective sample size determination. *Am. Stat.* **2001**, *55*, 187−193.
(39) Pocock, S. J. Clinical trials with multiple outcomes: a statistical perspective on their design, analysis, and interpretation. *Control Clin. Trials* **1997**, *18*, 530−545.
(40) Organisation for Economic Co-operation and Development. *OECD Guidelines for the Testing of Chemicals*, Section 4: Health Effects Test No. 443: Extended One-Generation Reproductive Toxicity Study; OECD Publishing: Paris, France, 2011.
(41) Organisation for Economic Co-operation and Development. *OECD Guidelines for the Testing of Chemicals*, Section 4: Health Effects Test No. 487: Genetic Toxicology: Rodent Dominant Lethal Test; OECD Publishing: Paris, France, 1984.
(42) Martin, M. J.; Judson, R. S.; Reif, D. M.; Kavlock, R. J.; Dix, D. J. Profiling chemicals based on chronic toxicity results from the U.S. EPA ToxRef Database. *Environ. Health Perspect.* **2009**, *117*, 393−399.
(43) Cochran, W. G.; Cox, G. M. *Experimental Designs*, 2nd ed.; Wiley: New York, 1957.
(44) Altman, D. G. Statistics in medical journals. *Stat. Med.* **1982**, *1*, 59−71.
(45) Gardner, M. J.; Altman, D. G. Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br. Med. J. (Clin. Res. Ed.)* **1986**, *292*, 746−750.
(46) Weller, A. C. *Editorial Peer Review: Its Strengths and Weaknesses*; Information Today: Medford, NJ, 2001.
(47) Cox, D. R.; Snell, E. J. *Applied Statistics: Principles and Examples*; Chapman and Hall: London, UK, 1981.
(48) Organisation for Economic Co-operation and Development. *Safety Evaluation of Foods Derived by Modern Biotechnology: Concept and Principles*; OECD Publishing: Paris, France, 1993.
(49) Kuiper, H. A.; Kleter, G. A.; Noteborn, H. P. J. M.; Kok, E. J. Substantial equivalence − an appropriate paradigm for the safety assessment of genetically modified foods. *Toxicology* **2002**, *181−182*, 427−431.
(50) U.S. Food and Drug Administration, Center for Drug Evaluation and Research. Guidance for Industry: Statistical Approaches to Establishing Bioequivalence, 2001; http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm070244.pdf.
(51) Committee for Proprietary Medicinal Products. Points to consider on switching between superiority and non-inferiority. *Br. J. Clin. Pharmacol.* **2001**, *52*, 223−228.
(52) Altman, D. G.; Bland, J. M. Absence of evidence is not evidence of absence. *Br. Med. J.* **1995**, *311*, 485.
(53) Altman, D. G.; Bland, J. M. Confidence intervals illuminate absence of evidence. *Br. Med. J.* **2004**, *328*, 1016−1017.
(54) Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B: Stat. Soc.* **1995**, *57*, 289−300.
(55) Payne, R. *A Guide to Regression, Nonlinear and Generalized Linear Models in GenStat*, 15th ed.; VSN International: Hemel Hempstead, UK, 2012.
(56) Nelder, J. A. Discussion on the papers by Wynn and Bloomfield, and O'Neill and Wetherill. *J. R. Stat. Soc. Ser. B: Stat. Soc.* **1971**, *33*, 244−246.
(57) Bryan-Jones, J.; Finney, D. J. On an error in 'Instructions to authors'. *Hortic. Sci.* **1983**, *18*, 279−281.
(58) Little, T. M. If Galileo published in HortScience [multiple range tests]. *Hortic. Sci.* **1978**, *13*.
(59) Preece, D. A. The design and analysis of experiments: what has gone wrong. *Utilitas Math. A* **1982**, *21*, 201−244.
(60) Perry, J. N. Multiple-comparison procedures: a dissenting view. *J. Econ. Entomol.* **1986**, *79*, 1149−1155.
(61) Nelder, J. A.; Wedderburn, R. W. M. Generalized linear models. *J. R. Stat. Soc. Ser. A: Stat. Soc.* **1972**, 370−384.
(62) Box, G. E. P.; Draper, N. R. *Empirical Model-Building and Response Surfaces*; Wiley: New York, 1987.
(63) Tukey, J. W. *Exploratory Data Analysis*; Pearson: Reading, MA, 1977.
(64) Johnson, R. A.; Wichern, D. W. *Applied Multivariate Stastical Analysis*, 6th ed.; Prentice Hall: New York, 2007.
(65) Shewry, P. Natural variability in grain composition in wheat and related cereals. *J. Agric. Food Chem.* **2013**, DOI: 10.1021/jf3043092.
(66) Harrison, J. M.; Breeze, M. L.; Harrigan, G. G. Introduction to Bayesian statistical approaches to compositional analyses of transgenic crops 1. Model validation and setting the stage. *Regul. Toxicol. Pharmacol.* **2011**, *60*, 381−388.
(67) U.S. Food and Drug Administration, Center for Devices and Radiological Health. Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials, 2010; http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071121.pdf.
(68) Kruschke, J. K. *An open letter*, 2010; http://www.indiana.edu/~kruschke/AnOpenLetter.htm.