Chapter 2: Formal concepts & principles and informal interpretations.

The standard use we make of frequentist inference is to throw light on the plausibility, or otherwise, of a hypothesis or hypotheses on the basis of data and with no input from prior knowledge about the issue. The proponents of this approach regard the lack of priors as giving their method greater objectivity than Bayesian inference since preconceptions are less able to influence the outcome. Without prior probabilities it is not possible to calculate a posterior probability for a hypothesis being true given the data observed, therefore frequentist inference uses a less direct approach.

In this chapter we describe in detail the ‘optimal’ theory of Neyman and Pearson[1] and contrast the formal properties of their approach with the informal, evidential, properties typically attributed to it. We define the concepts ‘sufficiency’ and ‘ancillarity’ and principles associated with them as well as the ‘likelihood principle’, which is at odds with frequentist inference.

2.1 Optimal inference.

The Neyman-Pearson theorem[2] identifies tests possessing a particular feature regarded by proponents of the theory as ‘optimal’. The theorem is written in the context of a comparison between two well-defined simple hypotheses, H and K. The approach is not symmetric in the two hypotheses. Suppose that, on the basis of our data, we must either reject the hypothesis H[3] (in favour of K) or accept H (as opposed to K). Then there are two possible errors that can be made. A *Type I error* has been made when we reject H despite the fact that H is actually true, and a *Type II error* if we accept H even though it is false. In Neyman-Pearson theory, the probability of a Type I error is

[1] For a detailed discussion and comparison of the different versions of frequentist inference (Fisherian, Neyman-Pearson and hybrid Fisher-Neyman-Pearson) in theory as well as practice, see Gigerenzer (1989) and Royall (1997).

[2] Kendall & Stuart, p. 166.

[3] I use the “accept/reject H” phrasing throughout this work because, despite all its problems, it is still the most common terminology employed. For the same reason I use the term null hypothesis for H and alternative hypothesis for K.

fixed by the analyst at a low level, in most cases not more than 5%. This value, denoted α, is the *significance level* of the test. The *rejection rule* specifies the data that causes H to be rejected and is based on the Neyman-Pearson theorem which partitions the sample space into two regions. The optimality of this method lies in the fact that the Neyman-Pearson rejection rule minimises the probability of Type II error for the given significance level (or, equivalently, maximises the *power* $=1-P(\text{Type II error})$). Any different rule (i.e. different dichotomous partition of the sample space) that produces the same significance level will produce a higher probability of Type II error. The significance level of the test can be thought of as the long-run failure rate when H is true; when α is small this failure rate is small (say 5%) and most of the time that H is true we will correctly accept H. Neyman-Pearson theory can be interpreted in a quality control sense as a way of ensuring that (in the long run) erroneous statements (or other actions) are not made more than a certain proportion of the time. Contrasting his approach to that of Fisher, Neyman frequently represented his theory as being pure decision theory with no evidential implications, however, it is not clear that he and Pearson were never more ambitious than this. Pearson in 1955 wrote that "from the start we shared Professor Fisher's view that in scientific enquiry, a statistical test is 'a means of learning'".[4] Whatever the theory's originators intended, it is a fact that test results are frequently given an evidential interpretation. The most common justification of this is as follows. Suppose that we reject H at the 5% level; we know that when H is true this will only happen 5% of the time, therefore it is tempting to imagine that H is probably not true, or, in other words, *that we have fairly strong evidence against H* (presumably relative to K). Since $P(\text{data}|\text{H})$ may be small without $P(\text{H}|\text{data})$ also being small, this reasoning is flawed[5], but how many psychology and biology students would attend our classes if they understood that tests do not extract evidence from data, and that confidence intervals do not *probably contain* μ? Most introductory textbooks contain statements that encourage this particular evidential interpretation. For instance, on hypothesis tests:

[4] Pearson, E. S. (1955). Statistical concepts in their relation to reality. *J. Roy. Statist. Soc., Ser. B*, **17**, 204-207. Quoted in Leymann (1993).

[5] For a much more in-depth exposition on these issues, see Royall (1997), especially pp. 41-76.

A significance test or hypothesis test enables us to check, with a measure of accuracy, *which of two such scenarios* [hypotheses] *is the most likely*[6] [my italics].

And, on confidence intervals:

…the fact that we had such a high degree of probability, prior to the performance of the experiment, that the random interval $(\bar{X}-2\sigma/\sqrt{n},\ \bar{X}+2\sigma/\sqrt{n})$ includes the fixed point (parameter) μ *leads us to have some reliance on the particular interval* $(\bar{x}-2\sigma/\sqrt{n},\ \bar{x}+2\sigma/\sqrt{n})$ [my italics]. This reliance is reflected by calling the known interval $(\bar{x}-2\sigma/\sqrt{n},\ \bar{x}+2\sigma/\sqrt{n})$ a 95.4 per cent confidence interval for μ[7].

One of the unsatisfactory features of the fixed α approach is that it splits the sample space into two regions without paying any attention to distinctions within each region. This seems quite inadequate if we want to use tests to assess evidence. According to the evidential interpretation, rejecting H at the 1% significance level provides stronger evidence against H than rejecting at the 5% significance level. For this reason many courses in statistics use the Fisherian concept of the p-value, usually in addition to significance level. There is an obvious relation between the two though they stem from different theoretical approaches. The rule 'Reject H if $x \le c_\alpha$' can be equivalently written as 'Reject H if p-value$(x) \le \alpha$'[8]; note also that the p-value(x) is the smallest value of α that would lead us to reject H on observing x. Instead of simply stating an accept/reject result the analyst may choose to state the exact p-value and hence (apparently) give a more exact measure of evidence. Oddly enough, despite the fact that computers have for a long time made the calculation of exact p-values simple, it is still common for p-values to be interpreted solely by reference to one or more predetermined significance levels. Thus we have the phenomenon of the 'starred p-values' where the p-value is stated, and, if significant ($\le$5%), highlighted by a number of stars signifying how significant the result is[9] (for example, one or two

[6] Smith, p. 515.

[7] Hogg & Craig, p.213.

[8] As long as $\exists x : \text{p-value}(x) = \alpha$ – always true when X is a continuous variable.

[9] See, for instance, Smith, p. 517. A version of this usage is the convention in psychology journals.

stars for p-values in the categories (1%,5%], or [0%,1%] respectively). The fact that this extra information is completely redundant suggests that these categories are the real basis for the inference, so the inference is still based on fixed levels to a large degree. (Unfortunately these levels no longer represent the long run failure rate since one star occurs not 5% of the time but $4\% (= 5\% - 1\%)$ of the time when H is true.) This type of approach is sometimes described as being a hybrid between the approaches of Neyman and Pearson and of Fisher[10], and indicates that we still feel a need to interpret the p-value via the 'error probability' concept.

The power of the test fulfils two functions. On the one hand it is the basis for a theory of optimality, on the other it is used to interpret the results of a particular test. Since data can be analysed in a number of different ways, some of which are clearly less efficient in their use of the available information than others, finding the *most powerful* test provides a means of choosing between them. However, a test may be the most powerful test and still have a fairly low power for a given experiment and hypotheses; when the two hypotheses are very close together (relative to the data variation) the power is always low. When the power is high and the data lies in the **accept H** region we may argue that we have evidence supporting H (or supporting H relative to K). The argument is exactly the same as that used to interpret the rejection of H as evidence against H, namely, when K is true, we know that we will accept H only a small proportion of the time (1-power), thus, finding that we have accepted H, we are inclined to imagine that K is probably not true. When the data is in the acceptance region for testing H versus K and the power is high (say $\geq 95\%$), it follows that a test of the null hypothesis K versus the alternative H would return a significant result at the 5% level. However, attitudes on this point vary, with some insisting that you can conclude nothing from a failure to reject H, others automatically interpreting it as evidence in favour of H and a sophisticated minority interpreting it as evidence in favour of H relative to K only when the power is high. The reformists in *psychology*, who insist that the power of a test must be reported, do so in order to facilitate this reasoning.[11]

[10] For an interesting discussion of whether it is plausible to see these two theories as one, see Lehmann (1993).

[11] See editorial in *Memory and cognition* (1993).

The Neyman-Pearson theorem and most powerful tests.

The Neyman-Pearson theorem identifies the *most powerful test* as that which is based on the value of the likelihood ratio. This region is called the *best critical* [i.e. rejection] *region* (BCR). Specifically (assuming that the likelihood ratio statistic is continuous under H), the best critical region for a test with significance level α is given by $\{x : LR(x) \leq k_\alpha\}$, where the identity $P_H(LR(X) \leq k_\alpha) = \alpha$ defines the value of k_α, i.e. k_α is that value which produces the required significance level. The power of the test is given by $P_K(LR(X) \leq k_\alpha)$ and is the highest power achievable (hence 'best').

If various different alternative hypotheses, K*ᵢ*, all give rise to the same BCR, then this BCR can be used as the basis for a test of H versus the disjunction of the K_i hypotheses. For example, if $X \sim N(\mu,1)$, the test of $\mathrm{H}{:}\mu = 0$ versus $\mathrm{K}_i{:}\mu = \mu_i$ has the same BCR (given α) for all values of μ_i which are positive. This BCR can therefore be used to test $\mathrm{H}{:}\mu = 0$ versus $\mathrm{K}{:}\mu > 0$. When the alternative hypothesis is composite in this way, the power is a function of its components; such a test is called *uniformly most powerful* because it is the most powerful test for each individual value μ_i in the set defined by K. Sometimes there is no most powerful test for a particular composite alternative. For the Normal case, any test of the form $\mathrm{H}{:}\mu = \mu_1$ versus $\mathrm{K}{:}\mu = \mu_i$, has a different BCR if $\mu_i > \mu_1$ than if $\mu_i < \mu_1$, so there is no most powerful critical region that can be used to test $\mathrm{H}{:}\mu = \mu_1$ versus $\mathrm{K}{:}\mu \neq \mu_1$. But, even here, a kind of optimality is achievable. A test is *unbiased*[12] if the power is greater than or equal to α for all values in the set defined by K. This is simply to say that we are more likely to reject H when K is true than when H is true, and this seems reasonable. This requirement excludes a number of tests from consideration, for instance, if we base a test of $\mathrm{H}{:}\mu = 0$ versus $\mathrm{K}{:}\mu \neq 0$ on the same critical region that worked for testing $\mathrm{H}{:}\mu = 0$ versus $\mathrm{K}{:}\mu > 0$, then for values of μ_i that are less than zero, the power will be less than α. When we exclude such tests we find that we are quite often able to identify a test that is uniformly most powerful among those tests that are

[12] Unbiased tests – not to be confused with unbiased estimators.

left. Such a test is called a *uniformly most powerful unbiased* test; many standard two-sided tests have this property.

2.2 Sufficiency & Ancillarity.

Sufficiency.

Data frequently includes information that is not relevant to the question at issue. For instance, if a series of observations are understood to come from identical and independently distributed random variables, then the order in which the observations appeared tells us nothing of interest – we could re-order the data with no loss of relevant information.

Suppose that we are interested solely in a parameter, $\theta \in \Theta$. The random variable X has a density, $f(x;\theta)$ dependent on θ (and, to simplify matters, suppose that everything else about the density is known), then a statistic, $s(X)$ is *sufficient for* $\theta \in \Theta$, if and only if it satisfies the following condition:

$$f(x;\theta) = g(s(x);\theta) \times h(x) \ \forall x, \ \forall \theta \in \Theta$$

This result is called the *factorisation theorem*.

Sufficient statistics are often regarded as containing all the information about $\theta \in \Theta$ that is present in x. This makes perfect sense in a Bayesian context since the posterior probabilities $P(\theta \mid X = x)$ and $P(\theta \mid s(X) = s(x))$ are always equal. In a frequentist context the argument is less direct, but is usually given as follows: since the conditional density $f(x \mid s(X) = s(x))$ is the same for all values of $\theta \in \Theta$, it cannot be used as the basis for making any kind of inference about $\theta \in \Theta$, it appears that all the information about θ that was present in x was actually present in $s(x)$ [13].

[13] See, for instance, Hogg & Craig, p. 343.

Sufficiency and Neyman-Pearson optimality are consistent, in a sense, as follows. If a test on θ is most powerful or uniformly most powerful then its critical region can be written in terms of any sufficient statistic for θ. This follows from the factorisation theorem since, if $s(X)$ is sufficient,

$$\{x : LR(x) \le k\}$$

can be written in the form

$$\{x : \frac{g(s(x);\theta_1)}{g(s(x);\theta_2)} \le k\}$$

where θ_1 and θ_2 are known constants for any test of specific hypotheses. (However Neyman and Pearson advocated the use of randomising variables in cases where a discrete X does not produce an appropriate significance level, and this causes the test result to depend on observations not incorporated in the sufficient statistic.)

Sufficient statistics are not unique and some are more efficient than others. As with all functions, they partition the domain into groups by attaching a label to each element of the domain (in this case the sample space $\mathfrak{X}(\Theta)$). A statistic which is sufficient for $\theta \in \Theta$ and which can be written as a function of each and every other sufficient statistic is called a *minimal sufficient* statistic for $\theta \in \Theta$; such statistics are viewed as maximally free of junk information. Every minimal sufficient statistic is a one-to-one function of any other minimal sufficient statistic, so they vary only in the labelling they give to the groups in the partition of the sample space. If we disregard the labelling, any minimal sufficient statistic is effectively unique.

Whether or not a particular statistic is sufficient depends on the particular parameter space, Θ, since certain statements are required to be true *for all* $\theta \in \Theta$. If s is sufficient for θ in the parameter space Θ, then clearly it will also be sufficient for $\theta \in \Theta'$ as long as $\Theta' \subseteq \Theta$. This is not true of minimal sufficiency where reducing the size of the parameter space will often make certain information in s irrelevant so that s will not be minimal sufficient for $\theta \in \Theta'$ even though it was minimal sufficient for $\theta \in \Theta \supseteq \Theta'$. In the following chapters we will often consider parameter spaces consisting of only *two* values, as is the case in a test of two simple hypotheses. In such cases the function of X that is a minimal sufficient statistic for θ in the more commonly used parameter spaces (for example, $\bar{X}$ for $\mu \in \mathbb{R}$) is no longer minimal sufficient; in fact, the likelihood ratio statistic is the minimal sufficient statistic for θ in a binary parameter space.

The sufficiency principle.

If it is true that a sufficient statistic contains all the relevant information about θ for a given context, then this should have implications for how we infer things or extract evidence from data. Suppose that x_1 and x_2 are two experimental observations on the random variable, X , and that $s(x_1) = s(x_2)$ where $s(X)$ is any sufficient statistic.[14] Then it must follow that x_1 and x_2 contain the same information about θ and we should make the same inference or conclusion about θ from x_1 as from x_2. Berger and Wolpert describe this as the weak sufficiency principle, we will abbreviate it to *sufficiency principle* (SP). The various hybrids of Fisher-Neyman-Pearson inference are mostly consistent with this principle (as is the inference of Fisher), and indeed it is one of the least controversial principles in statistics, agreeing with a large number of different theories.

The sufficiency principle does not dictate a particular mode of inference or analysis but simply describes a feature that any good inference should possess. Since the SP applies to all sufficient statistics, it applies, in particular, to the minimal sufficient statistic. To understand the implications of the SP in a particular case it is only necessary to apply it to the minimal sufficient statistic since this guarantees that it will be satisfied with respect to all other sufficient statistics. For example, if we have a fixed size random sample from a Normal population with known standard deviation, then the vector $X_1, \ldots, X_n$ is itself a sufficient statistic for the population mean μ and so the SP tells us that if two experimenters observe the same data, they should make the same inference. However the minimal sufficient statistic is the sample mean, $\bar{X}$; applying the SP to $\bar{X}$ we find that any two data sets with the *same mean* should give rise to the same inference. The fact that two identical data sets must give rise to the same inference follows automatically from this, but the opposite is not true.

[14] *Sufficient statistic* meaning always a sufficient statistic *for a specific parameter* θ.

The likelihood principle.

The concept of sufficiency applies only within the context of a single experiment. Now consider the case where two different experiments have been designed to yield information about the *same* parameter $\theta \in \Theta$. Experiment A observes the value of a random variable X, which has a density (likelihood) $f_A(\cdot;\theta)$, and Experiment B observes the value of a random variable Y, that has a density $f_B(\cdot;\theta)$. Suppose that x_0 is an observed value of X and y_0 is an observed value of Y.

Suppose further that, *for all values of θ in the parameter space*, the following *proportionality identity* holds:[15]

$$f_A(x_0;\theta) = k \times f_B(y_0;\theta)$$

for some (positive) constant k.

The *likelihood principle* (LP) states that, if this identity holds, then x_0 and y_0 contain the same information regarding θ (in Θ) and should thus give rise to the same inference about θ.

In essence, the LP asserts that the experimental result contains information about θ only through the likelihood function. This has some dramatic implications; for instance, a particular result, (say, 14 heads out of 20 independent tosses of a coin, where we are interested in $P(head) = p$) will be interpreted the same way regardless of the 'stopping rule' used in the experiment, for instance, regardless of whether we *tossed the coin 20 times* or *tossed the coin until we obtained 6 tails.* Some regard this as a point in favour of the LP; others are horrified by it.

[15] $L(\theta;x)$ is the likelihood function of θ. This assumes that x is fixed and θ varies, whereas the density $f(x;\theta)$ is the same function assuming that x varies and θ is fixed.

It is easy to show that the LP entails the SP. Suppose that for a single experiment $s(x_1) = s(x_2) = s_0$ where $s(X)$ is a sufficient statistic. Then, from the factorisation theorem, we know that, for all θ,

$$f(x_1;\theta) = g(s_0;\theta) \times h(x_1) \text{ and } f(x_2;\theta) = g(s_0;\theta) \times h(x_2)\,.$$

Thus, $f(x_1;\theta) = k \times f(x_2;\theta)$, where $k = h(x_1)/h(x_2)$ is a fixed constant (with respect to θ). It follows from the LP that we should infer the same about θ from x_1 as from x_2, in accordance with the SP.

The SP is sometimes called the weak likelihood principle[16], but its weakness is of a particular kind. The SP says exactly the same thing as the LP but its scope is limited to the case of a single experiment; it is like the difference between enacting a law at the state level and enacting the same law at the federal level. Nevertheless, methods that are consistent with the SP can be grossly in breach of the LP.

Most frequentist methods are consistent with the SP but all breach the LP by virtue of using tail-area probabilities. According to the LP, it is the value of the likelihood (density) *at the observed point* x that provides all the information. A tail-area (for instance, the p-value of x) depends not only on the value of the density at x, but also on the value of the density at all the other values 'more extreme than x' over which the density is integrated. Thus in a right-sided z-test of the hypothesis H: $\mu = \mu_1$, the p-value of an observation, x, is given by $\int_x^\infty f(t;\mu_1)dt$, involving the densities of many values other than x. We can imagine a different experiment, involving the same parameter μ, in which (for all μ) the random variable has a density that is different from $f(t;\mu)$ for all values of t *other than* $t = x$ (to make things simple, imagine that both variables have the same support). In this case the LP says that we should make the same inference from the observation x regardless of which experiment it came from, however the p-values of x, from the two experiments, will probably be different.

[16] Cox & Hinkley.

Recalling that Neyman-Pearson methods are based on the likelihood ratio, this may seem odd, for if the density of x is the same function of μ for both experiments, then the value of the likelihood ratio, $LR(x)$, must surely be the same for any μ_1 and μ_2. This is true, but, since $LR_1(\cdot)$ and $LR_2(\cdot)$ – the likelihood ratio functions for the two experiments – are different, the cut-off values for a test of given significance level (i.e. k_α) will usually also be different. The fact that there is no *general* connection between α and the critical likelihood ratio value (k_α), in frequentist inference, is another indication that the method is inconsistent with the LP.

Ancillarity.

We turn now to the concept of *ancillary statistic*. Fisher developed the concept[17] and eventually settled upon a fairly restrictive definition. A statistic (i.e. function of the data) is exactly ancillary only if, by itself, it contains no information about the parameter of interest $\theta \in \Theta$. To understand the purpose of this, consider the following example. We take a sample from a Normal population with unknown mean (the parameter of interest) but known variance and use a random mechanism (such as the toss of a coin or roll of a die) to choose the sample size n. Then N is a random variable that can take different values with probabilities dependent on the coin or die. Knowing the value of N tells us nothing about the value of the mean μ. But N has another property that Fisher required in ancillary statistics. Although N tells us nothing directly about μ, it is relevant in connection with other information. If we know the value of the sample mean, $\bar{x}$, then we also need to know n since the sample mean will be more reliable (as an estimate of μ) if n is large than if n is small. Thus we may say that N is indirectly informative about μ (or about the precision of $\bar{x}$ as an estimate of μ). In this case we note that the minimal sufficient statistic for $\mu \in \mathbb{R}$ is $(N, \bar{X})$ where $\bar{X}$ is also the maximum likelihood estimator of μ. This inspired Fisher's definition[18] of ancillary statistic as a statistic having the following two properties:

[17] Fisher.
[18] Basu (1964).

a) It has a distribution that is the same for all values of $\theta \in \Theta$; and
b) Together with the MLE of $\theta \in \Theta$, it comprises the minimal sufficient statistic for $\theta \in \Theta$.

(We have made Θ explicit in this definition; it is often missing on the assumption that the form of the parameter space is understood.)

An informal account of the conditional principle.

Fisher argued that inference on θ should be carried out *conditional upon* the observed value of the ancillary statistic. This position is easy to understand in the context of the sample size example. Suppose that N is determined by the toss of a fair coin so that $P(N=10)=P(N=100)=\frac{1}{2}$; i.e. the sample size is equally likely to be either 10 or 100. If the former, our data is a realisation of $X_1,\ldots,X_{10}$ and, if the latter, a realisation of $X_1,\ldots,X_{100}$. The sample space, $\mathfrak{X}$, is comprised of all possible realisations of both these kinds. As we saw above, the *unobserved* observations contained in $\mathfrak{X}$ play a crucial role in frequentist inference. This is as true for Fisher as for Neyman & Pearson. However, in the particular case of ancillary statistics, Fisher favoured removing that part of the sample space associated with unobserved values of the ancillary statistic. Thus, if the result of the coin toss is such that we use a sample size of *ten* (observed value of N is 10), the elements of $\mathfrak{X}$ of the form $X_1,\ldots,X_{100}$ (consistent with observing $N=100$) should be excluded from the sample space and the analysis carried out based on the reduced space containing only elements of the form $X_1,\ldots,X_{10}$. Thus the inference about μ, based on the observed sample mean, $\bar{x}$, will be the same as it would have been if the sample size n had been fixed at 10 in advance. This is justified as follows: since n turned out to be 10 (and this tells us nothing directly about μ), the fact that it *might have been* 100 is clearly irrelevant to any questions about μ. We can emphasise this by considering whether, given that n was 10, we would want to change the inference, based on $\bar{x}$, just because we discovered that the coin was biased instead of fair. (In Chapter 5 we will compare two principles derived from this general view.)

2.3 Evidential interpretations of Neyman-Pearson inference.

We have already touched on some potential problems with frequentist inference. Fisher believed that his approach was appropriate for scientists embarking on explorative research and criticised Neyman and Pearson for producing a method that was "represented by the doctrine that reasoning, properly speaking, cannot be applied to empirical data to lead to inferences valid in the real world".[19] Yet Fisher's own methods often lead to results that are numerically similar or even identical to those produced by Neyman and Pearson. There is quite a lot of evidence that Pearson, particularly, did not accept that Neyman-Pearson theory was so limited in application.[20] In our discussion, we assume that it may be reasonable to use frequentist methods to carry out inferences which have meaning in the real world – not just in quality control contexts; Birnbaum called this the "confidence concept of statistical evidence", adding "this concept is not part of the Neyman-Pearson theory, which denies any role to concepts of statistical evidence" *but* "[it is] the concept by which confidence limits and hypothesis tests are usually interpreted".[21] Thus we will discuss these methods in terms of their ability to provide us with evidence gleaned from the data and criticise them when they appear not to do so. Certainly, Neyman-Pearson hypothesis testing is widely used by those conducting general scientific research and is interpreted as providing evidence or inferences about the real world even if the mechanism by which it can be so interpreted is unclear. But there are ambiguities. If the test statistic lies in an α-level BCR, or equivalently, the p-value of the observed data is very small, is this evidence against the null hypothesis, H, and, if so, how strong? The p-value often seems to be unofficially regarded as a measure of evidence in favour of H. This is to some degree natural since small p-values seem

[19] Fisher, R. A. (1973), *Statistical methods and scientific inference, (3rd ed)*, Collins Macmillan, London, p. 7, quoted in Lehmann (1993), p. 1243.

[20] See Mayo.

[21] Birnbaum, A. (1970). Statistical methods in scientific inference, *Nature*, **225**, p. 1033, quoted in Giere, p. 8.

to suggest that H is not true, and the p-value gets smaller as x becomes less and less consistent with H (unlike the crude accept/reject function). The following passage is typical of the way in which students are introduced to hypothesis testing; it gives a largely uncontroversial and untechnical definition of the p-value and then moves off into the vaguer area of interpretation – note the use of the phrase "in the sense that" which allows an imprecise connection to be drawn between the two:

…the *p*-value … is a measure of how likely/unlikely it is to experience the observed data if the null hypothesis is indeed true. If the *p*-value is large, then the observed value of the test statistic is quite likely when H_0 is true; this, in turn, leads us to favour the null hypothesis in the sense that there is not enough evidence to reject it. If the *p*-value is small, then it is quite unlikely to obtain a test statistic of the magnitude observed when H_0 is true; this, in turn, leads us to favour the alternative hypothesis.[22]

But the p-value is nothing like a posterior probability of H, and small p-values can occur with large posterior probabilities of H.

Aside from the sufficiency principle (which is widely accepted among many schools of thought[23]), the nearest frequentism comes to a universal principle is the principle of 'maximising power' behind Neyman & Pearson's 'best' tests; but this is really an optimality property for design of experiments rather than a principle for interpreting the meaning of data; it is not feasible that all data which cause H to be rejected in favour of K at (say) the 5% level in a given highest power test necessarily have the same evidential meaning with respect to H and K. On the other hand, using p-values instead of a fixed significance level does not produce a general rule either, or at least, not a principle that is widely accepted, since there is disagreement about the meaning that can be attached to a p-value. What, for instance, is the significance – if any – of the sample size? In 1934 Fisher wrote, and thereafter maintained, that:

It is not true … that valid conclusions cannot be drawn from small samples; if accurate methods are used in calculating the probability [p-value], we thereby

[22] Smith, p. 517.

[23] "… no one known to me now rejects [the sufficiency principle]", Savage (1970), p. 401.

make full allowance for the size of the sample, and should be influenced in our judgement only by the value of the probability indicated.[24]

But, forty years later, ten influential applied statisticians (including D. R. Cox) claimed that the meaning that can be ascribed to any particular p-value depends on the sample size, arguing that:

A given *p*-value in a large trial is usually stronger evidence that the treatments really differ [i.e. against the null hypothesis] than the same *p*-value in a small trial of the same treatments would be.[25]

Lindley and Scott (1984) also believed that the sample size needs to be taken into account, but, according to them, its influence is in the other direction:

… the interpretation to be placed on the phrase 'significant at 5%' depends on the sample size: it is more indicative of the falsity of the null hypothesis with a small sample than with a large one[26]

and Mayo (1996) also takes this view.

Likelihood theorist, Royall, finds nothing strange about this since "we should not be surprised to find that a statistical procedure that purports to measure evidence, but in a way incompatible with the law of likelihood, is mired in paradox and controversy".[27]

[24] Fisher, R. A. (1934) *Statistical methods for research workers* (5th ed.), Oliver & Boyd, London, quoted in Royall, p. 70.

[25] Peto, R. et al. (1976) Design and analysis of randomised clinical trials requiring prolonged observation of each patient, I: Introduction and design, *British Medical J.*, **34**, 585-612, quoted in Royall, p. 71.

[26] Lindley, D. V. & Scott, W. F. (1984) *New Cambridge Elementary Statistical Tables*, C.U.P., London, quoted in Royall, p.71.

[27] Royall, p.71.