The Basics of Statistical Hypothesis Testing

Jesper Jerkert
Division of Philosophy, KTH

Contents

Introduction 1
What is statistical hypothesis testing? 2
John Arbuthnot's test 2
Problems in the outlined test procedure 4
Forming a rejection set 5
Fisherian hypothesis testing 8
Drawbacks and criticisms 10
The Neyman–Pearson contribution 17
Fisher vs. Neyman–Pearson 18
The role of the null 20
Bayesian hypothesis testing 22
Statistical hypothesis testing in the philosophy of science 24
References 25

Introduction

Those who have taken a first course in statistics might have learned the basics of statistical hypothesis testing. Normally the instruction in such courses is practically oriented: you get to know how to perform statistical hypothesis testing in a couple of typical situations, but time will not allow a full treatment of the rationale for such testing. Also, the teaching might conceal that there could be alternative ways of performing statistical hypothesis testing other than the method actually taught. What is being taught as standard statistical hypothesis testing could be a *bricolage* of ideas from different theorists, not necessarily adding up to a coherent whole upon scrutiny.

This text is intended to present statistical hypothesis testing in an argumentative and philosophical context. It will display and contrast various ideas of hypothesis testing against one another and will hopefully put the reader in a position to argue for and against different models of hypothesis testing. The ideal reader should be familiar with the mathematics typically encountered in simple statistical hypothesis testing, i.e. frequently used statistical distributions and basic theory of probability.

What is statistical hypothesis testing?

A hypothesis may be described as "a hunch, speculation, or conjecture proposed as a possible solution to a problem, and requiring further investigation of its acceptability by argument or observation and experiment" (Belsey, 1995). A hypothesis, then, is a statement that is (or should be) open to testing. According to generally acknowledged accounts of scientific progress, science proceeds by the acceptance or rejection of hypotheses in relation to the available evidence. The exact nature and scope of this process is, of course, subject to different opinions among philosophers. It is agreed, however, that statistical hypothesis testing is a subset of the general category of scientific inference. Statistical hypothesis testing is thus an important part of science. The adjective "statistical" signifies that the evidence is statistical in nature. However, since statistics is such a broad term ("the department of study that has for its object the collection and arrangement of numerical facts or data", according to *The Oxford English Dictionary*), this is not a sufficient condition for classifying an inference as an instance of statistical hypothesis testing. An additional necessary condition is that at least one hypothesis in the test is statistical in nature.

For example, the perennial philosophical problem of induction is usually not taken as an instance of statistical hypothesis testing, although the available evidence may well be considered statistical in nature ("this swan is white, that swan is white, another swan I observed was white as well, …"). A plausible reason for this is that the wanted conclusion, namely "All *F*s are *G*s", is not generally understood as a statistical hypothesis. On the other hand, the statement "80 % of all *F*s are *G*s" is a statistical hypothesis for sure, as is, e.g., the statement "when tossed, this coin will show heads and tails up with equal probability".

John Arbuthnot's test

John Arbuthnot[1] (1667–1735) was Queen Anne's physician during 1709–1714 and published a note called "An argument for Divine Providence, taken from the constant regularity observ'd in the births of both sexes" in the *Philosophical Transactions of the Royal Society* in 1710 (Stigler, 1986:225f). Arbuthnot claimed that the sex of a newborn might be represented by the throw of a two-sided die (or the flipping of a coin), the faces of which he took as equiprobable to turn up. However, data on christenings in London for the period 1629–1710, i.e. 82 consecutive years, showed that male births exceeded female births for each and every year. In other words, all 82 years were "boy years", if by this we mean years where boys formed a majority among newborns. The probability of having 82 consecutive boy years under the hypothesis of even chances, Arbuthnot argued, is $\left(\frac{1}{2}\right)^{82}$, a very tiny probability indeed. (In fact, this probability is approximately 2.1×10^{-25}.) Arbuthnot concluded that "it is Art, not Chance, that governs" (Stigler, 1986:226). Arbuthnot's argument was widely adopted by priests as evidence for divine intervention (Hacking, 1965:77). It should be noted, however, that Arbuthnot was a well-known satirist, friend of Jonathan Swift. He may well have had his tongue in cheek.

Arbuthnot's argument is considered one of the earliest examples of statistical hypothesis testing. Let us try to state in more general terms (and more carefully) the procedure

[1] His name is pronounced with stress on the second syllable: /ɑːˈbʌθnət/.

and logic of John Arbuthnot's test. He started with the formulation of a hypothesis. In this case the hypothesis (let's call it $H_{o,gen}$) was that there is an even chance whether a newborn will be a girl or a boy. Although not necessarily credible in the eyes of the researcher (Arbuthnot most certainly knew that boys were more numerous than girls among newborns), this is an easy hypothesis to formulate and an easy one to model mathematically. We may call it a *null hypothesis*, hence the subscript "o". The subscript "gen" is read "general", so that $H_{o,gen}$ is the general null hypothesis. The rationale for this name is that this hypothesis might be more general than the hypothesis we will actually use mathematically in the test. As for the term "null" and the exact nature of null hypotheses, we will return to this issue (see page 20).

Having stated the hypothesis in words, we need to model it mathematically in a way that can be used for comparison with the data we have (or will have). In Arbuthnot's case, the data concern the number of boy years in a series of 82 consecutive years. Therefore, a suitable hypothesis following from $H_{o,gen}$ is that approximately half of the years ought to be boy years in the series. We call this hypothesis H_o. Modeled more exactly, it states that d, understood as the number of boy years, is the outcome of a $\text{Bin}(82, \frac{1}{2})$ distribution, i.e. a binomial distribution with 82 independent trials, each with probability $\frac{1}{2}$.

I would like to stress the difference between $H_{o,gen}$ and H_o here. From the original (general) hypothesis $H_{o,gen}$ that there is an even chance whether a newborn will be a girl or a boy, many specific hypotheses useful for a test can be derived. For example, under $H_{o,gen}$ there should be around 50 boys among 100 newborns. But that particular piece of information is not of much help to us, since our actual data are not concerned with the number of boys among 100 newborns, but with the number of boy years in 82 consecutive years. Therefore, we need to spell out a hypothesis H_o that specifies the statistical behavior of the data we have (or will have). Hence we take H_o as explained in the preceding paragraph. The passing from $H_{o,gen}$ to H_o is justified in Arbuthnot's case, i.e. H_o follows with mathematical certainty from $H_{o,gen}$, but the proof was probably beyond Arbuthnot's knowledge (Hacking, 1965:76). It is useful to distinguish between $H_{o,gen}$ and H_o, for it is not always obvious that the H_o used in a test is a fair representation of $H_{o,gen}$. Examples will be given later in this text.[2]

Having stated H_o mathematically we now compare the actual outcome with what could be expected if H_o were true. More precisely, we calculate the probability of the outcome given H_o; using the formalism of probability theory this is $P(d|H_o)$, where d denotes the outcome (data) that we actually observed. (In this text, uppercase P denotes probability in general, whereas lowercase p will denote a particular probability, to be explained later.)

From the calculated probability, we decide whether to accept or reject H_o (and hence $H_{o,gen}$). In Arbuthnot's case, he decided that the probability $P(d|H_o)$ was so tiny that $H_{o,gen}$ must be rejected. If H_o follows from $H_{o,gen}$, as in Arbuthnot's case, rejecting $H_{o,gen}$ on the ground that H_o has been rejected is logically an instance of the rule *modus tollens*.[3]

[2] What I have here called the "general null hypothesis" ($H_{o,gen}$) could in some cases more aptly be dubbed a *theory*, or at least be considered part of a theory, so that the distinction between $H_{o,gen}$ and H_o is a distinction, more or less, between a theory and a hypothesis. Here is not, however, the proper place for discussing the relation between theories and hypotheses.

[3] Also known as "denying the consequent". This rule says that from *if P, then Q* and *not Q* we conclude *not P*. In Arbuthnot's case, from *if* $H_{o,gen}$ *then* H_o and *not* H_o it follows that *not* $H_{o,gen}$.

Hence, we can list the steps in John Arbuthnot's test as follows:

1. Formulate a hypothesis to be tested, $H_{o,gen}$.
2. Model the hypothesis mathematically so that it is relevant to the data that will be or has been collected. This hypothesis is called H_o.
3. Calculate $P(d|H_o)$, where d is the data actually observed.
4. Judging from the value of $P(d|H_o)$, decide whether to keep or to reject H_o (and hence $H_{o,gen}$).

This list describes quite accurately what John Arbuthnot did. But a little reflection will show that the steps provided in the list are not clear enough to be used successfully in a variety of similar situations. In fact, all steps need clarification.

Problems in the outlined test procedure

First, we could ask: Will there always be a single way of reasonably modeling H_o from a formulated $H_{o,gen}$? The answer is no. An early example of statistical reasoning is that of natural philosopher John Michell (1724–1793), who calculated that if the visible stars were scattered at random over the sky, much more rarely would they form double stars and clusters than what we can actually observe. Michell reported this in the *Philosophical Transactions of the Royal Society of London* in 1767. French mathematician Joseph Bertrand (1822–1900) has commented on Michell's writings. Bertrand agreed that there should be no clusters such as the *Pleiades* under the hypothesis of random distribution of stars, but pointed to the vagueness of the concept of closeness:

> "In order to make precise this vague idea of closeness, should we search for the smallest circle that contains the group? The greatest of the angular distances? The sum of all distances squared? The area of the spherical polygon the vertices of which coincide with some of the stars and that contains the other stars in its interior? All these quantities are for the Pleiades cluster smaller than what is probable. Which of them gives the measure of improbability?" (Bertrand, 1889:171, my translation).[4]

We have to decide exactly which test to use. In other words, we must decide upon a specific *test statistic*. A test statistic is a function taking data (in suitable form) as input and giving as output a numerical value that can be used to perform a test.[5] In order for the output value to be useful, we need to know that it can be regarded as the outcome of a specified statistical probability function under H_o. We can denote the test statistic as $T(d)$, being a function of the data d. In Arbuthnot's case, the test statistic is simply $T(d) = d$, where d is understood as the number of boy years. In Arbuthnot's case, we know that $T(d) = d$ would be an outcome of the $\mathrm{Bin}(82, \frac{1}{2})$ distribution if H_o is true.

Secondly, even if H_o follows from $H_{o,gen}$, we must make sure there is no other relevant information speaking against H_o as a realistic hypothesis. A clear example where this

[4] "Faut-il, pour préciser cette idée vague de rapprochement, chercher le plus petit cercle qui contienne le groupe? la plus grande des distances angulaires? la somme des carrés de toutes les distances? l'aire du polygone sphérique dont quelques-unes des étoiles sont les sommets et qui contient les autres dans son intérieur? Toutes ces grandeurs, dans le groupe des Pléïades, sont plus petites qu'il n'est vraisemblable. Laquelle d'entre elles donnera la mesure de l'invraisemblance?"

[5] Normally, the test statistic will give just one output value for each combination of input data. However, it is possible to use test statistics giving a vector (multi-valued) output. A vector output may be needed for tests of hypotheses involving both magnitudes and orderings of data. We will not consider such cases in this text.

issue was ignored is an article by Chadwick & Jensen (1971), a study of dowsing (i.e. the purported ability to find objects with the aid of a handheld forked rod or a bent wire). Chadwick & Jensen made voluntary test persons walk with a dowsing rod along a straight test path, located between two long lines of fruit trees. In one spot along the path, a large iron object had been hidden underground. This object locally affected the earth's magnetic field slightly. The researchers' idea was that this magnetic disturbance would result in more hits near this spot. The result, however, showed no accumulation of hits near the hidden iron object. The authors' $H_{o,gen}$ was that the participants had no ability to detect variations in the earth's magntic field by dowsing. The authors' H_o was that when the test path was divided into small sections, the hits should be randomly distributed over these sections as measured by the appropriate χ^2 statistic.

But this step from $H_{o,gen}$ to H_o is very dubious. Even if dowsing would not give any information in addition to what is perceived with the normal senses, we should *not* expect dowsers to get hits completely at random. When walking with a dowsing rod along lines of fruit trees, almost anything could trigger a reaction in the dowser's hands, causing the rod to bend: a beautiful fruit on a nearby tree, a peculiar stone, a wasted sweet paper on the ground, etc. There is really no reason for us to assume that the hits would be distributed completely at random under these circumstances, and hence the χ^2 test adopted by the researchers is irrelevant.

The main moral of this story is, I believe, that living things might not behave exactly according to laws of chance, unless factors contributing to non-chance behavior are ruled out. If your hypothesis involves humans or animals, you are justified in assuming that they will behave according to a statistical hypothesis only if you can rule out the presence of such disturbing factors. Sometimes, this can be quite difficult in practice.

Thirdly, there is something quite unclear about Arbuthnot's argument when he concluded that H_o should be rejected. In his case, the actual data d were the most extreme that could be conceived: not in a single year during 82 years did the girls form a majority among newborns. What would Arbuthnot have said if his data showed that boys were in majority for, say, 70 years out of 82? The probability of getting 70 boy years out of 82 under H_o is 1.7×10^{-11}. This is a very small probability, though admittedly not as small as the probability for zero boy years out of 82. What about the probability of having, say, 34 boy years out of 82? It is 0.027. That, too, is a small probability, but I would guess that few would reject H_o after getting 34 as the outcome of a hypothesized $\text{Bin}(82, \frac{1}{2})$ distribution. In other words, getting 34 out of 82 seems quite plausible under H_o although the probability is no more than 0.027.

Forming a rejection set

Thinking more carefully about these figures, it seems unreasonable to base the conclusion whether to accept or to reject H_o solely on the probability $P(d|H_o)$, for this probability is dependent on the possible range of d, which in turn is dependent on the sample size in the test (here: the number of years checked). For example, the probability of having 34 boy years out of 82 years is 0.027, but the probability of having 0 boy years in 5 consecutive years is 0.031. Although the latter probability is greater, many of us would view the latter outcome as less supported by the null hypothesis of an even chance than the former outcome. In doing this, we probably reason that 34 out of 82 is not very far removed from the statistical expected value (which is 41). Although the exact outcome 34 out of

82 is not in itself very probable under H_0, there are several possible outcomes that are even less probable (for example, 33 out of 82, or 32 out of 82, not to mention 10 out of 82). In the other scenario, 0 out of 5 is as far away from the expected value you can get, and there are no other outcomes that are less probable.

The morale is that it seems reasonable to consider a wider class of possibilities than merely the actual outcome. In other words, we need to select a *rejection set* of outcomes. When the value calculated from the test statistic $T(d)$ is in the rejection set, we reject H_0. We may call the rejection set w and the total outcome space W. Intuitively, then, the rejection set should have the following properties:

A. Under H_0, the probability of getting a random outcome in w should be low. Thus, w should consist of only a small portion of the total outcome space W (provided that each outcome is equiprobable).
B. w should contain only outcomes that deviate considerably from the statistical expected value under H_0.
C. The elements of w should in some sense be close to one another; w should not be formed, at least not exclusively, by scattered and isolated outcomes in W.[6]

Property A is reasonable because we don't want to reject H_0 lightly. We would like to reject H_0 only when the result at hand is contained in a small (and hence improbable) set of all possible outcomes. The number of possible outcomes in w divided by the number of all possible outcomes W could be called the *size* of the test, but is more commonly known as the *nominal significance level*.[7] Here we shall denote it p_0. (I write "the number of *possible* outcomes" because all permutations that lead to the same test statistic value must be considered as separate possible outcomes. For example, in John Arbuthnot's test there are $\sum_{k=0}^{82} \binom{82}{k} \approx 4.8 \times 10^{24}$ ways of selecting n years out of 82 years, where n could be any integer from 0 to 82, and hence this is the number of possible outcomes in W. Many of these of course correspond to the same test statistic value, since there are only 83 of them: $0, 1, 2, \ldots, 82$.) Put briefly, the nominal significance level is the probability that a random outcome x under H_0 is in w, that is $p_0 = P(x \subset w|H_0)$.

Property B says that the values of $T(d)$ giving rise to a rejection of H_0 ought to be "extreme" given H_0, i.e. the values should not be too close to what could be expected under H_0. This is quite self-evident as long as we are testing a hypothesis with a test statistic yielding a single value. For example, in Arbuthnot's test the statistic $T(d)$ gave a single value, *viz.* an integer from the set $\{0, 1, 2, \ldots, 82\}$.[8] We should note that even when the test statistic gives a single value, it is often reasonable to say that H_0 could be rejected if this value is unusually small *or* unusually great. This means that there are two separate subsets of w, one corresponding to deviations from the expected value in one direction and another corresponding to deviations in the opposite direction. In such cases, we say that the test is *two-sided* (or *two-tailed*). Returning to the Arbuthnot example again, it would seem fully reasonable to take, e.g., the following rejection set:

[6] It is, however, possible to find an exception: if the probability distribution modelling H_0 is unimodal, a reasonable rejection set may contain two values only, one in each tail of the distribution. The rejection set will then be formed exclusively by isolated outcomes.

[7] It is recommended to use the adjective "nominal", in order to distinguish it from the calculated significance level p, which will be presented below.

[8] In more complicated cases, the test statistic might not yield a single value (see footnote 5), and it could be impossible to order the outcomes in any natural way. If this is the case, we are unable to tell whether a particular result is more extreme than another.

$w = \{d : T(d) \leq 30 \text{ or } T(d) \geq 52\}$. The rejection set thus consists of the data such that $T(d) \leq 30$ or $T(d) \geq 52$. These are two separate areas of data if we look at the situation from the perspective of the probability distribution function of $T(d)$ under H_0, as shown in figure 1.

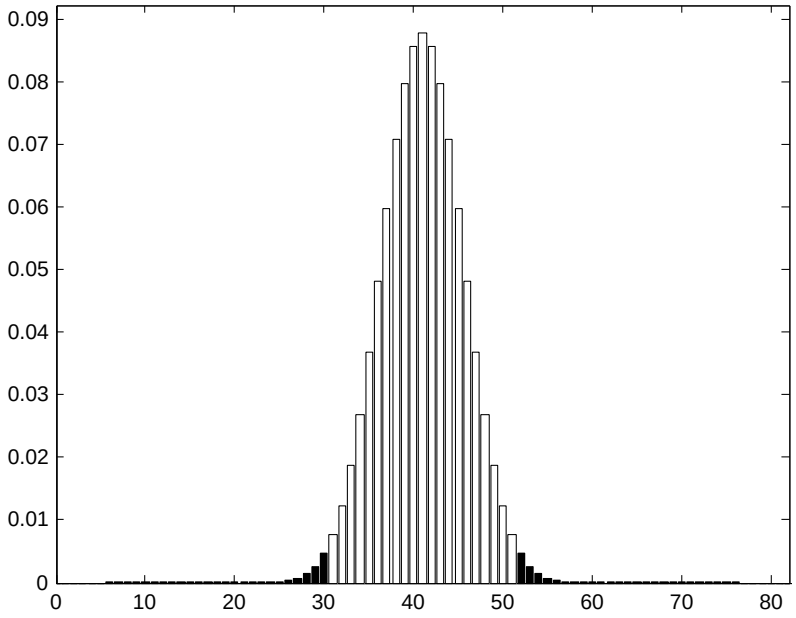

FIGURE 1 – The probability distribution function for $T(d)$ under H_0 in John Arbuthnot's test. For each integer along the x axis $(0, 1, 2, \ldots 82)$, the height of the bar shows the corresponding probability. The black bars correspond to a proposed rejection set (not the one actually used by Arbuthnot). The highest bar is the one for $d = 41$.

Speaking of deviations from the statistically expected value, it is useful to define D as the set of outcomes that includes the actual outcome d and all other outcomes that are less probable than d under H_0. The probability of having an outcome in D under H_0 is known as the *p value* or the *significance level*. In mathematical language, we can say that $p = P(D|H_0)$. Normally, this probability will correspond to two separate areas in a probability function graph for $T(d)$ under H_0, just as in figure 1, and will hence be relevant for a two-sided test. In one-sided tests, D must be defined as the set of outcomes that includes d and all other outcomes that are less probable than d under H_0 *and which deviate from the expected value under H_0 in the same direction as d* (the added clause for a one-sided test is in italics).

Property C, finally, is reasonable because there is something quite arbitrary with a rejection set that contains outcomes many of which are not close to the others. For example, let us roll a six-sided die three times. There are $6^3 = 216$ possible outcomes. We would like to test whether the die is fair or not. Suppose that the outcome of each toss is called d_i $(i = 1, 2, 3)$ and that we use the sum of the outcomes as our test statistic: $T(d) = \sum_i d_i$. Also suppose that we choose as our rejection set the outcomes $d = \{d_1, d_2, d_3\}$

being in

$$w = \{d : T(d) = 16 \text{ or } T(d) = 18\}.$$

In other words, w contains those outcomes d_1, d_2, and d_3 such that their sum is 16 or 18. Since there are 6 combinations giving the sum $T(d) = 16$ and 1 combination giving $T(d) = 18$, and since W contains a total number of 216 possible outcomes (i.e. 216 combinations $< d_1, d_2, d_3 >$ where any d_i may take the value $1, 2, \ldots, 6$), the nominal significance level is $p_o = \frac{7}{216} \approx 0.0324$. This value is not great, and hence property A is fulfilled. Furthermore, the values of $T(d)$ included in w deviate considerably from the expected value of $T(d)$ under H_o, which is $T(d) = 10.5$. Thus, property B is fulfilled as well. But property C is not fulfilled, since there is no explanation for why $T(d) = 17$ has been excluded from w. If $T(d) = 16$ and $T(d) = 18$ are both found to be fitting for inclusion in w, a very good reason should be presented for not including $T(d) = 17$ as well, but there seems to be no such reason. So w has a strange composition, unacceptable according to requirement C. (Another thing that could be discussed in this particular case is: why are only large values of $T(d)$ included? Wouldn't we suspect that the die is unfair also in the case of obtaining a very low value for $T(d)$, for example $T(d) = 3$?)

Fisherian hypothesis testing

Taking the above considerations into account, we can now make a new list of test steps, more elaborate than the previous one:

1. Formulate a hypothesis to be tested, preferably of *null* type, called $H_{o,gen}$. (We will return to the issue of what is meant by the term "null").
2. Model $H_{o,gen}$ mathematically so that it is relevant to the data d that will be collected. This mathematically modeled hypothesis is called H_o and should follow with the aid of logic and rational reasoning from $H_{o,gen}$. Also, make sure there is no other relevant information speaking against H_o as a valid null hypothesis for all possible outcomes. State the test statistic $T(d)$.
3. For $T(d)$, decide upon a rejection set w in the outcome space W, thereby establishing the nominal significance level (p_o).
4. Collect your data d and check whether $T(d)$ is in w or not. You should also calculate the p value of your data, i.e. $P(D|H_o)$, where D is the set of outcomes that includes d and all other outcomes that are less probable than d under H_o (and, for a one-sided test with a unimodal modelling of H_o, which deviate from H_o in the same direction as d).
5. If $T(d)$ is in w, or equivalently, if $p < p_o$, reject H_o and hence $H_{o,gen}$, otherwise accept $H_{o,gen}$. State your p value, if you calculated it in step 4.

A well-known advocate of this kind of statistical hypothesis testing was Ronald A. Fisher (1890–1962), who developed its theory in the 1920's and 1930's and disseminated his ideas of statistical hypothesis testing in highly influential textbooks that were issued in several editions until the 1950's.[9] We may therefore call the above list of steps a recipe

[9]Of course, there were forerunners to Fisher in various respects. For example, the American philosopher Charles S. Peirce (1839–1914) discussed a situation similar to a Fisherian hypothesis testing in 1878. A summary of Peirce's argument runs as follows (Peirce, 1878). In the 1870 US census the proportion of boys among white children aged under 1 year was 0.5082. The corresponding figure for colored toddlers was 0.4977. Could this difference be explained by chance? Comparing with a situation in which s balls are drawn from an urn with the

Face	Occurrences
1	23
2	18
3	15
4	21
5	26
6	17

TABLE 1 – Results of rolling a die 120 times.

for Fisherian hypothesis testing.[10]

Two examples

I will now give two examples, close to standard statistics textbook examples, in order to show how the conceptual apparatus presented in the list above can be applied in practice.

1. Testing the fairness of a die

We throw a six-sided die 120 times to test whether it is fair or not (i.e. whether the probability is $\frac{1}{6}$ for all faces or not). Our $H_{o,gen}$ is that the die is fair. Our H_o is that the data collected from 120 rolls will conform to the χ^2 distribution (with 5 degrees of freedom). The test statistic is $T(d) = \chi^2(d) = \sum_{i=1}^{n} \frac{(d_i - e_i)^2}{e_i}$, where d_i is the observed number of rolls showing the face i upwards ($i = 1, 2, 3, 4, 5, 6$), e_i is the expected number, and n is the number of faces (i.e. $n = 6$). In our case, $e_i = 20$ for all faces. As the rejection set w we select those values of the χ^2 distribution that form the top 5 % values. This is reasonable since larger deviations from the expected occurrences under H_o will yield a higher value of $\chi^2(d)$. From statistical tables we find that in our case we should reject H_o if the test statistic value $\chi^2(d)$ is greater than 11.07. In other words, the rejection set is

$$w = \{d : T(d) > 11.07\}.$$

It turns out that our actual data are as shown in table 1. We calculate $\chi^2(d)$ and get

$$\begin{aligned} \chi^2(d) &= \frac{3^2}{20} + \frac{(-2)^2}{20} + \frac{(-5)^2}{20} + \frac{1^2}{20} + \frac{6^2}{20} + \frac{(-3)^2}{20} \\ &= 4.2. \end{aligned}$$

Since the calculated $\chi^2(d)$ is less than the value required to reject H_o, we conclude that H_o can be kept. Thus, the data do not urge us to suspect that the die is unfair.

true proportion p white balls, Pierce found that 99 times out of 100 the error in the proportion of the sample will be no greater than $1.821\sqrt{2p(1-p)/s}$, and 9 999 999 999 times out of 10 000 000 000 no greater than $4.77\sqrt{2p(1-p)/s}$. Using $p = \frac{1}{2}$ Pierce found, with $s_{white} = 1\,000\,000$ and $s_{colored} = 150\,000$, that a combined error as great as that really observed should be found by chance only once in 10 000 000 000 censuses. Hence the observed difference is very unlikely to be due to chance. Although Peirce did not use the wording "rejection of a hypothesis", it may well be how he thought about the situation.

[10] The exact wording does not follow Fisher. Rather, the list is an interpretation of how he might have argued had he been asked to provide a list of test steps.

2. Comparing body lengths in two male populations

As a second example, consider two groups of adult males. We wish to find out whether body lengths are equal or different in the two groups. But the groups are so large that we have to base our verdict on information not from the entire populations but from smaller samples. From the first group, called A, we randomly draw $n_A = 25$ males and find that their mean length is $\bar{d}_A = 182.2$ cm with the standard deviation $s_A = 6.6$ cm. From the second group, B, we pick $n_B = 27$ males by random and find their mean body length to be $\bar{d}_B = 178.5$ cm with standard deviation $s_B = 6.9$ cm. If we assume that the individual body lengths are outcomes of normal distributions with a common standard deviation, a fitting H_0 is that our data $d = \{n_A, \bar{d}_A, s_A, n_B, \bar{d}_B, s_B\}$ will conform to the so-called t distribution with $n_A + n_B - 2$ degrees of freedom if combined according to the test statistic formula

$$t(d) = \frac{\bar{d}_A - \bar{d}_B}{\sqrt{\frac{(n_A-1)s_A^2+(n_B-1)s_B^2}{n_A+n_B-2}\left(\frac{1}{n_A}+\frac{1}{n_B}\right)}}.$$

We know that this test statistic behaves according to the t distribution under H_0. The t distribution is symmetrical around zero. In order to perform a two-sided test – meaning that if the difference between $\bar{d}_A$ and $\bar{d}_B$ is great enough, it may lead to a rejection of H_0 irrespective of whether $\bar{d}_A$ is greater than $\bar{d}_B$ or vice versa – we take as our rejection set w the top 2.5 % and the bottom 2.5 % of the values of the t distribution, in our case with $n_A + n_B - 2 = 50$ degrees of freedom. From statistical tables we find that if $|t(d)|$ exceeds 1.676 we should reject H_0. In other words, our rejection set is

$$w = \{d\colon |t(d)| > 1.676\}$$

Using our empirical data, we get

$$\begin{aligned} t(d) &= \frac{182.2 - 178.5}{\sqrt{\frac{24\cdot(6.6)^2+26\cdot(6.9)^2}{50}\left(\frac{1}{25}+\frac{1}{27}\right)}} \\ &\approx 1.973. \end{aligned}$$

This value exceeds 1.676. Our conclusion must be that we reject H_0. There is thus a significant ($p < 0.05$) difference in body lengths between the two populations, according to the evidence from our samples.

Drawbacks and criticisms

Even though the Fisherian list of test steps is better than the previous ones — and fully functioning, judging from the two worked examples given above — there is still considerable room for critical questions.

The logic of rejection and the credibility of hypotheses

For example, we could ask: What is the best way of formulating a null hypothesis? Why a null hypothesis in the first place? We shall, however, postpone these questions until a later section (starting on page 20).

Instead, let us start with this question: What is the rationale and logic of an inference using the Fisherian list? Fisher himself, discussing John Michell's calculations showing that the visible stars were not dispersed at random, wrote:

> "The force with which such a conclusion is supported is logically that of a simple disjunction: *Either* an exceptionally rare chance has occurred, *or* the theory of random distribution is not true" (Fisher, 1956:39).

So, says Fisher, when we reject H_0 (in this case a hypothesis of random distribution), we do this because either it is false, or it is true but an "exceptionally rare" chance has occurred.[11] Reasonable as it may seem, however, Fisher's assertion is not entirely correct. Even if we say that we test a single hypothesis, we are not testing it in isolation. When we say that we test a single hypothesis H_0, in reality this means that we test H_0 alongside a set of auxiliary hypotheses $H_{A,1}, H_{A,2}, \ldots, H_{A,k}$. When the test tells us to reject H_0, we should really interpret this as an exhortation to reject at least one of the hypotheses $H_0, H_{A,1}, H_{A,2}, \ldots, H_{A,k}$. There are always auxiliary hypotheses involved, for example concerning the proper functioning of apparatus or the comparability of groups at baseline.[12] Since we have not modeled these auxiliary hypotheses mathematically, we cannot say for sure that the situation of rejecting H_0 is accurately described by Fisher's disjunction.

If we reject H_0 (and hence, normally, $H_{0,\text{gen}}$), the test cannot tell us which hypothesis to adopt instead of the rejected one. In other words, by rejecting H_0 we have not gained support for any other, specific hypothesis. The explanation for this is that H_0 could be wrong for so many different reasons. But since the test was concerned only with the accepting or rejection of H_0 (the only hypothesis mathematically modeled), the test says nothing about the credibility of other hypotheses.

This limitation of testing a single hypothesis is often misunderstood, sometimes even by statisticians. In a well-known Swedish textbook in statistics, a parapsychological experiment is taken as an example of hypothesis testing: a person claims to be able to tell whether heads or tails will turn up in coin tossing. This, the claimant says, is done through extrasensory perception (ESP). $H_{0,\text{gen}}$ is that the claimant does not have an ESP ability. For a series of 12 tossings, H_0 is that the number of correct answers comes from the $\text{Bin}(12, \frac{1}{2})$ distribution. A calculation shows that the null hypothesis should be rejected at the 5 % significance level if the claimant is right in 10 cases or more.[13] The book states:

> "Hence, if the person answers correctly 10 times or more, we should say that he has an ESP ability, but not so if he doesn't" (Blom et al., 2005:321, my translation).[14]

This is plain wrong. If the person answers correctly in 10 trials or more out of 12, we are justified in rejecting H_0 at the 5 % significance level, but we are not automatically permitted to endorse the alternative hypothesis of an ESP ability, for there are other

[11] One could question that the wording "exceptionally rare" is appropriate. A nominal significance level of 0.05 is often used. Personally, I would not judge a result that could appear by chance with a probability of 0.05 as "exceptionally rare".

[12] Cf the final section, starting on page 24.

[13] The corresponding p value is 0.0193. Being right in only 9 cases or more corresponds to $p = 0.0730$, exceeding the stipulated 5 % significance level.

[14] Original text: "Om personen svarar rätt minst 10 gånger bör man alltså påstå att han har ESP, men inte annars."

alternative hypotheses that are fully compatible with the result (e.g. the test subject is cheating, or the coin is not a proper randomizer, or the test person simply was lucky) and the test does not urge us to select any particular alternative. A test involving a single hypothesis H_o cannot assist us in selecting which alternative hypothesis is the best; it cannot even say whether any particular alternative hypothesis (e.g. the claimant is cheating) is more probable than H_o, nor whether different alternative hypotheses are more likely than one another (e.g., the test subject is cheating *vs.* the coin is not a proper randomizer). Only a test involving at least two hypotheses, all of which are statistically modeled, can tell us that one hypothesis is more likely than another, given the outcome.

Even if we do *not* reject H_o (and $H_{o,gen}$), there are critical questions to be asked. If we should retain H_o according to the test, we still do not know how much faith we should have in it, for no weight is given to our initial belief or disbelief in H_o and $H_{o,gen}$. According to the list of test steps, the test should be performed identically irrespective of whether the tested hypothesis is judged to be very probable (e.g., no paranormal powers exist) or as highly unlikely (e.g., this drug has no effect on a given disease). It might be felt that a good test should take such assessments into account. However, this is not possible within the test paradigm discussed so far.

Whether we accept or reject H_o (and $H_{o,gen}$), we are asked to compute the p value, if possible. In all kinds of research papers involving statistics, p values are given and discussed. But it can be doubted that the p value is important. A very common misunderstanding concerning p is that it denotes the probability of a hypothesis. As explained above, this is not right. The p value is about $P(D|H_o)$, that is the probability of a set of outcomes D given the null hypothesis H_o. It is thus about the probability of certain *outcomes*, not the probability of any hypothesis. Although Fisher himself did not mix up probabilities of outcomes with probabilities of hypotheses, he made other disputable statements on the proper interpretation of p values:

> "When a prediction is made, having a known low degree of probability, such as that a particular throw with four dice shall show four sixes, an event known to have a mathematical probability, in the strict sense, of 1 in 1296, the same reluctance will be felt towards accepting this assertion, and for just the same reason, indeed, that a similar reluctance is shown to accepting a hypothesis rejected at this level of significance" (Fisher 1956:43).

In the case of the dice-throwing prediction, there is (according to Fisher's example) a *known* probability of having a certain outcome. In other words, we *know* the probability distribution that governs the behaviour of the four dice. In the case of the rejection of a hypothesis H_o, we do *not know* whether the hypothesis is true or not, or else there would be no point in testing H_o in a formal hypothesis test. This important difference is ignored in Fisher's statement.

When people are mistaken as to what p means, they often seem to think that p denotes $P(H_o|d)$, i.e. the probability that H_o (and hence, we assume, $H_{o,gen}$) is true given the actual data d. And we can all agree that it would be very interesting to estimate this probability. That probability, however, is not obtainable within the test framework presented so far. The only feasible way of obtaining $P(H_o|d)$ would be to use Bayes's theorem, which requires our initial belief in H_o to be stated as a prior probability $p(H_o)$. This method will be discussed in the section starting on page 22.

A general criticism against the use of significance levels is that it dichotomizes a continuous scale: a result is either significant or not; the p value is either lower than

the predetermined nominal significance level or it is not. Results that are statistically significant are normally given much greater attention than non-significant results. But the exact nominal significance level is of course arbitrarily set. A result corresponding to $p = 0.06$ may be very interesting although it is not labeled "significant". And vice versa: statistical significance is not the same as material or practical significance (or, in medical settings, clinical significance). That a result is statistically significant means that there is a deviation from the null hypothesis, but such a deviation may be very small and still statistically significant if the sample size is large. In fact, any departure from null will be significant at a pre-assigned p_0 level provided the sample size is large enough. It could therefore be argued that significant results with small sample sizes are more interesting than significant results from tests with large sample sizes, for significant results from small samples will more often correspond to a larger effect. On the other hand, any evidence against the null generally carries greater weight as the sample size increases (Rosenkrantz, 1973:314).

Things that happened or that might have happened

Still other criticisms can be directed against the test paradigm presented so far. These criticisms are related to questions regarding what happened in the test as opposed to what might have happened.

A quite general criticism in this vein is the following. When we decide whether to accept or reject H_0, we do this by checking whether the empirical data d is in the rejection set w or not. (Alternatively, we decide upon the fate of H_0 by calculating $p = P(D|H_0)$, an operation exactly equivalent to checking whether d is in w or not, provided that w has been constructed according to requirements A–C presented above.) Suppose that d is indeed in w, but that d is among the more moderate – i.e., not extreme – outcomes included in w. We then reject H_0. We may ask why the more extreme outcomes in w, outcomes that actually did not occur, should be part of our argument for rejecting H_0. Or, using the alternative p value calculation procedure, we may ask why outcomes that did not occur appear in the calculation of p. Since these outcomes did not occur, how come they play such an important role in the hypothesis testing? Wouldn't it be more suitable to perform a test that takes into account only the actual data, not hypothetical outcomes that did not occur and that are possibilities according to a hypothesis we might not even believe in?

This criticism questions the soundness of forming a region of rejection in the first place. In a well-known example due to Fisher, a lady says that by drinking tea with milk she will be able tell whether the milk or the tea infusion was added first to the cup. The lady is put to test and asked to taste the tea from eight cups ordered randomly, where tea was added first in four cups and milk was added first in the remaining four. The lady is asked to divide the eight cups into two sets of four cups each, hopefully agreeing with the treatments they have received. There are 70 ways of choosing four objects out of eight, disregarding the ordering. (Generally, there are $\frac{n!}{(n-k)!k!}$ ways of selecting k objects out of n if the order is disregarded.) What if the lady picks three correct cups and one wrong? Fisher writes:

> "In the present instance '3 rights and 1 wrong' occurs 16 times, and '4 right' occurs once in 70 trials, making 17 cases out of 70 as good or better than that observed. The reason for including cases better than that observed becomes obvious on considering

> what our conclusions would have been had the case of 3 right and 1 wrong only 1 chance, and the case of 4 right 16 chances of occurrence out of 70. The rare case of 3 right and 1 wrong could not be judged significant merely because it was rare, seeing that a higher degree of success would frequently have been scored by mere chance" (Fisher, 1951:15).

Fisher seems to cover what I have called properties A and B for a good rejection set, but it is doubtful whether he also covers C. Whether Fisher's argument is good enough is a question about which each reader may form an opinion.

As another example of a possible criticism arising from reflections on what might have happened, consider the following story.[15] Suppose that we wish to test whether the average IQ score in a particular population is higher than the general population average or not. From the particular population we select five people at random. Our general null hypothesis $H_{0,\text{gen}}$ is that there is no difference between the particular population and the general population. For the general population we assume that the individual IQ's are outcomes of the $\mathcal{N}(100, 15^2)$ distribution, i.e. a normal (Gaussian) distribution with expected value 100 and variance 15^2 (this is known from large IQ tests previously performed). A suitable H_0, then, is that the mean value for our population of five people is an outcome of the $\mathcal{N}(100, 15^2/5)$ distribution. Our test statistic is the arithmetic mean $T(d) = \bar{d} = \left(\sum_{i=1}^{5} d_i\right)/5$. With $p_0 = 0.05$ and a one-sided test (remember, we were only interested in whether the particular population had a *higher* average IQ than the general population) we find that we should reject H_0 if the test statistic $\bar{d}$ (the arithmetic mean) in our group of five people is larger than 111.0. The rejection set w is then the set of ordered individual IQ values $< d_1, d_2, \ldots, d_5 >$ such that their arithmetic mean $\bar{d}$ is larger than 111.0:

$$w = \{< d_1, d_2, \ldots, d_5 > : T(d) > 111.0\}.$$

Suppose that we obtain the empirical test statistic value $\bar{d} = 115$. This leads to a rejection of H_0. So far, so good.

Now suppose we receive a letter from the company from which we bought the tests and grading equipment, stating: "We have found out that our computer program was partly malfunctioning on the day of your test. Any mean score $\bar{d}$ below 100 was reported as 100. For scores $\bar{d}$ above 100 the program produced the correct results." It may seem that we need not worry, since we obtained a mean score $\bar{d}$ above 100. But this is a potential matter of dispute, for the letter indicates that our null hypothesis H_0 is inaccurate: it is *not* true that we should have expected $\bar{d}$ to be an outcome of a $\mathcal{N}(100, 15^2/5)$ distribution under $H_{0,\text{gen}}$. According to the $\mathcal{N}(100, 15^2/5)$ distribution, values below 100 are fully possible, but according to the information from the company, values below 100 were in fact impossible on the day of the test.

So, you might say, we used an erroneous H_0 and hence the test is invalid. Or you might say that the test is still valid, because the new information doesn't matter.

The question is to what extent what *might have happened* but actually *did not happen* casts doubt upon the test procedure. We *actually* got the result $\bar{d} = 115$ and there is no reason to suspect that it is wrong. Also, we know from numerous earlier studies that individual IQ measurements in the general population can be regarded as outcomes of a $\mathcal{N}(100, 15^2)$ distribution and hence that a fitting H_0 for testing the mean value in a group of five people is that this value comes from a $\mathcal{N}(100, 15^2/5)$ distribution. The letter from

[15] The scenario has been inspired by Efron (1978:236f).

the company did not change any of these facts. Still, the letter from the company tells us that *if* we had got a result $\bar{d} = 100$ from the computer program it would have been irrelevant to compare this figure with a hypothesized $\mathcal{N}(100, 15^2/5)$ distribution. The letter tells us that the $\mathcal{N}(100, 15^2/5)$ distribution is not an appropriate H_0 instantiation for all possible outcomes that $\bar{d}$ could assume. The question is: does it matter, now that we actually got $\bar{d} = 115$?

It might be argued that it doesn't, for the following reason: If we had known about the malfunctioning computer program before the test, we could have adjusted H_0 accordingly. We could have assumed that the $\mathcal{N}(100, 15^2/5)$ distribution was correct only for values larger than 100. This would have lead to a rejection set w identical to the rejection set actually used. Since w and the p_0 associated with it would remain unaffected, we conclude that it doesn't matter. This seems reasonable.

Now let us imagine a slightly different story. Instead of getting $\bar{d} = 115$, assume we got $\bar{d} = 100$ and then received the letter from the company as above. With $\bar{d} = 100$ we would accept H_0. Again, the letter would not change w nor p_0, and we can be certain that the result we got ought not lead to a rejection of H_0. So again we could argue that the test in the narrowest sense – i.e, whether to accept or reject H_0 – is as certain as it could be. The difference is: this time we cannot trust the result $\bar{d} = 100$. Maybe the real mean value is less than 100. This also means that we cannot compute the exact p value for the actual result.

Or consider another alteration to the original story. Instead of a one-sided test, we perform a two-sided test with $p_0 = 0.05$. Calculations show that for such a test, H_0 is rejected if $\bar{d} < 86.9$ or $\bar{d} > 113.1$. Assume that we get the empirical result $\bar{d} = 115$ and that we receive the letter from the company as above. Our result is again clearly in the rejection set w and is unaffected by the information from the company. But this time, half of the set w is affected: it is simply impossible to reject H_0 by getting a $\bar{d}$ value lower than 86.9. Does this invalidate the test as a whole?

A final version of the story might be the following. We perform a one-sided or two-sided test as above, we get $\bar{d} = 115$ and then receive a letter from the company stating: "On the day of your test, results above $\bar{d} = 115$ were misrepresented; unfortunately, we don't know exactly how. Results up to and including $\bar{d} = 115$ are correct." Although the result we actually got is correct, we may hesitate to accept the test, for we no longer know the exact size of w.

At which point in the different variants of the story recounted above does the test become invalid (if at all)? It is not trivial to give an answer, and we will not pursue the question here. One thing should be noted, though. Even when the result we actually got remained unaffacted by the faulty program, the letter from the company could influence the validity of the test. How is that possible? The answer is (arguably) that the whole outcome space is relevant to the interpretation of the test. Although the faulty program affects only outcomes that did not occur, this is relevant to the interpretation of the whole test.

Temporal aspects

There is also an interesting temporal aspect in the IQ story: could information that we learn *after* the test invalidate it? On a very general level, the answer is yes, of course. We could learn for example that we made a miscalculation. That would invalidate the

test. But in the case of the IQ test, we have learned (in all variants of the story) that H_o was not an appropriate hypothesis for all possible outcomes. We have viewed this as a genuine problem, though of course we could have "solved" it by simply changing H_o after the collection of data. Changing H_o after the collection of data is not, however, seen as appropriate in the test scheme under consideration. One important reason for this is that the actual data may influence the choice of hypothesis, opening up possibilities of cheating by picking a hypothesis already known to fit the data.

This is also related to the topic of optional stopping, i.e. the test is terminated at a point not decided in advance. In fact, if we allow tests where the number of trials has not been decided in advance, it is fully possible for a given set of data to be judged by very different hypotheses. Suppose we would like to test a hypothesis regarding the probability that a coin lands heads. We could imagine two different plans for performing the experiment (Mayo, 1981:185). (1) We decide in advance to keep tossing the coin until 10 heads are obtained. It turns out that we need to toss the coin 25 times. (2) We decide in advance that we should toss the coin 25 times. It turns out that we get 10 heads. In (1), we must formulate a H_o for the distribution of the number of tosses required. In (2), H_o must specify the statistical behavior of the number of heads, provided that we toss the coin 25 times. These are different hypotheses. Hence, the result may lead to different conclusions.[16] One might therefore ask critically: Should not a given result always lead to the same conclusion irrespective of how the hypothesis has been phrased?

Summary of criticisms

We have discussed several problems, or potential problems, with the hypothesis testing model used so far. These can be summarized as follows:

- When we reject a null hypothesis, the test cannot tell us which alternative hypothesis is more credible.
- Even when we retain the null hypothesis, the test does not tell us how much credibility we should ascribe to it.
- The test gives us a value of $P(D|H_o)$, also known as the p value, but this tells us very little. An arguably more interesting figure would have been $P(H_o|d)$.
- Is the justification offered so far for the selection of w good enough?
- The test takes into account hypothetical outcomes that did not occur, and could be sensible to issues of timing and new information that should not, it might be felt, have any influence.

Some of these criticisms are quite sweeping. Perhaps a few of them should be dismissed simply on the ground that trying to meet them would ruin the possibilities of doing meaningful tests. In other words, they might be the price to be paid in order to perform reasonable tests at all. This text is too short to discuss all the problems above in detail. But two more models for statistical hypothesis testing will be presented briefly, the Neyman–Pearson model and the Bayesian approach. They deal with various aspects of the encountered problems.

[16] Cf. the discussion starting on page 20 about nulls and how a given set of data can support different and even inconsistent nulls.

The Neyman–Pearson contribution

Alongside Ronald Fisher, the most important figures in the recent history of statistical hypothesis testing are Jerzy Neyman (1894–1981) and Egon S. Pearson (1895–1980). They co-authored a number of statistical papers in which they tried to solve some (by no means all) of the problems listed above.

So far, all tests discussed have concerned only one explicit (mathematically modeled) hypothesis each. This hypothesis has been called H_0. (In addition, we have noted that there may be auxiliary hypotheses involved, but these have not been formulated or modeled explicitly.) According to Ronald Fisher, a single hypothesis is all you need in order to perform a statistical test (Hacking 1965:82f). The Neyman–Pearson (N–P) theory emphasizes the need for alternative hypotheses.

A clear description of N–P theory can be found in the introductory section of Neyman & Pearson (1936). Condensed into one sentence, the theory says: "Arrange your test so as to minimize the probability of error" (Neyman & Pearson, 1967:203).

Neyman and Pearson note that there are two kinds of error that are relevant in tests of a hypothesis H_0 (against an alternative hypothesis H_1). The error of the first kind or *type I error* is to reject H_0 although H_0 is true. The error of the second kind or *type II error* is to accept H_0 although H_0 is false. These errors, Neyman and Pearson say, will normally lead to very different consequences and should therefore be distinguished.

For example, assume that a manufacturer of lamps wishes to test whether a batch of 1500 lamps adhere to the conditions laid down in a written specification regarding their initial efficiency as light producers. The manufacturer measures the initial efficiency in an appropriate way for a few lamps selected at random from the batch, assuming that the tested lamps will represent the whole population of lamps. If the manufacturer concludes from the test that the lamps are OK, but a high proportion is in fact faulty, the manufacturer's reputation may suffer from allowing a faulty batch of lamps to go on the market. On the other hand, concluding from the test that the batch of lamps is not good enough, when in fact it *is* good enough, will result in higher expenses (from the destruction of good lamps or from the arrangement of further tests). We see that the consequences of the two possible errors are quite different from one another.

Formally, Neyman and Pearson proceed like this: Let $x_1, x_2, \ldots, x_n$ be a system of variables the values of which can be found through observation. A system of actual values $x_1, x_2, \ldots, x_n$ can be represented by a point E in the n-dimensional space W, which is called the sample space.[17] E is called the sample point. The variables $x_1, x_2, \ldots, x_n$ are random variables if for every subset w in W there is a number $P(E \in w)$ that represents the probability that E is in w. We assume that a function $p(x_1, \ldots, x_n) = p(E)$ is defined in W such that the value of $P(E \in w)$ can be found by calculating the integral $\int_w p(E)$. A test of the hypothesis H_0 can be seen as equivalent to rejecting H_0 when E falls within a specified set w, called the *critical region*. The ratio w/W is called α; i.e. the probability is α that a sample point E picked randomly from W is in w. α is identical to the type I error.

The type I error level is determined in advance by the researcher. Of course, it should be quite low. The type II error (the probability of accepting H_0 although an alternative hypothesis H_1 is true) may be called β. This error should be kept low as well, but the null hypothesis H_0 is normally selected so that it is more important to keep α low than to

[17] In the list of Fisherian test steps, I called this the "outcome space".

keep β low. The quantity $1 - \beta$ is the probability of correctly rejecting H_0 and is called the *power* of the test. A good test, Neyman and Pearson say, is one in which the power is maximized for a given level of α.

Testing H_0 against a two-sided, non-specific alternative hypothesis H_1 (i.e., the expected value under H_1 may be less than *or* greater than the expected value under H_0), will lead to two-sided tests. On the other hand, testing two specific hypotheses against each other, or a specific H_0 against a one-sided H_1, will normally lead to one-sided tests with respect to H_0, for the only important difference between the hypotheses will usually be their expected values. If, for example, the expected value under H_0 is less than the expected value under H_1, only values in the upper tail of the probability density distribution for H_0 will be included in the rejection set w. Although outcomes falling into the lower tail of H_0 are just as unlikely as values in the upper tail under H_0, they are even less likely under H_1 (unless the H_1 distribution has a much larger variance than the H_0 distribution) and will therefore not lead to the rejection of H_0.

When we discussed the formation of a rejection set under H_0, we applied conditions A–C (above, page 6). In fact, these conditions can be subsumed under a common argument, according to N–P theory, namely the argument of minimizing error (and maximizing power). So this is one problem that N–P theory can claim to be able to solve, out of those listed on page 16.

By allowing two hypotheses to be mathematically modeled and tested against each other, N–P theory is also in a position to recommend another specific hypotheses when the null hypothesis (H_0) is being rejected. This, too, can be seen as an answer to one of the problems listed on page 16. However, any intelligent proponent of N–P theory must admit that H_0 and the alternative H_1 may not be equally likely from an informed point of view. For example, H_0 may be the hypothesis that a person will correctly guess which side of a die will show up with probability $\frac{1}{6}$ (i.e., no paranormal ability), whereas H_1 may be that the probability of being right is $\frac{1}{2}$ in each case, which would amount to a considerable paranormal ability (or cheating, or a badly performed experiment). When we get a result that tells us to reject H_0 we may still hesitate to adopt H_1, given our previous knowledge of the (non-)existence of paranormal powers. From this perspective, the problem of selecting the most credible hypothesis remains largely unsolved even within the N–P paradigm. But when there is no *a priori* aversion towards an alternative hypothesis, N–P theory can do the job of selecting the hypothesis best supported by data.

Fisher vs. Neyman–Pearson

There are great similarities between Fisher's and Neyman–Pearson's theories of testing. The N–P theory is perhaps a little clearer when you read the original articles. This is mainly due to Fisher's habit of teaching by example rather than by laying down firm principles. Someone has noted that to Fisher, *significance* was the most central term; indeed, he coined the phrase "test of significance". To Neyman and Pearson, it has been argued that *hypothesis* was the key concept. But this difference is of course not very informative.

To appreciate philosophical differences, however, we need look no further than the probabilities denoted α and p, respectively. Often the quantity that I have denoted p_0 in connection with Fisher is called α in statistical textbooks, thereby intermingling Fisherian and N–P concepts. Indeed, the p_0 denotation is my own invention, and many textbooks

just say that the p value is to be calculated and compared with the predetermined error level α. But it could be argued that a Neyman–Pearson type I error probability α is *not* the same entity as Fisher's p value. The former is a specified error probability, the latter is the probability of obtaining a result at least as extreme as the one actually obtained, given H_o. They are not identical in terms of their meaning and philosophical justification. The p value is intended to measure evidence (or rather, evidence against a hypothesis) and is calculated after the collection of data. It alone is sufficient in order to come to a decision in the Fisherian paradigm: if p is small enough, we reject H_o. The quantity α, on the other hand, is a pre-determined (and hence not measured) error level which must be understood in a frequentist sense: if H_o is true and the test is imagined to be repeated indefinitely with data drawn from the same population, then we will reject H_o incorrectly with probability less than or equal to α in the long run. But to Fisher, the p value was to be understood exclusively as a measure of evidence against a hypothesis (namely, H_o). He explicitly repudiated any frequentist interpretations of p:

> "On the whole the ideas (*a*) that a test of significance must be regarded as one of a series of similar tests applied to a succession of similar bodies of data, and (*b*) that the purpose of the test is to discriminate or 'decide' between two or more hypotheses, have greatly obscured their understanding, when taken not as contingent possibilities but as elements essential to their logic. (…) Though recognizable as a psychological condition of reluctance, or resistance to the acceptance of a proposition, the feeling induced by a test of significance has an objective basis in that the probability statement on which it is based is in fact communicable to, and verifiable by, other rational minds. The level of significance in such cases fulfils the conditions of a measure of the rational grounds for the disbelief [in H_o] it engenders. It is more primitive, or elemental than, and does not justify, any exact probability statement about the proposition" (Fisher, 1956:42f).

Also, in the N–P paradigm α is mentioned in connection with the other possible error probability, β. But according to Fisher, there is no need for hypotheses other than H_o, and therefore β has no meaning in his theory.

In terms of measurement of credibility, Fisher preferred $P(d|H)$ or $P(D|H)$. The latter is the significance. The former he called "likelihood", thereby using a word that up until then had been considered an exact synonym for "probability" (Halldén 2003:126). Neyman–Pearson thought that the fraction $\frac{P(D|H)}{P(D|\neg H)}$ was more interesting than just $P(D|H)$ (Neyman & Pearson, 1928).

The differences between Fisher on the one hand and Neyman and Pearson one the other have been assessed in diverse ways by different authors. There is no doubt that the philosophical underpinnings (and implications) differ. To what extent statisticians should care about this is a matter of debate. Hubbard & Bayarri (2003) express annoyance that Fisherian and N–P concepts are often intermingled without philosophical reflection. By contrast, Lehmann (1993) emphasizes the similarities and argues that from a practical standpoint the theories are complementary rather than contradictory. (For a fascinating historical account of Fisher's and Neyman's work, see Lehmann, 2011).

The role of the null

In this text, we have mentioned several times that hypotheses being tested are often of *null* type. But we have not explained what this means, though the reader might have formed some ideas about it from the presented examples.

What should be clear is that there seems to be an asymmetry between the null hypothesis (H_0 or the more general $H_{0,\text{gen}}$) and any alternative hypotheses in that H_0 is given the benefit of doubt and the error associated with rejecting H_0 incorrectly is controlled and held at a well-defined (low) level, namely α. Actually, there are three main ways of justifying such an asymmetry (Godfrey-Smith, 1994).

First, there is a *semantic* justification, giving attention to the meaning of the term "null". The idea is that at least some hypotheses are "natural nulls"; they state that there is no difference between groups, nothing is going on, there is no effect, etc. So we could take hypotheses of "no effect" as nulls. If we do so, it is reasonable to view the type I error as more serious simply because of Occam's razor: rejecting H_0 falsely calls for a more complex model than is needed. Therefore, there should be a bias in favor of the simpler null hypothesis. Of course, α must be smaller than β (in the Neyman–Pearson terminology) for this bias to be established.

Secondly, there is a *pragmatic* justification associated with the writings of Neyman and Pearson. They argue that one of the decisions associated with hypothesis testing (accepting H_0 or rejecting H_0 in favor of an alternative hypothesis) is usually more serious than the other, in the sense that it leads to less wanted consequences for whoever performs the test. The argument is then simply a definition:

> "The error that a practising statistician would consider the more important to avoid (which is a subjective judgment) is called the error of the first kind" (Neyman, 1976:161).

Error of the first kind (type I error) is the error associated with the faulty rejection of H_0. So Neyman and Pearson say that the hypothesis the rejection of which is the most serious among available hypotheses is to be the null. There is no guarantee that the hypothesis judged to be the null according to the semantic view is the null according to the pragmatic view, too.

A third justification could be called *doxastic* (a term meaning 'related to belief'). According to this view, researchers have different attitudes towards different hypotheses. When H_0 is rejected, this is seen as an important knowledge gain. But when H_0 is kept, this step is more seen as a suspended judgment; the researcher tentatively holds H_0. This view has been advanced by Isaac Levi (1962). In the framework of a doxastic justification for the selection of a null, any hypothesis could be regarded as the null as long as the rejection of this hypothesis would be seen as an important knowledge gain. There is no straight-forward mapping to the semantic and pragmatic justification views above.

If one consults statistics textbooks one will find a mixture of justifications for the selection of null hypotheses (Godfrey-Smith, 1994). The Neyman–Pearson pragmatic justification seems to have been widespread in the 1950's and 1960's, when Neyman's and Pearson's conception of how to perform statistical hypothesis testing was dominant. Today, the semantic justification is likely to be at least as common as the pragmatic justification. Some textbooks simply avoid giving any particular justification for nulls, for example the Swedish textbook already mentioned, stating no more than this:

> "We would like to test some *null hypothesis* H_0 regarding the distribution. The null hypothesis amounts to a certain specification of the distribution" (Blom et al., 2005:322, my translation).[18]

Does it really matter which hypothesis is labeled the null? Yes, it does. Since the null hypothesis is conventionally given the benefit of doubt, it can be quite simple to state several hypotheses that would be consistent with a given result. Whichever was called the null will then be supported by the data. Here's a very simple example: Suppose we toss a coin 40 times to test the probability of getting heads. One null hypothesis might be that the number of heads is an outcome of a binomial distribution Bin(40, 0.5). Another null hypothesis candidate could be that the result is an outcome of a Bin(40, 0.55) distribution. Now suppose that we get the empirical result $d = 21$. This result is consistent with both hypotheses with $\alpha = 0.05$ (and also for considerably smaller values of α). Hence, whichever hypothesis was selected as the null would be supported by the outcome of the test.

The example above involves hypotheses where a parameter to be modeled, θ, is assumed to have one specific point value. It is perhaps not surprising that different point values can be supported by the same set of data. However, it is fully possible to give an example where two hypotheses state two non-overlapping ranges for a parameter and both still get support from the same empirical set of data (Rosenkrantz, 1973:315): We toss a coin 100 times in order to test the probability of getting heads up in each toss, which we call θ. Suppose we had taken $0.45 \leq \theta \leq 0.55$ as our null hypothesis. For $\alpha = 0.05$ calculations assuming a binomial distribution would show that this hypothesis should be rejected if $d < 37$ or $d > 63$. Suppose that our empirical result from 100 tosses is $d = 50$. This result will then make us accept the hypothesis $0.45 \leq \theta \leq 0.55$. On the other hand, had we taken $\theta > 0.55$ as our null hypothesis, calculations would show that for $\alpha = 0.05$ we should reject this hypothesis if $d < 47$. Thus, the actual data $d = 50$ would make us accept the hypothesis $\theta > 0.55$. Hence, both $0.45 \leq \theta \leq 0.55$ and $\theta > 0.55$ are supported by the data although these hypotheses are jointly inconsistent.

Nulls in natural and social sciences

Sometimes it is argued that nulls have different roles in different fields of inquiry. There is some truth in this assertion. For example, within psychology it seems to be widely held that nulls are usually not interesting, and that only rejections of the null are worth publishing. On the other hand, in evolutionary biology, statistical tests supporting the (already well-corroborated) theory of evolution are considered worth publishing (Godfrey-Smith, 1994:285, 289).

Another perhaps more striking difference between academic fields may be whether the null is a point hypothesis (i.e., a parameter assumes a certain value according to the null) or a directional hypothesis (i.e., a parameter deviates from a value in a specified direction). The former seems to be more common in the natural sciences than in the social sciences.

[18] Original text: "Vi vill pröva en viss *nollhypotes* H_0 rörande fördelningen. Nollhypotesen innebär att man på något sätt specificerar hur fördelningen ser ut."

Bayesian hypothesis testing

It is not difficult to find situations where Bayes's theorem appears to be useful in hypothesis testing. Here is a typical textbook example (Borel, 1914:96ff).

Two urns A and B contain four balls each. Urn A holds 3 white balls and one black; B holds one white ball and 3 black. One ball is drawn randomly from one of the urns and is found to be white. What is the probability that the white ball was drawn from A? We write H_A for the hypothesis that the ball was drawn from A and H_B for the hypothesis that the ball was drawn from B. The available evidence (the drawn ball being white) is denoted E. We want to find $P(H_A|E)$. According to Bayes's theorem, we have

$$P(H_A|E) = \frac{P(E|H_A)P(H_A)}{P(E)},$$

where the denominator can be rewritten, using the law of total probability, so that we get

$$P(H_A|E) = \frac{P(E|H_A)P(H_A)}{P(E|H_A)P(H_A) + P(E|H_B)P(H_B)}.$$

In order to use this formula, we must know the values of $P(H_A)$ and $P(H_B)$, i.e. the probabilities of selecting urn A or urn B when the ball is drawn. If we lack any information to the contrary, it may seem reasonable to take $P(H_A) = P(H_B) = \frac{1}{2}$. Plugging probabilities into the formula will then give

$$P(H_A|E) = \frac{\frac{3}{4} \cdot \frac{1}{2}}{\frac{3}{4} \cdot \frac{1}{2} + \frac{1}{4} \cdot \frac{1}{2}} = \frac{3}{4}.$$

Similarly, we get $P(H_B|E) = \frac{1}{4}$, of course. In other words, it is three times more likely that the ball was drawn from A than from B. Using Bayes's theorem, we have judged one hypothesis to be more likely than another. There were two hypotheses exhausting all possibilities, and since we we had no indication that one of the urns was more likely to be selected than the other, it was reasonable to use $P(H_A) = P(H_B) = \frac{1}{2}$.

Hardly anyone would object to the reasoning above. And still, this simple example contains an embryo for performing hypothesis testing in a way quite different from the suggestions due to Fisher and Neyman–Pearson. For if we can use Bayes's theorem to find $P(H_A|E)$, why couldn't we use the very same theorem to find $P(H_0|d)$ in a large number of situations? After all, the evidence E must be understood as identical to what has been called the data d in the Fisherian and N–P settings. According to Bayes's theorem, then, we find that

$$P(H_0|d) = \frac{P(d|H_0)P(H_0)}{P(d)}$$

or

$$P(H_0|d) = \frac{P(d|H_0)P(H_0)}{P(d|H_0)P(H_0) + P(d|H_1)P(H_1)}$$

in the expanded form, where H_0 and H_1 must be understood as exhaustive alternatives ($P(H_0) + P(H_1) = 1$), i.e. H_1 is the negation of H_0. If there are more than two hypotheses exhausting all possibilities, the denominator will have to be expanded accordingly.

No one denies the general validity of Bayes's theorem, so the formulae above are correct and can be used to find $P(H_0|d)$. But to get $P(H_0|d)$, we have to enter values

of $P(d|H_o)$, $P(d|H_1)$, $P(H_o)$, and $P(H_1)$ according to the bottom formula. The first two – $P(d|H_o)$ and $P(d|H_1)$ – should be easy to compute, provided we have formulated H_o and H_1 mathematically. Also, $P(H_1)$ is simply $1 - P(H_o)$ if H_1 is the negation of H_o. This leaves only $P(H_o)$ to be stated. But since we want to compute $P(H_o|d)$, is it at all feasible to state $P(H_o)$?

Actually, it is. But we then have to view $P(H_o)$ as a more or less subjective estimate reflecting our knowledge before we have learned about the data d. In a Bayesian framework, $P(H_o)$ would then be called the *prior* or (in Latin) *a priori* probability. The probability $P(H_o|d)$ that we get after having taken d into account is called the posterior or *a posteriori* probability. In the example with the urns, above, $P(H_A) = P(H_B) = \frac{1}{2}$ were our prior probabilities.

Within the framework of Bayesian hypothesis testing, then, we are able to produce probabilities of hypotheses given the collected data, that is $P(H_o|d)$, by inserting a value of $P(H_o)$ that we believe to be reasonable. This is nice, since $P(H_o|d)$ seems more interesting than the p value $P(D|H_o)$. But the price to be paid is quite high: we need to understand probabilities as subjective measures of the credibility of hypotheses. Anyone unwilling to support this particular understanding of probability must be reluctant to adopting Bayesian hypothesis testing.

And even if personal probabilities are accepted, there are other obstacles in the Bayesian framework (Efron, 1978:236). How are we to find a prior value or distribution when there is no obvious one? What about the situation where an unknown parameter is known to be a physical constant with no random character and we must come up with a prior probability distribution for this parameter in a Bayesian test where the value of the parameter is to be determined experimentally, is such a Bayesian prior reasonable at all? Also, we could imagine situations where we wish to create a Bayesian prior expressing ignorance, but this ignorance can be assigned quite different numerical values depending on how the available alternatives are phrased; e.g., the probability of an object being blue is $\frac{1}{2}$ if the only alternative is not-blue, but so is the probability of the same object being red when contrasted with not-red, or black when contrasted with not-black. But each of $P(\text{blue})$, $P(\text{red})$, and $P(\text{black})$ cannot be $\frac{1}{2}$ since their sum exceeds one. Even for continuous priors, there is a general problem of finding the correct one for a given parameter θ. For example, what appears to be a flat, non-committal prior in θ may not be so flat in, say, $f(\theta) = \theta^2$. The flatness, then, will depend on what function involving the unknown parameter one is interested in.

It is clear that the Bayesian approch to hypothesis testing has its own problems. All the same, in principle it solves a lot of problems that plague the other approaches presented earlier (cf. the list on page 16). Bayesianism is not just a theory of hypothesis testing, but a theory of how new evidence is (and should be) incorporated in the enterprise of knowledge and science. Bayesianism has some followers among philosophers and statisticians, but in numbers of supporters among practicing scientists it is very far from the popularity of Fisher, Neyman–Pearson, or a merged Fisher/N–P theory. On the other hand, the popularity of Bayesian methods has increased during the last decades due to more powerful computers making these methods applicable in an expanding number of fields.

Statistical hypothesis testing in the philosophy of science

Last but not least, I would like to consider a few connections between statistical hypothesis testing and well-known concepts and theories in the general philosophy of science. Although I have made a few philosophical points so far, I have not tried to frame statistical hypothesis testing in a general philosophy of science curriculum. In fact, such a framing is both possible and interesting.

First, we could note that there is no substantial difference between statistical and "ordinary" (non-statistical) hypothesis testing with respect to the difficulty of proving an isolated hypothesis wrong. As pointed out in many textbooks in the philosophy of science, it seems to be impossible to test a hypothesis in isolation, for there are always some auxiliary hypotheses that we tacitly assume to be true, and that in conjunction with the explicit hypothesis (H_0 or whatever it is called) form the set of hypotheses that is really put to test. This impossibility is sometimes referred to as the "Duhem–Quine thesis".[19]

The concept of *underdetermination* is well-known from the philosophy of science. In this context, underdetermination denotes the fact that any set of data can be accounted for by several hypotheses or theories. This has its direct equivalence in statistical hypothesis testing; as we have seen, every test statistic value $T(d)$ can be accounted for by several hypotheses. This means that every statistical hypothesis is underdetermined by the available data.

Another area where statistical hypothesis testing meets general philosophy of science is in the possibility of a connection to Karl Popper's falsificationism. Ronald Fisher has stated:

> "[I]t should be noted that the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis" (Fisher 1935:19).[20]

Note that this statement was made already in 1935, only one year after the publication of the German edition of Karl Popper's seminal work *Logik der Forschung* (a rewritten English version, *The Logic of Scientific Discovery*, appeared in 1959). Clearly, there is a similarity between Popper's falsificationism and Fisher's statement, but the general understanding is that this is just a coincidence, at least in terms of origin. Nonetheless, one could argue that "[s]tatistical tests are (…) based implicitly on methodological falsifiability, and their introduction and widespread adoption by statisticians provides striking corroboration of the value of Popper's approach" (Gillies, 1995:110f).

On the other hand, the differences between the Fisher and the Neyman–Pearson approaches are not greater than allowing another commentator to state the following: "In fact, one might fairly regard Neyman–Pearson theory as the statistical embodiment of Popperian methodology" (Rosenkrantz, 1973:320). So perhaps one should not try to marry Fisherian *or* N–P theory with falsificationism, but accept that both approaches bear resemblance with the Popperian idea.

[19] Pierre Duhem (1861–1916) was a French physicist, mathematician, and philosopher of science. W. V. O. Quine (1908–2000) was an American philosopher.

[20] Also in later editions, e.g., Fisher (1951:16).

Statistical hypothesis testing has secured a prominent place in the methodological arsenal of many natural sciences, whereas in the social sciences (in particular, economics and sociology), the proper role of statistical hypothesis testing has been widely debated (Giere, 1972; Morrison & Henkel, 1970). One of the points repeatedly made in this debate is the following (Giere, 1972:173). In social science, it can be argued that very often one knows that there is at least *some* small difference between any two parameters (e.g., between two groups). Therefore, a null hypothesis stating no difference is not credible in the first place, and even if we should accept the null hypothesis according to the test result, we could believe that it would have been rejected had we taken a larger sample. So why perform a test at all? This argument has been advanced as a general criticism against statistical hypothesis testing in the social sciences. But the argument is easy to refute, at least within the Neyman–Pearson paradigm where two hypotheses are stated: we are not interested in just *any* difference, but only in differences large enough to be detectable and important.

Still, a general point remains: in the social sciences (and also frequently in medicine) one seldom believes that a null hypothesis stating no difference is completely accurate. Since the objects of study are human beings – and we know that humans are not often exactly alike – we do not expect group differences to vanish completely. Perhaps this is related to what has been described as a principal difference between different fields on inquiry, namely that in e.g. psychology one tests directional hypotheses related to low content theories, whereas in physics one tests point hypotheses related to high content theories (Giere, 1972:176f).

Finally, it could be noted that statistical reasoning is of interest to philosophers trying to come to grips with the classical problem of induction. Reasoning very closely related to that commonly adopted in statistical hypothesis testing has been invoked to solve the problem of induction (Williams, 1947). This problem, however, is beyond the scope of the present text.

References

Belsey, Andrew (1995). "Hypothesis", in Ted Honderich (ed.), *The Oxford Companion to Philosophy*, Oxford: Oxford University Press, 385.

Bertrand, Joseph (1889). *Calcul des probabilités*. Paris: Gauthier-Villars et fils.

Blom, Gunnar et al. (2005). *Sannolikhetsteori och statistikteori med tillämpningar*, 5th edition. Lund: Studentlitteratur.

Borel, Émile (1914). *Le Hasard*. Paris: Librairie Félix Alcan.

Chadwick, D. G. & Jensen, L. (1971). *The Detection of Magnetic Fields Caused by Groundwater and the Correlation of Such Fields with Water Dowsing*, Progress Report 78:1, Utah Water Research Laboratory.

Efron, Bradley (1978). Controversies in the foundations of statistics, *American Mathematical Monthly* 85(4), 231–246.

Fisher, Ronald A. (1925). *Statistical Methods for Research Workers*, 1st edition. Edinburgh: Oliver & Boyd.

Fisher, Ronald A. (1935). *The Design of Experiments*, 1st edition. Edinburgh: Oliver & Boyd.

Fisher, Ronald A. (1951). *The Design of Experiments*, 6th edition. Edinburgh: Oliver & Boyd.

Fisher, Ronald A. (1956). *Statistical Methods and Scientific Inference*, 1st edition. Edinburgh: Oliver & Boyd.

Giere, Ronald N. (1972). The significance test controversy, *The British Journal for the Philosophy of Science* 23(2), 170–181.

Gillies, Donald (1995). Popper's Contribution to the Philosophy of Probability. In: Anthony O'Hear (ed.), *Karl Popper: Philosophy and Problems*, Cambridge: Cambridge University Press, 103–120.

Godfrey-Smith, Peter (1994). Of Nulls and Norms, *PSA* (*Philosophy of Science Association*) *1994*, 1, 280–290.

Hacking, Ian (1965). *Logic of Statistical Inference*. Cambridge: Cambridge University Press.

Halldén, Sören (2003). *Från vardagsvett till statistisk beviskonst*. Nora: Nya Doxa.

Hubbard, Raymond & Bayarri, M. J. (2003). Confusion Over Measures of Evidence (p's) Versus Errors (α's) in Classical Statistical Testing, *The American Statistician* 57(3), 171–178.

Lehmann, E. L. (1993). The Fisher, Neyman–Pearson Theories of Testing Hypotheses: One Theory or Two? *Journal of the American Statistical Association* 88, 1242–1249.

Lehmann, E. L. (2011). *Fisher, Neyman, and the Creation of Classical Statistics*. New York: Springer.

Levi, Isaac (1962). On the Seriousness of Mistakes, *Philosophy of Science* 29(1), 47–65.

Mayo, Deborah (1981). Testing statistical testing. In: J. C. Pitt (ed.), *Philosophy in Economics: Workshop on Testability and Explanation in Economics, Virginia Polytechnic Institute and State University, 1979*. Dordrecht: D. Reidel Publishing Company, 175–203.

Morrison, Denton E. & Henkel, Ramon E. (eds) (1970). *The Significance Test Controversy: A Reader*. Chicago: Butterworth & Co.

Neyman, Jerzy (1976). The Emergence of Mathematical Statistics: A Historical Sketch with Particular Reference to the United States. In: D. B. Owen (ed.), *On the History of Statistics and Probability*, New York: Marcel Dekker, 149–193.

Neyman, Jerzy & Pearson, Egon S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference, *Biometrika* 70A, 175–240, 263–294. Also in Neyman & Pearson (1967), 1–98.

Neyman, Jerzy & Pearson, Egon S. (1933). On the Problem of the most Efficient Tests of Statistical Hypotheses, *Philosophical Transactions of the Royal Society of London, Series A*, 231, 289–337. Also in Neyman & Pearson (1967), 140–185.

Neyman, Jerzy & Pearson, Egon S. (1936). Contributions to the theory of testing statistical hypotheses. I. Unbiassed critical regions of type A and type A_1, *Statistics Research Memoirs* 1, 1–36. Also in Neyman & Pearson (1967), 203–239.

Neyman, Jerzy & Pearson, Egon S. (1967). *Joint Statistical Papers*. Berkeley: University of California Press.

Peirce, Charles S. (1878). The probability of induction, *Popular Science Monthly*. Also reprinted as chapter 13 in *Philosophical Writings of Peirce*, Dover 1955.

Rosenkrantz, R. D. (1973). The significance test controversy, *Synthese* 26, 304–321.

Stigler, Stephen M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*. Cambridge (Mass.): The Belknap Press of Harvard University Press.

Williams, Donald C. (1947). *The Ground of Induction*. Cambridge MA: Harvard University Press.