DISCUSSION

ON PRIOR PROBABILITIES OF REJECTING STATISTICAL HYPOTHESES*

HERBERT KEUTH

University of Mannheim

1. In [5], Paul E. Meehl summarized his article as follows[1]:

> Because physical theories typically predict numerical values, an improvement in experimental precision reduces the tolerance range and hence increases corroborability. In most psychological research, improved power of a statistical design leads to a prior probability approaching ½ of finding a significant difference in the theoretically predicted direction. Hence the corroboration yielded by "success" is very weak, and becomes weaker with increased precision. 'Statistical significance' plays a logical role in psychology precisely the reverse of its role in physics. This problem is worsened by certain unhealthy tendencies prevalent among psychologists, such as a premium placed on experimental "cuteness" and a free reliance upon *ad hoc* explanations to avoid refutation. ([1], p. 103)

1.1. Meehl assumes that a *physical theory*, T predicts *point values*, e.g. a mean μ_0. When such a theory is subjected to a "significance test," μ_0 is compared with an observed mean $\bar{x}_0$. So the theory is tested by testing a *point-null hypothesis* like $H_0{:}\mu = \mu_0$, which it implies. *Psychological theories T*, on the other hand, used to predict only that there was *some difference* $\mu_0 - \mu_1 = \delta_{01}$ between the means μ_0 and μ_1 of two population distributions. They implied *inexact hypotheses* like $H_{\mathrm{I}}{:}\mu_0 \neq \mu_1$ or $H_{\mathrm{I}}{:}|\mu_0 - \mu_1| = \delta_{01} > 0$, which were indirectly tested by testing their point-alternatives $H_0{:}\mu_0 = \mu_1$. (We choose the unfamiliar Roman subscript 'I', because we need '1' to designate a point-alternative to H_0.) In recent years psychologists became more interested in theories T which at least imply *directional hypotheses* like $H_{\mathrm{II}}{:}\mu_0 > \mu_1$ or $H_{\mathrm{II}}{:}\mu_0 - \mu_1 = \delta_{01} > 0$. Such hypotheses in their turn are tested by testing the corresponding directional null hypotheses $H_{02}{:}\mu_0 \leq \mu_1$ or $H_{02}{:}\mu_0 - \mu_1 = \delta_{01} \leq 0$.

1.2. The difference between testing point-null hypotheses and directional hypotheses is supposed to produce a "*methodological paradox*," which is stated in a weaker and in a stronger version.

1.2.1. Let us first turn to the *weaker* one:

> In physics T implies H_0. Therefore increasing precision or power of the test will lead to decreasing probability of accepting H_0 and T tentatively along with it, given that T lacks verisimilitude. In the social sciences the situation is precisely the reverse. ([1], p. 113)

* Received September, 1972.

[1] Meehl's article has been positively commented on by well-known contributors to [6], and also in [8].

Now 'versimilitude' is not a term of probability theory or of statistics but stems from Popperian philosophy of science. Thus far, Popper has not furnished a tenable definition. We may, however, easily interpret Meehl's statement as saying that increasing precision will lead to decreasing probability of accepting H_0, given that it is false. This is true e.g. of hypotheses on means; but *generally* we can assert only that the probability remains constant or decreases. However, in the social sciences the situation is by no means "reversed." The probability of accepting an inexact hypothesis H_{I}, given that it is false, is identical with the significance level, and the probability of accepting a directional hypothesis H_{II}, given that it is false, approaches the significance level as its upper limit. (For details see below section *2.*)

1.2.2. Now we turn to the *stronger* version. It says that in physics, "if the theory has negligible versimilitude, the logical probability of its surviving such a test is negligible" ([1], p. 113). Meehl uses 'logical probability' simply as a synonym of 'prior probability' and does not refer to any calculus of logical probability or partial implication as produced by Carnap. But the probability he refers to is of course relative or "a posteriori." He obviously means that the probability of accepting H_0 (and T tentatively along with it) given H_0 is false, approaches *zero*, if the precision of the test grows perfect. Again this is true e.g. of hypotheses on means, but it cannot be *generally* asserted. While there is no stronger statement on the probability of accepting an *inexact* hypothesis H_{I}, given it is false, Meehl claims that the probability of accepting a *directional* hypothesis H_{II}, "*even if the theory* [which implies the hypothesis] *is totally without merit*," approaches $p = 0.5$, when the power grows perfect ([1], p. 111). If the level of significance is, as usual, set at 0.01 or 0.05, the probability of accepting H_{II}, given it is false, cannot be $p = 0.5$. But the *absolute* probability of accepting H_{II} can still be 0.5. However, its computation presupposes material assumptions as to the prior probabilities of the various alternatives.

1.2.3. Though Meehl considers making "prior-probability assumptions concerning the actual distribution of true differences in the whole vast world of psychological experimental contexts" ([1], p. 111) illegitimate, he thinks he can consider any point-null hypothesis "(quasi-)always false." He cannot, however, assign every point hypothesis zero prior probability; and even this assumption does not imply a probability approaching 0.5 of accepting directional hypotheses. But Meehl thinks he can generate this probability by means of a *two-urn model* and possibly an additional random process:

> Let us now conceive of a large "theoretical urn" containing counters designating the indefinitely large class of actual and possible substantive theories concerning a certain domain of psychology . . . Let us conceive of a second urn, the "experimental-design" urn, containing counters designating the indefinitely large set of possible experimental situations which the ingenuity of man could devise. . . . Since the point-null hypothesis H_0 is [quasi-] always false, almost every one of these experimental situations involves a non-zero difference on its output variable (parameter). Whichever group we (arbitrarily) designate as the "experimental" group and the "control" group, in half of these experimental settings the true value of the dependent variable difference (experimental minus control) will be positive, and in the other half negative.
>
> We now perform a random pairing of the counters from the "theory" urn with the counters from the "experimental" urn, and arbitrarily stipulate—quite irrationally—that a

> "successful" outcome of the experiment means that the difference favors the experimental group [$\mu_E - \mu_C > 0$]. . . . by assuming counterfactually that there is *no connection whatever between our theories and our experimental designs* (the two-urn idealization), thereby fixing the expected frequency of successful refutations of the directional null hypothesis H_{02} at $p = 1/2$ for experiments of *perfect power*; it follows that, as the power of our experimental designs and significance tests is increased . . ., we approach $p = 1/2$ as the limit of our expected frequency of "successful outcomes," i.e. of attaining statistically significant experimental results in the theoretically predicted direction. ([1], pp. 110–111)

Before we examine Meehl's urn model in section *3*, we must elaborate in section *2* the presuppositions of computing the absolute and relative probabilities of accepting a hypothesis.

2. When computing the probability of obtaining certain experimental outcomes, we have to make assumptions on the true population distribution. Usually we do not need a complete specification. We need only know (or conjecture) the distribution is of such a kind that the corresponding sampling distribution of the parameter we are interested in may be adequately approximated by the normal distribution, provided the sample size is large. As the Central-Limit Theorem shows, any population with finite variance generates sampling distributions of the mean which may be so approximated. If we choose a significance level and decide that we want the smallest region of acceptance possible at this level, we are ready to test e.g. $H_0 : \mu = \mu_0$.

2.1. If we now *increase n* and keep everything else constant, we get a region of acceptance which is a proper subset of the previous one. Let $A_n(H_0)$ be the region of acceptance of H_0 when the sample contains n elements and $R_n(H_0)$ correspondingly its region of rejection. Then $A_{n+m}(H_0) \subset A_n(H_0)$ and, as $R_n(H_0)$ is the complement of $A_n(H_0)$, $R_n(H_0) \subset R_{n+m}(H_0)$. However, a general statement on the effect of enlarging the sample size, which does not presuppose any specific population distribution and does not refer to any specific parameter, cannot but assert that $A_{n+m}(H_0) \subseteq A_n(H_0)$, i.e. the *region of acceptance of H_0 remains constant or decreases with increasing n*. Therefore $P(A_{n+m}(H_0)) \leq P(A_n(H_0))$ and $P(R_n(H_0)) \leq P(R_{n+m}(H_0))$, i.e. *the absolute probability of accepting H_0 remains constant or decreases* while the absolute probability of rejecting H_0 remains constant or increases.

On the other hand the risk of a type I error, the *probability $P(R_n(H_0)/H_0)$ of rejecting H_0 though it is true* has, by fiat, *not changed*, since the significance level which expresses this risk, has been kept constant. The risk of type II errors, however, the *probability $P(A_n(H_0)/H_{i\neq 0})$ of accepting H_0 though it is false* and some alternative $H_{i\neq 0}$ is true instead, *remains constant or decreases* for any $H_{i\neq 0}$. As the region of acceptance narrows, parameter values falling into the difference set of the previous and the present region of acceptance will no longer suggest the acceptance of H_0. Correspondingly the *power* of the test, the *probability $P(R_n(H_0)/H_{i\neq 0})$ of rejecting H_0, given it is false, remains constant or increases* when the region of acceptance of H_0 narrows, as $P(R_n(H_0)/H_i) = 1 - P(A_n(H_0)/H_i)$. Whether it changes and how much it changes again depends on the various alternative population distributions and on the sampling distributions of the respective parameters which they generate. In case of normal sampling distributions we can make stronger statements. All equalities in the last paragraph may be eliminated and e.g. $P(A_{n+m}(H_0)) \leq P(A_n(H_0))$ replaced by $P(A_{n+m}(H_0)) < P(A_n(H_0))$. We may even assert that the *risk of type II errors* approaches *zero* and the *power unity*, if n grows infinitely large.

The "absolute" or "prior" probability $P(R_n(H_0))$ of rejecting H_0 may be computed by means of the Theorem on Total Probability:

$$P(R_n(H_0)) = P(R_n(H_0)/H_0)P(H_0) + \cdots + P(R_n(H_0)/H_m)P(H_m).$$

The computation, however, presupposes not only the relative probabilities of rejecting H_0, but also the prior probabilities $P(H_i)$. As the event H_i occurs, if and only if the hypothesis designated by 'H_i' is true, we must, for each i, know the probability that the hypothesis H_i is true. We need not here cross the border to the philosophy of science in order to find out what a prior probability of an hypothesis being true might mean and which values it might assume. In any case we cannot assign zero prior probability $P(H_i) = 0$ to *all* hypotheses H_i, because then any term of the sum would be zero and so would $P(R_n(H_0))$. But this would lead to a contradiction, as the same would of course hold for any term of the sum $P(A_n(H_0)/H_0)P(H_0) + \cdots + P(A_n(H_0)/H_m)P(H_m)$ and so for $P(A_n(H_0))$. But $A_n(H_0)$ is the complement of $R_n(H_0)$, so that $P(A_n(H_0)) = 1 - P(R_n(H_0))$ and consequently $0 = 1 - 0$. Therefore *equal* prior probabilities can only be assigned to a *finite* number of alternatives. If we assigned equal nonzero prior probability to a *countable* number of alternatives, their sum would be infinite instead of unity. For *continuous* random variables the situation is similar. Their values cannot all have *zero* prior probability density and cannot all have *equal* nonzero probability density.

2.2. Let us now turn to a type of statistical hypotheses which, as Meehl reports, used to prevail in psychology, viz. *inexact* hypotheses stating that there is some difference between the parameters characterizing two populations, e.g. $H_I: \mu_0 \neq \mu_1$. As such hypotheses do not indicate point values, they do not generate sampling distributions. Therefore they can only be indirectly tested by testing their complements which necessarily indicate point values, as e.g. $H_0: \mu_0 = \mu_1$.

Anything stated above on point hypotheses is true of these point hypotheses too. But $A_n(H_0) = R_n(H_I)$. Therefore $A_{n+m}(H_0) \subset A_n(H_0)$ implies $R_{n+m}(H_I) \subset R_n(H_I)$ and this in its turn implies $A_n(H_I) \subset A_{n+m}(H_I)$, as $R_n(H_I)$ is the complement of $A_n(H_I)$. But if we disregard the fact that in our example H_I asserts the inequality of two *means* and make a general statement, we have to replace the proper inclusions by simple or improper inclusions. So, *when the sample size grows, the region of acceptance of H_I remains constant or grows*. Consequently $P(A_n(H_I)) \leq P(A_{n+m}(H_I))$, i.e. *the prior probability of accepting H_I remains constant or grows* as well. Again the probability of rejecting H_0, given that H_0 is true, or of *accepting H_I, given that H_I is false*, has by fiat *not changed*, as the level of significance has been kept constant: $P(A_{n+m}(H_I)/H_0) = P(A_n(H_I)/H_0)$. On the other hand the probability of *accepting H_I, given H_I is true*, as one of the hypotheses H_i from which H_I follows is true, *remains constant or increases*, i.e. for any $i \neq 0$, $P(A_{n+m}(H_I)/H_i) \geq P(A_n(H_I)/H_i)$. In case of a *normal* sampling distribution the power $P(R_n(H_0)/H_i)$ *approaches unity* for any $i \neq 0$ and so does the probability $P(A_n(H_I)/H_i)$ of *correctly accepting H_I*. But unless we know the prior probabilities $P(H_i)$, we cannot compute the absolute probability $P(A_n(H_I))$ of accepting H_I. Can we call this the result of a *more lenient test*? Obviously nothing is wrong with increasing the chances of accepting a hypothesis when it is in fact true.

2.3. Meehl's central topic, however, is significance tests of *directional* hypotheses like $H_{II}:\mu_0 > \mu_1$. The directional alternative of H_{II} is $H_{02}:\mu_0 \leq \mu_1$, which specifies the necessary point value $\mu_0 = \mu_1$. (Meehl uses 'H_1' where we use 'H_{II}'.) Anything said in *2.2.* on inexact hypotheses may be transferred without restriction to directional hypotheses. Only one additional point has to be mentioned. With directional hypotheses there is only *one* limit of the regions of acceptance and rejection which may "move" when n changes, while with hypotheses like $\mu_0 = \mu_1$ ($\mu_0 \ngtr \mu_1$ and $\mu_0 \nless \mu_1$) and $\mu_0 \neq \mu_1$ ($\mu_0 > \mu_1$ or $\mu_0 < \mu_1$) there are *two*, one for each of the relations given in parentheses. But this does not change anything about the (topological) relations between the regions of acceptance for different n, nor does it change anything about the relations of their probabilities. If, in the formulae given in *2.2.*, we substitute 'H_{02}' for 'H_0', 'H_{II}' for 'H_I' and 'for any H_i which implies H_{II}' for 'for any $i \neq 0$', we get analogous theorems valid for directional hypotheses. In particular the probability $P(A_n(H_{II})/H_i)$ of *correctly accepting the directional hypothesis H_{II} approaches unity*—not 0.5—for each H_i which implies H_{II}, when the sampling distribution is *normal*.

3. Let us now examine Meehl's *urns*. The number of counters which they contain has to be infinite. Otherwise, e.g. the "theoretical urn" could not even hold counters representing all rational values which some random variable can take. But as $n \div \infty$ is zero and $\infty \div \infty$ undefined, the probability that a counter representing a specific theory or experimental design will be drawn is zero, if the theory or design is represented by a finite number of counters and otherwise undefined.

If we assume that the "theoretical urn" contains only counters designating directional hypotheses, we might be tempted to argue as follows. To each hypothesis $H_{II}:\mu_1 - \mu_2 > 0$ there is a hypothesis $H_{III}:\mu_1 - \mu_2 < 0$. (Meehl uses '$H_2$' where we use '$H_{III}$'.) The urn is the union of the set of all counters designating hypotheses H_{II} and of the set of all counters designating hypotheses H_{III}. Therefore the probability of drawing a counter from either set is unity. But the sets exclude each other. Therefore the sum of their probabilities is unity as well. As there is a one-to-one mapping between the sets, they are equipollent. So the probability of drawing a counter from one of them can only be 0.5. But this reasoning involves a fallacy, because from each of the two sets we can extract infinite disjoint subsets, which may be mapped on the original sets and on each other. Therefore they must all have the same cardinal number. But as they are disjoint, the Axiom of Total Additivity demands that the sum of their probabilities be unity. So they cannot have a probability of 0.5.

Suppose from the "theoretical urn" we have drawn a counter representing the hypothesis that the capacity to recall nonsense syllables is greater when a hand dynamometer is squeezed. From the "experimental design urn" we may then draw counters representing experimental designs like weighing test persons, or measuring the time a satellite travels until it reaches Jupiter, or, . . ., or finally counting the nonsense syllables that a person squeezing a hand dynamometer recalls and comparing their number to that of the syllables recalled by a person not squeezing it. The Special Theorem on Multiplication now shows that, provided the drawings from the two urns are independent, the chances of choosing a certain theory and at the same time an experimental design permitting a test of this theory are zero—if both the theory and the relevant design are represented on a finite number of counters—otherwise they are undefined. So, random pairing of counters from the two urns will result in *zero or undefined probability of subjecting a theory to a relevant test*. Correspondingly the *probability of subjecting a theory to a relevant test and accepting it is zero or undefined*; and this, in its turn, implies that the *probability of accepting a hypothesis, given it has been subjected to a relevant test is, in any case, undefined.* So we have to discard the two-urn model. After all, in order to avoid a "logical connection between our theories and the direction of the experimental outcomes" ([1], p. 111) we have to choose *unbiased* tests, but this does not mean that the choice of the tests has to be *statistically independent* of the choice of the theories.

4. Meehl mentions still another random process. He speaks of "assign[ing], in a strictly random fashion, the names 'experimental' and 'control' to the two groups which a given experimental setup treats in two different ways" ([1], p. 110). What can this mean?

Selecting the test subjects randomly and randomly attributing the chosen subjects to the group which is subjected to the test (usually called the "experimental group") or to the group which is not (usually called "control group") is a presupposition of the application of probability theory which need not be mentioned here, as it has *no effect* on the probability of one or the other outcome of the experiment. If, however, we administer the experimental treatment to one randomly selected group, measure the random variable in this and in the control group and *afterwards* randomly exchange both the labels 'experimental' and 'control' and also the measured values along with them, we at best get equal probabilities of drawing the right or wrong *conclusions* from our test.

Probably the idea behind it is the same as that of "arbitrarily assign[ing] one of the two directional hypotheses H_1 or H_2 [H_{II} or H_{III} in our terminology] to each theory" ([1], p. 111). As a theory either implies a hypothesis or does not imply it, we cannot arbitrarily assign some hypothesis to a given theory. But we might as well test a theory T_{II} which implies H_{II} as some theory T_{III} which implies H_{III}. The probability of choosing either of them might be 0.5. However, the random process which determines this probability can, for the reasons given above, not consist in drawing from the "theoretical urn." The decision might be taken by drawing from a two-element urn or by flipping a coin.

4.1. Will knowledge of the probability $P(C(H_{II}))$ of subjecting H_{II} to a test permit to compute the *probability* $P(A(H_{II})/C(H_{II}))$ *of accepting* H_{II}, *given it has been chosen for test*? As $P(A(H_{II})/C(H_{II})) = P(A(H_{II}) \cap C(H_{II}))/P(C(H_{II}))$, the computation presupposes knowledge of the probability $P(A(H_{II}) \cap C(H_{II}))$ of choosing H_{II} for test and accepting it as a result of the test. If choosing and accepting are independent, $P(A(H_{II}) \cap C(H_{II})) = P(A(H_{II})) \cdot P(C(H_{II}))$. But in order to compute this product we have to know $P(A(H_{II}))$.

4.2. Possibly Meehl mistook the probability of randomly choosing a hypothesis for test for its probability of being true. This would account for his statement that "if we randomly assign one of the two directional hypotheses H_1 or H_2 [here H_{II} or H_{III}] to each theory, that hypothesis will be correct half of the time" ([1], p. 111); and for his statement "that the effect of increased precision . . . is to yield a probability approaching ½ of corroborating our substantive theory by a significance test, *even if the theory is totally without merit*" ([1], p. 111), for

$$P(A(H_{II})) = P(A(H_{II})/H_{II})P(H_{II}) + P(A(H_{II})/H_0)P(H_0) + P(A(H_{II})/H_{III})P(H_{III}).$$

If $P(H_{II}) + P(H_{III}) = 0.5$, $P(H_0) = 0$, $P(A(H_{II})/H_{III})$ approaches $\alpha/2 = 0.025$ and $P(A(H_{II})/H_{II}) = 1 - \beta$ approaches 1, then $P(A(H_{II}))$ approaches 0.5125.

If either H_{II} or H_{III} is true and the probability of choosing either of them is 0.5, then the probability of choosing the true one is of course also 0.5, but this is not the probability that H_{II} is true or the probability that H_{III} is true. Perhaps this is more easily recognized, when the example is innocent. Suppose we have an urn containing two balls marked A and B. A is red, while B is blue. Then the probability of drawing A is identical with the probability of drawing the red ball, but it is not the probability that A is red. Only this probability, however, would correspond to the probability that H_{II} is true, and this is what we need in order to compute the probability of accepting H_{II} or H_{III}.

4.3. Let us for the moment assume that we have to decide whether to take H_{II} or H_{02} as the test hypothesis. Then $P(C(H_{II})) = P(C(H_{02})) = 0.5$ implies that the *probability of choosing the true alternative* is 0.5 as well. As H_{II} and H_{02} exclude each other and either of them must be true, $P(H_{II} \cup H_{02}) = P(H_{II}) + P(H_{02}) = 1$. The true alternative is chosen, when H_{II} is true and is chosen or when H_{02} is true and is chosen. As our randomly choosing a hypothesis is independent of its being true,

$$\begin{aligned} & P((C(H_{II}) \cap H_{II}) \cup (C(H_{02}) \cap H_{02})) \\ &= P(C(H_{II})) \cdot P(H_{II}) + P(C(H_{02})) \cdot P(H_{02}) \\ &= P(C(H_{II})) \cdot (P(H_{II}) + P(H_{02})) \\ &= P(C(H_{02})) \cdot (P(H_{II}) + P(H_{02})) \\ &= \quad 0.5 \quad \cdot \quad 1. \end{aligned}$$

The result is independent of the prior probabilities of H_{II} and of H_{02} and of the power of the test. It remains valid for *any* hypothesis and its complement, even for H_I and H_0. We need only substitute H_I for H_{II} and H_0 for H_{02} in the above proof.

The *probability of choosing and accepting the true alternative* can be computed, if the probability of accepting the true alternative is known. But the computation of the latter presupposes knowledge of the individual prior probabilities of each of the alternatives, e.g. $P(H_{II})$ and $P(H_{02})$, while we only know the probability of their union, e.g. $P(H_{II} \cup H_{02})$.

Yet, we may compute the *probability of choosing an alternative and accepting the chosen one* as easily as we computed the probability of choosing the true alternative.

As the region of acceptance of a hypothesis and that of its complement exclude each other, either of them will be accepted, so that e.g. $P(A(H_{II}) \cup A(H_{02})) = P(A(H_{II})) + P(A(H_{02})) = 1$. The chosen alternative is accepted, when H_{II} is chosen and accepted or when H_{02} is chosen and accepted. As our randomly choosing a hypothesis is independent of our accepting it,

$$\begin{aligned} & P((C(H_{II}) \cap A(H_{II})) \cup (C(H_{02}) \cap A(H_{02}))) \\ &= P(C(H_{II})) \cdot P(A(H_{II})) + P(C(H_{02})) \cdot P(A(H_{02})) \\ &= P(C(H_{II})) \cdot (P(A(H_{II})) + P(A(H_{02}))) \\ &= \quad 0.5 \quad \cdot \quad 1. \end{aligned}$$

Again the result is independent of the power of the test and of the prior probabilties of the respective alternatives, and it is valid for *any* hypothesis and its complement. Therefore it is also independent of the fact that "theories . . . in the so called 'soft' fields . . . are not quantitatively developed to the extent of being able to generate point-predictions" ([1], p. 113) and cannot give rise to a methodological paradox.

But we must not mistake the probability of choosing an alternative and accepting the chosen one for the *probability of choosing a specific alternative and accepting it*; nor must we mistake it for the *probability of accepting a specific alternative*; and, *a fortiori*, we must not mistake it for the *probability of accepting a specific alternative*, "*even if it is totally without merit.*"

4.4. Now we can turn to the test of H_{II} against H_{III}. It amounts to a modified test of the point-null hypothesis H_0. Its decision rule says that H_{II} is accepted, if the difference e.g. of sample means $\bar{x}_E - \bar{x}_C$ is greater than zero and falls outside the region of acceptance of H_0. Correspondingly H_{III} is accepted, if the difference is less than zero and falls outside the region of acceptance of H_0.

4.4.1. Therefore $P(A_n(H_{II})/H_0 \cup H_{III})$ is the *probability of falsely accepting* H_{II}. As $P(R_n(H_0)/H_0) = \alpha$, it approaches $\alpha/2$ as its upper limit, and correspondingly for H_{III}. On the other hand the *probability* $P(A_n(H_{II})/H_{II})$ *of correctly accepting* H_{II} approaches unity as the power of the test grows perfect, and correspondingly for H_{III}.

4.4.2. This time $P(C(H_{II})) = P(C(H_{III})) = 0.5$, but as $P(H_{II} \cup H_0 \cup H_{III}) = 1$, so $P(H_{II} \cup H_{III}) \leq 1$ and hence $P(C(H_{II})) \cdot (P(H_{II}) + P(H_{III})) \leq 0.5$. However, if we assume with Meehl that $P(H_0) = 0$, then $P(H_{II} \cup H_{III}) = 1$ and hence $P(C(H_{II})) \cdot (P(H_{II}) + P(H_{III})) = 0.5$. Therefore, $P((C(H_{II}) \cap H_{II}) \cup (C(H_{III}) \cap H_{III})) = 0.5$, i.e. the *probability of choosing the true test hypothesis from among* H_{II} *and* H_{III} *is* 0.5.

On the other hand the *probability of choosing the test hypothesis from among* H_{II} *and* H_{III} *and accepting the chosen one* depends on the power of the test. As $P(A(H_{II}) \cup A(H_0) \cup A(H_{III})) = 1$, so $P(A(H_{II})) + P(A(H_{III})) \leq 1$ and hence $P(C(H_{II})) \cdot (P(A(H_{II})) + P(A(H_{III}))) = P(C(H_{III})) \cdot (P(A(H_{II})) + P(A(H_{III}))) \leq 0.5$. But if we assume with Meehl that H_0 is false and that the test approaches perfection, so that $P(A(H_0)/H_{II} \cup H_{III})$ approaches zero, then $P(A(H_0))$ approaches zero too. In this case, however, $P(A(H_{II})) + P(A(H_{III}))$ approaches unity and $P((C(H_{II}) \cap A(H_{II})) \cup (C(H_{III}) \cap A(H_{III})))$ *approaches* 0.5. The result remains valid, when H_0,

H_{II} and H_{III} are replaced by $H_{04}: \mu_E - \mu_C = n$, $H_{IV}: \mu_E - \mu_C > n$ and $H_V: \mu_E - \mu_C < n$ respectively, where n can take any value.

So, if there is anything paradoxical, it is the procedure by which, according to Meehl, hypotheses on the value which some random variable can take are chosen for test. Out of the possibly infinitely many different hypotheses only one and its negation are preselected. Each of them is then given a probability of 0.5 of being finally chosen as the test hypothesis. As Meehl himself calls his model "preposterous" ([1], p. 110), we can only wonder why he makes his evaluation of statistical hypotheses depend on it.

Summary. Meehl's statement "In most psychological research, improved power of a statistical design leads to a prior probability approaching 1 ÷ 2 of finding a significant difference in the theoretically predicted direction" is without foundation. The computation of prior probabilities of accepting or rejecting a hypothesis presupposes knowledge of the prior probabilities that this hypothesis or any of its conceivable alternatives is true. As we do not have such knowledge, we cannot give any numerical values of prior probabilities of accepting or rejecting hypotheses in any statistical test procedure. Only topological statements are possible, as for example: when the region of acceptance of a hypothesis narrows, the prior probability that it will be accepted remains constant or increases. These statements are implied by the axioms of Probability Theory and so do not presuppose any knowledge of reality.

REFERENCES

[1] Carnap, R. *The Logic Foundations of Probability*. (2nd ed.). Chicago: University of Chicago Press, 1962.
[2] Feller, W. *An Introduction to Probability Theory and Its Applications*. (3rd ed.). Vol. I. New York: John Wiley & Sons, 1968.
[3] Kleene, S. C. *Introduction to Metamathematics*. New York: Van Nostrand, 1952.
[4] Kolmogorov, A. N. *Foundations of the Theory of Probability*. (2nd ed.). New York: Chelsea Publishing Co., 1956.
[5] Meehl, P. E. "Theory Testing in Psychology and Physics: A Methodological Paradox." *Philosophy of Science* 34 (1967): 103–115. (Reprinted in [6].)
[6] Morrison, D. and Henkel, R. E. (eds.). *The Significance Test Controversy: A Reader*. Chicago: Aldine Publishing Co., 1970.
[7] Popper K. R. *Conjectures and Refutations*. London: Routledge and Kegan Paul, 1963.
[8] Woodson, C. E. "Parameter Estimation vs. Hypothesis Testing." *Philosophy of Science* 36 (1969): 203–204.