STATISTICS IN MEDICINE
Statist. Med. 2002; **21**:2437–2444 (DOI: 10.1002/sim.1072)

LETTER TO THE EDITOR

A comment on replication, *p*-values and evidence

S.N. Goodman, *Statistics in Medicine* 1992; **11**:875–879

From: *Stephen Senn*
Department of Epidemiology and Public Health
Department of Statistical Science
University College London
1-19 Torrington Place
London WC1E 6BT, U.K.

Some years ago, in the pages of this journal, Goodman gave an interesting analysis of 'replication probabilities' of *p*-values. Specifically, he considered the possibility that a given experiment had produced a *p*-value that indicated 'significance' or near significance (he considered the range $p = 0.10$ to 0.001) and then calculated the probability that a study with equal power would produce a significant result at the conventional level of significance of 0.05. He showed, for example, that given an uninformative prior, and (subsequently) a resulting *p*-value that was exactly 0.05 from the first experiment, the probability of significance in the second experiment was 50 per cent. A more general form of this result is as follows. If the first trial yields $p = \alpha$ then the probability that a second trial will be significant at significance level α (and in the same direction as the first trial) is 0.5.

I share many of Goodman's misgiving about *p*-values and I do not disagree with his calculations (except in slight numerical details). I also consider that his demonstration is useful for two reasons. First, it serves as a warning for anybody planning a further similar study to one just completed (and which has a marginally significant result) that this may not be matched in the second study. Second, it serves as a warning that *apparent* inconsistency in results from individual studies may be expected to be common and that one should not overreact to this phenomenon.

However, I disagree with two points that he makes. First, he claims that 'the replication probability provides a means, within the frequentist framework, to separate *p*-values from their hypothesis test interpretation, an important first step towards understanding the concept of inferential meaning' (p. 879). I disagree with him on two grounds here: (i) it is not necessary to separate *p*-values from their hypothesis test interpretation; (ii) the replication probability has no direct bearing on inferential meaning. Second he claims that, 'the replication probability can be used as a frequentist counterpart of Bayesian and likelihood models *to show that p-values overstate the evidence against the null-hypothesis*' (p. 875, my italics). I disagree that there is such an overstatement.

SIGNIFICANCE TESTS AND HYPOTHESIS TESTS

In my opinion, whatever philosophical differences there are between significance tests and hypothesis test, they have little to do with the *use* or otherwise of *p*-values. For example,

Received June 2000
Accepted June 2001

Lehman in *Testing Statistical Hypotheses* [1], regarded by many as the most perfect and complete expression of the Neyman–Pearson approach, says

> 'In applications, there is usually available a nested family of rejection regions corresponding to different significance levels. It is then good practice to determine not only whether the hypothesis is accepted or rejected at the given significance level, but also to determine the smallest significance level $\hat{\alpha} = \hat{\alpha}(x)$, *the significance probability* or *p-value*, at which the hypothesis would be rejected for the given observation'. (reference [1], p. 70, original italics).

Similarly, just as using exact p-values is not incompatible with hypothesis testing we have no less an authority than Fisher himself that simply noting whether a result is significant or not is not incompatible with significance testing. In *Statistical Methods for Research Workers* [2], he takes Student's famous analysis [3] of the Cushny and Peebles data [4] and whereas Student, who was a Bayesian, calculated the exact p-value, Fisher, who was not, merely notes that the result is significant at the 1 per cent level.

In fact, it was Fisher, together with Yates, who introduced the practice of tabulating critical values corresponding to percentage points [5], thus showing that, if anything, Fisher felt it was less important to use exact p-values than Bayesians like Student and Karl Pearson. In short, it is not the use of p-values that distinguishes Fisher's system from that of Neyman and Egon Pearson – that is simply a choice between a dichotomy and (preferably) a continuous measure – it is rather the issue as to whether inferences or decisions are made.

Now consider a Fisherian conductor of significance tests and a Neymanite conductor of hypothesis tests. Suppose that each tests the same null hypothesis using the same statistic on the same data. Each can calculate a p-value for his or her different reasons: the Fisherian to make an inference; the Neymanite to permit others to come to a decision. This p-value will be the same. Suppose that the p-value is 0.05 and so (just) conventionally significant. Goodman's concept of a replication probability is not part of either of these systems of inference. It is not a likelihood as used by Fisher. It is not a power as used by Neyman and Pearson. However, it is closely related to the Bayesian idea of a predictive probability, and by stepping out of these two systems and into a third, you can calculate a Bayesian probability that the Fisherian will observe $p < 0.05$ next time an identical experiment is run and that the Neymanite will observe a result 'significant at $\alpha = 0.05$'. These two probabilities are identical, given the same prior, and thus, as should be obvious, cannot be the first or indeed any 'step to separate p-values from their hypothesis test interpretation' (Goodman, p. 879).

Similarly, although it is quite true that the probability of replicating a significant result is higher, other things being equal, given that one has merely noted on the first occasion 'result significant at the 5 per cent level' rather than $p = 0.05$ [6], this is a trivial mathematical consequence of the fact that the average p-value from all trials that are significant at the 5 per cent level is less than 0.05. It is thus a phenomenon of the same sort as the following. The average diastolic blood pressure, on re-measuring, of men who have been selected for treatment because their diastolic BP is *higher* than 100 mmHg will be higher, other things being equal, than that for a group whose diastolic BP on first measurement was *exactly* 100 mmHg.

REPLICATION PROBABILITIES AND INFERENTIAL MEANING

Replication probabilities are not of direct relevance to inferential meaning. They confuse the issue of making inferences. This is because we make inferences primarily about hypotheses or about the state of nature and not about future samples. However, in the context of parametric inference, a replication probability is a reflection of two things. The first is the likely or probable or reasonable (depending upon one's point of view) value of an unknown parameter. The second is the probabilistic distribution of a future test statistic given a particular value of the unknown parameter, but the second, for example, depends on the size of trial one happens to choose next time around. Goodman has considered the case where the second trial is the same size as the first. This may be a natural choice but it is not inferentially necessary. It would be absurd if our inferences about the world, having just completed a clinical trial, were *necessarily* dependent on assuming the following:

1. We are now going to repeat this experiment.
2. We are going to repeat it only once.
3. It must be exactly the same size as the experiment we have just run.
4. The inferential meaning of the experiment we have just run is the extent to which it predicts this second experiment.

DO p-VALUES OVERSTATE THE EVIDENCE AGAINST THE NULL HYPOTHESIS?

There are two answers to this. The first depends on accepting that the p-value is what frequentists say it is. Thus, if Fisherian, it is the probability of observing a result as extreme or more extreme than the result observed under the null hypothesis. If Neymanite, it is the most stringent possible type I error rate that one could entertain and still reject the null hypothesis. If this is accepted then the answer must be 'no'. The p-value does not *overstate* the evidence against the null hypothesis. The Bayesian criticism would then be that it does not actually *state* anything of relevance at all. The second answer depends on granting the Bayesian claim that p-values are interpreted by everybody as if they were Bayesian posterior probabilities. In that case it is not necessarily true that p-values overstate the evidence. The reason is that p-values can correspond to Bayesian posterior probabilities of a particular kind.

A simple example is in Student's famous paper mentioned above [3]. In analysing the data collected by Cushny and Peebles [4], and speaking of differences in hours of sleep gained comparing two treatments, Student writes, 'The mean value of this series is +1.58, while the S.D. is 1.17, the mean value being +1.35 times the S.D. From the table the probability is 0.9985, or the odds are about 666 to 1 that 2 is the better soporific'. Student was in the habit of calculating his standard deviation using the divisor n, so that we have to divide by $\sqrt{(n-1)}$ to obtain the standard error in modern terms. In his case n is 10, so we obtain $1.17/\sqrt{9} = 0.39$ and hence a t-value of 4.05 on 9 degrees of freedom [7]. I find a left-hand probability corresponding to this of 0.9986, so that Student is remarkably accurate in his calculations. The right-hand probability of $1 - 0.9986 = 0.0014$ is the one-tailed p-value and we could, if we wanted to, follow Student in giving it a Bayesian interpretation as the probability that treatment 2 is worse than treatment 1.

Bayesian statements of the sort that Student makes *must* demonstrate a Martingale property if the Bayesian uttering them is to be coherent. That is to say that, for a Bayesian

to demonstrate coherence in using such statements he has to *expect*, having observed (for example) $p = 0.0014$ that at whatever stage in the future he is asked the question 'what is the probability that 2 is worse than 1?', however much experimentation has been performed in the meantime, his answer will be 0.0014. Here *expect* has to be understood in the strict statistical sense. Put more formally, if p_1 is the probability *now*, at time point 1, that the treatment is effective and P_2 is a random variable between 0 and 1 describing the possible probability statements that will be issued *in future* at time point 2, then $E_1(P_2) = p_1$, where E_1 is expectation at time point 1. If this property did not apply, it would simply mean that future and current probabilities did not form a coherent set.

Suppose we consider the following three questions that a Bayesian might put having obtained a posterior probability of 0.05 that treatment 2 was worse than treatment 1:

Q1. What is the probability that in a future experiment, taking that experiment's results on its own, the results would favour 1 rather than 2?
Q2. What is the probability, having conducted this experiment, and pooled its results with the current one, the results would favour 1 rather than 2?
Q3. What is the probability that having conducted a future experiment and then calculated a Bayesian posterior using a uniform prior and the results of this second experiment alone, the *probability* that 2 would be worse than 1 would be less than or equal to 0.05?

Now, the answer to Q1, since the experiment is not infinitely large, is somewhat greater than 0.05. (If the trial is very small the probability will be nearly 0.5 because the result will be nearly all due to chance; for an infinitely large experiment the result will reflect which treatment is truly best and this probability is only 0.05 for treatment 1.) For the same reason, the answer to Q2 is somewhat less than 0.05. (Obviously, if the experiment is very small it will not be able to outweigh to any degree the current results so that the probability of this event will be nearly zero.) Neither of these two questions, however, is analogous to the question that Goodman asks about *p*-values. The analogous question is Q3. This question is not a question about the *confirmation* of a result; it is a question about the repetition of a probability associated with a result. In fact, if the experiment to come is the same size as the one that has been run, the answer to Q3 is 0.5, exactly Goodman's result for the *p*-value (see Appendix).

Figure 1 plots these replication probabilities against relative precision (on the log scale) of the second trial given that the first trial has a one-sided *p*-value of 0.05 (or equivalently a Bayesian posterior probability of the sort discussed above). Large trials are represented by low values on the abscissa and small trials by high values. A trial with exactly the same precision as the first trial is represented by the point 1. It can be seen that the value of Q3, which is Goodman's replication probability, is 0.5 for this case. (Assumptions and derivations are given in the Appendix.)

We do not, however, need this replication probability to be higher than 0.5 to believe that the efficacy of the treatment is probable. A long series of trials, 50 per cent of which were significant at the 5 per cent level, would be convincing evidence that the treatment was effective. Given an uninformative prior, followed by one significant result at the 5 per cent level, what the replication probability shows is that the probability that any one of the remaining trials chosen at random from the series is significant is 50 per cent (given that the results of the other trials are not known). Because the probabilities are not independent we

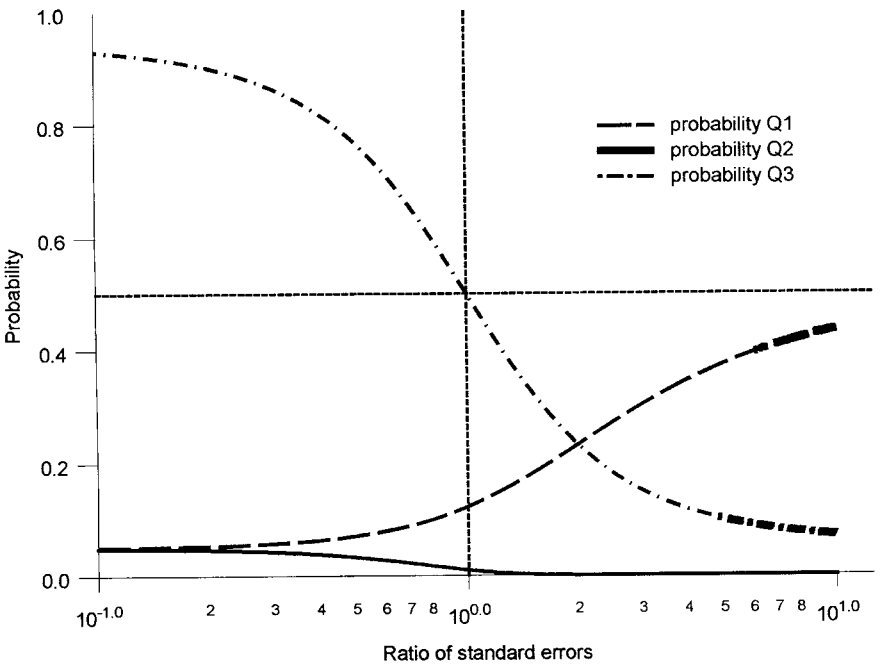

Figure 1. Probabilities for three questions as a function of the ratio of the standard error of the second trial to that of the first and given that $p = 0.05$ in the first trial. Note that the ratio is a log scale and that low values correspond to high relative precision of the subsequent trial.

cannot easily go further than this. However, what we can say is that the probability (that is to say our subjective assessment) that a very large meta-analysis would show that the treatment was effective would be 95 per cent.

THE TRUE PROBLEM WITH *p*-VALUES

The uninformative prior that Goodman considers causes no difficulties for *p*-values at all. The problem is rather that the 'uninformative' prior is rarely appropriate. In general, however, it is not possible to survive as a Bayesian on uninformative priors. It is a key feature of Jeffreys's approach to scientific inference, for example, that he recognized that some way had to be found for down-weighting the effect of higher order terms [8]. To give another example, this approach is essential in dealing with carry-over in analysing cross-over trials [9].

If, in testing the effect of a treatment in a clinical trial, we have a lump of probability on the treatment effect being zero, then, as is well known from the Jeffreys–Good–Lindley paradox, a *p*-value overstates the evidence against the null [10, 11]. So, of course, do Bayesian posterior statements of the sort made by Student [3].

The important distinction here is Cox's distinction between precise and dividing hypotheses [12]. In the Neyman–Pearson framework, the former corresponds to testing $H_0 : \tau = 0$ against

$H_1 : \tau \neq 0$, whereas the latter corresponds to testing $H_0 : \tau \leqslant 0$, against $H_1 : \tau > 0$. The Bayesian analogue of the first case it is to have a lump of probability on $\tau = 0$. Where such a probability is appropriate, then from a Bayesian perspective, the p-value will have most unfortunate properties.

It is important to realize, however, that the reason that Bayesians can regard p-values as overstating the evidence against the null is simply a reflection of the fact that Bayesians can disagree *sharply* with each other. For example, suppose in fact that we have two Bayesians who agree before seeing some data that the probability that the treatment is beneficial is 0.5. Given that the treatment is effective they have the same conditional prior distribution as to *how* effective it will be. However, one of them, the 'pessimist', believes that if not beneficial it may be *harmful*. On the other hand the other, the 'optimist' believes that if not beneficial it will be *harmless*. After running the trial the pessimist now believes with probability 0.95 that the treatment is beneficial, whereas the optimist now believes with probability 0.95 that it is useless.

The reason is that the result of the trial is marginally positive. For the optimist such a result could have easily arisen under the 'null', which is concentrated on zero. In fact, if most of the prior belief under the alternative corresponds to large treatment benefit, a moderate observed benefit is more likely under the null than under most of the alternative. Hence, the optimist is now inclined to believe the null. For the pessimist, however, such a result is even less likely under the 'null' than under the alternative, since both stretch away towards infinity from zero but the point estimate is in the alternative region. Hence, the pessimist is now inclined to believe the alternative hypothesis.

IN CONCLUSION

Although it is interesting to consider the repetition property of p-values, it is false to regard this as being relevant to separating p-values from their hypothesis test interpretation and it is false to regard the modest probability shown as being regrettable. It is desirable. Suppose it were the case that a low p-value brought with it a very high probability that it would be repeated. This would then imply that there was a very high probability that a meta-analysis using the current trial and the future one would produce an even lower p-value. This would mean that an anticipated result would have (nearly) the same inferential value as an actual one. One the contrary, the desirable property is that the meta-analysis should *confirm* the p-value of the current trial. (We can hope that it will do more but must also fear that it will do less.) However, this requires us to expect that the result of the new trial combined with the result we already have will leave us where we were. To think otherwise is to make exactly the same mistake a physician makes in writing of a result, 'the result failed to reach significance, $p = 0.08$ because the trial was too small'.

To expect that a future trial will be significant given that the current has yielded $p = 0.05$ is to expect that the future p-value will be at least as small as the current one. However, if we have an expectation that further experimentation will produce an even smaller p-value than the current trial, this is not because the p-value overstates the evidence against the null. On the contrary, it can only be because our prior belief is such that we consider it *understates* it. In other words, our prior belief enables us to recognise the p-value as being too pessimistic.

Goodman's replication property of p-values is interesting but this property should not be misinterpreted. There are many reasons for moving medical researchers on from their current excessive reliance on p-values, and Goodman, in a series of other papers, has given some persuasive arguments [13, 14]. If, however, we abandon p-values *because* of their failure to satisfy with high probability the repetition property of the *probability*, then this is a false reason to do so and, in moving to Bayesian methods as an alternative, we are likely to be disappointed.

APPENDIX: REPETITION PROBABILITIES FOR VARIOUS EVENTS

We assume that all nuisance parameters are known and that we are considering Normally distributed continuous outcomes. The observed treatment difference from the first experiment is $m_1 > 0$ with variance σ_1^2. The posterior distribution of the true mean is $\mathrm{N}(m_1, \sigma_1^2)$. The as yet unobserved mean of the second experiment is M_2 and will be observed with variance σ_2^2. We let $\lambda = \sigma_2/\sigma_1$ be the ratio of the two standard errors. The predictive distribution of this mean is $\mathrm{N}(m_1, \sigma_1^2 + \sigma_2^2)$. The posterior density of the true mean after carrying out the second experiment and observing $M_2 = m_2$, will be

$$\mathrm{N}\left(\frac{m_1\sigma_2^2 + m_2\sigma_1^2}{\sigma_1^2 + \sigma_2^2}, \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}\right)$$

Let Z be a standard Normal variable, let $\Phi(.)$ be the Normal distribution function and let $z_1 = m_1/\sigma_1$ be the observed standardized score for the first experiment:

Q1. We require $P(M_2 < 0)$.
Q2. We require

$$P\left(\frac{m_1\sigma_2^2 + M_2\sigma_1^2}{\sigma_1^2 + \sigma_2^2} < 0\right)$$

Q3. We require $P\{[P(\mu > 0|M_2) > P(\mu > 0|m_1)]|m_1\}$.

For Q1 we have

$$P(M_2 < 0) = P\left(Z < \frac{-m_1}{\sqrt{(\sigma_1^2 + \sigma_2^2)}}\right) = P\left(Z < \frac{-z_1}{\sqrt{(1 + \lambda^2)}}\right) \tag{A1}$$

Clearly, for an infinitely large experiment we have $\sigma_2^2/\sigma_1^2 = \lambda^2 = 0$, from which it follows that (A1) is simply $P(Z < -z_1)$, the posterior probability of inferiority as a result of having observed m_1. In other words, this is the 'p-value' for the first experiment. For a very small experiment we have very large λ and so (A1) is $P(Z < -\delta)$ where $\delta \approx 0$ so that this probability is close to 0.5.

For Q2

$$P\left(\frac{m_1\sigma_2^2 + M_2\sigma_1^2}{\sigma_1^2 + \sigma_2^2} < 0\right) = P\left(M_2 < -\frac{m_1\sigma_2^2}{\sigma_1^2}\right) = P\left(Z < -\frac{m_1\sqrt{(\sigma_1^2 + \sigma_2^2)}}{\sigma_1^2}\right)$$

$$= P(Z < -z_1\sqrt{(1 + \lambda^2)}) \tag{A2}$$

Now, for an infinitely large experiment so that $\sigma_2^2/\sigma_1^2 = \lambda^2 = 0$, (A2) reduces to $P(Z < -z_1)$ which is simply the posterior probability after the first experiment that treatment 2 is inferior to treatment 1: the p-value again. On the other hand, if we have a small experiment and σ_2^2 is large, we have $P(Z < -\gamma)$ where γ is large so that this probability is close to zero.

Now, for Q3 we have

$$P\left(1-\Phi\left(\frac{M_2}{\sigma_2}\right)<1-\Phi\left(\frac{m_1}{\sigma_1}\right)\right)=P\left(\Phi\left(\frac{M_2}{\sigma_2}\right)>\Phi\left(\frac{m_1}{\sigma_1}\right)\right)=P\left(\frac{M_2}{\sigma_2}>\frac{m_1}{\sigma_1}\right)$$
$$=P\left(Z>\frac{\frac{m_1\sigma_2}{\sigma_1}-m_1}{\sqrt{(\sigma_1^2+\sigma_2^2)}}\right)=P\left(Z>\frac{z_1(\lambda-1)}{\sqrt{(1+\lambda^2)}}\right) \quad \text{(A3)}$$

Where the second experiment is of the same precision as the first, so that $\lambda = 1$, this reduces to $P(Z > 0) = 0.5$. As the second experiment increases in size and λ approaches zero, this approaches $P(Z > -z_1) = 1 - \alpha$, where α is the posterior probability that the treatment is not effective. On the other hand, if the second experiment is small so that λ is very large, the probability approaches $P(Z > m_1/\sigma_1) = \alpha$.

REFERENCES

1. Lehmann EL. *Testing Statistical Hypotheses*. Chapman and Hall: New York, 1994.
2. Fisher RA. Statistical methods for research workers. In *Statistical Methods, Experimental Design and Scientific Inference*, Bennet JH (ed.). Oxford University: Oxford, 1925.
3. Student. The probable error of a mean. *Biometrika* 1908; **6**:1–25.
4. Cushny AR, Peebles AR. The action of optimal isomers. II. Hyoscines. *Journal of Physiology* 1905; **32**:501–510.
5. Fisher RA, Yates F. *Statistical Tables for Biological Agricultural and Medical Research*. Longman: Harlow, 1974. (First published Oliver and Boyd: Edinburgh, 1938.)
6. Royall RM. The effect of sample size on the meaning of significance tests. *American Statistician* 1986; **40**: 313–315.
7. Senn SJ, Richardson W. The first t-test. *Statistics in Medicine* 1994; **13**:785–803.
8. Jeffreys H. *Theory of Probability*. Clarendon Press: Oxford, 1961.
9. Senn SJ. Consensus and controversy in pharmaceutical statistics (with discussion). *Statistician* 2000; **49**: 135–176.
10. Lindley DV. A statistical paradox. *Biometrika* 1957; **44**:187–192.
11. Bartlett MS. A comment on D.V. Lindley's statistical paradox. *Biometrika* 1957; **44**:533–534.
12. Cox DR. The role of significance tests. *Scandinavian Journal of Statistics* 1977; **4**:49–70.
13. Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. *Annals of Internal Medicine* 1999; **130**:995–1004.
14. Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of Internal Medicine* 1999; **130**:1005–1013.

