A COMMENT ON REPLICATION, P-VALUES AND EVIDENCE

STEVEN N. GOODMAN

Johns Hopkins University School of Medicine, Department of Oncology, Division of Biostatistics, 550 N. Broadway, Suite 1103, Baltimore MD 21205, U.S.A.

SUMMARY

It is conventionally thought that a small p-value confers high credibility on the observed alternative hypothesis, and that a repetition of the same experiment will have a high probability of resulting again in statistical significance. It is shown that if the observed difference is the true one, the probability of repeating a statistically significant result, the 'replication probability', is substantially lower than expected. The reason for this is a mistake that generates other seeming paradoxes: the interpretation of the post-trial p-value in the same way as the pre-trial \( \alpha \) error. The replication probability can be used as a frequentist counterpart of Bayesian and likelihood methods to show that p-values overstate the evidence against the null hypothesis.

INTRODUCTION

There are few concepts more basic to medical research than the need for replication of results to establish the credibility of a hypothesis. For many researchers, replication means the repetition of a statistically significant result in a study of adequate power. Another central concept is that the credibility of a result increases as its associated p-value decreases. In this commentary we will explore the connection between this notion of replication and the p-value. We will show that, for a given p-value, the probability of replicating a significant result, under certain assumptions, is smaller than many expect. Understanding why this is so provides insight into the problems posed by the use of p-values within a hypothesis testing framework, and indirectly into the difficulties inherent in interpreting p-values as measures of evidence.

We start by exploring the logical connection between p-values and replication of statistically significant results. Using the Neyman–Pearson (NP) hypothesis testing framework for designing and interpreting experiments, we control the probabilities of making a type I or type II error. The structure of the hypothesis test guarantees that a true null hypothesis is unlikely to generate significant results, and conversely, if the alternative hypothesis is true, that studies of adequate power will usually result in significant findings.

Another frequentist approach to inference is to use a p-value as a measure of the degree of conflict of the data with the null hypothesis. For any observed effect, it is argued that the smaller the associated p-value, the more evidence there is against the null hypothesis, making the hypothesis more likely to be false. In addition, invoking hypothesis test logic, the more likely the null hypothesis is to be false based on one experiment, the more probable that a repetition of that experiment will result in a small p-value, and a significant result.

In practice, if a p-value is low enough to be considered significant, the observed effect is often claimed to be 'real', with the subsequent thought that a repetition of the experiment should, with
high probability, produce a similar statistical verdict. In the medical literature, the failure of an experiment to repeat the statistical significance achieved by previous studies often causes concern. A quantitative analysis will show that this concern is based on a fundamental misunderstanding of the p-value.

**PROBABILITY OF REPLICATING SIGNIFICANT RESULTS**

Suppose an intervention is evaluated by measuring the effect of applying it to one of two groups. To simplify calculations, we will presume that the observed difference x has a normal distribution, and that the standard errors of the observed effect under the null (μ = 0) and alternative (μ ≠ 0) hypotheses are equal. The experiment is designed to have size α (two-sided) and power 1 − β with respect to a difference Δ of interest. All p-values are two-sided and are based on a fixed sample size. Z_{α/2} is the positive critical value, and for most calculations we will use the conventional α = 0.05, Z_{0.025} = 1.96.

Assume that this experiment produces a 'statistically significant' result (p < α). The probability of repeating this result cannot be calculated without the unknown true effect, but we will start by assuming that the experimental result is correct; that is, that the measured difference x is equal to the unknown treatment effect μ. If we repeat this experiment under identical conditions (similar groups, same sample size, same intervention), what is the probability of observing another statistically significant result in the same direction as the first?

This probability is equal to the power of the study with respect to the observed difference. On a Z-score scale, if the observed difference is the true one, the distribution of future results will be normal with unit variance centred on the observed result $Z_x$ (Figure 1). The probability of repeating a significant result in the same direction is the area under that curve beyond $Z_{α/2}$, or $\Phi(|Z_x| - Z_{α/2})$. For a p-value of 0.01 and α = 0.05, $Z_x = 2.58$, $\Phi(2.58 - 1.96) = 0.73$. The probabilities associated with other p-values are shown in Table I. Note that because we have specified that the experiment be replicated, these results are independent of the power of the original study.

The second column in Table I shows Bayesian results under the more realistic scenario that we do not know that $μ = x$, and our uncertainty is modelled as a prior probability of $μ$ that is locally uniform, that is the distribution of $μ$ prior to the second experiment is proportional to its likelihood function in the first experiment, namely

$$ F(z_x) = \int_{-\infty}^{\infty} \int_{Z_{α/2}}^{\infty} \varphi(z - |z_x|) \varphi(y) \, dy \, dz, \quad (1) $$

where $z_x$ is the observed Z-score in the first experiment, $F(z_x)$ is the probability of $|Z| > Z_{α/2}$ (sign ($Z$) = sign $(z_x)$) upon repetition of an experiment in which $z_x$ was observed, and $\varphi(\cdot)$ is the standard normal density function.

**DISCUSSION**

The replication probabilities in Table I are much lower than most people expect, have a weak relationship to the size of their associated p-values, and are not in accord with the informal credibility assigned to the null hypothesis when these p-values are observed.¹ They show, for example, that if an experiment produces a correct result with $p = 0.01$, there is still greater than a one in four chance that a repetition of this experiment will not reach statistical significance at $α = 0.05$. We do not achieve a 95 per cent probability of replication until $p = 0.00032$, and 99 per
Figure 1. Shaded area is the probability of repeating a significant result in the same direction as the observed $x$, when $\mu = x$ (formula in text)

Table I. Probability of statistical significance ($p < 0.05$) upon repetition of an experiment as a function of the $p$-value of the first experiment, under two different distributions for the true effect size $\mu$ (see text)

<table>
<thead>
<tr>
<th>$P$-value of initial experiment</th>
<th>$\mu$ equals the difference $x$ observed in the first experiment</th>
<th>$\mu$ has a uniform prior distribution before the first experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.37</td>
<td>0.41</td>
</tr>
<tr>
<td>0.05</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>0.03</td>
<td>0.58</td>
<td>0.56</td>
</tr>
<tr>
<td>0.01</td>
<td>0.73</td>
<td>0.66</td>
</tr>
<tr>
<td>0.005</td>
<td>0.80</td>
<td>0.71</td>
</tr>
<tr>
<td>0.001</td>
<td>0.91</td>
<td>0.78</td>
</tr>
</tbody>
</table>

That this is a best-case scenario is confirmed by the Bayesian calculations, which more accurately reflect the fact that there is still uncertainty about the true difference after the first experiment. They show that when $p < 0.05$, the replication probabilities are even lower.

These data can be considered from another perspective, noting that in the standard normal sample size calculation, the $Z$-score of the difference $\Delta$ of interest is set equal to $Z_{\alpha/2} + Z_p$. When $\alpha = 0.05$, the alternative with 80 per cent power ($\Delta_{90}$) will have a $Z$-score of $1.96 + 0.84 = 2.8$, and for $\Delta_{90}$, $Z = 3.24$. The corresponding $p$-values are, respectively, 0.005 and 0.0012. Therefore, the $p$-value must be less than those small values for the observed difference to exceed the difference of interest. This highlights an interesting inconsistency of some researchers. When $\Delta_{90}$ is chosen to represent the minimum important difference (as opposed to the expected
difference) and a point estimate with $0.05 > p > 0.0012$ is obtained, meaning that a difference smaller than the minimum has been observed, few dismiss the observed effect as clinically unimportant.

If these results are surprising it may be because much of our intuition about the meaning of $p$-values is founded on their similarity to pre-trial error rates ($\alpha$). This is reinforced by their definition in many texts as 'observed error rates'. However, this similarity is only superficial, making the intuition wrong. We can show this by rephrasing the original question in terms of a pure hypothesis test, to demonstrate that the replication probability can be related to a true pre-trial error rate.

In an NP hypothesis test, there is no formal role for a $p$-value, and we are supposed to 'accept' or 'reject' the null hypothesis depending on whether the observation falls into a critical region, defined by the study design and the pre-trial $\alpha$ and $\beta$ errors. The precise location of the observation within the critical region (as indicated by an exact $p$-value) is not relevant. Therefore, in rephrasing the original question we must omit any information about the result other than its 'significance'. We now ask: if an experiment is performed, and we know that the null hypothesis was correctly rejected at a given $\alpha$ (that is, the alternative hypothesis is true), what is the probability of repeating this result in the same direction? If our hypothesis test is designed to choose between $H_0: \mu = 0$ and $H_1: \mu \geq \Delta_{1-\beta}$, the truth of $H_1$ implies, by definition, a replication probability of at least $1 - \beta$. Thus, by hiding from ourselves the actual observed difference, and knowing only whether (not where) a result fell in a 'significant' region, the replication probability when the alternative is true is equal to or greater than the pre-trial power with respect to that alternative, and the dissonance between our intuition and the numbers is eliminated. If the alternative is $H_1: \mu > 0$, its truth implies only that the replication probability exceeds $\alpha/2$, the 'power' against the null hypothesis.

The difference between knowing the exact result or a dichotomous 'significance' verdict has long been known, at least qualitatively. R. A. Fisher, a proponent of $p$-values as a basis for inference, was acutely aware that defining the $p$-value (significance test) as an 'observed alpha level' invited a connection with hypothesis tests ('acceptance procedures') that greatly confused their interpretation:

The conclusions drawn from [significance] tests constitute the steps by which the research worker gains a better understanding of his experimental material. . . . More recently, indeed, a considerable body of doctrine has attempted to explain, or rather to reinterpret, these tests on the basis of quite a different model, mainly as a means to making decisions in an acceptance procedure. The differences between these two situations seem to the author many and wide.  

Fisher was trying to make the distinction between controlling errors in the long run, the goal of hypothesis testing, and inference in the individual case, which he claimed the $p$-value could be used for. Neyman and Pearson also recognized this difference, although they stated that inference in the individual case was impossible using a frequency definition of probability. The textbook definitions of the $p$-value as an 'observed' type I error rate, with no discussion of the problem inherent in combining pre-trial (error rate) and post-trial ('observed') concepts in a single index, implicitly deny this distinction. This has produced confusion on issues as basic as how the interpretation of the $p$-value should be affected by sample size.

This confusion did not originate in the textbooks, since even Fisher himself was not able to make clear, if one rejected the 'observed error rate' interpretation, exactly how one was to use the $p$-value's magnitude as a measure of inferential meaning. We have used the replication
probability because it is a frequentist index that indirectly reflects such meaning. The fact that under favourable assumptions it is still substantially lower than many would guess from its associated p-value suggests that informal assessments of the evidential strength of p-values are usually exaggerated.

When we step out of the frequentist model, the connections between replication probability, inferential meaning and the p-value become easier to analyse. Fisher noted (elaborated by Barnard\textsuperscript{9}) that when the result is reported as a hypothesis test dichotomy the power curve is equivalent to the likelihood function. This is similar to the point that the replication probability is equal to the power against an alternative not adjacent to the null. Berger\textsuperscript{9,10} and others\textsuperscript{11-14} have used likelihood and Bayesian methods to show that the p-values overstate the evidence against the null hypothesis, and that the inferential meaning of a result which might be anywhere within a critical region is profoundly different from one whose precise place within the critical region is known. These kinds of arguments have not had a great impact on medical research, possibly because they involve non-frequentist notions of probability that most researchers either reject or poorly comprehend. However, all of these phenomena are reflected in the easily understood replication probability.

Unfortunately, because p-values and the critical regions of hypothesis tests are both tail area probabilities, they are easy to confuse. This confusion blurs the division between concepts of evidence and error for the statistician, and obscures it completely for nearly everyone else. 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.

REFERENCES