Power and Replication in Case-Control Studies

Jean-Marc Lalouel and Andreas Rohrwasser

The case-control study design, a common staple of epidemiology, is increasingly used to test for genetic association. The simplicity of the design accounts for both its appeal and its limitations. Too often, however, apparent controversy arises for lack of appreciation of basic tenets underlying statistical testing. Power and replication are two concepts most commonly ignored in evaluating such studies. We review the basic principles of statistical testing, recall simple means to calculate power, and provide numerical examples pertaining to the association between angiotensinogen and essential hypertension. Am J Hypertens 2002;15:201–205 © 2002 American Journal of Hypertension, Ltd.

Key Words: Association, case-control study, power, replication, genetics, hypertension, angiotensinogen.

It seems that no case-control study can be published without rapidly prompting qualifications, reservations, or outright disbelief from observers in the field. This should not be surprising, however, as it is the business of statistics to measure degree of belief, or disbelief, in competing hypotheses on the basis of data. A statistical test is constructed to evaluate the relative merit of the two hypotheses, in effect providing an objective measure of uncertainty.¹

A compounding challenge of statistical inference is that, in addition to the natural uncertainty about the hypotheses in competition, the necessity of sampling—rather than observing entire populations—introduces uncertainty of a different nature, namely, that of making inferences from samples to populations. The statistical test serves to assess the relative merit of competing hypotheses in the sample; to what reference population this inference extends remains to be established. The significance of sampling bias in limiting inference from samples to populations has been commonly discussed in the context of case-control studies, and this issue will not be revisited here.

A most obvious mechanism to validate a hypothesis is to subject it to further test through independent replication. As multiple studies and tests accumulate, it soon becomes necessary to evaluate the overall support for the hypotheses in competition. Methods have been developed to perform such an evaluation, including metaanalysis, but such attempts may be of limited practical utility without standardized review of individual patient data.² However simple it may be, the concept of replication remains exceedingly critical in such an exercise. Whether a study constitutes an appropriate replication of another is not the object of the test, and yet it may drastically affect its validity.

In what follows, our focus will be on two critical elements of statistical inference: elementary principles of statistical testing, particularly power, and the concept of replication. A specific hypothesis—the association between angiotensinogen and essential hypertension—is used to illustrate the practical significance of these concepts in the biomedical sciences.

Test of Association Genotypes and Alleles

Tests of association between disease and marker are performed by contrasting either genotypic or allelic frequencies between cases (C) and controls (R). The geneticist usually favors comparing allelic frequencies on the grounds that, in the absence of an association, sampling genotypes from a random-mating population is equivalent to sampling gametes, or alleles, from the gene pool that gave rise to the population. Depending on allelic frequency and penetrance, however, higher significance may be achieved when dominance or recessivity warrants pooling genotypic classes. Although our focus will be on comparison of allelic frequencies, the principles outlined extend to other situations.

The Test and the Decision

A test of association is constructed to evaluate the relative merit of two competing hypotheses, namely: $H_0$: $T$ occurs at equal frequency in cases and controls; $H_1$: $T$ occurs at different frequencies in the two groups.

Without evidence to the contrary, the first hypothesis,
also called the null hypothesis, must be contemplated. The first step in ascertaining relative support for the hypotheses in competition is to compute the familiar test of association from the two-way table of observed allelic frequencies in cases and controls. By construction, the statistic is directly proportional to sample size. The sample frequencies of allele T in cases and controls, although directly observed, depart from their true, unknown population values, as a result of sampling fluctuations. Under the null hypothesis—and a few ancillary conditions pertaining to convergence in law of the test statistic—the test follows the known probability distribution of a $\chi^2$ variable with one degree of freedom.

To proceed with hypothesis testing, then the value of $\chi^2$ with probability $P$ of being exceeded can be computed from this distribution or read off a reference table. Owing to chance alone, the statistic may exceed the value of 3.84 with a probability of 0.05 (5%), or the value 6.64 with a probability of 0.01 (1%).

The decision to favor one or the other hypothesis involves a risk of error. The Neyman-Pearson theory of statistical testing attempts to provide an objective assessment of this risk. It recognizes that two types of error can be performed when a decision is reached on the basis of a statistical test, referred to as type I and type II errors, respectively (Table 1).

Type I error occurs when a difference judged significant on the basis of the sample estimates does not actually exist in the population. Even when the null hypothesis ($H_0$) is true, there is a 5% chance that the statistic will appear significant at the .05 level—the assumed risk—due to chance fluctuations alone. Evidently, greater confidence in the decision reached can be achieved by requiring a more stringent significance level or by providing replication in one or more independent samples.

Type II error occurs when the null hypothesis is retained although it is false. The possibility of this type of error is not as well appreciated as the type I error by the average user of a statistical test. In simple language, it means that, although a difference may truly exist between cases and controls, it may not be detected as a statistically significant difference in a sample selected from the population. Evidently, everything else being the same, the probability of such an error depends on the actual magnitude of the difference in the population and on the size of the sample.

In the initial report of a significant association, the type I error is the primary concern because the test suggests the existence of a statistical difference. The central issue is the probability that this difference arises by chance alone. In an attempt at replication, however, the type II error becomes the dominant concern if the test statistic does not reach significance. A common logical fallacy consists of interpreting such an outcome as “negative evidence against the existence of an association.” When such finding is contrasted with the result of a previous report that provided evidence in support of a significant association (where the null hypothesis $H_0$ was rejected), it is common for the investigators to conclude that the data from the two studies are “conflicting,” thereby creating an apparent “contradiction” or “controversy.” We will see that, in most practical instances, negative evidence cannot be substantiated by a statistical test, and seemingly discordant results are in fact logically consistent and actually expected from statistical theory.

### Power

If $\beta$ stands for the probability of retaining $H_0$ when it is false ($H_1$ is true), or type II error, its complement, $1 - \beta$, or power, defines the probability of rejecting $H_0$ when it is false ($H_1$ is true). The probability that a test identifies as significant a given difference in the population frequency of a marker between cases and controls can be estimated by calculating the power of the test. For a given difference in marker frequency between cases and controls in the population, and for a specified sample size, power is the probability that the statistical test achieves significance at a defined level, say .05 or .01. Thus, an experiment with 80% power at the .05 level indicates that a true difference will be detected with a probability of 80%, or equivalently in four of five samples of similar size drawn from the same population. It is generally recognized that, unless it affords power of at least 80%, a study yielding a statistic that does not reach significance at the 0.05 level must be considered inconclusive. Such a study does not provide adequate information to select either one of the two competing hypotheses with sufficient confidence.

### Power Calculation

In spite of the fundamental importance of power in assessing a statistical outcome, it is usually not evaluated in reports that most critically depend on it, namely those in which the statistical test does not reach significance. Although the formula required is a bit more involved that the one used to generate the test, it can be performed with a scientific calculator having the appropriate trigonometric function; for calculators lacking this feature, a series expansion can be used. This is given in the Appendix, together with a numerical example.

### Replication

When the hypothesis of an association is advanced on the basis of a statistical test (ie, when the null hypothesis $H_0$ is

<table>
<thead>
<tr>
<th>Actual</th>
<th>Decision</th>
<th>Diagnosis</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$ is true</td>
<td>Keep $H_0$</td>
<td>OK</td>
<td>$1 - \alpha$</td>
</tr>
<tr>
<td>$H_0$ is true</td>
<td>Accept $H_1$</td>
<td>Type I error</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>$H_1$ is true</td>
<td>Keep $H_0$</td>
<td>Type II error</td>
<td>$\beta$</td>
</tr>
<tr>
<td>$H_1$ is true</td>
<td>Accept $H_1$</td>
<td>OK</td>
<td>$1 - \beta$</td>
</tr>
</tbody>
</table>
rejected and the alternative $H_1$ is retained), it can be submitted to further tests through replication in independent samples. Because population differences may obscure the conclusions, such replication is usually sought in a population of similar extraction; in the context of a genetic hypothesis, this means a similar ethnic background. Evidently, when environmental exposure and lifestyle significantly contribute to disease development, the epidemiology of the replicating population must also be representative of that of the original report.

For a study to constitute a true replication of a previous report, it must satisfy a minimum number of obvious criteria: 1) cases and controls should be defined strictly as in the original report, using the same inclusion and exclusion criteria; 2) the same sampling frame should be applied; and 3) the same, or a similar, reference population should be examined. As will be seen later, true replication is almost never achieved, leaving open unmeasured ambiguity in the actual significance of the statistical differences observed.

### Illustration: Angiotensinogen and Hypertension

#### Initial Finding

An association between allele T235 of the angiotensinogen gene (AGT) and essential hypertension was initially suggested as part of a report documenting genetic linkage in hypertensive siblings and association of T235 with both essential hypertension and plasma angiotensinogen concentration in two independent samples of whites ascertained in Salt Lake City, Utah, and Paris, France. Thereafter, the association of AGT T235 with various phenotypes and in various populations was subjected to multiple tests. Because tests reached significance in some but not all instances, the subject was commonly deemed controversial. We shall see, however, that such spread should be anticipated of any true association.

### The Relevance of Power

It is common to qualify data as controversial, contradictory, or to refer to positive and negative results when multiple tests of a given hypothesis do not consistently achieve statistical significance. For example, one report listed 56 topics considered contradictory because not all case-control studies examining the same topic provided significant support for an association. Such qualification is not appropriate, however. Even if independent samples were drawn from a population where a true association was known to exist, statistical significance would be expected in some, but generally not in all samples. As emphasized earlier, the lower the power, the greater the probability that the association will remain undetected.

To illustrate this point, we have calculated in Table 2 the power to detect as significant a given difference in the frequency of T235 between cases and controls, namely 11%, 9%, or 8% as reported in the initial report of Jeunemaitre et al., and two other studies by Schmidt and Johnson and their colleagues. We have included in Table 2 only the white studies reviewed in the meta-analysis of Kunz et al. that met minimum replication criteria (see below). It is clear that, overall, lack of significance correlates with low statistical power as a result of small sample size. Evidently, when sample size is low, the only appropriate conclusion is to consider the study inconclusive. Another clear point that appears in Table 2 is the rapid decline in power of $\beta_{11}$ as the true difference in marker frequency between cases and controls decreases in the population sampled.

It is worth noting that negative evidence against an initial finding, where a null hypothesis $H_0$ was rejected with a probability of error $\alpha$, would require rejection of the hypothesis of a true difference $H_1$ in subsequent replication attempts with a probability at least equal to $\alpha$. As the significance supporting rejection of $H_0$ in Jeunemaitre et al. was .001, this would require power of at least .999 to provide negative evidence against the initial finding, or a total of 3248 cases and controls in equal proportions.

<table>
<thead>
<tr>
<th>Source</th>
<th>$n$</th>
<th>$n_C$</th>
<th>$n_R$</th>
<th>$p_C$</th>
<th>$p_R$</th>
<th>$\chi^2$</th>
<th>$\alpha$</th>
<th>$1-\beta_{11}$</th>
<th>$1-\beta_9$</th>
<th>$1-\beta_8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jeunemaitre 1992$^3$</td>
<td>894</td>
<td>430</td>
<td>464</td>
<td>0.47</td>
<td>0.36</td>
<td>11.10</td>
<td>0.001</td>
<td>0.95</td>
<td>0.86</td>
<td>0.81</td>
</tr>
<tr>
<td>Johnson 1996$^7$</td>
<td>1510</td>
<td>778</td>
<td>732</td>
<td>0.39</td>
<td>0.31</td>
<td>8.77</td>
<td>0.01</td>
<td>1.00</td>
<td>0.97</td>
<td>0.95</td>
</tr>
<tr>
<td>Jeunemaitre 1997$^5$</td>
<td>1682</td>
<td>954</td>
<td>728</td>
<td>0.47</td>
<td>0.38</td>
<td>12.60</td>
<td>0.001</td>
<td>1.00</td>
<td>0.98</td>
<td>0.96</td>
</tr>
<tr>
<td>Schmidt 1995$^5$</td>
<td>900</td>
<td>438</td>
<td>462</td>
<td>0.48</td>
<td>0.40</td>
<td>5.38</td>
<td>0.03</td>
<td>0.96</td>
<td>0.86</td>
<td>0.81</td>
</tr>
<tr>
<td>Hingorani 1996$^{11}$</td>
<td>810</td>
<td>436</td>
<td>374</td>
<td>0.42</td>
<td>0.45</td>
<td>0.62</td>
<td>n.s.</td>
<td>0.94</td>
<td>0.83</td>
<td>0.77</td>
</tr>
<tr>
<td>Fornage 1995$^{10}$</td>
<td>598</td>
<td>208</td>
<td>390</td>
<td>0.41</td>
<td>0.40</td>
<td>0.04</td>
<td>n.s.</td>
<td>0.83</td>
<td>0.68</td>
<td>0.62</td>
</tr>
<tr>
<td>Jeunemaitre 1993$^{16}$</td>
<td>452</td>
<td>272</td>
<td>180</td>
<td>0.44</td>
<td>0.38</td>
<td>1.50</td>
<td>n.s.</td>
<td>0.75</td>
<td>0.60</td>
<td>0.54</td>
</tr>
<tr>
<td>Bennett 1993$^9$</td>
<td>374</td>
<td>184</td>
<td>190</td>
<td>0.42</td>
<td>0.39</td>
<td>0.24</td>
<td>n.s.</td>
<td>0.70</td>
<td>0.54</td>
<td>0.49</td>
</tr>
<tr>
<td>Barley 1994$^{17}$</td>
<td>280</td>
<td>132</td>
<td>148</td>
<td>0.51</td>
<td>0.49</td>
<td>0.06</td>
<td>n.s.</td>
<td>0.59</td>
<td>0.45</td>
<td>0.40</td>
</tr>
<tr>
<td>Caulfield 1994$^{15}$</td>
<td>254</td>
<td>126</td>
<td>128</td>
<td>0.51</td>
<td>0.49</td>
<td>0.06</td>
<td>n.s.</td>
<td>0.55</td>
<td>0.42</td>
<td>0.38</td>
</tr>
</tbody>
</table>

$n$ = total number of alleles tested; $n_C$, $n_R$ = number of alleles tested in cases and controls respectively; $p_C$, $p_R$ = frequency of T235 in cases and controls respectively; $\chi^2$ = test of association; $\alpha$ = significance of the test; $1-\beta_{11}$ = power of test when the true difference ($x$) of T235 frequency in cases and controls for $x = 0.11$, $0.09$, $0.08$, respectively; n.s. = not significant.
The Meaning of Replication

Attempts at replication should reproduce the original definition of cases and controls. In the study by Jeunemaitre et al., cases presented with clinically defined hypertension, and exclusion criteria included diabetes or causes of secondary hypertension. Furthermore, all cases had at least one sibling with essential hypertension meeting similar criteria. Controls were individuals with normal blood pressure (BP) and no diagnosis of hypertension. Only studies based on clinically defined hypertension were included in Table 2. Only one study included affected first-degree relatives as part of the selection. The additional selection criterion that both parents be hypertensive, in the face of etiologic heterogeneity and dominance/recessivity relationships, may be distinctly different from a criterion applied only to siblings, however.

Evidently other factors than power may have affected the outcome of the test in the studies listed in Table 2. In the study by Fornage et al., for example, where no difference in the frequency of T235 between cases and controls was observed, it is interesting to note that this difference was significant between men and women in the normotensive controls (0.47 vs 0.36, \( \chi^2 = 4.8, P < .05 \)) as well as in the entire sample (0.46 vs 0.37, \( \chi^2 = 5.32, P < .03 \)), suggesting another confounding factor at play between the two groups. The reason for lack of significance despite good power in the study by Hingorani et al. remains less evident; they have proposed that their study may be more representative of the wider population of hypertensive subjects normally managed by family physicians.

Reports not included in Table 2, their intrinsic merit notwithstanding, used definitions of cases arguably quite distinct from that of Jeunemaitre et al. A number of studies used casual, single measurement of BP as phenotype, clearly not an equivalent of clinically defined hypertension. In Tiret et al., hypertensive cases were selected among survivors of myocardial infarction. In the study conducted by Kiema et al. in Finland, cases and controls were recruited indirectly through examination of Social Insurance registers. Cases were defined by use of the proxy variable of a reimbursement rate of 75% or greater for hypertension medication. As clearly stated by them, the likelihood of qualifying for this reimbursement rate critically depends on such factors as target organ damage, family history of cardiovascular disease and sudden death, hyperlipidemia, or diabetes (an exclusion criterion in the initial study of Jeunemaitre et al.). Controls were selected randomly from the same register, attempting to exclude subjects with the right to reimbursement for hypertension medication. Yet, Kiema et al. indicated that more than 10% of these controls were on chronic medication affecting BP. Surprisingly, more than half of these controls had a systolic BP in excess of 140 mm Hg, and on the basis of the data reported it can be estimated that less than 50% of these controls would qualify as normotensive subjects, with systolic BP < 140 mm Hg and diastolic BP < 90 mm Hg.

Conclusion

Hypotheses advanced on the basis of a statistical test can be rigorously subjected to independent replication, but this requires care to ensure that the same outcome is under examination. The results of the tests must be interpreted cautiously with appropriate reference to the Neyman-Pearson theory of statistical tests. Power is of critical relevance in such exercises. A good understanding of these concepts is essential in assessing scientific evidence.

If a single point was to be made to conclude this discussion, it is that variability in the statistical significance of replication attempts is simply not controversial. Rather, it is an expected feature of hypothesis testing in finite samples. Any additional source of heterogeneity, whether in population, sampling, or definition of cases and controls, can only further compromise our ability to make inferences from sampled observations.

Appendix

Power Calculation in a Two-Way Contingency Tables

Power Calculation

Let \( p_C \) and \( p_R \) denote the frequency of allele T in cases and controls from the reference population, respectively. Likewise, define \( n_C \) and \( n_R \) as the number of genes tested in cases and controls, respectively. If \( n_C = n_R \), define \( n = n_C \), otherwise calculate the geometric mean:

\[
   n = 2n_C n_R / (n_C + n_R).
\]

Using the arcsine transformation \( \theta = 2 \arcsin(p^{1/2}) \), calculate \( \theta_C \) and \( \theta_R \), and \( h = \theta_C - \theta_R \). The arcsine function is also denoted inverse sine, or \( \sin^{-1} \), on most calculators, and evidently in the present situation the calculator should be set to use radian rather than degree. If not available on a calculator, a Taylor’s series approximation can be used: \( \sin^{-1} p = p + p^3/6 + 3 p^5/40 + 5 p^7/112 \). Compute:

\[
   z_{1-\alpha} = h(n/2)^{1/2} - z_{1-\alpha}.
\]

The quantity \( z_{1-\alpha} \) gives the probability that a normal variable with zero mean and unit variance be smaller than \( 1-\beta \). Having computed the former, the latter, or power, is obtained by reference to a table of the normal distribution. Given reported values of \( p_C \) and \( p_R \), the formula can be used to calculate the expected power of a planned study for any given values of the sample sizes \( n_C \) and \( n_R \). The reader may consult a standard statistical textbook such as Cohen for more details.
A Numerical Example

To illustrate how power was computed in our synopsis table, we will consider the study by Caulfield et al.\textsuperscript{15} We want to calculate the power to detect in this study a specified difference in the frequency of the T235 marker in cases and controls among whites. We take as best gene frequency estimates those derived from the initial report by Jeunemaitre et al.,\textsuperscript{3} namely $p_C = 0.47$ and $p_R = 0.36$. The number of alleles tested in the study by Caulfield et al.\textsuperscript{15} were $n_C = 126$ and $n_R = 128$ in cases and controls, respectively (see Table 2). Following our section on power calculation, we compute:

\[
\begin{align*}
\theta_C &= 2 \arcsin \left( \frac{p_C^{1/2}}{2} \right) = 1.5108 \\
\theta_R &= 2 \arcsin \left( \frac{p_R^{1/2}}{2} \right) = 1.2870 \\
h &= \theta_C - \theta_R = 0.2238 \\
n &= \frac{2n_Cn_R}{(n_C + n_R)} = 126.99
\end{align*}
\]

To calculate power at a significance level of $\alpha = 0.05$, we obtain $z_{1-\alpha}$ from a table of the normal distribution $z_{1-0.05} = 1.65$ (the value of $t$, in the table, for which the shaded area under the curve is equal to 0.95). Using formula (2), we obtain:

\[
z_{1-\beta} = 0.1824 \cdot 7.9581 - 1.65 = 0.1330.
\]

This gives the value of $t$ to be referred to in the table of the normal distribution to obtain the power $1-\beta = 0.55$.

References