Estimation From Samples

Lisa M. Sullivan

Circulation 2006;114;445-449
DOI: 10.1161/CIRCULATIONAHA.105.600189
Circulation is published by the American Heart Association. 7272 Greenville Avenue, Dallas, TX 72514
 Print ISSN: 0009-7322. Online ISSN: 1524-4539

Estimation From Samples

Lisa M. Sullivan, PhD

Estimation is the process of determining a likely value for a population parameter (eg, the true population mean or proportion) based on a random sample. In practice, a sample is drawn from the target population, and sample statistics (eg, the sample mean or sample proportion) are used to generate estimates of the unknown parameter. The sample should be representative of the population, ideally with participants selected at random from the population. Because different samples can produce different results, it is necessary to quantify the sampling error or variation that exists among estimates from different samples.

Appropriate classification of the key study variable, also referred to as the outcome or end point, as continuous or discrete is critically important in estimation and in other statistical applications. Once a variable or outcome is correctly classified, other issues such as the number of comparison groups and whether those groups are independent or dependent (ie, matched or paired) affect the determination of the appropriate estimation technique.

Two types of estimates can be produced for any population parameter: point estimates and confidence interval (CI) estimates. A point estimate for a population parameter is a single-valued estimate of that parameter. A CI estimate is a range of values for a population parameter with a level of confidence attached (eg, 95% confidence that the interval contains the unknown parameter). The CI starts with the point estimate and builds in a margin of error that incorporates the confidence level and the sampling variability or standard error. CIs are presented below for different types of variables. Sample size determination and issues related to interpretation and precision follow.

The Basics

A CI is a range of values that are likely to cover the true population parameter, and the general form is point estimate±margin of error. The point estimate is determined first. For estimating a population mean or proportion, the point estimates are the sample mean or sample proportion, respectively. Next, a level of confidence is selected that reflects the likelihood that the CI contains the true, unknown parameter. Usually, confidence levels of 90%, 95%, and 99% are chosen, although theoretically any likelihood can be selected. The confidence level is often represented as $100(1-\alpha)\%$, where α is the level of significance in a 2-sided test of hypothesis. For example, a 2-sided test with $\alpha=0.05$ corresponds to a $100(1-0.05)=95\%$ confidence level. For large samples, the 95% CI takes the following form:

(1) Point estimate±1.96 SE (point estimate)

where 1.96 is the value from the standard normal distribution reflecting 95% probability and SE (point estimate) is the standard error of the point estimate. The value 1.96 is determined from probability tables or statistical algorithms.[1] In the standard normal distribution, 95% of the area under the curve lies between −1.96 and 1.96. Many textbooks use the notation $Z_{1-\alpha/2}$, where $1-\alpha/2$ is the lower tail area. For example, $Z_{0.975}=1.96$.[1] When the sample size is small ($n<30$), an appropriate value is selected from the t distribution (as opposed to the standard normal distribution). The t value is determined in a similar fashion based on the desired confidence level (eg, 95%, 99%), as well as the exact sample size (smaller samples will have larger t values translating to larger margins of error).[1,2] The standard error can be estimated from the sample and depends on the sampling method used, the estimation technique used, the sample size, and the variability of the characteristic being estimated. Nonsampling error also is a factor that affects the precision of an estimate. Nonsampling error includes error resulting from nonresponse or loss to follow-up. Unfortunately, nonsampling error often is impossible to quantify; however, it should be considered when estimates are interpreted from samples.

Sample Data

Data in Table 1 were measured on participants in the Framingham Heart Study at Offspring Examinations 4 and 5.[3] The examinations were conducted between April 1987 to September 1991 and January 1991 to August 1995, respectively. A total of 4019 participants attended examination 4; 3799 attended examination 5; and a total of 3626 attended both. Means and standard deviations are presented for continuous variables, and numbers and percents of participants responding affirmatively or with the condition of interest are presented for dichotomous variables. These data are used to illustrate estimation techniques in the sections that follow.

Estimates for Continuous Outcomes

It is of interest to estimate the mean of a continuous variable in a single population, the difference in means when there are 2 independent populations, and the mean difference when there are 2 dependent, matched, or paired populations. The last can arise in designs in which each participant is measured

From the Department of Biostatistics, Boston University School of Public Health, Boston, Mass.

Correspondence to Lisa M. Sullivan, PhD, Boston University, School of Public Health, Department of Biostatistics, 715 Albany St, Boston, MA 02118. E-mail lsull@bu.edu

(*Circulation.* **2006;114:445-449.**)

Circulation **is available at http://www.circulationaha.org** **DOI: 10.1161/CIRCULATIONAHA.105.600189**

TABLE 1. Description of Participants Attending Framingham Offspring Examinations 4 and 5

Characteristic	Examination 4* (n=3626)	Examination 5† (n=3626)
Dates	Apr 1987–Sep 1991	Jan 1991–Aug 1995
Mean (SD) age, y	51.4 (10.0)	55.6 (10.0)
Male, n (%)	1724 (47.6)	1724 (47.6)
Mean (SD) SBP, mm Hg	126.5 (18.6)	126.3 (18.8)
Mean (SD) DBP, mm Hg	79.1 (10.0)	74.5 (10.0)
On antihypertensive therapy, n (%)	638 (17.6)	703 (19.5)
With diabetes, n (%)	241 (6.7)	333 (9.2)
With prevalent CVD, n (%)	269 (7.4)	358 (9.9)

*n=4019 attended examination 4 (n=3626 attended both examinations 4 and 5).

†n=3799 attended examination 5 (n=3626 attended both examinations 4 and 5).

twice (eg, in a crossover trial in which each participant is measured under 2 different treatments or at 2 different points in time such as in the Framingham study) or when matched pairs are formed and each member of each pair is measured.

Estimating the Mean in a Single Population

If a continuous outcome is measured in a single sample, the CI for the mean of that variable in the population is given by the following:

$$(2) \qquad \overline{X} \pm t_{1-\alpha/2} SE(\overline{X})$$

Where $\overline{X}$ is the mean of the characteristic in the sample, $t_{1-\alpha/2}$ is the value from the t distribution with lower tail area equal to $1-\alpha/2$ reflecting the desired confidence level (eg, for large samples and 95% confidence, $t_{0.975}=1.96$), and $SE(\overline{X})$ is the standard error or standard deviation of the sample mean. An estimate of the SE is as follows: $SE(\overline{X}) = \frac{s}{\sqrt{n}}$, where s is the standard deviation of the outcome of interest. Equation 2 is appropriate when either the sample size is large or the outcome of interest is approximately normally distributed. If the outcome is highly nonnormal and the sample size is small, then a transformation (eg, natural log) might be appropriate to promote normality before computing the CI.

Example 1

Estimate the mean systolic blood pressure (SBP) in the early 1990s based on data collected in the Framingham Heart Study. Using data measured at examination 5, we can construct a CI estimate using Equation 2 as follows: $126.3 \pm 1.96 \frac{18.8}{\sqrt{3626}}$, 126.3 ± 0.61, or (125.7 to 126.9). The margin of error is extremely small here because of the large sample size.

Estimating the Difference of Means in 2 Independent Populations

If a continuous outcome is measured in 2 independent samples, the CI for the difference in means in the respective populations is given by the following:

$$(3) \qquad (\overline{X}_1 - \overline{X}_2) \pm t_{1-\alpha/2} SE(\overline{X}_1 - \overline{X}_2)$$

where $\overline{X}_1$ and $\overline{X}_2$ are the means of the characteristic in the independent samples, $t_{1-\alpha/2}$ is the value from the t distribution reflecting the desired confidence level, and $SE(\overline{X}_1 - \overline{X}_2)$ is the standard deviation of the difference in sample means. $SE(\overline{X}_1 - \overline{X}_2) = Sp\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$, where Sp is the pooled estimate of the standard deviation of the outcome (assuming that the standard deviations in the populations are similar) computed as the weighted average of the standard deviations in the samples $\left(Sp = \sqrt{\frac{(n_1-1)s_1^2 + (n_1-1)s_2^2}{n_1+n_2-2}}\right)$. The null (or no effect) value of the CI for the difference in means[3] is zero. If a 95% CI for the difference in means does not include zero (the null value), then the difference in means is statistically significant at the 5% level of significance. The next articles in this statistical series will discuss hypothesis testing in detail.

Example 2

Estimate the difference in mean blood pressures in persons taking and not taking antihypertensive therapy. Using data collected at examination 5, we know that there are n=703 persons (19.5%) on antihypertensive therapy, whereas the remaining n=2909 are not. Their mean (SD) SBPs are 137.9 (19.8) and 123.5 (17.5) mm Hg, respectively (data not shown in Table 1). A 95% CI for the difference in mean SBP between persons taking and not taking antihypertensive therapy is given by the following: $(137.9-123.5) \pm 1.96(0.76)$, 14.4 ± 1.49, or (12.9 to 15.9). This difference is statistically significant at the 5% level of significance (because the CI does not include zero) with persons taking antihypertensive therapy having a mean SBP that is 14.4 U higher, on average, than the mean for persons not taking antihypertensive therapy.

Estimating the Mean Difference in Matched or Paired Populations

If a continuous outcome is measured twice in a single sample or if samples are matched or paired and the characteristic is measured on each participant in each matched pair, the CI for the mean difference of that variable is given by the following:

$$(4) \qquad \overline{X}_d \pm t_{1-\alpha/2} SE(\overline{X}_d)$$

where $\overline{X}_d$ is the mean difference score (where differences are computed on each participant or between members of a matched pair), $t_{1-\alpha/2}$ is the value from the t distribution reflecting the desired confidence level, $SE(\overline{X}_d)$ is the standard deviation of the mean difference and equivalent to $\frac{S_d}{\sqrt{n}}$, and S_d is the standard deviation of the difference scores. The null value in the CI for the mean difference is zero.

Example 3

Estimate change in systolic and diastolic blood pressure (DBP) over 4 years using data collected in the Framingham Heart Study. To construct the CIs for the mean differences in SBP and DBP, difference scores must be computed for each participant. This is done here by subtracting the SBPs and

DBPs measured at examination 4 from those measured at examination 5. A positive difference reflects an increase over time; a negative difference reflects a decrease. The means (standard deviations) of the difference scores for SBP and DBP are −0.11 (14.78) and −4.49 (9.06), respectively. The SBPs did not change much over time, whereas DBPs decreased by 4.49 U on average over time. The standard deviations are the standard deviations of the difference scores. A 95% CI estimate of the mean difference in SBP over 4 years is given by the following: $-0.11 \pm 1.96 \frac{14.78}{\sqrt{3626}}$, -0.11 ± 0.48, or (−0.59 to 0.37). The change in SBP over time is not statistically significant at the 5% level. A 95% CI estimate of the mean difference in DBP over 4 years is given by: $-4.49 \pm 1.96 \frac{9.06}{\sqrt{3626}}$, -4.49 ± 0.29, or (−4.78 to −4.20), which is statistically significant at the 5% level (CI does not include zero).

In estimation and other statistical inference applications, it is critically important to appropriately identify the unit of analysis. Units of analysis are independent entities. In the 1-sample and 2-independent-samples applications, participants are the units of analysis. In the 2-dependent-samples application, the pair is the unit (and not the number of measurements, which is twice the number of units).

Estimates for Dichotomous Outcomes

When the outcome of interest is dichotomous, estimates of the population proportion are produced. If there is a single population, estimates of the proportion in that population are produced; if there are 2 independent populations, estimates of the difference in proportions or the ratio of proportions are produced.

Estimating the Proportion in a Single Population

If a dichotomous outcome is measured in a single sample, it is of interest to generate an estimate of the proportion in the population based on data observed in the sample. The CI is given by the following:

$$\hat{p} \pm Z_{1-\alpha/2} SE(\hat{p}) \quad (5)$$

where $\hat{p}$ is the sample proportion, $Z_{1-\alpha/2}$ is the value from the Z distribution reflecting the desired confidence level (eg, for 95% confidence $Z_{0.975}=1.96$), and $SE(\hat{p})$ is the standard deviation of the sample proportion, which is $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. The above is appropriate for large samples, defined in these applications as at least 5 participants in each of the 2 response categories. When there are <5 positive or negative responses, then exact methods based on the binomial distribution as opposed to a normal approximation should be used to estimate the population proportion.[4]

Example 4

Estimate the prevalence of cardiovascular disease (CVD) in 1992. Using data collected at examination 5, we know that n=358 (9.9%) of the sample had prevalent CVD. A 95% CI estimate for the prevalence of CVD is given by 0.099±0.055 or (0.089 to 0.108), equivalent to (8.9% to 10.8%).

Estimating the Difference in Proportions in 2 Independent Populations

Several measures are used to compare proportions in 2 independent populations. The absolute difference, sometimes called the risk difference or excess risk, is computed by taking the difference in proportions between comparison groups and is similar to the estimate of the difference in means for a continuous outcome.[3] The risk difference can be interpreted as the excess risk of outcome associated with the characteristic that defines the groups. The relative risk, also called the risk ratio, is another useful measure to compare proportions between 2 independent populations. It is computed by taking the ratio of proportions. Generally, the reference group (eg, unexposed persons, persons without a risk factor, or persons assigned to the control group in a clinical trial setting) is considered in the denominator of the ratio. The relative risk is often thought to be a better measure of the strength of an effect than the risk difference because it is relative to a baseline or comparative level.

The CI for the difference in risks is given by the following:

$$(\hat{p}_1 - \hat{p}_2) \pm Z_{1-\alpha/2} SE(\hat{p}_1 - \hat{p}_2) \quad (6)$$

where $\hat{p}_1$ and $\hat{p}_2$ are the sample proportions in the independent samples, $Z_{1-\alpha/2}$ is the value from the Z distribution reflecting the desired confidence level, and $SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$. This formula is appropriate for large samples, defined as at least 5 participants in each of the 2 response categories in each sample. The null value in the CI for the risk difference is zero.

Example 5a

Estimate the difference in incidence of CVD over 4 years between persons with and without diabetes. There are n=3357 participants free of CVD at offspring examination 4 (n=3626−269 with prevalent CVD). Each participant is followed up for 4 years for the development of CVD. There are 24 of 271 (8.9%) incident CVD events in participants with diabetes and 65 of 3086 (2.1%) incident CVD events in participants free of diabetes (data not shown in Table 1).

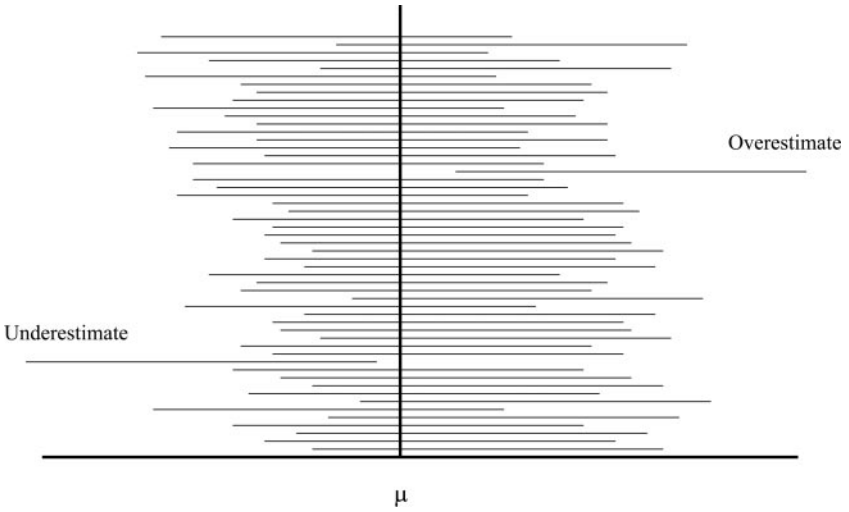
A 95% CI for the difference in risks of incident CVD between diabetics and nondiabetics is (0.089−0.021)± 1.96(0.017), 0.068±0.033, or (0.035 to 0.101), equivalent to (3.5% to 10.1%). The excess risk in incident CVD attributable to diabetes is between 3.5% and 10.1%, which is statistically significant at the 5% level. Note that the above does not account for other confounding factors such as age, sex, blood pressure, and smoking.

A point estimate for the population relative risk is given by $\hat{RR} = \frac{\hat{p}_1}{\hat{p}_2}$. The relative risk is a ratio and is not normally distributed. The natural log (Ln) of the $\hat{R}$ R is approximately normally distributed and is used to produce the CI:

$$\ln(\hat{RR}) \pm Z_{1-\frac{\alpha}{2}} \sqrt{Var[\ln(\hat{RR})]} \quad (7)$$

The variance is[5]

$$Var[\ln(\hat{RR})] = \sqrt{\frac{(1-\hat{p}_1)}{n_1\hat{p}_1} + \frac{(1-\hat{p}_2)}{n_2\hat{p}_2}}.$$



Interpreting CIs (95% CI for μ).

The limits of the $100(1-\alpha)\%$ CI for the population relative risk are produced by taking the antilog of the upper and lower limits produced by the above. The null value in the CI for the relative risk is 1. If a 95% CI for the relative risk does not include 1, the risks are statistically significantly different at the 5% level of significance.

Example 5b

Estimate the relative risk of incident CVD over 4 years between persons with and without diabetes. A point estimate for the relative risk of incident CVD in diabetics compared with nondiabetics is $\frac{0.089}{0.021} = 4.2$. A 95% CI for the relative risk of incident CVD for diabetics compared with nondiabetics is computed as follows: $\mathrm{Ln}\left(\frac{0.089}{0.021}\right) \pm 1.96\ (0.230)$, equivalent to 1.44 ± 0.45 or (0.99 to 1.90). Taking the antilog of each limit (eg, $e^{0.99}=2.69$) produces the CI for the relative risk: 2.69 and 6.68, which is statistically significant at the 5% level. Again, this relative risk is not adjusted for other confounding factors.

In some study designs, it is not possible to estimate a relative risk (eg, the case-control study). This and other issues related to estimating relative risks and odds ratios, as well as issues related to interpretation, can be found in Rothman and Greenland's book.[6]

Interpreting the Confidence Level

The Figure shows 40 different 95% CIs for the mean of a population, μ. In theory, when a 95% confidence level is used, 38 (95% of 40) CIs will cover or include the true mean μ. In practice, 1 random sample is selected, and a single CI is produced. The interval may or may not cover the true mean; the observed interval may overestimate or underestimate μ. The 95% CI is the likely range of the true, unknown parameter. It is important to note that a CI does not reflect the variability in the unknown parameter; rather, it provides a range of values that is likely to include the unknown parameter.

Sample Size Determination for Estimation

If the goal of an analysis is to generate an estimate of an unknown population parameter, the number of participants required to ensure a prespecified level of precision should be determined before data are collected. Formulas to determine sample sizes required to estimate a mean, the difference between means, the mean difference, a proportion, and the difference in proportions are given in Table 2. To implement the formulas for means, an appropriate estimate of the population standard deviation (σ) is required. Suitable estimates are based on values reported from other comparable studies, historical data, or a pilot study. The estimate of the

TABLE 2. Sample Sizes Required to Estimate Population Parameters

Parameter	Sample Size(s)	Inputs*
Mean	$n=\left(\frac{Z_{1-\alpha/2}\sigma}{E}\right)^2$	σ=Standard deviation of the outcome of interest
Difference in means	$n_i=2\left(\frac{Z_{1-\alpha/2}\sigma_p}{E}\right)^2, i=1,2$	σ_p=Common or pooled or standard deviation
Mean difference	$n_{pairs}=\left(\frac{Z_{1-\alpha/2}\sigma_d}{E}\right)^2$	σ_d=Standard deviation of the difference scores
Proportion	$n=p(1-p)\left(\frac{Z_{1-\alpha/2}}{E}\right)^2$	p=Estimate of the true population proportion
Difference in Proportions	$n_i=2p_p(1-p_p)\left(\frac{Z_{1-\alpha/2}}{E}\right)^2, i=1,2$	p_p=Estimate of the common or pooled population proportion

*$Z_{1-\alpha/2}$ reflects the confidence level that will be used in the analysis (usually $Z_{1-0.05/2}=1.96$); E is the prespecified margin of error.

standard deviation should always be conservative. Specifically, when estimating the standard deviation, one should err of the side of selecting a larger value that will produce a larger sample size. Should the standard deviation prove to be smaller than anticipated, a larger sample size will ensure the desired precision. However, if the standard deviation proves to be larger than anticipated, the sample size will not be adequate to ensure the desired precision. To estimate the difference between means, an estimate of the common or pooled standard deviation (σ_p) is required. In a clinical trial setting, the standard deviation of the outcome in an untreated or placebo arm is often used. To implement the formulas for proportions, an estimate of the population proportion (p) is required. A suitable estimate can be derived from other studies and again should be conservative. The sample size is maximized for p=0.5; thus, if there are no available data, then p=0.5 should be used to produce the most conservative (largest) sample size.

Example 6

Determine the sample size required to estimate SBP to within 5 U of the true value with 95% confidence. The formula to determine sample size is given by $n=\left(\frac{Z_{1-\alpha/2}\sigma}{E}\right)^2$. In this example E=5 U and $Z_{0.975}=1.96$. An appropriate estimate of the standard deviation (σ) also is needed. On the basis of the data in Table 1, an appropriate estimate is 18.8 (the larger value). To ensure that the CI for SBP has a margin of error not exceeding 5 U, a total of 55 participants are needed: $n=[1.96(18.8)/5]^2=54.3$ (always round up). If the desired margin of error was 4 U, a sample of size 85 would be required. If the goal of the analysis was to estimate the proportion of hypertensive patients in a population, the sample size would be determined by $n=p(1-p)\left(\frac{Z_{1-\alpha/2}}{E}\right)^2$. Suppose that the plan was to develop a 95% CI and that a margin of error not exceeding 3% was desired. A sample size of 1068 would be required: $n=0.5(1-0.5)(1.96/0.03)^2$; without data on the prevalence of hypertension, p=0.5 was used. If data were available on prevalence, eg, p=0.20 in a similar population, 683 participants would be required.

Precision

Precision refers to reproducibility and addresses the likelihood of observing similar results when a study or experiment is repeated. In estimation, precision is quantified by the margin of error in the CI. A larger margin of error produces a wider interval and indicates less precision. A study may report a relative risk of incident disease of 5.0, suggesting a 5-fold increase in risk of disease in one comparison group compared with the other. However, the 95% CI may be from 0.5 to 12.7. Because the interval includes 1, the null value, there is no statistically significant difference between groups in terms of risk. The wide CI suggests that the study is small. A second study may report a relative risk of 1.8 with a 95% CI of 1.7 to 2.0. The effect is smaller, but the estimate of effect is precise. The factors that affect the precision of an estimate include the level of confidence (Higher levels of confidence result in wider intervals), the variability of the point estimate (More variability results in wider intervals), and the sample size (Smaller samples result in wider intervals). Of these components, the sample size is controlled most easily by the investigator. It is important to perform sample size computations before mounting studies to ensure that resultant CIs will be adequately precise to address the research question.

Reporting Confidence Intervals

Most medical journals request that CIs be provided. A CI is a range of likely values for an unknown population parameter. In analyses that compare means or proportions, it is often of interest to assess whether the observed difference provides sufficient evidence to conclude that there is a difference in the population at a preselected level of significance. These assessments are often summarized by actual significance levels or probability values. Although probability values are important, they address only statistical significance. To assess clinical significance, the magnitude of the difference is important, as is the range of plausible values for the difference. CIs are particularly useful when a difference between groups fails to reach statistical significance. Nonsignificant *P* value indicates no statistically significant difference, but the CI provides additional data that might be useful, for example, in planning future studies.[7]

Disclosures

None.

References

1. D'Agostino RB, Sullivan LM, Beiser A. *Introductory Applied Biostatistics*. Belmont, Calif: Duxbury–Brooks/Cole; 2004.
2. Altman DG, Machin D, Bryant TN, Gardner MJ. *Statistics with Confidence*. 2nd ed. Boston, Mass: BMJ Books; 2000.
3. Kannel WB, Feinleib M, McNamara PM, Garrison RJ, Castelli WP. An investigation of coronary heart disease in families: the Framingham Offspring Study. *Am J Epidemiol*. 1979;110:281–290.
4. Newcomb RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Stat Med*. 1998;17:857–872.
5. Agresti A. *Categorical Data Analysis*. 2nd ed. New York, NY: John Wiley & Sons; 2002.
6. Rothman KJ, Greenland S. *Modern Epidemiology*. 2nd ed. Philadelphia, Pa: Lippincott-Raven Publishers; 1998.
7. Sterve JA, Smith GD. Sifting the evidence: what's wrong with significance tests? *BMJ*. 2001;322:226–231.