# Transposed Conditionals, Shrinkage, and Direct and Indirect Unbiasedness

*Stephen Senn*

**Abstract:** Unbiasedness as conventionally understood is not a necessary property of good inferences. Such unbiasedness is "direct"—it guarantees that, on average, an estimate equals the thing it is estimating (the parameter). Strange as it may seem, this does not mean that the parameter is on average equal to its estimate. This would require the very different property of inverse unbiasedness. When this phenomenon is understood, shrinkage of results can be seen to be a necessary fact of life.

(*Epidemiology* 2008;19: 652–654)

John Ioannidis[1] is to be congratulated on promoting an awareness of the fallibility of research findings. I agree with his conclusions that true associations are inflated (on average) but I have a somewhat different view as to why, which I shall try explain.

In my opinion much of the confusion regarding the reliability of research findings is due to basic misunderstandings of statistics and in particular of the weasel-word "bias." A trivial fact of logic that applies to syllogisms and one that, by extension, applies to probabilities and other inferential statements, is that you cannot transpose conditionals. Thus it is true that a minority of women are suffering from breast cancer but it is false that a minority of breast cancer sufferers are women. Many who are perfectly aware of this logic when they are required to think about it, nevertheless completely overlook its crucial relevance to statistical inference.

Consider for example a randomized clinical trial in which true diastolic blood pressure, $\chi$, at baseline is measured with error $\epsilon$ so that $X = \chi + \epsilon$ is the observed blood pressure. Suppose we knew that the true mean difference, $\Delta$, between patients (in mm Hg) was $\delta$. What does this tell us about the observed difference, $D$, in these patients? If measurement is unbiased and independent of $\chi$, so that errors are randomly above or below the true value with an expected difference of zero, then the answer is that we expect the observed difference to be $\delta$ so that $E[D \mid \Delta = \delta] = \delta$.

Now, however, suppose that we were to reverse the question and ask, given that the observed difference $D$ (in mm Hg) is $d$, what do we expect the true difference to be? It turns out that the answer is less than $d$, so that $E[\Delta \mid D = d] < d$. This can easily be explained in terms of regression coefficients. The regression of observed on true is the covariance of true and observed, $\sigma_{\chi X}$, divided by the variance of true, $\sigma^2_\chi$. Under the model, these 2 values are the same and so $\sigma_{\chi X} = \sigma^2_\chi$. Hence the regression is 1.0. On the other hand if we regress true on observed we take the same covariance but divide by the variance of observed, $\sigma^2_X$, which, due to measurement error, is greater than the variance of true; thus $\sigma^2_X > \sigma^2_\chi$ and therefore $\sigma^2_X > \sigma_{\chi X}$. Hence we have a regression to the mean.

I note, by the by, that failure to understand this distinction has led to incorrect claims that, when covariates are measured with error, analysis of covariance in randomized clinical trials is biased due to attenuation of the regression effect between outcome and covariate. However, this overlooks the other regression (of true baseline difference on observed baseline difference), and it turns out that the two cancel.[2]

Now consider a large collection of true treatment effects; it may be helpful to consider the context of a microarray as an example. These effects are measured with perfectly unbiased "studies," which are, of necessity, of limited precision. The variance of the estimates will exceed the variance of the true effects. It thus follows that even if all the experiments are perfectly unbiased—so that, on average, conditional on a given true treatment effect the observed effect, is equal to the effect it is estimating—the reverse does not hold. We can expect that any true effect will be closer to the mean of all effects than is the observed effect.

Note that this applies even if all the scientists involved are experimental saints who measure everything in a perfectly unbiased way. The point is that unbiasedness is a property that applies to the "forward expectation" of observed effects given true effects, but not therefore to the "backward expectation" of true effects given observed effects. To assume that the one implies the other is to make a comparable error of logic to assuming that because Greeks are with 100% probability European, Europeans are with 100% probability Greek.

Submitted May 12, 2008; accepted May 30, 2008.
From the Department of Statistics, University of Glasgow, UK.
***Editors' Note:*** *Related articles appear on pages 649, 655, and 657.*
Correspondence: Stephen Senn, Department of Statistics, 15 University Gardens, University of Glasgow, UK, G12 8QQ. E-mail: stephen@stats.gla.ac.uk.

ISSN: 1044-3983/08/1905-0652
DOI: 10.1097/EDE.0b013e318181b3e3

**TABLE 1.** Mean Treatment Effects: True Values, Raw Means, and Shrunk Means, Based on 100,000 Simulations

| Type of Study | Number | Raw Mean | | | | Shrunk Mean | | |
|---|---|---|---|---|---|---|---|---|
| | | True | Pilot | Both | Sequential | Pilot | Both | Sequential |
| Stopped | 25,228 | 1.16 | 1.38 | 1.23 | 1.38 | 1.15 | 1.15 | 1.15 |
| Complete | 74,772 | –0.39 | –0.46 | –0.42 | –0.42 | –0.39 | –0.39 | –0.39 |
| Both | 100,000 | 0.00 | 0.00 | 0.00 | 0.036 | 0.00 | 0.00 | 0.00 |

See text for further details.

Direct expectation is what is relevant to what one might call direct bias (or unbiasedness as the case may be) and inverse expectation is relevant to what might be called inverse bias (or unbiasedness). Consideration of the latter is natural to Bayesians but maybe frequentists would do well to think about it more.

I do not deny that some of the features (an obsession with significance, selective reporting, conflicts of interest) that Ioannidis discusses may have some relevance to the phenomenon to which he draws attention. Nonetheless, I do not believe that they are necessary. "Regression to the truth,"[3] is to be expected anyway.

In particular, I think that the role of statistical significance and early stopping as a cause of this phenomenon is overstressed, as I shall now try demonstrate using the results of a simple simulation.

Consider a case where we have a large population of treatments effects $\tau$ with variance $\gamma^2$ (which, without loss of generality, we may assume equal to 1) and mean $E[\tau] = 0$. Now suppose that we have some basic experimental unit, which might be patients measured more than once (as, say, in a cross-over trial) or centers with 2 patients, 1 on treatment and control (as in a parallel group trial)—each of which is capable of providing an estimate with variance $\gamma^2$. We can consider such a basic experimental unit as being the minimum that would provide an estimate of any interest. A Bayesian interpretation of the set-up I am describing is that the prior information about a given treatment effect, as incorporated in $\gamma^2$, is about equivalent to what an absolutely minimal experiment would show. Different amounts of prior information could be considered appropriate in practice and different minimal experiments could be envisaged, but this set-up will be sufficient to illustrate the relevant points.

Now suppose, however, that we have carried out an experiment that constitutes an $n$-fold replication of the basic unit. For example we might have $n$ centers. Bayesian theory suggests that any unbiased estimate, $\hat{\tau}$, of a given effect $\tau$ should be shrunk by a factor $n/(n + 1)$ to produce a shrunk estimate $\hat{\tau}_s = \hat{\tau}\dfrac{n}{n+1}$. Although $\hat{\tau}_s$ is not unbiased in the classic forward sense, because $E[\hat{\tau}_s | \tau = \tau'] \neq \tau'$, it is unbiased in the Bayesian backward sense because $E[\tau | \hat{\tau}_s = \hat{\tau}_s'] = \tau\hat{\tau}_s'$.

Now suppose we run an experimental trial in which we study $n_1$ units in a first pilot stage and are prepared in principle to study $n_2$ units in a second stage, but that we will stop the study if the first-stage results are "interesting." Inference will be based on $n_1$ results for a stopped study and on $n = n_1 + n_2$ for a completed study.

Table 1 shows the results of 100,000 simulations for such a scheme with $n_1 = 5$, $n_2 = 10$ and stopping based on conventional one-sided significance at the 5% level after stage 1. Note that the figure labeled "Both" is, in the case of a stopped study, hypothetical, since the data from the second $n_2$ patients would not be obtained. The standard errors of the means from this table are all less than 0.004. (All these means, by the way, can be calculated theoretically, but the simulation has some attraction as a demonstration.)

From the raw means it can be seen that, averaged over both types of trials, the sequential design is indeed biased. However it is also obviously true that if we choose to conduct a pilot study without possibility of continuing, then for all such pilot studies where the result is significant the "bias" is the same as for a sequential trial. Thus the sequential trial has the same bias as a significant pilot study run in a nonsequential manner. Similarly, we could in principle examine any nonsequential trial in $n = n_1 + n_2$ patients to see whether it would have stopped had it been run sequentially. Obviously again, trivially, the complete study gives the same result as the sequential study.

It thus follows that a sequentially run trial can always be matched with results from a nonsequentially run trial that would show exactly the same degree of bias provided that the results and their precision are available. Under such circumstances the fact that the trial is sequential is irrelevant to any judgment of bias.

Now consider the shrunk means in Table 1 produced using the shrinkage factors $\beta_{pilot} = {}^{n_1}/_{(n_1+1)} = {}^{5}/_{6}$ and $\beta_{complete} = {}^{n}/_{(n+1)} = {}^{15}/_{16}$. It can now be seen that, whether stopped or not, the true treatment effect is correctly estimated on average.

Of course all of this will be perfectly obvious to Bayesians. I am not, however, suggesting that we use Bayesian statistics for reporting trials. This would get us in a terrible mess. (For the same reason, adjusting *P*-values for

multiplicity is also undesirable.) However, it does suggest that we should use Bayesian eyeglasses for looking at trials—or, if one prefers, we should view all results with skepticism.

Basically, this point of view suggests that the problem with sequential trials is the reverse coin of their advantage. They are cheaper because they are on average, smaller, and by the same token they collect less information and therefore require stronger shrinkage—but none of this is due to the fact that they are sequential. Thus, while I applaud Ioannidis's determination to make us recognize the fallibility of research findings, I do not follow him in blaming it on the cult of significance—lamentable though that cult may be. Nor do I regard it as being necessary to seek the explanation in the less-than-perfect behavior of scientists. Random variation means that results are less-than-perfectly reliable, and this simple fact is enough to suggest a degree of skepticism when interpreting anything.

## ABOUT THE AUTHORS

*STEPHEN SENN is currently Professor of Statistics at the University of Glasgow. He has also worked for the pharmaceutical industry in Switzerland and the National Health Service in England. He describes his book, "Dicing with Death" (Cambridge, 2003) as an attempt to explain to a skeptical public that statistics as a subject is far more interesting than 99.9% of humanity appreciates.*

## REFERENCES

1. Ioannidis JPA. Why most discovered true associations are inflated. *Epidemiology*. 2008;19:640–648.
2. Senn SJ. In defence of analysis of covariance: a reply to Chambless and Roeback [letter; comment]. *Stat Med.* 1995;14:2283–2285.
3. Pocock SJ, White I. Trials stopped early: too good to be true? [comment] [see comments]. *Lancet*. 1999;353:943–944.