May 30, 2025

Below you will find

- an *article* from 1985, written by a psychology researcher, entitled **"Variance Explanation Paradox: When a Little is a Lot"**
- a 1986 draft of my *response* **"At Variance: With Oneself and With Others"**, which I wrote while on sabbatic leave, but never submitted anywhere
- '*slides*' (such as were possible with 1980s' technology) from a seminar, entitled **"Discussion"**, I gave on it it in my department.

I don't have the readable-today source files, so I am unable to fix the way some of the Greek symbols are displayed, but you will still see what I was doing. And you will be able, from the Figure legends, to imagine what the (no longer readable) figures showed.

I often used the baseball (and cancer) examples in my subsequent teaching to emphasize the fundamental difference between random variables that take on just two possible values, and those on a full interval scale – and how the idea of variance explanation does not transfer well from what we teach in first courses on regression to subsequent ones involving binary regression.

Readers might wish to follow up on this, or repeat the survey in class to see if today's psychology and statistics students – and researchers – have a better understanding of 'variance explanation' !

Sincerely...

James Hanley

webpage: https://jhanley.biostat.mcgill.ca

email: `james.hanley@mcgill.ca`

Psychological Bulletin
1985, Vol. 97, No. 1, 129-133

0033-2909/85/$00.75

A Variance Explanation Paradox: When a Little is a Lot

Robert P. Abelson
Yale University

Concerning a single major league at bat, the percentage of variance in batting performance attributable to skill differentials among major league baseball players can be calculated statistically. The statistically appropriate calculation is seriously discrepant with intuitions about the influence of skill in batting performance. This paradoxical discrepancy is discussed in terms of habits of thought about the concept of variance explanation. It is argued that percent variance explanation is a misleading index of the influence of systematic factors in cases where there are processes by which individually tiny influences cumulate to produce meaningful outcomes.

It is generally accepted that percentage of variance explained is a good measure of the importance of potential explanatory factors. Correlation coefficients of .30 or less are often poor-mouthed as accounting for less than 10% of the variance, a rather feeble performance for the influence of a putatively systematic factor. In analysis of variance contexts, the percentage of variance explanation is embodied in the omega-squared ratio of the systematic variance component to the total of the systematic and chance variance components. It, too, is often small; when it is, this is a source of discouragement for the thoughtful investigator.

Psychologists sometimes tend to rely too much on statistical significance tests as the basis for making substantive claims, thereby often disguising low levels of variance explanation. It is usually an effective criticism when one can highlight the explanatory weakness of an investigator's pet variables in percentage terms.

Having been trained, like all of us, in the idiom of variance explanation, I have always believed that when levels of variance explanation are extremely small, then the variables involved are really quite unimportant (however much one may lament the fact in a given case). However, I have been led to reexamine this notion.

A colleague and I recently had an argument in which we took opposing views of the role of chance in sports events. I claimed that many games of baseball and football are decided by freaky and unpredictable events such as windblown fly balls, runners slipping in patches of mud, baseballs bouncing oddly off outfield walls, field goal attempts hitting the goalpost, and so on. Even without obvious freakiness, I claimed, the ordinary mechanics of skilled actions such as hitting a baseball are so sensitive that the difference between a home-run swing and a swing producing a pop-up is so tiny as to be unpredictable, thus requiring it to be considered in largely chance terms.

My colleague argued that chance characterizations of sports events ignore the obvious fact that good teams usually win, that even under freaky circumstances (wind, mud, and so on) skilled players will better overcome difficulties than mediocre players, and furthermore that the visual–motor coordination of skilled athletes is subject to causal analysis.

Without trying to resolve in any serious way the deeper issues involved in the meanings of causation and chance in sports events, a straightforward statistical question can be raised: What percentage of the variance in

Willa Dinwoodie Abelson, Fred Sheffield, Allan Wagner, and Rick Wagner provided helpful comments on an earlier draft of this article. I wish also to thank the faculty and graduate students of the Yale University Psychology Department for exposing themselves to potential collective embarrassment by filling out the questionnaire.

Requests for reprints should be sent to Robert P. Abelson, Box 11A Yale Station, New Haven, Connecticut 06520.

athletic outcomes can be attributed to the skill of the players, as indexed by past performance records? This variance explanation question is analogous to those that characterize psychological investigations, but arises in a context where there exist strong intuitions (among sports fans, at least). A comparison of intuition with fact might therefore prove interesting.

To elicit intuitions, the athletic performance in question must be concretized. A simple performance with which most Americans are familiar, and for which copious records exist, is batting in baseball. The simplest event to consider is whether or not the batter gets a hit in a given official time at bat. It is possible to calculate statistically the proportion of the variance of this event (getting or not getting a hit) explained by skill differentials between batters.

Calculation of Variance Explanation

Let the dependent variable be $X = 1$ for a hit and $X = 0$ for no hit, and conceptualize the data matrix as in Table 1. Columns represent different batters. Rows represent different times at bat in, say, 5 years of at bats for each batter, a period long enough to give a reliable indication of the batters' true averages. The number of at bats might as well be taken as equal for all batters: The subsequent calculation is not affected by this factor.

Much as in the usual analysis of variance fashion, Equation 1 decomposes the entire set of X_{ji} in Table 1 into a true mean B_i for the ith batter and an error component e_{ji} for the jth occasion for the ith batter:

$$X_{ji} = B_i + e_{ji}. \tag{1}$$

The variance components σ_B^2 and σ_e^2 attaching to the two terms give the ingredients necessary to answer our variance explanation question. The former represents the variability of true batting averages, the latter the variability of performance given the batting average.

Both components depend on the distribution of true batting averages. Let the mean of the distribution of B_i be μ_B and the standard deviation σ_B. To compute the within-batter variance, σ_e^2, consider a batter with true

Table 1
Hypothetical Data Matrix for Batting Outcomes

	Batters								
At bats	1	2	3	·	·	i	·	·	·
1	0	0	1	—	—	—	—	—	—
2	1	0	0	—	—	—	—	—	—
3	0	0	0	—	—	—	—	—	—
·	—	—	—	—	—	—	—	—	—
·	—	—	—	—	—	—	—	—	—
j	—	—	—	—	—	X_{ji}	—	—	—
·	—	—	—	—	—	—	—	—	—
Batting average	.282	.214	.301	—	—	—	—	—	—

Note: 0 = no hit; 1 = hit.

average, B_i. On occasions when this batter gets a hit, $X_{ji} = 1$, and from Equation 1, $e_{ji} = 1 - B_i$. When the batter fails, $X_{ji} = 0$, and $e_{ji} = -B_i$. The first type of event happens on the proportion B_i of all occasions, the second type of event on the proportion $(1 - B_i)$. Weighting the squares of the e_{ji} by these proportions, the result is

$$\begin{aligned}\sigma_{e(i)}^2 &= B_i(1 - B_i)^2 + (1 - B_i)(-B_i)^2 \\ &= B_i(1 - B_i)[(1 - B_i) + B_i] \\ &= B_i(1 - B_i). \end{aligned} \tag{2}$$

(This is simply the formula for the variance associated with a binomial event around a true proportion B_i; I have rederived it in order to be explicit.)

Now consider the fact that because batting averages differ, the error variance is not the same for all batters. To obtain a summary value for σ_e^2, Equation 2 must be averaged over all values of B_i, weighted by the probability $p(B_i)$ of their occurrence.

$$\begin{aligned}\sigma_e^2 &= \sum_i B_i(1 - B_i)p(B_i) \\ &= \sum_i B_i p(B_i) - \sum_i B_i^2 p(B_i). \end{aligned} \tag{3}$$

The respective terms on the right are by definition the raw first and second moments of the distribution of B_i. That is,

$$\begin{aligned}\sigma_e^2 &= \mu_B - (\mu_B^2 + \sigma_B^2) \\ &= \mu_B(1 - \mu_B) - \sigma_B^2. \end{aligned} \tag{4}$$

Hence, the omega-squared ratio for proportion of variance attributable to skill is:

$$\omega^2 = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_e^2} = \frac{\sigma_B^2}{\sigma_B^2 + \mu_B(1 - \mu_B) - \sigma_B^2}$$

$$= \frac{\sigma_B^2}{\mu_B(1 - \mu_B)}. \quad (5)$$

Finally, realistic values are needed for σ_B and μ_B to substitute in Equation 5. These parameters of the distribution of true batting averages of course differ somewhat from year to year and league to league. However, the bulk of the distribution of observed batting averages of major league regulars in a given year typically lies between the low .200s and the low .300s. This suggests parameters such as $\mu_B = .270$ and $\sigma_B = .025$. These values yield

$$\omega^2 = \frac{(.025)^2}{(.270)(.730)} = .00317.$$

In other words, the percentage of variance in any single batting performance explained by batting skill is about one third of 1%.

What's Going on Here?

One's first reaction to this result is incredulity. My personal intuition was jarred by this result, which seems much too small. To check my own intuition against those of others, I circulated a one-item questionnaire to all graduate students and faculty in the Department of Psychology at Yale University. This group was chosen not simply for convenience, but because they would be familiar with the concept of variance explanation. Respondents were asked to refrain from answering if they knew nothing about baseball or the concept of variance explanation. Participants were asked to imagine a time at bat by an arbitrarily chosen major league baseball player, and to estimate what percentage of the variance in whether or not the batter gets a hit is attributable to skill differentials between batters.

The median of the 61 estimates of the variance attributable to skill was 25%, an overestimate of the calculated estimate by a factor of 75. The estimates of over 90% of the sample were too high by a factor of at least 15. Only 1 person gave an underestimate. I also posed the skill variance question to colleagues outside of Yale (some of whom are well known for their statistical acumen) and commonly received answers around 20% or 30%. The outcome of the statistical calculation, .3%, is indeed surprising.

Another attack on the paradox is to look for flaws in the statistical calculation. One thing to consider is the sensitivity of Equation 5 to variations in the parameters σ_B and μ_B. The term $\mu_B(1 - \mu_B)$ does not change appreciably with small variations in μ_B; the value for ω^2, in other words, would be nearly the same if I took $\mu_B = .265$ or .260 or .275 rather than .270. The ratio is more sensitive, though, to variations in σ_B. If σ_B were more than .025, then ω^2 would of course be bigger. However, .025 is, if anything, a generous estimate. If lifetime batting averages are taken as more indicative of true ability than season-by-season averages their standard deviation would be used for σ_B. Calculated from data in James's (1983) baseball abstract, the mean lifetime average was .268 and the standard deviation of lifetime averages for all major league regulars active in 1983 was .021. Even if I generously inflated this estimate to include nonregular players—even if I, say, doubled it to .042—the omega-square for skill variance would still be below 1%.

Could Equation 5 ever give a large value for ω^2? Yes, if every batter batted either 1.000 or .000 (i.e., either perfect or perfectly awful), then $\sigma_B^2 = \mu_B(1 - \mu_B)$, and $\omega^2 = 1$, as one would expect. This extreme situation contrasts sharply with reality. (Indeed, a way to understand the paradox is to realize that in the major leagues, skills are much greater than in the general population. However, even the best batters make outs most of the time.)

So the paradox remains. When I told my colleague the result of the calculation, he said, "You mean to tell me that the difference between George Brett and Len Sakata doesn't amount to anything?" This comment places the burden of the skill variance on extreme exemplars. The statistical calculation, of course, includes players of all levels of ability, most of them nearly average. Also, the comment appeals to the long-run differences in ability, whereas the calculation refers to the single at bat, a much chancier proposition. Thus, the paradox may arise in part because

the intuitive way of conceptualizing the question is intrinsically different from the appropriate statistical formulation, as in the phenomena discussed by Kahneman and Tversky (1982) and by Nisbett, Krantz, Jepson, and Fong (1982).

Is the statistical formulation therefore somehow unfair or irrelevant (Cohen, 1981)? Hardly. The single at bat is a perfectly meaningful context. I might have put the question this way: As the team's manager, needing a hit in a crucial situation, scans his bench for a pinch hitter, how much of the outcome variance is under his control? Answer: one third of 1%. Qualification: This assumes that the standard deviation of batting averages against a given pitcher is the same as the standard deviation of batting averages in general.

One might also argue that, in this framework, the manager may be able to choose someone two standard deviations above average and definitely avoid someone two standard deviations below average. By so doing, he would effectively double the standard deviation, and thus quadruple the skill component of variance. Even at that, the percentage of variance explanation would be only about 1.3%. In variance explanation terms, the difference between, say, George Brett and Len Sakata really is of small consequence. To appreciate why this is so and perhaps alleviate one's sense of paradox, it may be helpful to picture this comparison as in Table 2.

In Table 2 the rows represent batters with widely different skill levels, and the columns represent the outcome variable of getting a hit or not. The entries represent projected frequency of each outcome per 1,000 at bats. Even though hits are almost 50% more frequent for the .320 than for the .220 batter, the correlation between skill and outcome is not very sizable. The phi coefficient calculated from Table 2, for example, is .113. Taking the square of this as an estimate of variance explanation yields 1.3%.

Table 2
Correlation Between Skill and Outcome

	Outcome	
Skill of batter	Hit	No hit
Well above average	320	680
Well below average	220	780

Note. 1,000 at bats per batter.

Larger Implications

I have given an example from a nonpsychological context in which the percentage of variance actually explained by an independent variable (skill) is pitifully small, whereas "everyone knows" that the variable in question has substantial explanatory power. The paradox probably does not depend on some peculiarity of the intuitions of psychologists. The public cannot reasonably be asked the exact question about variance explanation, but it is a safe guess that skill is considered relatively important by the typical baseball fan.

What does the baseball paradox suggest for the usual standards for conceptualizing variance explanation? If one-third percent indicates such a trivial degree of explanation as to be virtually meaningless, should differential batting skill then be dismissed as an explanatory variable in baseball? Or should one instead be more suspicious of variance explanation as an index of systematic influence, and revise the notions surrounding less than 1% of variance explanation?

The answer lies in the type of example under consideration. The baseball example, as it turns out, exaggerates the paradox. The baseball case may take advantage of the "illusion of control" (Langer, 1975), by which skill influences are exaggerated at the expense of chance influences. Beyond that, however, there is a sound basis for the belief that systematic differences in batting averages are nontrivially predictive of success in baseball, in ways not captured by the statistical calculation. First, the individual batter's success is appropriately measured over a long season, not by the individual at bat. Second, a team scores runs by conjunctions of hits, so a team with many high-average batters is more likely to stage rallies than a team with many low-average batters. Thus, team success over a long season is influenced by average batting skill far more than is individual success in the single at bat because the effects of skill

cumulate, both within individuals and for the team as a whole.

The statistical effects of cumulation are well known, although they are usually discussed in methodological contexts, such as the psychometrics of reliability of measurement or the prediction of behavior from attitude measures (Epstein, 1979). The message here is that it is the *process* through which variables operate in the real world that is important. In the present context, the attitude toward explained variance ought to be conditional on the degree to which the effects of the explanatory factor cumulate in practice. Some examples of potentially cumulative processes are educational interventions, the persuasive effects of advertising, and repeated decisions by ideologically similar policy makers. In such cases, it is quite possible that small variance contributions of independent variables in single-shot studies grossly understate the variance contribution in the long run.

Thus, one should not necessarily be scornful of miniscule values for percentage variance explanation, provided there is statistical assurance that these values are significantly above zero, and that the degree of potential cumulation is substantial. On the other hand, in cases where the variables are by nature nonepisodic and therefore noncumulative (e.g., summary measures of personality traits), no improvement in variance explanation can be expected.

In sum, the large intuitive overestimation of the variance in batting outcome explained by skill is not simply an error in the appreciation of statistics. It reflects an intuition that skill does matter. Indeed it does, in the long run, albeit not very consequentially in the single episode. The baseball paradox is thus a model for similar paradoxes that may arise in psychological contexts.

References

Cohen, L. J. (1981). Can human irrationality be experimentally demonstrated? *The Behavioral and Brain Sciences, 4,* 317–331.

Epstein, S. (1979). The stability of behavior: I. On predicting most of the people most of the time. *Journal of Personality and Social Psychology, 37,* 1097–1126.

James, B. (1983). *The Bill James baseball abstract.* New York: Ballantine Books.

Kahneman, D., & Tversky, A. (1982). Variants of uncertainty. *Cognition, 11,* 143–157.

Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology, 32,* 311–328.

Nisbett, R. E., Krantz, D. H., Jepson, C., & Fong, G. T. (1982). Improving inductive inference. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 445–459). New York: Cambridge University Press.

Received February 14, 1984
Revision received June 19, 1984 ■

DRAFT... 1-5-86

COMMENTS INVITED

At Variance: With Oneself and With Others

James A. Hanley,

Department of Epidemiology and Biostatistics,

McGill University,

Montréal, PQ, H3A 1A2

Canada

Abstract

Abelson's paper "A variance expanation paradox: when a little is a lot" uses analysis of variance to calculate the percentage of variance in outcome of a single at bat in baseball that can be attributable to skill. He finds that the small percentage so explained is "seriously discrepant with intuitions about the influence of skill in batting performance" (he and his colleagues greatly overestimated this percentage).

The way in which he elicited these percentages from his colleagues makes it difficult to interpret their overestimates; I argue that the inherent unpredictability of binary outcomes could have been elicited by using more neutral and less distracting settings or by casting the question in other forms. However, the author's own reactions to the results of the formal analysis of variance are more disturbing: he looks for flaws in the calculation; performs a sensitivity analysis; considers only the best and the worst batters; tries to explain the resulting answer by appealing to another equivalent index of correlation; and only mentions statistical cumulation at the very end. These reactions suggest that the full statistical implications of working with binary variables and with intra-individual variation are

not fully understood or appreciated. I discuss some of these implications and outline some measurement principles to be followed if the real task in variance explanation is to be given a serious chance of succeeding.

INTRODUCTION

A recent paper "A variance expanation paradox: when a little is a lot" by Abelson examines the concept of variance explanation. He calculates the extent to which differentials in skill (as measured by players' longterm batting averages) explain variance in the possible outcomes (hit, no hit) when each baseball batter has one time at bat. He shows that if batters were chosen at random from the 220 to 320 range, the percentage variance in the outcomes attributable to skill is a mere one third of 1%, much less than he or his colleagues had thought. If players were deliberately chosen from the two extremes, the variance explained would still only amount to 1.3%.

One suspects that his survey results might be sensitive to the way in which he posed the question to his colleagues. Had he used "batting averages" instead of "skill differentials between players", or had he spoken of "the variance in whether or not different randomly selected players get hits", or had he asked them to estimate how much of the variance is unexplained, he might have received quite different estimates. However, my main concern is not with his colleagues' estimates, which could have many interpretations, but with his own reactions to his findings from a formal analysis of variance that players' batting averages are poor predictors of the outcomes of individual at bats. His first reaction was to look for flaws in the calculation: he performed a sensitivity analysis, then restricted attention to the best and the worst batters. The result in the latter scenario was still paradoxical and so "perhaps to alleviate one's sense of paradox", he calculates another index of correlation, the phi coefficient; unfortunately, all this does is confirm the "paradox" with the same low estimate of variance explanation. The clues to the paradox, the fact that the outcomes were binary (rather than binomial) and the absence of statistical cumulation, were only mentioned at the end.

If his reactions are typical, they suggest that the full statistical implications of working with binary variables and with intra-individual variation are not fully understood or appreciated. In this paper, I will discuss some of these implications and outline some measurement principles to be followed if the real task in variance explanation is to be given a serious chance of succeeding. The paper is divided into three parts I: Alternative ways of showing the large role

of chance (defined as lack of predictability) in Abelson's example II a discussion of the special nature of binary variables and III consideration of interindividual variation in general, and binary variation in particular, and what they mean for research design. At the end, I will briefly discuss how luck and probabilities dominate these same issues, at a more macroscopic (population) level, and how they pose difficult challenges in epidemiology and in interventions to prevent disease.

I: ASSESSING PREDICTABILITY IN MORE NEUTRAL SETTINGS AND/OR WITH ALTERNATIVE FORMULATIONS

The same question in other similar situations

Consider three analogous examples: (1) suppose a student's academic grade is simply the percentage of correct answers to 2000 multiple choice questions in 4 years of examinations; randomly select one of the 2000 and record whether the answer was correct. (2) choose a random 1 minute 'window' in each basketball player's playing career and record whether the player scored in this interval. (3) choose one spin of the roulette wheel in a casino and record whether the players (collectively) or the house win.

Randomly sampling a tiny portion of each academic or scoring record points up the overwhelming role of sampling variablity or "chance" and the little room for real individuality or systematic variation. The individual occasions examined are microscopic and the one outcome per individual cannot possibly be regarded as typical of the individual's full record; thus, one should not expect to be able to explain them. Even sports commentators, who like to dissect outcomes, and who are sometimes at a loss for something to say between at bats, would not attempt to analyse this variance.

Why then did Abelson, in spite of his contention that such variations must be "considered in largely chance terms", still overestimate the explained variance? Because, I believe, neither he nor his colleagues in their research investigations would ever attempt to study interpersonal variation in some behaviour using such a small portion of each person's behaviour. (It is also possible that his colleagues took skill to mean something different from what he did)

The same issue, but posed differently

Consider two players; one has a longrun batting average of 0.320 and the other 0.220. We wish to know who had which average and can use the raw data from the two series of at bats (see Table 1). How many at bats (randomly sampled) from each series should we analyse before safely deciding which series belongs to whom? If we choose m=1 at bat from each player, we stand a 60% chance of getting concordant answers and of not being able to make any decision (other than by tossing a coin); if we are lucky enough to actually obtain some variance, i.e. discordant answers, the probability is approximately 40% that it is the poorer batter who had the hit and that we will be incorrect* . With samples of m=5, 10, 25 and 100 at bats we decrease the probability of an incorrect decision decreases to approximately 35, 30, 20 and 5% respectively.

This simple example shows that in order to measure the relationship between player characteristics (height, handedness, ...) and their batting performances, one must measure each person's batting performance with sufficient precision that different players are (at least) correctly ranked (ideally, correctly spaced) along the "batting average" scale. Otherwise, inadequate measurements jumble their correct order and attenuate, or make it difficult to see, any real relationships.

(II) BINARY VARIABLES

They are inherently unpredictable; low indices of determination or correlation do not lie

Compared to the more commonly used interval or ordinal measures, binary outcomes are inherently much less predictable (one need only contrast how much more wrong one can be in predicting whether the next day will rain than in predicting the maximum temperature). The low summary measures, borrowed from analysis of variance, of the strength of relationship reflect this unpredictability. However, one should expect this to happen. The reason stems from the fact that whereas most measured data tend to pile up towards or close to their mean and so are reasonably well predicted by it, binary, or "0/1" data, by definition, pile up away from the mean. Thus, whereas the coefficient of variation for measured data is often a good deal less than 20-50%, that of binary data is much higher, e.g. 300% if the average is .100, 150% if it is near the the typical .270 batting average and 100% when the average is .500#. A further contrast is that in

measured data, the standard deviation and the mean are two distinct quantities, while in binary data the mean determines the variance and vice versa. [Incidently, it is strange that researchers who are quite wary of analysing bimodal or highly skewed data by anova methods, would use such techniques with binary data, the ultimate in overdispersion].

If one uses, or is forced to use, a binary variable to characterise the outcome in each individual, discrimination or full separability, of individuals is limited. The implications of this limited variability often surprise researchers who analyse binary outcomes for the first time, particularly if only a small minority of the outcomes are positive (or negative). Moreover, one's concept of degrees of freedom must be revised when dealing with binary outcomes: the real degrees of freedom are not the numbers of individuals studied but the number of minority outcomes.

The inherent unpredictability of individual binary outcomes, such as the results of single at bats, can be illustrated using the rationale underlying another index, . This coefficient, varying between 0 and 1, and having the same formulation as the $\sum 2$ statistic, was developed for genetics studies to measure how much 'closer' related individuals were to each other than to other unrelated individuals in respect to height, weight, blood pressure etc. Values close to 1 indicate that related individuals are quite clustered with respect to a characteristic and values near zero mean that within intra-family or "intra-class" variation is almost as wide as it is in the population at large. In the batting context, the interclass correlation answers the question: knowing the outcome of an at bat, will the outcome of a second randomly chosen at bat of this same batter be more like the first than a randomly selected at bat of another batter? One can see, by sampling from Table 2, that it will not.

The reason why the "percentage variance explained" is so low in Abelson's example is readily, and even more strikingly, seen if, as one should, one "plots his data". In the top left panel of Figure 1, the batting averages (Abelson's proxy for their skills) of 21 players are plotted on the x axis against the binary outcome (1=hit, 0=did not hit), of a single randomly chosen at bat per player, (y-axis). The batting averages (the Bi in Abelson's notation) were chosen uniformly from the .220 to .320 range (this way, the regression coefficient and correlation are stronger and are more precisely estimated than if one choses the Bi in relation to their natural (centrally tending)

distribution); each binary outcome was given a value of 1 or 0 depending on whether a corresponding computer generated random number was below or above Bi. As the graph shows, knowledge of each player's batting average is of little discriminatory value in a single instance. For those who still need to calculate the r2 before conceding, the percentage explained was indeed small, 0.05. This disappointing finding is not based on a freak data pattern, but as one can empirically verify, is possibly better than average. Of the 10 such plots I produced, 4 of them produced negative slopes; the pattern shown is the second best positive one. The figure also explains, more forcefully than does Abelson's Table 2, why correlation coefficients calculated from two binary variables, or even a binary and a measured one, are low even if there is a strong relationship between their averages: it is imposible for a straight line to be near to the four corners of the data!

III INTRA-INDIVIDUAL OR INTER-INDIVIDUAL?

(What variance does one seek to explain anyway?)

Although Abelson dealt with the single at bat, he likened his question to those commonly asked in psychological investigations, i.e. whether interpersonal variation in some personal characteristic or "outcome" is related to interpersonal variations in other, "explanatory", factors. By definition, then, one's interest is in explaining typical (characteristic/general/average) behaviour rather than that in any one specific randomly chosen instance. Given that this is a common research task, how does one characterise each individual?

When one can, repeat assessments. A proportion is an average too!

Some characteristics are strictly binary, and any number of repeated assessments in the same individual should give the same unchanging answer; examples include whether one was born in June and whether delivered by caesarian section. Judgements about other characteristics, characteristics which we think of as binary, such as whether one is blue eyed, or male, or born before term, or weighing less than 2,500 g at birth, or a "Type A" personality, are subject to some variance, depending on how or when assessments are made, and who makes them. Still others, such as batting performance, intelligence, academic performance, one's placement on a

psychological scale or ranking in interviews, are more quantitative; nevertheless, they rely on an "averaging" or other "summarization" of a series of items or components (at bats, examination questions, questionnaire items, votes of interviewers). In practice, such quantitative measures are derived after examining/consulting a limited, but presumably an adequate and representative, portion of the domain of "components".

How many? Signal vs Noise

How many (m) intra-individual components one should sample depends on the ratio of the intra-individual to inter-individual variation. In this "signal-to-noise" ratio, the true variation between individuals constitutes the signals or the target of the study; the necessity to characterise an individual on the basis of a sample constitutes the noise. Both the $\sum 2$ statistic and the intraclass correlation comprise these components, with signal = ßB and noise = ße , except that they combine them in the form

$\sum 2$ = signal2/(signal2 + noise2)

in order to produce a coefficient bounded between 0 and 1.

Assessing m, rather than just one, intra-individual components reduces the noise variance by a factor of m so that the associated $\sum 2$ statistic, which we can denote by $\sum m2$ becomes

$\sum m2 = ßB2 / (ßB2 + ße2 / m)$

To appreciate the need for repeated measurements, consider the effect of intra-individual variability in two different abilities, respiratory function as measured by the Forced Expiratory Volume in one second (FEV1), and batting performance in baseball. In the former, inter- and intra-individual variation are of the order of ßB = 0.5l and ße = 0.2l, a signal to noise ratio of 2.5; in the latter, one can use Abelson's values of ßB = 0.025 and ße = 0.443, a signal-to-noise ratio fifty times weaker. Figure 2 contrasts these two situations: if, as with with FEV1, the signal to noise ratio is appreciable, assessing an individual a small number of times (3-5) is more than sufficient to characterise that individual; if repeated assessments produce highly variable answers (relative to small signals), a large number of assessments of each individual are needed. For instance, the proportion of hits in 400 randomly selected at bats

is still only fair as a proxy for a player's overall average (∑4002 = 55%, r=0.74). For those who prefer to see than to believe formulae, this same message was evident in Figure 1, where even with a sample of m=100 at bats, the performance in this sample is only correlated r2=0.33 (r=0.57) with the true averages. The r2's in Figure 1 are somewhat higher those in Figure 2 because the individuals were deliberately chosen to be more spread out on the x axis i.e. ßB ≈ 0.030.

One way to understand what ∑m2 means is to examine its consequences for designing studies which compare performance of two groups of individuals, e.g. left- and right-handed players. If one uses their full batting records, the sensitivity (power) of such a study depends inversely on the quantity ßB2 / n, where n is the number of players studied from each group and the numerator ßB2 represents the variation in true batting averages between different players of the same handedness. If, instead of using each player's true average, one only used a sample of m of each player's at bats, the numerator is increased to ßB2 + ße2 / m. Thus, if for example ∑m2 =0.25, it can be interpreted as follows: a study of n=100 players/group, and using m at bats/player, has the same statistical power as a study of n=25 players/group which uses the entire batting record of each player. Put another way: the inverse of ∑m2 is the factor by which sample size n must be increased to account for the imperfect measurement of each player's overall performance. Calculating ∑m2 for different m's allows one to compare heir relative efficiency.

One can decrease (ßB2 + ße2 / m) / n by increasing n and m. In practice, cost and other constraints limit both n and m or force one to strike a balance between the two. Textbooks on sampling give methods for making the most efficient choice.

The need for multiple assessments of an individual is not confined to variables that are binary, but applies also to any variables which show a sizeable component of intra-individual variation. One author has recently arguedPete: do you remember what it was? that a test-retest correlation of r=0.6 in measuring physiologic hyperreactivity compromises a study that uses only one assessment/individual. One must either seek to increase the reliability of this single measure, or if one cannot find and control the sources of the variation, assess it more than once.

Comparing larger units: intra- and inter-population variation

The paradox discussed by Abelson has an important parallel in epidemiology, which studies the etiology of disease in populations (a population is analogous to a baseball player and the individuals in it analogous to his different at bats). Doll and Peto put it very clearly:

"the determinants of who will and who will not get cancer can be divided into three categories, not only the usual "nature" (genetic makeup) and "nurture" (what people do or have done to them) but also "luck" (the play of chance) ... Nature and nurture affect the probability that each individual will develop cancer and luck then determines which individuals will actually do so.& However, although for each single individual the role of luck is enormous, in a population of a hundred thousand or more, the role of luck is smaller and consequently in the comparison of national cancer rates, only nature and nurture are important".

Likewise, Rose explains: "I find it increasingly helpful to distinguish two kinds of etiological question. The first seeks the causes of cases, and the second seeks the causes of incidence. 'Why do some individuals have hypertension?' is quite a different question from 'Why do some populations have much hypertension, whilst in others it is rare?' The questions require different kinds of study and they have different answers".

Because individual probabilities of developing a particular form of cancer are relatively low, even in those considered to be at higher risk, prevention methods aimed at individuals have some serious drawbacks; preventive actions offer only a small benefit to each individual, since, as Rose says, "most of them were going to be all right anyway, at least for many years". He calls this the Prevention Paradox: a preventive measure which brings much benefit to a population offers little to each participating individual. An analysis, such as in Abelson's Table 2, of who does and does not develop lung cancer would find that whereas it is many times more likely in the smoker than in the non-smoker, the proportion of individual variation in outcomes explainable by their smoking habits is only a few percent@. For rarer cancers, or lower relative risks, the proportion of variance explained is even smaller, and the task of convincing the individual to lower the risk all the greater.

DISCUSSION

Abelson admits that the results of his survey may have been strongly influenced by the way he framed his question and even by the very choice of subject matter. However, his main point is not so much that we may be poor judges of r2, but that "one should not be scornful of miniscule values for percentage variance explanation, provided there is statistical assurance that these values are statistically above zero and that the degree of potential cumulation is substantial".

I have argued that when we have the choice, we should not try to measure explained variance using a microscope, and then hope (if we are lucky enough to find it) that the effect will cumulate. Rather we should cumulate first, making it unnecessary to use a microscope, and then try to explain what we can see. Abelson was probably well aware of the analogies with the need for repeated assessment to produce a stable and characteristic measure; however, his emphasis on the power of cumulation (as practised in advertising, education,..) rather than on averaging (as practised to measure a characteristic more precisely). However, his remarks might lead others to others to think that there was less need to make repeat assessments of intra-individual behaviour. As I hope this paper has shown, intra-indivual variation is a powerful "leveler" and dilutor of real inter-individual variation, all the more so when the within-individual elements are binary. I urge researchers to first try very hard todetermine where each study individual really stands relative to others, and only then to ask why? In some instances, intra-individual variation can be controlled for by assessing individuals at the same time, or with the same observers, etc; in order to avoid its insidious effects, repeated assessment of the remaining uncontrollable or unexplainable interindividual variation offers the only alternative.

Table 1

Figure Legends

Figure 1: Each player's average in m randomly selected at bats (y axis) plotted against the player's longrun average (x axis). The four panels correspond to m=1,5,25 and 100. Each panel represents a 'median' data pattern from among 10 panels generated. Even with large m, the correlations are weak, and many players would be misranked.

Figure 2: Average percentage of variance explained when individuals' longrun performances are used to predict the average of m assessments of batting success and Forced Expiratory Volume (FEV1). Because of the much smaller signal-to-noise ratio, the average of 5 repeated FEV1's has an $r > 0.95$ correlation with the true average; in contrast, more than m=1000 at bats per person are needed to achieve the same r. In both examples, the r or r2 depends not just on the intra-individual variability but also on the range of inter-individual variation being studied. There is also the implicit assumption that the true batting average is made up of a very large number of at bats (relative to m) so that the finite sampling correction is not needed.

* Prob(A,B both hit)=0.32 x 0.22 = 0.0704; Prob(both miss)=0.68 x 0.78 = 0.5304; Prob(A hits, B does not)=0.32 x 0.78 = 0.2496; Prob(A does not, B does)=0.68 x 0.22 = 0.1496;

Incidentally, if one thinks about it, this accords with experience. More often than not, in a critical at bat, even a highly esteemed pinch hitter fails to hit; sometimes, a lesser player shows him up by producing the important hit. If, as will commonly happen, both fail to hit or both hit, the contrast is not made; but if one hits and the other does not, there is quite a good chance (as high as 20-40% as mentioned above) that it is the lesser player who hit. With this common reversal, it is no surprize that so little of this variance is predictable.

Indeed one could question the very use of variances and standard deviations to describe binary data. with measured data, the majority of the values commonly lie less than 1 SD from the mean. In contrast, when the average binary value is 0.5, all of the data lie exactly one SD away;

Snedecor and Cochran

Doll R & Peto R. The causes of cancer: quantifiable estimates of the proportion of avoidable cancers in the US today.

& One could say about baseball players that each one is born with a certain batting average, but they each does after all have to produce the right number of hits and misses in order to realise it. It would be very boring baseball if it became any more predictable than that.

Rose G. Sick individuals and sick populations. International Journal of Epidemiology, 14: 32-38, 1985.

@ for example, assuming that non-smokers (half the population) have a 1% probability of developing lung cancer and that smokers have a 9% probability, the $f2$ coefficient is 3.4%.

AT VARIANCE:
WITH ONESELF AND WITH OTHERS

DISCUSSION OF

"A VARIANCE EXPLANATION PARADOX:

WHEN A LITTLE IS A LOT"

Question on variance explanation in baseball

You are ineligible if you	
(a)	know nothing about baseball or
(b)	know nothing about the concept of variance explanation or
(c)	have read Abelson's paper

Imagine a time at bat of an arbitrarily chosen major league baseball player.

Estimate what percentage of the variance in whether or not the batter gets a hit is attributable to skill differentials between players:

_____ %

The concept of variance explanation (Abelson)

- good measure of importance of potential explanatory factors

r ≤ 0.30 often "poor-mouthed" (≤ 10% of the variance")

- in anova contexts, % variance explained (ω^2) is a central concept

$$\omega^2 = \sigma^2_{\text{systematic}} / (\sigma^2_{\text{systematic}} + \sigma^2_{\text{chance}})$$

often small and discouraging

- trained that small % variance explained ==> variables quite unimportant

-recently led to consider this concept

Argument: the role of chance in sports events

Abelson:

- many football & baseball games decided by freaky & unpredictable events
    - windblown fly balls
    - runners slipping in patches of mud
    - baseball bouncing oddly off outfield wall
    - field goal attempts hitting goalpost
- even without obvious freakiness,

    mechanisms of skilled actions (eg hitting baseball) so sensitive that

    Δ between home-run and pop-up swing so tiny as to be unpredictable

Argument: the role of chance in sports events

Colleague:

- cannot be chance
    - good teams usually win
    - more skilled players overcome freaky conditions better
    - visual-motor coordination subject to causal analysis

straightforward statistical question

how much do differentials in skill* explain variance in the possible outcomes (hit, no hit) when each baseball batter has one time at bat?

*as measured by players' longterm batting averages

- Can compare intuition with mathematical calculation
- Copious data

Results of Survey

(n= 61 graduate students & faculty, Department of Psychology, Yale)

Median estimate:	25%	(high by 75 x)
90% of estimates above:	5%	(high by 15 x)

only 1/61 underestimated

Answer:

Calculation

Y_i : outcome of a random at bat of batter i

μ_i : true mean (nbatting average) for batter i e.g. $\mu_i = 0.289$

e_i : random binary outcome with mean μ_i i.e. e_i = 0(no hit) or 1(hit)

So

$$Y_i = \mu_i + e_i$$

$$\omega^2 = \sigma^2_{\text{systematic}} / (\sigma^2_{\text{systematic}} + \sigma^2_{\text{chance}})$$

$\sigma^2_{\text{chance}} = \mu_i(1-\mu_i)$ i.e. binary variance "averaged" over batters

$\sigma^2_{\text{systematic}}$ = variance of μ_i 's over batters

$\sigma^2_{systematic}$ = variance of μ_i 's over batters = ???

"The bulk of the distribution of μ_i 's of major league regulars in a given year typically lies between the low .200s and the low .300s"

so, range of μ_i 's: .220 to .320

average μ_i : .270

$\sigma^2_{systematic:}$: $.025^2$ = .000625

i.e. .100 = 4 x σ

$\sigma^2_{chance} = \mu_i(1-\mu_i)$ i.e. binary variance "averaged" over batters = ???

$\mu_i(1-\mu_i)$ ranges from

$.220 \times .780 = .171600 = .41^2$

to

$.320 \times .680 = .217600 = .46^2$

average is approximately .196475

So

$\omega^2 = \sigma^2_{systematic} / (\sigma^2_{systematic} + \sigma^2_{chance})$

= .000625 / (.000625 + .196475)

= .003

ω^2 = 1/3 of 1%

Sensitivity analysis

batters chosen	% outcome variance attributable to skill
at random from 220-320 range	0.3%
from two extremes i.e. 220 and 320	1.3%
uniformly from 000-320 range	6.3%
uniformly from 000-1000 range	33.3%

Implications... à la Abelson

- baseball e.g. exaggerates paradox
  illusion of control
  (ie skill influences are exaggerated at expense of chance infleuence)

- systematic Δs in μ_i 's are non-trivially predictive of success in baseball
  batter judged over entire season
  team scores by conjunction of runs

- statistical effects of cumulation well known
  psychometrics... reliability of measurement
  prediction of behaviour from attitude measures
  must consider if effects of factor cumulate
    - educational interventions
    - advertising

- OK to have small r's if
  $\rho >> 0$
  potential for substantial cumulation
  (not useful if X is summary measure)

- difficult to interpret their overestimates

- elicit inherent unpredictability by other means.

looks for flaws in the calculation
performs a sensitivity analysis
considers only the best and the worst batters
tries to appeal to another equivalent index of correlation
only mentions statistical cumulation at the very end

- discuss some of these implications
- outline some measurement principles needed

  to give variance explanation a chance of succeeding.

I other ways to show large role of chance in Abelson's example

[chance = lack of predictability]

II special nature of binary variables

III interindividual variation in general and binary variation in particular

what they mean for research design

how luck and probabilities dominate these same issues

challenges in epidemiology & 1° prevention (at population level)

The same Question in other similar situations: 3 analogous examples

- GPA = % correct answers to 2000 multiple choice questions

  ==> randomly select 1 of the 2000 and see if answer was correct

- basketball player's playing career

  ==> does player score in randomly choosen 1 minute 'window'

- gambling casino

  does the house win in one randomly chosen spin of the roulette wheel

  tiny portion of each academic/scoring record
  overwhelming role of sampling variablity or "chance"
  little room for real individuality or systematic variation

survey results might be sensitive to the way he posed question

"batting averages" vs "skill differentials between players"

"variance in whether/not different randomly chosen players get hits"

"estimate how much of the variance is unexplained"

The same issue, but posed differently

Consider 2 players A and B

one player has a longrun average of $\mu_1 = 0.320$

the other has a longrun average of $\mu_2 = 0.220$

who had which average ???

can use the raw data from the two series of at bats

How many at bats (randomly sampled) from each series should we analyse before safely deciding which series belongs to whom? i.e. m=???

If we choose m=1 at bat/player

60% chance of getting concordant answers

if lucky enough to obtain some variance

≈ 37% chance that it is the poorer batter who had the hit

at bats (sample size)	m =	5	10	25	100
prob. of incorrect decision		35%	30%	20%	5%

to assess relationship b/w player characteristics and batting performances

- must measure performance with sufficient precision that

  different players are (at least) correctly ranked (spaced)

  along the "batting average" scale.

- Inadequate measurements jumble correct order

  attenuate any real relationships.

inherently unpredictable: low r's and ω^2 's do not lie

- much less predictable than interval or ordinal measures

  predicting rain tomorrow
  vs
  predicting the maximum temperature

- Low r^2 borrowed (from anova) reflect this

- One should expect this to happen

  most measured data pile up towards or close to mean

  binary ("0/1") data, by definition, pile up away from the mean

coefficient of variation

for measured data...	often a good deal less than 20-50%
for binary data...	much higher
	e.g. 300% if μ = .100
	150% if μ is .270
	100% if μ = .500

limited r^2 with 0/1 variable

surprises those analysing binary outcomes for 1st time

dramatic if only a small minority of outcomes are +ve (-ve)

concept of degrees of freedom must be revised

(real degrees of freedom are the number of minority outcomes)

intra-class correlation:

ranges from 0 to 1, developed for genetics studies

how much 'closer' related individuals are than unrelated individuals

In the batting context interclass correlation answers question:

knowing outcome of an at bat, will outcome of a 2nd randomly chosen at bat of same batter be more like 1st than a random at bat of another batter?

Another clue: plot the data

This disappointing finding is not based on a freak data pattern, but as one can empirically verify, is possibly better than average. Of the 10 such plots I produced, 4 of them produced negative slopes; the pattern shown is the second best positive one. The figure also explains, more forcefully than does Abelson's Table 2, correlation coefficients calculated from two binary variables, or even a binary and a measured one, are low even if there is a strong relationship between their averages: it is imposible for a straight line to be near to the four corners of the data!

commonly asked research question

is interpersonal variation in personal characteristic/"outcome" related to interpersonal variations in other, "explanatory", factors ?

-interest is in explaining typical behaviour
(characteristic /general/ average)

rather than that in any one specific randomly chosen instance.

Q: how does one characterise each individual?

A: When one can, repeat assessments (A proportion is an average too!)

characteristics

strictly binary: repeat assessments should give the same answer

e.g. ? born in June ? delivered by caesarian section.

should be but... : depends on how, when, who, ..

e.g. blue eyed; male; born before term;

birthweight < 2500g; "Type A" personality;

quantitative : rely on "averaging" or "summarizing" series of items

(at bats, exam questions, questionnaire items, votes of interviewers)

e.g. batting performance; intelligence; academic performance

location on psychological scale; ranking in interviews

.... derived after examining/consulting a sample of item domain

How many (m) intra-individual components to sample? Signal vs Noise

depends on ratio of intra-individual to inter-individual variation

true variation between individuals ==> signal/target of study

noise <== must characterise individual using sample of components

$\omega^2 = \text{signal}^2/(\text{signal}^2 + \text{noise}^2)$

Assessing m > 1 components reduces noise variance by a factor of m
$\omega_m^2 = \text{signal}^2 / (\text{signal}^2 + \text{noise}^2 / m)$

Effect of intra-individual variability in two different abilities

	respiratory function (FEV_1)	batting performance
inter-individual variation	σ_b = 0.5	0.025
intra-individual variation	σ_e = 0.2	0.443
ratio signal:nois	2.5	18

Figure 2 contrasts these two situations:

FEV_1: a small m (3-5) is sufficient to characterise individual

batting: a large m needed

average/400 only fair guide to player's overall average

$\omega_{400}^2 = 55\%$

another look at what ω_m^2 means

<u>designing studies to compare performance of two groups of individuals</u>

e.g. left- and right-handed players.

- using full batting records ($m=\infty$) from n players per group

  power depends inversely on the quantity σ_b^2 / n

  σ_b^2 : variation in batting averages among Rh (or Lh) players

- using 1at bat/player increases the "noise" to $\sigma_b^2 + \sigma_e^2 / m$
- using m at bats reduces noise to $\sigma_b^2 + \sigma_e^2 / m$

e.g., if $\omega_m^2 = 0.25$

a study of n=100 players/group, and using m at bats/player

has the same statistical power as

a study of n=25 players/group which uses entire record of each player

- One can decrease $(\sigma_b^2 + \sigma_e^2 / m) / n$ by increasing n and m

  In practice, cost and other constraints limit both n and m

- need for m>1 applies to any variable with sizeable σ_e^2

Comparing larger units: intra- and inter-population variation

parallel in epi: population = player; individuals = different at bats

Doll and Peto (Causes of Cancer)

> "the determinants of who will and who will not get cancer can be divided into three categories, not only the usual "nature" (genetic makeup) and "nurture" (what people do or have done to them) but also "luck" (the play of chance)
>
> Nature and nurture affect the probability that each individual will develop cancer
>
> luck then determines which individuals will actually do so
>
> However, although for each single individual the role of luck is enormous, in a population of a hundred thousand or more, the role of luck is smaller and consequently in the comparison of national cancer rates, only nature and nurture are important".

Rose (Sick individuals and sick populations. Int J Epi, : 32-38, 1985)

> "I find it increasingly helpful to distinguish two kinds of etiological question. The first seeks the causes of cases, and the second seeks the causes of incidence. 'Why do some individuals have hypertension?' is quite a different question from 'Why do some populations have much hypertension, whilst in others it is rare?' The questions require different kinds of study and they have different answers".

:

- individual probabilities of a particular type of cancer relatively low

- prevention methods aimed at individuals have serious drawbacks

(preventive actions offer only small benefit to each individual)

"most were going to be all right anyway, at least for many years".

preventive measure which brings much benefit to a population offers little to each participating individual

An analysis of variance of who does and does not develop lung cancer

many times more likely in the smoker than in the non-smoker

but the proportion of individual variation in outcomes explainable by their smoking habits is only a few percent

e.g. assuming

half the population smokes

non-smokershave a 1% probability of developing lung cancer

smokers have a 9% probability

the ω^2 coefficient is still only 3.4%

Abelson

> results of his survey may have been strongly influenced by framing and by the very choice of subject matter

However, his main point

> not so much that we may be poor judges of r^2
>
> but that "one should not be scornful of miniscule values for percentage variance explanation, provided there is statistical assurance that these values are statistically above zero and that the degree of potential cumulation is substantial".

Hanley

- when we have the choice
  - should not try to measure explained variance using a microscope
  - should , making it unnecessary to use a microscope,
    - try to explain what we can see

Abelson emphasized power of cumulation rather than on averaging

but, ... lest his remarks be misinterpreted, I would re-emphasize:
- intra-indivual varn. is powerful "leveler" of real inter-individual varn.
- (all the more so when the within-individual elements are binary)

so, we should urge researchers to
- first try very hard to determine each study individual stands relative to others
- and only then to ask ?

In some instances intra-individual variation can be controlled for
- by assessing individuals at the same time
- with the same observers

in order to avoid its insidious effects, repeated assessment of the remaining uncontrollable or unexplainable interindividual variation offers the only alternative.

Table 1

Figure 1: Each player's average in m randomly selected at bats (y axis) plotted against the player's longrun average (x axis). The four panels correspond to m=1,5,25 and 100. Each panel represents a 'median' data pattern from among 10 panels generated. Even with large m, the correlations are weak, and many players would be misranked.

Figure 2: Average percentage of variance explained when individuals' longrun performances are used to predict the average of m assessments of batting success and Forced Expiratory Volume (FEV_1). Because of the much smaller signal-to-noise ratio, the average of 5 repeated FEV_1's has an $r > 0.95$ correlation with the true average; in contrast, more than m=1000 at bats per person are needed to achieve the same r. In both examples, the r or r^2 depends not just on the intra-individual variability but also on the range of inter-individual variation being studied. There is also the implicit assumption that the true batting average is made up of a very large number of at bats (relative to m) so that the finite sampling correction is not needed.