Statistical Science
2003, Vol. 18, No. 3, 342–345

John W. Tukey as Teacher

Stephan Morgenthaler

John W. Tukey was an atypical intellectual for our times, a thinker of surprising inventiveness. He spent his whole academic career at Princeton University, almost from the start dividing his time between Bell Labs and the University. Each term he taught courses for undergraduates and doctoral students. He cherished teaching and was beloved and highly respected by his students.

He had many interests outside of science, but was such a consummate statistician that we thought it might bring him closer to you, the reader, if we offer you a glimpse of one of his courses.

A COURSE ON COMBINING DATASETS

In the spring term of 1982, John W. Tukey taught a graduate course entitled "Combining Datasets." At that time, John's teaching method was firmly established. John wrote transparencies by hand and Eileen Olszewski, for many years his secretary at Princeton, typed them for distribution to the students. During lectures John used two projectors with each slide being first discussed and then moved to the second projector.

Combining Datasets was announced as follows:

> The purpose of this course will be to review methods of combining results. This is, the simplest problems of data analysis—treating single batches, regression, and analysis of variance—all involve a single (often internally structured) body of data. The next natural step is to put together—to combine—the results of such separate analyses.
>
> We do this in many ways, using as little, from each individual data body, as an apparent direction and as much as an estimated amount; together with an estimated variance for that estimate, and an indicated number of degrees of freedom for that estimated variance. We will try to work our way through the most general of these methods, beginning with the simplest and adding complexity step by step.

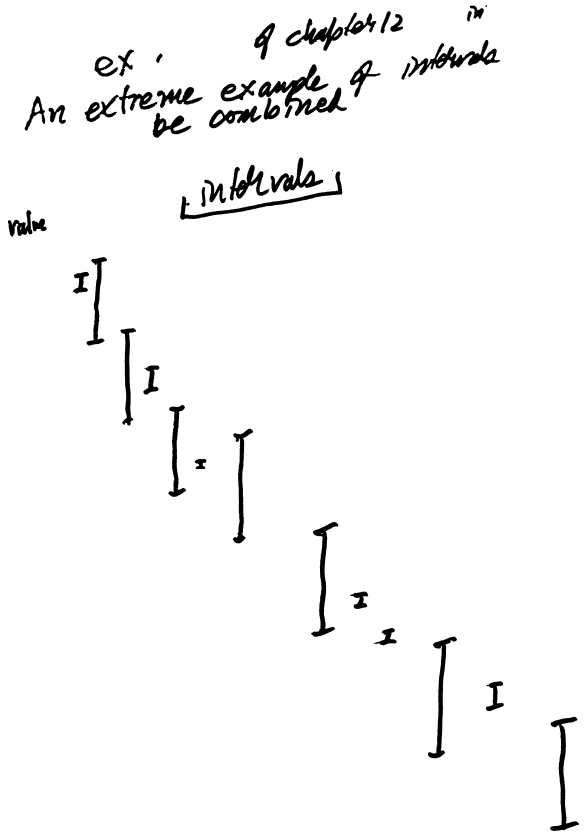
The chapter headings included:

> General outline; Combining independent results and assessing their significance; Combination of directionalities and indications of directionality; Combining tests of fit; Combining values to get a value; Combining values to get an interval; Combining intervals to get a value: group by group (scared, Paull-2, Paull-2F and PL combinations); Externally weighted combination of intervals to get an interval; Which combination when?

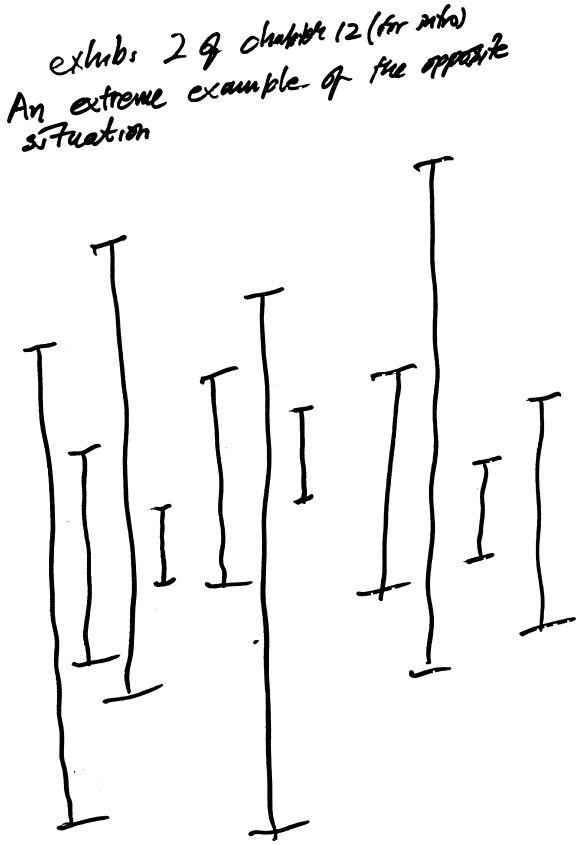
Chapter 12 discussed a statistical problem, the combination of intervals to get a value, that had fascinated John over a long period and the remainder of this section consists essentially of an extract of the course notes.

> In this chapter and the next, we start from intervals—essentially values with estimated uncertainties—and combine them to get a value. Essentially, then, our problem is to choose the weights with which our values are to be combined, in the first instance on the basis of the given interval lengths (and in the second, if we go over to resistive combination, with a view to a more precise result).
>
> The crucial consideration, which we met casually above, but which now drives and determines our choice of method, is the question of whether the values to be combined estimate the same thing or whether each value to be combined estimates a different thing—and we want to estimate a summary of the different things that are to be combined. Exhibit 1 illustrates one extreme, where it is clear that both:
>
> - the intervals are NOT estimating the same thing, and

Stephan Morgenthaler is Professor, Institute of Mathematics, Swiss Federal Institute of Technology (EPFL), 1015 Lausanne, Switzerland (e-mail: stephan.morgenthaler@epfl.ch).



EXIBIT 1.



EXIBIT 2.

- the individual estimates are not of the same precision.

As a result, we would be foolish indeed to let the choice of weights depend upon the interval lengths. While this might give us a combination of minimum variance, its interpretation would be at best blurred, and probably quite unreasonable. In such a case, the weights ought to depend upon the intrinsic importance of the targets of the several series, which may often mean that they should be equal.

There are intermediate cases, which will be the subject of the following chapter.

Exhibit 2 illustrates the situation to be treated in the present chapter, where there is no reason—either from the data or from our understanding of the subject-matter situation—to believe that different series estimate different values. [Ha! Ha!]

Here, if there is no reason to weight the individual-series values to reflect some known intrinsic importance (and there is no reason to believe that length of interval is related to value of center?), we will want to weight the values for the individual series in a way that reflects their apparent precisions (accuracies, perhaps).

Just how to reflect is the main topic of this chapter.

What NOT to Do!

The most natural—to the unthinking—way to reflect the interval lengths is also just what we SHOULD NOT DO. This is to treat the s's underlying the interval lengths as if they were σ's, which means weighting like $1/s_i^2$.

Why is this dangerous—and a loser on average? We have repeatedly stressed two points:

- most s^2 are poor estimates of their σ^2—some high, some low,
- the worst thing one can do, when weighting, is to put a big weight on what deserves a small one.

If we weight by $1/s_i^2$, some s_i^2 will be much too small, so that $1/s_i^2$ will be much too big, so that we will have done just exactly what we should not!

Partial Weighting

For the last quarter century, the standard response to this challenge has been partial weighting as expounded in Cochran's paper of 1954 in which after a careful inquiry Cochran confirmed the usefulness of the following procedure:

- find a naive weight for each series;
- order the series by these naive weights;
- take $1/2$ to $2/3$ of the series with the highest naive weights (no recipe for how to pick the exact fraction);
- for the series in this low-variability group, use the mean of all their naive weights as their weight, or, better, do this with the reciprocals;
- for the remaining series, use their naive weights.

This procedure of "partial weighting" meets the major difficulties head on, and overcomes them. So long as most series will tend to have about the same accuracy, partial weighting will ensure that no individual series gets a catastrophically high weight. It is not surprising that it has been a standard for so many years.

If we want to look for difficulties, we are likely to focus upon:

1. the absence of a recipe for the size of the lower group, and
2. the possible misweighting of the series NOT in the lower group.

Grouped Weighting

Let us plan to take the natural variation of the naive weights (which we suppose to be $1/s_i^2$'s) into account. How can we do this? In the overutopian case, where the series summaries come from individual measurements that follow a Gaussian distribution, each s_i^2 is distributed like a multiple of χ_f^2/f. We could certainly try to allow for this much variability.

In the real world, where series summaries come from individual measurements that follow some stretched tail distribution, each s_i^2 will be more widely dispersed than a multiple of χ_f^2/f. We get part way there by allowing for χ^2/f variability. Even part way is good! (We return, in the next section, to a way of going further.)

The Order Statistics of χ_f^2/f. The simplest way to deal with an *unknown* multiple of χ_f^2/f is to focus on *ratios*, and adjust by *division*. Suppose we have a number of s_i^2, for convenience based on a common number of df. How might we bring them more nearly to a common value?

If they are indeed from the distribution of

$$k\chi_f^2/f$$

for a common k, and if we order them, the ordered values will behave like order statistics from $k\chi_f^2/f$, or, equivalently, as k times the order statistics of a sample from χ_f^2/f, so that it is natural to look at

$$s_i^2/c(i|n, f)$$

where $c(i|n, f)$ is a typical value for the ith order statistic of n from χ_f^2/f. These ought all to behave as though they were estimating k—which in our context is the common σ^2 from which all s_i^2 came.

A reasonable approximation, good enough for many purposes, to the $c(i|n, f)$ can be had from

$$c(i|n) = \text{Gau}^{-1}\left(\frac{3i-1}{3n+1}\right)$$

and

$$c(i|n, f) = \left|1 - \frac{2}{9f} + c(i|n)\sqrt{\frac{2}{9f}}\right|^3, \qquad f \geq 3,$$

so that we have numbers for $c(i|n, f)$ when we want them.

One way to proceed would be to replace the naive weight

$$w_i^* = \frac{1}{s_i^2}$$

by the pushed-back weights

$$W_i^* = \frac{1}{s_i^2/c(i|n, f)} = \frac{c(i|n, f)}{s_i^2}.$$

Doing this would be a real gain, since the largest weights would be pulled down, but this does not feel like the place we should stop.

If, for example, we had $n = 10$, $f = 6$ and

$$s_i^2 = 4, 7, 8, 9, 10, 15, 18, 20, 200, 500,$$

the values of $s_i^2/c(i|n, 6)$ would be, roughly, respectively, 13.2, 15.2, 13.6, 12.7, 12.1, 15.6, 16.3, 15.5, 130.4, 252.9, we see that we have increased the weights (decreased the reciprocal weights, just given) on the two series with $s_i^2 = 200$ and 500 without any real justification, because, even after adjustment, it is clear that these two series were more variable than the others.

Accordingly, instead of being divided by

$$c(9|10, 6) = 1.534$$

and

$$c(10|10, 6) = 1.977$$

they deserve, if anything, to be divided by

$$c(1|2, 6) = .623$$

and

$$c(2|2, 6) = 1.231$$

giving 321 and 406, respectively, instead of 130 and 253.

The course goes on to develop a rule to decide on the natural lower group on the basis of the observation that misweighting by factor of no more than 2 is almost harmless.

FINAL REMARKS

The material reprinted here is quite typical for John's course notes. He put the emphasis on problems, on general approaches and on a few specific methods he considered to be reasonable. The historical development of an area and the systematic discussion of all the relevant ideas was not his style. In private discussion with him or in asking questions during his lectures, however, the students got a glimpse of the depth of his knowledge. Because his courses were not of the standard type, he rarely recommended a textbook. In the tradition of Princeton's mathematics department, he expected students to read related material on their own and to ask questions if something remained unclear. John also seldomly put exercices to the students. His notes explained the computations necessary for implementing the methods and he expected the readers and especially his reasearch students to have understood the formulas and algorithms, if necessary by performing them on a few simple examples by hand. This expectation was not always met and this could make conversations with him faintly surreal. One could be drawn into a discussion of a new idea without fully understanding it, because it would have been necessary to reread some course material beforehand.

Teaching took a large amount of John's time and has remained one of his loves throughout his career. He met with his graduate student once a week and always tried hard not to miss this appointment. My impression was that he gave as much time to each as he thought was needed, but his personal interest in the topic the student worked on also had an impact. He was at times keen on seeing the results of a new idea.