#076

Reports from the Research Laboratories
of the
Department of Psychiatry
University of Minnesota

Detecting Latent Clinical Taxa, II: A Simplified Procedure, Some Additional Hitmax Cut Locators, A Single-Indicator Method, and Miscellaneous Theorems.[1]

by

Paul E. Meehl

August 15, 1968

Report Number PR-68-4

[1] This research was supported in part by a grant from the National Institute of Mental Health, United States Public Health Service, Research Grant # M4465.
[Pagination in this posted version differs from the original publication.]

Notes added for this digital copy

Subscripts 't' for taxon and 'c' for complement groups were used in later publications. This research report uses 's' (schizotype),'n' (nonschizotype), and 't' (total sample).

Pagination in this version differs from the original. Page locations in references to the previous report PR-65-2 have been deleted because the posted version of that has changed pagination as well. A few mis-typings were corrected in equations and are noted by accompanying comments.

For convenience, a Contents list has been added.

Contents

Detecting Latent Clinical Taxa, II: A Simplified Procedure, Some Additional Hitmax Cut Locators, A Single-Indicator Method, and Miscellaneous Theorems

Paul E. Meehl

In a previous contribution to this research report series (Meehl, 1965 *[PR-65-2]*) I suggested a new method of identifying latent clinical taxa for the presumed dichotomous case, and proved a number of theorems arising from an idealized model of the latent situation, in which it is assumed that the intra-taxon covariances between pairs of fallible quantitative indicators are zero (or, more generally, and adequate for most of the derivations involved, that they are at least approximately equal as between the two latent taxa). A Monte Carlo study of one stage of the proposed sequential procedure, namely, that locating the hitmax cut on a single indicator by plotting the curve of the covariance of another pair of indicators as a function of the first indicator, was sufficiently encouraging with regard to accuracy and sample size required to justify further inquiry into the method's value (Seth, 1965). The present brief report (1) provides a more direct approach, and one presumably less subject to sampling instability, than that proposed in the earlier report, (2) provides additional hitmax cut locators, (3) suggests an alternative method which, by making the assumption of approximately normality within taxa, permits the desired inferences to be drawn from data involving only a single indicator-variable, (4) presents several new theorems concerning the latent situation, and (5) offers some tentative suggestions about lines of further development *[this is section 4g]*.

The basic ideas and developments, as well as notation, of the previous report will be presupposed in what follows, so it is desirable to have ready access to that report when reading this one. *[In a 1981 reprinting of PR-65-2, Meehl advised: Sections 4-5-6 and Appendix are now obsolete, being replaced by the procedure in Meehl: Psychodiagnosis (1973) Chapter 12 "MAXCOV-HITMAX" where the hitmax interval covariance test (described on pp. 28-29 of the original PR-65-2) is elevated from the role of a consistency test to that of a main estimator. See also Meehl and Golden "Taxometric methods" in P.C. Kendall and J.N. Butcher (eds.) Handbook of research methods in clinical psychology (1982) and references cited therein.—LJY]*

1. A more direct method of estimating the latent parameters

In PR-65-2, while several alternative methods of estimating certain quantities were proposed, the basic procedure consisted of the following steps:

A. Locating the hitmax cut on a given indicator by finding the interval of its distribution within which the covariance of another indicator-pair is maximized (Section 3 of PR-65-2).

B. Locating cuts on a pair of variables which equate the (latent) valid positive and valid negative rates, by utilizing a relationship between covariances and squares of observed frequencies above and below cuts and the latent positive and negative rates (Section 4).

C. Relying on the assumption that the latent valid positive and valid negative rates have been successfully equated by this preceding procedure (B), the observed frequency of tallies in a fourfold table determined by such rate-equating cuts can then be expressed in terms of latent values, including the unknown latent base-rate *P*, in the form of a system of three quadratics (Section 5).

D. Given an estimate of the latent base-rate *P*, we can write two equations in the unknown latent means for each indicator, one equation expressing the grand mean in terms of the latent means and base-rates, the other equation expressing the mean of cases lying in the hitmax interval in terms of the same pair of unknown latent means with coefficients [1/2, 1/2] since these latter are the frequencies of the two taxa in the hitmax interval. Solving this pair of equations gives us the latent means (Section 6).

E. Possessing estimates of the latent means on one indicator enables us to estimate the latent frequencies within any class interval of another indicator, since the manifest mean within such an interval is a weighted composite of the latent means, the weights being proportional to the frequencies of the two taxa in the interval (Section 7).

While these computations are not excessively onerous, and while the method includes an extensive list of consistency tests for corroborating the latent model (Section 9), the method to be proposed here is much shorter, more straightforward, relies upon estimates of fewer latent quantities, and has the special advantage that it dispenses with the system of three quadratics, which can surely be supposed to be excessively subject to random sampling fluctuations, and

therefore to demand a larger sample size for application of the method than will typically be available.

Essentially, the straightforward method elevates one of the consistency tests, the "hitmax interval covariance test" (Section 9-c) from the status of a mere consistency test to a main method of estimating the desired latent values.

In PR-65-2 it was shown that the (*xy*)-covariance of a mixed group of schizotypes and nonschizotypes, given the assumption of negligible intrataxon covariance, depends only upon the amount of "mixture," as shown in equation (3) in PR-65-2, thus

[1] $$\text{cov}(xy) = \text{K}pq = \Delta\bar{x}\Delta\bar{y}\,pq = (\bar{x}_\text{s} - \bar{x}_\text{n})(\bar{y}_\text{s} - \bar{y}_\text{n})pq$$

Having used this expression for the observed (*xy*)-covariance in a mixed batch to show that the hitmax cut on an input variable, say *w*, can be located by finding the *w*-interval within which the (*xy*)-covariance is a maximum (since Equation [1] is maximized when $p = q = 1/2$), we infer that the observed value of the manifest covariance at its maximum (in the hitmax interval where $p = q = 1/2$) is 1/4 times the product of the latent mean differences. We can therefore use the statistics of that hitmax interval to write directly

[2] $$\text{K} = \Delta\bar{x}\Delta\bar{y} = 4\,\text{cov}(xy)_{w_\text{h}}$$

and solve for the constant K.

But since Equation [1] holds for all *w*-intervals, knowing the value of K provides a direct method for calculating the latent schizotype-nonschizotype ratio in any *w*-interval, not only at hitmax. In any *w*-interval, the covariance of the output indicators *x* and *y* being an observable statistic, we have

$$\text{K}p_i q_i = \text{cov}(xy)_i$$

$$p_i(1 - p_i) = \frac{\text{cov}(xy)_i}{\text{K}}$$

[3] $$p_i^2 - p_i + \frac{\text{cov}(xy)_i}{\text{K}} = 0$$

so that all we need do, after once solving for the latent constant K in the hitmax interval, is then to apply repeatedly the quadratic roots formula to Equation [3] within each *w*-interval to get the proportion p_i of schizotypes in that interval.

Multiplying the p_i and corresponding q_i [= 1 – p_i] thus obtained for each *w*-interval by the absolute (manifest) frequency N_i within each *w*-interval yields the latent absolute frequencies N_{si} and N_{ni} for the interval.

The result of repeating this process over the whole range of *w*-intervals is the inferred pair of latent frequency-functions $f_s(w)$ and $f_n(w)$, from which the latent means $\bar{x}_s$, $\bar{x}_n$, $\bar{y}_s$, and $\bar{y}_n$ can now be, so to speak, computed directly. And of course the sum of all of the schizotypal and nonschizotypal absolute frequencies over all intervals gives us the total number N_s of schizotypes and N_n of nonschizotypes in the finite sample, and therefore yields a direct estimate of the latent base-rate *P* and its complement *Q*.

The base-rate *P* can also be inferred more directly from hitmax covariance and grand covariance (i.e., without the intermediate step of estimating each interval proportion p_i), as explained in Section 2-c below.

2. Some further theorems and procedures

a. *An additional hitmax cut locator, employing the derivative of the difference between an output variable's sum above and below a sliding cut*

Consider the sum (not average) of values of an output variable *y* for patients lying above an arbitrary cut on an input indicator variable *x*, and the corresponding sum of *y* values below that arbitrary *x*-cut. Then we show that the rate of change of the difference between these *y*-sums with respect to the cumulative frequency N_b below the *x*-cut is equal to minus twice the *y*-mean in the interval surrounding the cut, when the cut is the hitmax cut.

Proof

[4] $$\sum^{N_a} y_a = \sum^{H_s} y_a + \sum^{M_n} y_a = H_s \bar{y}_s + M_n \bar{y}_n$$

[5] $$\sum^{N_b} y_b = \sum^{H_n} y_b + \sum^{M_s} y_b = H_n \bar{y}_n + M_s \bar{y}_s$$

Define a difference function of the cumulative frequency (up to a sliding cut on *x*)

[6] $$G(N_b) = \sum^{N_a} y_a - \sum^{N_b} y_b$$
$$= \left(H_s \bar{y}_s + M_n \bar{y}_n\right) - \left(H_n \bar{y}_n + M_s \bar{y}_s\right)$$

Differentiating with respect to N_b

$$\frac{d\mathrm{G}}{dN_b} = \frac{dH_s}{dN_b}\overline{y}_s + \frac{dM_n}{dN_b}\overline{y}_n - \frac{dH_n}{dN_b}\overline{y}_n - \frac{dM_s}{dN_b}\overline{y}_s$$

$$= \left(\frac{dH_s}{dN_b} - \frac{dM_s}{dN_b}\right)\overline{y}_s - \left(\frac{dH_n}{dN_b} + \frac{dM_n}{dN_b}\right)\overline{y}_n$$

$$= \left(-\frac{dM_s}{dN_b} - \frac{dM_s}{dN_b}\right)\overline{y}_s - \left(\frac{dH_n}{dN_b} + \frac{dH_n}{dN_b}\right)\overline{y}_n$$

[7] $$= -2\left(-\frac{dM_s}{dN_b}\overline{y}_s + \frac{dH_n}{dN_b}\overline{y}_n\right)$$

these derivatives being everywhere positive and $\frac{d\mathrm{G}}{dN_b}$ being everywhere negative.

At hitmax cut (and nowhere else) we have also

[8] $$\frac{dM_s}{dN_b} = \frac{dH_n}{dN_b} = \frac{1}{2}$$

so at hitmax we have, putting [8] in [7],

[9] $$\frac{d\mathrm{G}}{dN_b} = -2\left(\frac{1}{2}\overline{y}_s + \frac{1}{2}\overline{y}_n\right) = -\left(\overline{y}_s + \overline{y}_n\right)$$

But we know that, for the cases lying within the hitmax interval of x, the mean y is, in latent terms,

[10] $$\overline{y}_{hx} = \frac{1}{2}\left(\overline{y}_s + \overline{y}_n\right)$$

So from [9]-[10] we infer that at the hitmax cut and within the interval containing that cut,

[11] $$\frac{d\mathrm{G}}{dN_b} = -2\overline{y}_{hx}$$

In practice, we would plot the graph of the quantity $(\sum y_a - \sum y_b)$ against the abscissa N_b and estimate the slope of this graph at intervals, let us say graphically. We also plot $\overline{y}_x$ at various N_b values (i.e. within various x-intervals) and there should appear a value of N_b (corresponding to a value of x) at which point the slope of the first graph equals twice the value of the second graph. (Physical intuition assures us that there will only be one such place, although I have not shown analytically that the derivative of the difference of sums equals $-2\overline{y}_x$ at only one point).

b. ***An additional hitmax cut locator, employing the difference between an output* variable's *mean* (*or, if a qualitative sign, its* (+)*-rates*) *above and below a sliding cut***

If a family of indicators are correlated solely or mainly by virtue of each indicator's being discriminative of the latent taxa, it seems intuitively that when the quantitative members of this family are treated dichotomously as clinical "signs" by locating a cut, above which the quantitative indicator is called "positive" and below which it is called "negative," then a very poor choice of such a cut should have, in general, a tendency to reduce the statistical tendency of such "signs" to go together. For example, suppose a neurologist were dealing with the clinical taxonomy of meningitis versus non-meningitis, and the two indicator-variables under consideration were temperature and neck-stiffness. A moderate-to-high temperature elevation being associated with meningitis, and marked pain on anteroflection of the neck also being a clinical sign of meningitis, the combination of signs "high temperature with marked stiff neck" would occur together much more often than by chance in a clinical population of meningitic versus non-meningitic patients. But if our clinician were so unwise or unlucky as to have chosen a very low cutting score on these two variables, such as any temperature above 99.0° and any sign or complaint, however slight, of stiff neck or reluctance to flex the neck, then considerable numbers of patients without meningitis but with other milder infectious conditions (even including the common cold) would show one or both of the "signs," and two untoward results would be expected. First, the manifest correlation between the two clinical indicators would be reduced; secondly, in terms of the latent situation, the identification of the taxon of interest, namely meningitis, would be poorer. This general line of thought suggests that another possible approach to the hitmax cut location would be some kind of maximizing of the "agreement" between two signs, this agreement being hopefully higher when each sign is optimally (or near-optimally) located. Unfortunately, one's intuitions here do not seem to be confirmed, at least by any definition of "concordance" or "agreement" that I have investigated, and I have tried half-a-dozen such without success. Thus, for example, one cannot show that a concordance index defined as being the proportion of total cases found in the concordant cells of a fourfold table (i.e., the sum of cases in the first and third quadrants) is a maximum when each variable is cut optimally; and, in fact, it can be shown that the hitmax cuts will maximize the concordance of this joint-sign table only under the special condition that the latent valid positive and valid negative rates on each indicator are symmetrical which will rarely be the case in practice; or, at

least, however often it may be approximately true, we have no way of knowing whether it is true or not. The basic reason why intuitively likely possibilities, such as the sum of *y*-positives above and *y*-negatives below an *x*-cut, cannot be maximized to find the hitmax cut on *x* is the asymmetry between $(p_s - q_s)$ and $(p_n - q_n)$. It may be that I have simply not been ingenious enough to hit upon the right (somewhat arbitrary) index for measuring concordance, and the reader is invited to try his hand at it, if his general intuitions are the same as mine that some such hit-maximizing concordance index must exist.

However, one intuitively plausible manifest criterion turned out to be an almost perfect hitmax cut locator every time it was applied to "fake data," including situations of considerable base-rate asymmetry and departures from normality. Paper and pencil numerical runs suggested that the proposed criterion almost always located the hitmax cut correctly, and on the rare occasions when it did not do so, it was never found to be in error by more than one class-interval (unit integral increments in *x*). The proposed index of (*xy*)-concordance is the difference between $\bar{y}_{ax}$ (mean of *y* above the *x*-cut) and $\bar{y}_{bx}$ (mean of *y* below the *x*-cut).[2] If, instead of dealing with quantitative values of *y* we have already chosen an arbitrary *y*-cut on some other basis, or the *y*-variable as it is presented to us clinically exists only as a qualitative "sign," maximizing the difference of the *y*-means above and below an *x*-cut is algebraically equivalent to maximizing the difference between the proportion of *y*(+) cases ["sign-positive"] lying above and below the *x*-cut.

It turns out, however, that when we express this index in terms of the latent quantities, its maximum value is not exactly identical with the value at the hitmax cut, shown as follows: Define

$$\bar{y}_{ax} = \frac{\sum^{N_{ax}} y_{ax}}{N_{ax}} \qquad \text{and} \qquad \bar{y}_{bx} = \frac{\sum^{N_{bx}} y_{bx}}{N_{bx}}$$

and simplify notation by dropping the *x*-subscripts which occur throughout. Then the proposed concordance-index is

[12] $$y_d(x) = \bar{y}_a - \bar{y}_b = \frac{\sum y_a}{N_a} - \frac{\sum y_b}{N_b}$$

---

[2] *This subsequently became the MAMBAC (Mean Above minus Mean Below A sliding Cut) procedure for detecting taxonicity and estimating latent parameters (Meehl & Yonce, 1994).—LJY*

which in terms of latent variables is

$$y_{\mathrm{d}}(x) = \frac{H_{\mathrm{s}}\overline{y}_{\mathrm{s}} + M_{\mathrm{n}}\overline{y}_{\mathrm{n}}}{N_{\mathrm{a}}} - \frac{H_{\mathrm{n}}\overline{y}_{\mathrm{n}} + M_{\mathrm{s}}\overline{y}_{\mathrm{s}}}{N_{\mathrm{b}}}$$

$$= \frac{H_{\mathrm{s}}}{N_{\mathrm{a}}}\overline{y}_{\mathrm{s}} + \frac{M_{\mathrm{n}}}{N_{\mathrm{a}}}\overline{y}_{\mathrm{n}} - \frac{H_{\mathrm{n}}}{N_{\mathrm{b}}}\overline{y}_{\mathrm{n}} - \frac{M_{\mathrm{s}}}{N_{\mathrm{b}}}\overline{y}_{\mathrm{s}} \qquad [13]$$

The latent difference between the schizotype and nonschizotype y-means being a constant unaffected by our choice of x-cut, define

$$\Delta\overline{y} = \overline{y}_{\mathrm{s}} - \overline{y}_{\mathrm{n}} = \text{A constant} > 0 \qquad [14]$$

and substitute

$$\overline{y}_{\mathrm{s}} = \overline{y}_{\mathrm{n}} + \Delta\overline{y}$$

in [13] to get

$$\overline{y}_{\mathrm{d}}(x) = \frac{H_{\mathrm{s}}}{N_{\mathrm{a}}}\left(\overline{y}_{\mathrm{n}} + \Delta\overline{y}\right) + \frac{M_{\mathrm{n}}}{N_{\mathrm{a}}}\overline{y}_{\mathrm{n}} - \frac{H_{\mathrm{n}}}{N_{\mathrm{b}}}\overline{y}_{\mathrm{n}} - \frac{M_{\mathrm{s}}}{N_{\mathrm{b}}}\left(\overline{y}_{\mathrm{n}} + \Delta\overline{y}\right)$$

$$= \left(\frac{H_{\mathrm{s}}}{N_{\mathrm{a}}} + \frac{M_{\mathrm{n}}}{N_{\mathrm{a}}}\right)\overline{y}_{\mathrm{n}} - \left(\frac{H_{\mathrm{n}}}{N_{\mathrm{b}}} + \frac{M_{\mathrm{s}}}{N_{\mathrm{b}}}\right)\overline{y}_{\mathrm{n}} + \frac{H_{\mathrm{s}}}{N_{\mathrm{a}}}\Delta\overline{y} - \frac{M_{\mathrm{s}}}{N_{\mathrm{b}}}\Delta\overline{y}$$

$$= \frac{N_{\mathrm{a}}}{N_{\mathrm{a}}}\overline{y}_{\mathrm{n}} - \frac{N_{\mathrm{b}}}{N_{\mathrm{b}}}\overline{y}_{\mathrm{n}} + \frac{H_{\mathrm{s}}}{N_{\mathrm{a}}}\Delta\overline{y} - \frac{M_{\mathrm{s}}}{N_{\mathrm{b}}}\Delta\overline{y}$$

$$= \frac{H_{\mathrm{s}}}{N_{\mathrm{a}}}\Delta\overline{y} - \left(1 - \frac{H_{\mathrm{n}}}{N_{\mathrm{b}}}\right)\Delta\overline{y}$$

$$= \Delta\overline{y}\left(\frac{H_{\mathrm{s}}}{N_{\mathrm{a}}} + \frac{H_{\mathrm{n}}}{N_{\mathrm{b}}} - 1\right) \qquad \Delta\overline{y} > 0 \qquad [15]$$

which is maximized by maximizing the term in parentheses, hence, by maximizing the sum of the proportion of hits above and the proportion of hits below the x-cut.

Since it is not true in general that $(H_{\mathrm{s}} + H_{\mathrm{n}}) \longrightarrow \text{Max}$

when $\left(\frac{H_{\mathrm{s}}}{N_{\mathrm{a}}} + \frac{H_{\mathrm{n}}}{N_{\mathrm{b}}}\right) \longrightarrow \text{Max}$

we cannot show that maximizing the latter will locate hitmax exactly. But we can show that it is a very good approximation, the error in analytic maximization being smaller than we can expect from sheer sampling irregularities, and usually less than a class-interval increment in x. The proof, while not recondite, is unfortunately somewhat tedious.

[Quasi]-Proof

It will be convenient to take derivatives with respect to N_b [= cumulative frequency of manifest (mixed) distribution up to x-cut] rather than the usual derivative w.r.t. the abscissa-variable x, of which N_b is an increasing monotonic function.

We define

[16] $$h_a = \frac{H_s}{N_a}$$

[17] $$h_b = \frac{H_n}{N_b}$$

Hit-rates above and below cut, respectively

[18] $h = h_a + h_b$ Sum of these two hit-rates

To maximize the sum of hit-rates above and below the cut (obviously a maximum if an extremum, from the physical situation) we set

$$\frac{dh}{dN_b} = \frac{d}{dN_b}\left(h_a + h_b\right) = 0$$

[19] $$\frac{dh_a}{dN_b} = -\frac{dh_b}{dN_b}$$

which we want to prove is approximately true at the hitmax cut. (It is obvious that these rates of change have opposite sign, everywhere.)

Left-hand side is

$$\frac{dh_a}{dN_b} = \frac{d}{dN_b}\left(\frac{H_s}{N_a}\right) = \frac{N_a \frac{dH_s}{dN_b} - H_s \frac{dN_a}{dN_b}}{N_a^2}$$

[20] $$= \frac{1}{N_a}\left(\frac{dH_s}{dN_b} + \frac{H_s}{N_a}\right) = \frac{1}{N_a}\left(\frac{dH_s}{dN_b} + h_a\right)$$

Right-hand side is, neglecting minus-sign,

$$\frac{dh_b}{dN_b} = \frac{d}{dN_b}\left(\frac{H_n}{N_b}\right) = \frac{N_b \frac{dH_n}{dN_b} - H_n \frac{dN_b}{dN_b}}{N_b^2}$$

[21] $$= \frac{1}{N_b}\left(\frac{dH_n}{dN_b} - \frac{H_n}{N_b}\right) = \frac{1}{N_b}\left(\frac{dH_n}{dN_b} - h_b\right)$$

At hitmax,

$$\frac{dH_s}{dN_b} = -\frac{1}{2} \qquad \text{and} \qquad \frac{dH_n}{dN_b} = +\frac{1}{2}$$

which values being substituted in [20] - [21] give

[22] $$\frac{dh_a}{dN_b} = \frac{1}{N_a}\left(\frac{H_s}{N_a} - \frac{1}{2}\right)$$

[23] $$\frac{dh_b}{dN_b} = \frac{1}{N_b}\left(\frac{1}{2} - \frac{H_n}{N_b}\right)$$

Putting these values into [19], we want to show that

$$\frac{1}{N_a}\left(\frac{H_s}{N_a} - \frac{1}{2}\right) \simeq -\frac{1}{N_b}\left(\frac{1}{2} - \frac{H_n}{N_b}\right)$$

$$\left(\frac{H_s}{N_a^2} - \frac{H_n}{N_b^2}\right) - \frac{1}{2}\left(\frac{N_b - N_a}{N_a N_b}\right) \simeq 0$$

[24] $$\left(\frac{h_a}{N_a} - \frac{h_b}{N_b}\right) - \frac{1}{2}\left(\frac{N_b - N_a}{N_a N_b}\right) \simeq 0$$

which does not simplify further. If one puts in numerical values appropriate to the physical context, where the *h*'s are decimals and the *N*'s are numbers of order 10^2, it is obvious from the algebraic structure of this error-term that it will be very small. I have not been able to concoct numerical combinations where this error >.01, and usually it is smaller than that by one or two orders of magnitude.

From geometric intuition it is pretty clear that when marked asymmetries $N_b \gg N_a$ obtain, the $h_b : h_a$ imbalance will be in the same direction, which exerts a "compensatory" influence in the error-expression. But even with these asymmetries large and (I think impossibly) reversed, such as $200 = N_b \gg N_a = 50$ and $.60 = h_b < h_a = .90$, one gets an error of only .0075.

However, it may not be sufficient that the error in finding the derivative's zero is absolutely small, because our concern is with the mis-location of an abscissa-value, via a mis-location of the optimal cumulative frequency-value N_b. That is, we need to know how large an error ΔN_b in

finding the N_b-value is produced by the error in mis-calling the hitmax cut as the cut which zeros the derivative $\frac{dh}{dN_a}$. This amounts to the question, "How much does N_b change per change in $\frac{dh}{dN_b}$?" We have therefore to deal with the second derivative.

(It is convenient to use hitmax cut-values as substitutions, since these simplify; so the error is being measured in terms of what should be, not where we are.)

We want to know the error in x-location due to approximating the zero of $\frac{dh}{dN_b}$ as if it were at hitmax, which it is not precisely. To get this x-error we must get the error in N_b.

[25] Let $v = \frac{d}{dN_b}\left(h_a + h_b\right)$

Then we are going to approximate the error ΔN_b by the differential, so we want to evaluate

[26] $\frac{dN_b}{dv}\Delta v = \Delta N_b$ approximately.

So we want to evaluate

$$\frac{dN_b}{dv} = \frac{1}{\frac{dv}{dN_b}} = \frac{1}{\frac{d}{N_b}\left[\frac{d\left(h_a + h_b\right)}{dN_b}\right]}$$

[27] $$= \frac{1}{\frac{d^2\left(h_a + h_b\right)}{dN_b^2}}$$

The denominator (second derivative of sum of hit-rates above and below x-cut with respect to cumulative frequency) is

$$\text{Den} = \frac{d}{dN_b}\left[\frac{1}{N_a}\left(\frac{dH_s}{dN_b} + \frac{H_s}{N_a}\right)\right] + \frac{d}{dN_b}\left[\frac{1}{N_b}\left(\frac{dH_n}{dN_b} - \frac{H_n}{N_b}\right)\right]$$

which at hitmax is

[28] $$\frac{d}{dN_b}\left[\frac{1}{N_a}\left(-\frac{1}{2} + h_a\right)\right] + \frac{d}{dN_b}\left[\frac{1}{N_b}\left(\frac{1}{2} - h_b\right)\right]$$

which expands and re-arranges to give

$$\frac{1}{N_a}\frac{d}{dN_b}\left(\frac{H_s}{N_a}\right)-\frac{1}{N_b}\frac{d}{dN_b}\left(\frac{H_n}{N_b}\right)+\frac{h_a-\frac{1}{2}}{N_a^2}+\frac{h_a-\frac{1}{2}}{N_b^2}$$

$$=\frac{1}{N_a}\left(\frac{N_a\frac{dH_a}{dN_b}-H_s\frac{dN_a}{dN_b}}{N_a^2}\right)-\frac{1}{N_b}\left(\frac{N_b\frac{dH_n}{dN_b}-H_n\frac{dN_b}{dN_b}}{N_b^2}\right)+\frac{h_a-\frac{1}{2}}{N_a^2}+\frac{h_a-\frac{1}{2}}{N_b^2}$$

which at hitmax is

$$=\frac{1}{N_a}\left(\frac{N_a\left(-\frac{1}{2}\right)-H_s(-1)}{N_a^2}\right)-\frac{1}{N_b}\left(\frac{N_b\left(\frac{1}{2}\right)-H_n(+1)}{N_b^2}\right)+\frac{h_a-\frac{1}{2}}{N_a^2}+\frac{h_a-\frac{1}{2}}{N_b^2}$$

which simplifies to

[29] $$=\frac{2}{N_a^2}\left(h_a-\frac{1}{2}\right)+\frac{2}{N_b^2}\left(h_b-\frac{1}{2}\right)$$

This is the denominator of $\frac{1}{\frac{dN_b}{dv}}$

so its reciprocal is the rate of change of N_b with respect to the first derivative $\frac{d}{dN_b}\left(h_a+h_b\right)$. So we have, at hitmax,

[30] $$\frac{dN_b}{dv}=\frac{1}{\frac{2}{N_a^2}\left(h_a-\frac{1}{2}\right)+\frac{2}{N_b^2}\left(h_b-\frac{1}{2}\right)}$$

The error in the derivative v is

[31] $$\Delta v=\left(\frac{h_a}{N_a}-\frac{h_b}{N_b}\right)-\frac{1}{2}\left(\frac{N_b-N_a}{N_aN_b}\right)\neq 0$$

because it should be at zero and the above is its actual value at hitmax, from Equation [30].

So the error N_b is, approximating by the differential,

/

[32] $$\Delta N_{\mathrm{b}} \simeq \frac{dN_{\mathrm{b}}}{dv}\Delta v = \frac{\left(\frac{h_{\mathrm{a}}}{N_{\mathrm{a}}} - \frac{h_{\mathrm{b}}}{N_{\mathrm{b}}}\right) - \frac{1}{2}\left(\frac{N_{\mathrm{b}} - N_{\mathrm{a}}}{N_{\mathrm{a}}N_{\mathrm{b}}}\right)}{\frac{2}{N_{\mathrm{a}}^2}\left(h_{\mathrm{a}} - \frac{1}{2}\right) + \frac{2}{N_{\mathrm{b}}^2}\left(h_{\mathrm{b}} - \frac{1}{2}\right)}$$

substituting from [30] and [31].

This simplifies to

[33] $$\Delta N_{\mathrm{b}} \simeq \frac{dN_{\mathrm{b}}}{dv}\Delta v = \frac{N_{\mathrm{a}}N_{\mathrm{b}}}{2}\left[\frac{N_{\mathrm{b}}\left(h_{\mathrm{a}} - \frac{1}{2}\right) - N_{\mathrm{a}}\left(h_{\mathrm{b}} - \frac{1}{2}\right)}{N_{\mathrm{b}}^2\left(h_{\mathrm{a}} - \frac{1}{2}\right) + N_{\mathrm{a}}^2\left(h_{\mathrm{b}} - \frac{1}{2}\right)}\right]$$

which is easily shown (by plugging in "bad and unlikely" numerical values) to be satisfactorily small.

Numerical Examples

A "bad" (large-error) setup would be if $N_{\mathrm{b}} \gg N_{\mathrm{a}}$ but $h_{\mathrm{a}} \gg h_{\mathrm{b}}$, unlikely [impossible?] to occur because large h_{a} will, as pointed out above, be associated with the larger N_{a}, and this makes the numerator of [33] self-corrective toward smallness. But suppose we have the bad situation

$$N_{\mathrm{a}} = 50 \qquad h_{\mathrm{a}} = .90$$
$$N_{\mathrm{b}} = 200 \qquad h_{\mathrm{b}} = .60$$

Substituting these values in the expression for ΔN_{b} yields an error

$$\Delta N_{\mathrm{b}} = \frac{(200)(50)}{2}\left[\frac{(200)(.40) - (50)(.10)}{(40{,}000)(.40) - (2{,}500)(.10)}\right] = 23.8$$

so that even this extreme setup is off only some 24 cases in the cumulative frequency N_{b}, which would not shift the cut more than one class-interval (given sample size appropriate for the method's use, and operating in the middle region of the manifest x-distribution.) Note: I am inclined from geometry to believe that these numerical values are impossible, but have not been able to prove it analytically.

The usual situation would be for the $N_{\mathrm{a}} : N_{\mathrm{b}}$ and $h_{\mathrm{a}} : h_{\mathrm{b}}$ asymmetries to go in the same direction. Suppose the values are [I chose these arbitrarily, not ad hoc]

$$N_{\mathrm{a}} = 50 \qquad h_{\mathrm{a}} = .60$$
$$N_{\mathrm{b}} = 200 \qquad h_{\mathrm{b}} = .90$$

Then the error in ΔN_b is

$$\Delta N_b = \frac{(200)(50)}{2}\left[\frac{(200)(.10)-(50)(.40)}{(40{,}000)(.10)-(2{,}500)(.40)}\right] = 0.0000$$

A more symmetrical case, with both the hit-rates and total frequencies nearer to equality above and below,

$$N_a = 100 \qquad h_a = .70$$
$$N_b = 150 \qquad h_b = .80$$

gives us

$$\Delta N_b \simeq \left[\frac{(150)(.20)-(100)(.30)}{(150)^2(.20)-(100)^2(.30)}\right] = 0.0000$$

Or, taking a "weaker" indicator, with less discriminating power than one of the better MMPI keys, suppose

$$N_a = 100 \qquad h_a = .65$$
$$N_b = 150 \qquad h_b = .75$$

the error in cumulative frequency is

$$\Delta N_b = \frac{(100)(150)}{2}\left[\frac{(150)(.15)-(100)(.25)}{(22{,}500)(.15)+(10{,}000)(.25)}\right]$$

$$= 7500\left(\frac{-2.5}{5875}\right) \qquad < 4 \text{ cumulative cases,}$$

which would very rarely displace the cut into an adjoining x-interval.

I conclude that the maximum of sum of hit-rates above and below an x-cut is achieved by a cut very close to hitmax, and therefore the x-cut which maximizes the manifest statistic $y_d(x) = \overline{y}_a(x) - \overline{y}_b(x)$ is approximately the hitmax cut, give or take an error of one class-interval at most.

If the y-indicator is a dichotomous "sign," this procedure amounts to finding the x-cut such that the y^+-rates above and below the x-cut differ maximally. The latent situation is then represented by

[34] $$p^+_{ya}(x) - p^+_{yb}(x) = \Delta\overline{y}\left(\frac{H_s}{N_a} + \frac{H_n}{N_b} - 1\right) = \left(p_{sy} - p_{ny}\right)\left(\frac{H_s}{N_a} + \frac{H_n}{N_b} - 1\right)$$

[35] $$= \Delta p_y \left(\frac{H_{\mathrm{s}}}{N_{\mathrm{a}}} + \frac{H_{\mathrm{n}}}{N_{\mathrm{b}}} - 1 \right)$$

since the "mean value" of a dichotomous indicator is simply the proportion of cases where it is present. This relation shows further that if the "output" indicator y is initially continuous but treated dichotomously (by locating a y-cut), the preliminary y-cut which maximizes the maximum attainable for $p^{+}_{y\mathrm{a}}(x) - p^{+}_{y\mathrm{b}}(x)$ using the method of this section is that y-cut which yields, in the latent situation, the largest y-"validity" in one sense, to wit, the largest difference between the valid positive rate $p_{\mathrm{s}y}$ and the false positive rate $p_{\mathrm{n}y}$. So that a joint search for sliding cuts on both x and y which chooses the cut-pair $(x_{\mathrm{c}}, y_{\mathrm{c}})$ by maximizing the difference in (y^{+})-"sign" proportions above and below the x-cut has the simultaneous property that

1. $(p_{\mathrm{s}y} - p_{\mathrm{n}y})$ is maximized, exactly
2. $(H_{\mathrm{s}x} + H_{\mathrm{n}x})$ is maximized, approximately

c. *Estimating base-rate P most directly from two covariances and K*

In Section 1 the base-rate P is estimated by summing the estimated latent interval frequencies $n_{\mathrm{s}i}$ and $n_{\mathrm{n}i}$ over all intervals. These latent interval-frequencies had in turn been estimated via the constant $\mathrm{K} = (\bar{x}_{\mathrm{s}} - \bar{x}_{\mathrm{n}})(\bar{y}_{\mathrm{s}} - \bar{y}_{\mathrm{n}})$, inferred from the value of the hitmax interval covariance. But we can bypass the summation procedure and estimate P directly, once K is obtained. The grand covariance of the manifest distribution depends upon K and P only, that is, in latent terms,

[36] $$\mathrm{cov}_{\mathrm{t}} = PQ(\bar{x}_{\mathrm{s}} - \bar{x}_{\mathrm{n}})(\bar{y}_{\mathrm{s}} - \bar{y}_{\mathrm{n}})$$

[37] $$= \mathrm{K}P(1-P)$$

a quadratic in P, since K has been estimated from the hitmax statistic $\mathrm{cov}_{\mathrm{h}}(xy)$ and $\mathrm{cov}_{\mathrm{t}}(xy)$ is an observed statistic of the manifest (mixed) distribution. Solving this quadratic we have

$$P = \frac{\mathrm{K} \pm \sqrt{\mathrm{K}^2 - 4\mathrm{K}\,\mathrm{cov}_{\mathrm{t}}(xy)}}{2\mathrm{K}}$$

[38] $$= \frac{1}{2} \pm \sqrt{\frac{1}{4} - \frac{\mathrm{cov}_{\mathrm{t}}(xy)}{\mathrm{K}}}$$

Since $0 < \dfrac{\mathrm{cov}_{\mathrm{t}}(xy)}{\mathrm{K}} < 1/4$ both roots are real and lie between 0 and 1, so the selection of root cannot be made without additional considerations, i.e., we have to decide whether the schizotype

base-rate is $> 1/2$ or $< 1/2$. For this reason the more laborious cumulative method of Section 1 may be preferable, using the present method as a consistency test. However, we are practically certain to make the correct choice if we decide so that the $P : Q$ asymmetry is in the same direction as the $N_a : N_b$ asymmetry, because (as is shown in sub-section 2d following) the ratio of [(Hit-rate above) – 1/2] to [(Hit-rate below) – 1/2] is very nearly proportional to the manifest ratio $N_a : N_b$ of frequencies above and below the hitmax cut, from which it follows that the base-rate asymmetry is in the same direction as the asymmetry of cases above and below the cut. Only in a situation very close to $P = Q$ and $N_a = N_b$ would this choice be wrong; and in such situations, of course, the two roots will be so close together (both $\simeq 1/2$) that choosing wrongly is a pragmatically unimportant mistake.

d. ***A relation between the manifest frequencies above and below hitmax cut, and the latent hit-rates.***

A somewhat surprising near-proportionality obtains between the deviation of hit-rate from 1/2 and the manifest (mixed-taxon) frequency determined by a hitmax cut, although geometric intuition dispels some of the oddity. We define two hit-functions

[39] $$u_{ax} = h_{ax} - 1/2 = \frac{H_{sx}}{N_{ax}} - 1/2$$

[40] $$u_{bx} = h_{bx} - 1/2 = \frac{H_{nx}}{N_{bx}} - 1/2$$

dropping the x-subscript from here on, and noting that these hit-rates are proportions of the manifest frequencies N_{ax} and N_{bx}, <u>not</u> the "valid positive" and "valid negative" rates $p_s(x)$ and $p_n(x)$, whose denominators are the true taxon frequencies N_s and N_n.

Then we show that, to a very good approximation,

[41] $$\frac{u_a}{u_b} \simeq \frac{N_a}{N_b}$$

at the hitmax cut.

<u>Proof</u>

<u>Lemma</u>: We first obtain a (non-approximative) relation between the derivatives of the hit-functions $u_a(x)$, $u_b(x)$ and the manifest frequencies, namely,

[42] $$\frac{u'_{\mathrm{a}}(x)}{u'_{\mathrm{b}}(x)} = -\frac{N_{\mathrm{b}}}{N_{\mathrm{a}}}\frac{u_{\mathrm{a}}(x)}{u_{\mathrm{b}}(x)}$$

when $x = x_{\mathrm{hitmax}}$.

The hit-function derivatives are

$$u'_{\mathrm{a}}(x) = \frac{d}{dx}\left(h_{\mathrm{a}} - 1/2\right) = \frac{d}{dx}\frac{\left(H_{\mathrm{s}}\right)}{\left(N_{\mathrm{a}}\right)} = \frac{N_{\mathrm{a}}H'_{\mathrm{s}} - H_{\mathrm{s}}N'_{\mathrm{a}}}{N_{\mathrm{a}}^2}$$

Eq $u'_{\mathrm{a}}(x)$ =...
In numerator $H_{\mathrm{s}}N'_{\mathrm{a}}$ was originally mis-typed as $H_{\mathrm{a}}N'_{\mathrm{a}}$

$$= \frac{-f_{\mathrm{s}}N_{\mathrm{a}} - H_{\mathrm{s}}(-f_{\mathrm{t}})}{N_{\mathrm{a}}^2}$$

$$= \frac{-f_{\mathrm{s}}N_{\mathrm{a}} + 2f_{\mathrm{s}}H_{\mathrm{s}}}{N_{\mathrm{a}}^2} \qquad \text{Since } f_{\mathrm{t}} = 2f_{\mathrm{s}} \text{ at hitmax}$$

$$= \frac{f_{\mathrm{s}}}{N_{\mathrm{a}}}\left(2\frac{H_{\mathrm{s}}}{N_{\mathrm{a}}} - 1\right)$$

[43] $$u'_{\mathrm{a}}(x) = 2\frac{f_{\mathrm{s}}}{N_{\mathrm{a}}}u_{\mathrm{a}}$$

Similarly for the hit-function below,

$$u'_{\mathrm{b}}(x) = \frac{d}{dx}\left(h_{\mathrm{b}} - 1/2\right) = \frac{d}{dx}\frac{\left(H_{\mathrm{n}}\right)}{\left(N_{\mathrm{b}}\right)} = \frac{N_{\mathrm{b}}H'_{\mathrm{n}} - H_{\mathrm{n}}N'_{\mathrm{b}}}{N_{\mathrm{b}}^2}$$

$$= \frac{N_{\mathrm{b}}f_{\mathrm{n}} - H_{\mathrm{n}}(f_{\mathrm{t}})}{N_{\mathrm{b}}^2}$$

$$= \frac{N_{\mathrm{b}}f_{\mathrm{n}} - 2f_{\mathrm{n}}H_{\mathrm{n}}}{N_{\mathrm{b}}^2} \qquad \text{Since } f_{\mathrm{t}} = 2f_{\mathrm{n}} \text{ at hitmax}$$

$$= \frac{f_{\mathrm{n}}}{N_{\mathrm{b}}}\left(1 - \frac{2H_{\mathrm{n}}}{N_{\mathrm{b}}}\right)$$

[44] $$u'_{\mathrm{b}}(x) = -2\frac{f_{\mathrm{n}}}{N_{\mathrm{b}}}u_{\mathrm{b}}$$

Dividing [43] by [44] and recalling that $f_{\mathrm{s}} = f_{\mathrm{n}}$ at hitmax,

[45] $$\frac{u'_{\mathrm{a}}}{u'_{\mathrm{b}}} = -\frac{N_{\mathrm{b}}}{N_{\mathrm{a}}}\frac{u_{\mathrm{a}}}{u_{\mathrm{b}}}$$

which is in itself an interesting fact about the hitmax cut.

Given this lemma, the proof is immediate. In Section 2 (b) above we saw that the derivatives $\frac{d}{dN_b}(u_a)$ and $\frac{d}{dN_b}(u_b)$ are equal except for algebraic sign, to a very close approximation, at the hitmax cut. Hence, since

[46] $$u'_a = \left(\frac{d}{dN_b}u_a\right)\left(\frac{dN_b}{dx}\right) \qquad \text{Chain Rule}$$

[47] $$u'_b = \left(\frac{d}{dN_b}u_b\right)\left(\frac{dN_b}{dx}\right) \qquad \text{Chain Rule}$$

we can write, at hitmax,

[48] $$u'_a \simeq -u'_b$$

very nearly. Substituting [48] in [45] we get the desired

[49] $$\frac{u_a}{u_b} \simeq \frac{N_a}{N_b}$$

at the hitmax cut.

This can be obtained directly from the general expressions for derivatives of hit-rates above and below with respect to cumulative frequency, Equations [20]-[21] which expressed in terms of $u_a = (h_a - 1/2)$ and, $u_b = (h_b - 1/2)$ tell us that, everywhere,

[50] $$\frac{du_a}{dN_b} = \frac{1}{N_a}\left[u_a + \left(\frac{dH_s}{dN_b} + \frac{1}{2}\right)\right]$$

[51] $$\frac{du_b}{dN_b} = -\frac{1}{N_b}\left[u_b + \left(\frac{dH_n}{dN_b} - \frac{1}{2}\right)\right]$$

so at hitmax, where

[52] $$\frac{dH_s}{dN_b} = -\frac{1}{2}$$

[53] $$\frac{dH_n}{dN_b} = +\frac{1}{2}$$

we have

[54] $$\frac{du_a}{dN_b} = \frac{u_a}{N_a}$$

[55] $$\frac{du_{\mathrm{b}}}{dN_{\mathrm{b}}} = -\frac{u_{\mathrm{b}}}{N_{\mathrm{b}}}$$

and therefore, at hitmax,

[56] $$\frac{du_{\mathrm{a}}}{dN_{\mathrm{b}}} = -\frac{N_{\mathrm{b}}}{N_{\mathrm{a}}}\frac{u_{\mathrm{a}}}{u_{\mathrm{b}}}$$

as well as [49].

From [50] and [52] we see that when a cut lies above hitmax, we have

[57] $$\left|\frac{du_{\mathrm{a}}}{dN_{\mathrm{b}}}\right| < \frac{u_{\mathrm{a}}}{N_{\mathrm{a}}}$$

that is, the rate of change of hit-function u above the cut is numerically smaller than the ratio of this function to frequency above the cut. Correspondingly from [51] and [53] the derivative of the hit-function u_{b} below the cut with respect to cumulative frequency has an absolute value larger than the below-cut ratio,

[58] $$\left|\frac{du_{\mathrm{b}}}{dN_{\mathrm{b}}}\right| > \frac{1}{N_{\mathrm{b}}}\left[u_{\mathrm{b}} - \left(\frac{dH_{\mathrm{n}}}{dN_{\mathrm{b}}} - \frac{1}{2}\right)\right]$$

So for any cut above hitmax, the above-cut hit-function is changing slower than $\frac{u_{\mathrm{a}}}{N_{\mathrm{a}}}$ and the below-cut hit-function is changing faster than $\frac{u_{\mathrm{b}}}{N_{\mathrm{b}}}$. This fact should permit an easy proof that the hitmax cut also maximizes the sum of hit-functions $u = u_a + u_{\mathrm{b}} = h_{\mathrm{a}} + h_{\mathrm{b}} - 1$, but I have as yet been unsuccessful in constructing such a proof.

e. *A theorem concerning the sum of sums of squares above and below a cut in terms of the latent mean difference on the output indicator and the latent hit-rates on the input indicator*[3]

Consider a sliding cut x_{c} on an input indicator x, and the associated statistics on an output indicator y determined by various choices of the x-cut. Intuition suggests that, on the intra-taxon independence assumption, the dispersions (by some appropriate measure) of the output-variable y when calculated separately upon the cases falling above and those falling below the x-cut should, in a general way, tend to be small when the x-cut is optimally located. That this must be

---

[3] *This is the consistency test "Minimizing $SS_b + SS_a$" described in Meehl & Yonce (1996, p. 1135).—LJY*

roughly true seems clear from the fact that the *y*-variance in any subpopulation is composed of the components of variance contributed by the intra-taxon sources of variation, whether systematic or random (including measurement error), plus a component of variance assignable to the between-taxon mean difference. Thus if we imagine an input indicator so highly valid when optimally cut that it separates the schizotypes and nonschizotypes quasi-perfectly, then practically all the cases lying above the cut will be schizotypic and practically all the cases lying below the cut will be nonschizotypic. Hence, the variance on the output indicator *y* when calculated on the above-cut cases will receive a negligible contribution from the sum of squares attributable to taxon difference, being based upon a sum of squares which reflects (almost) solely the *y*-variance of schizotypes. And similarly that quantity calculated upon cases lying below the *x*-cut will be dependent almost solely on the residual *y*-variance characteristic of a "pure" group of nonschizotypes. Hence it seems that as the amount of "mixture" in a subpopulation (i.e., its taxonomic heterogeneity, its being composed of more nearly equal proportions of the two latent taxa) increases, the dispersion on an output variable will, in a general way, tend to increase.

This intuitive reasoning leads to the suggestion that one might derive an additional hitmax cut locator by relying on the notion of a minimum pooled sum of squares on an output indicator for the sub-populations falling above and below a sliding cut on an input indicator. That one should attempt to locate the hitmax cut on *x* by minimizing the sum of sums of squares on *y* above and below a sliding *x*-cut rather than by minimizing the sum of the variances is rather obvious, since the latter quantity is intrinsically "corrected" for differing frequencies above and below the cut, whereas the hitmax cut must take account of the difference in frequencies $N_a \neq N_b$.

However, it can be shown that finding the minimum sum of sums of squares on *y* does <u>not</u> precisely locate the *x*-hitmax, except in the special case where the hit-rates above and below the *x*-cut are exactly equal. And we know that such a relationship, while it will be approximated rather closely except in the case of extreme distribution distortions or marked base-rate asymmetries, will not in general hold precisely. I believe however, that minimizing this sum of sums of squares achieves a hitmax cut location which is satisfactorily close, i.e., within a single class interval, and the grounds for supposing this are presented in section 4-f below. Since the demonstration is not completely satisfactory, I have not presented the method as an additional hitmax cut locator in the present section. But the first step in attempting such a proof turns out to be a theorem regarding latent values which is of some intrinsic interest, and which might provide

a starting point for further developments. The theorem says that the sum of the sums of squares of an output indicator within the subpopulations lying above and below a sliding cut on an input indicator will be minimized when the cut is so chosen as to minimize a latent quantity which is equal to the frequency above the cut times the hit-rate above the cut times its complement, plus the frequency below the cut times the hit-rate below the cut times its complement.

Proof

Consider an input variable x on which we are locating a sliding cut, each value of which determines dispersion statistics on an output variable y, taken about the manifest y-means above and below the x-cut.

Define:

p_a = Proportion of above-cut cases that are schizotypes = hit-rate above cut $= h_a = \frac{H_s}{N_a}$

q_a = $1 - p_{sa}$ = Proportion of above-cut cases that are nonschizotypes = miss-rate above cut
$= 1 - h_a = 1 - \frac{H_s}{N_a} = \frac{M_n}{N_a}$

p_b = Proportion of below-cut cases that are schizotypes = miss-rate below cut
$= 1 - h_b = 1 - \frac{H_n}{N_b} = \frac{M_s}{N_b}$

q_b = Proportion of below-cut cases that are nonschizotypes = hit-rate below cut $= h_b = \frac{H_n}{N_b}$

Then from the general formula for variance of a mixed population, given the means and variances of the two sub-populations (latent taxa) the y-variances above and below are expressible in latent terms thus:

[59] $$\sigma_a^2 = p_a\sigma_s^2 + q_a\sigma_n^2 + p_a\left(\overline{y}_s - \overline{y}_a\right)^2 + q_a\left(\overline{y}_n - \overline{y}_a\right)^2$$

[60] $$\sigma_b^2 = p_b\sigma_s^2 + q_b\sigma_n^2 + p_b\left(\overline{y}_s - \overline{y}_b\right)^2 + q_b\left(\overline{y}_n - \overline{y}_b\right)^2$$

and multiplying by the frequencies above and below

[61] $$N_a\sigma_a^2 = N_a p_a\sigma_s^2 + N_a q_a\sigma_n^2 + N_a p_a\left(\overline{y}_s - \overline{y}_a\right)^2 + N_a q_a\left(\overline{y}_n - \overline{y}_a\right)^2$$

[62] $$N_b\sigma_b^2 = N_b p_b\sigma_s^2 + N_b q_b\sigma_n^2 + N_b p_b\left(\overline{y}_s - \overline{y}_b\right)^2 + N_b q_b\left(\overline{y}_n - \overline{y}_b\right)^2$$

so adding [61]-[62] to get sum of manifest within-groups sums-of-squares above and below we have

[63] $$SS_{\mathrm{a}} + SS_{\mathrm{b}} = \left(N_{\mathrm{a}}p_{\mathrm{a}} + N_{\mathrm{b}}p_{\mathrm{b}}\right)\sigma_{\mathrm{s}}^2 + \left(N_{\mathrm{a}}q_{\mathrm{a}} + N_{\mathrm{b}}q_{\mathrm{b}}\right)\sigma_{\mathrm{n}}^2$$
$$+N_{\mathrm{a}}p_{\mathrm{a}}\left(\overline{y}_{\mathrm{s}} - \overline{y}_{\mathrm{a}}\right)^2 + N_{\mathrm{a}}q_{\mathrm{a}}\left(\overline{y}_{\mathrm{n}} - \overline{y}_{\mathrm{a}}\right)^2$$
$$+N_{\mathrm{b}}p_{\mathrm{b}}\left(\overline{y}_{\mathrm{s}} - \overline{y}_{\mathrm{b}}\right)^2 + N_{\mathrm{b}}q_{\mathrm{b}}\left(\overline{y}_{\mathrm{n}} - \overline{y}_{\mathrm{b}}\right)^2$$

[64] $$= \left(H_{\mathrm{s}} + M_{\mathrm{s}}\right)\sigma_{\mathrm{s}}^2 + \left(M_{\mathrm{n}} + H_{\mathrm{n}}\right)\sigma_{\mathrm{n}}^2$$
$$+H_{\mathrm{s}}\left(\overline{y}_{\mathrm{s}} - \overline{y}_{\mathrm{a}}\right)^2 + M_{\mathrm{n}}\left(\overline{y}_{\mathrm{n}} - \overline{y}_{\mathrm{a}}\right)^2$$
$$+M_{\mathrm{s}}\left(\overline{y}_{\mathrm{s}} - \overline{y}_{\mathrm{b}}\right)^2 + H_{\mathrm{n}}\left(\overline{y}_{\mathrm{n}} - \overline{y}_{\mathrm{b}}\right)^2$$

The first two terms of this, being $\left(N_{\mathrm{s}}\sigma_{\mathrm{s}}^2 + N_{\mathrm{n}}\sigma_{\mathrm{n}}^2\right)$, are constant regardless of the *x*-cut. To minimize (SS_a + SS_b) is therefore to minimize the sum of the last four terms, call it S:

[65] $$\mathrm{S} = H_{\mathrm{s}}\left(\overline{y}_{\mathrm{s}} - \overline{y}_{\mathrm{a}}\right)^2 + M_{\mathrm{n}}\left(\overline{y}_{\mathrm{n}} - \overline{y}_{\mathrm{a}}\right)^2 + M_{\mathrm{s}}\left(\overline{y}_{\mathrm{s}} - \overline{y}_{\mathrm{b}}\right)^2 + H_{\mathrm{n}}\left(\overline{y}_{\mathrm{n}} - \overline{y}_{\mathrm{b}}\right)^2$$

Evaluating the first parenthetical term,

$$\overline{y}_{\mathrm{s}} - \overline{y}_{\mathrm{a}} = \overline{y}_{\mathrm{s}} - \left(p_{\mathrm{a}}\overline{y}_{\mathrm{s}} + q_{\mathrm{a}}\overline{y}_{\mathrm{n}}\right)$$
$$= \overline{y}_{\mathrm{s}} - p_{\mathrm{a}}\overline{y}_{\mathrm{s}} - q_{\mathrm{a}}\overline{y}_{\mathrm{n}}$$
$$= \overline{y}_{\mathrm{s}}\left(1 - p_{\mathrm{a}}\right) - \overline{y}_{\mathrm{n}}q_{\mathrm{a}}$$
$$= q_{\mathrm{a}}\left(\overline{y}_{\mathrm{s}} - \overline{y}_{\mathrm{n}}\right)$$

[66] $$\overline{y}_{\mathrm{s}} - \overline{y}_{\mathrm{a}} = q_{\mathrm{a}}\Delta\overline{y}$$

Analogously for the other three parentheses, we have

[67] $$\overline{y}_{\mathrm{n}} - \overline{y}_{\mathrm{a}} = -p_{\mathrm{a}}\Delta\overline{y}$$

[68] $$\overline{y}_{\mathrm{s}} - \overline{y}_{\mathrm{b}} = q_{\mathrm{b}}\Delta\overline{y}$$

[69] $$\overline{y}_{\mathrm{n}} - \overline{y}_{\mathrm{b}} = -p_{\mathrm{b}}\Delta\overline{y}$$

Substituting [66] - [69] into [65] we obtain

[70] $$\mathrm{S} = H_{\mathrm{s}}q_{\mathrm{a}}^2\Delta\overline{y}^2 + M_{\mathrm{n}}p_{\mathrm{a}}^2\Delta\overline{y}^2 + M_{\mathrm{s}}q_{\mathrm{b}}^2\Delta\overline{y}^2 + H_{\mathrm{n}}p_{\mathrm{b}}^2\Delta\overline{y}^2$$

[71] $$= \Delta\overline{y}^2\left[H_{\mathrm{s}}q_{\mathrm{a}}^2 + M_{\mathrm{n}}p_{\mathrm{a}}^2 + M_{\mathrm{s}}q_{\mathrm{b}}^2 + H_{\mathrm{n}}p_{\mathrm{b}}^2\right]$$

which is minimized by minimizing the bracket.

The bracketed quantity, a function of cumulative cases below, is more simply expressible in terms of the hit-rates and frequencies above and below, thus:

[72] $$\mathrm{b}(N_\mathrm{b}) = H_\mathrm{s} q_\mathrm{a}^2 + M_\mathrm{n} p_\mathrm{a}^2 + M_\mathrm{s} q_\mathrm{b}^2 + H_\mathrm{n} p_\mathrm{b}^2$$

and multiplying terms by $\frac{N_\mathrm{a}}{N_\mathrm{a}}$ or $\frac{N_\mathrm{b}}{N_\mathrm{b}}$ as desired, we have

[73] $$\begin{aligned}\mathrm{b}(N_\mathrm{b}) &= N_\mathrm{a}\frac{H_\mathrm{s}}{N_\mathrm{a}} q_\mathrm{a}^2 + N_\mathrm{a}\frac{M_\mathrm{n}}{N_\mathrm{a}} p_\mathrm{a}^2 + N_\mathrm{b}\frac{M_\mathrm{s}}{N_\mathrm{b}} q_\mathrm{b}^2 + N_\mathrm{b}\frac{H_\mathrm{n}}{N_\mathrm{b}} p_\mathrm{b}^2 \\ &= N_\mathrm{a} p_\mathrm{a} q_\mathrm{a}^2 + N_\mathrm{a} q_\mathrm{a} p_\mathrm{a}^2 + N_\mathrm{b} p_\mathrm{b} q_\mathrm{b}^2 + N_\mathrm{b} q_\mathrm{b} p_\mathrm{b}^2 \\ &= N_\mathrm{a} p_\mathrm{a} q_\mathrm{a}(q_\mathrm{a} + p_\mathrm{a}) + N_\mathrm{b} p_\mathrm{b} q_\mathrm{b}(q_\mathrm{b} + p_\mathrm{b}) \\ &= N_\mathrm{a} p_\mathrm{a} q_\mathrm{a} + N_\mathrm{b} p_\mathrm{b} q_\mathrm{b}\end{aligned}$$

[74] $$\mathrm{b}(N_\mathrm{b}) = N_\mathrm{a} h_\mathrm{a}(1 - h_\mathrm{a}) + N_\mathrm{b} h_\mathrm{b}(1 - h_\mathrm{b})$$

which is to be minimized. The expression for the sum of sums of squares of *y* reflecting dispersion of *y* within cases falling above and below the *x*-cut is then

[75] $$SS_\mathrm{a} + SS_\mathrm{b} = N_\mathrm{s}\sigma_\mathrm{s}^2 + N_\mathrm{n}\sigma_\mathrm{n}^2 + \Delta\bar{y}^2\left[N_\mathrm{a} h_\mathrm{a}(1 - H_\mathrm{a}) + N_\mathrm{b} h_\mathrm{b}(1 - H_\mathrm{b})\right]$$

which is a minimum when the sliding *x*-cut is located so as to minimize the bracket. It is evident that this cut cannot, in general, be exactly the hitmax cut, because the bracket expands as

[76] $$\mathrm{b}N_\mathrm{b} = N_\mathrm{a} h_\mathrm{a} + N_\mathrm{b} h_\mathrm{b} - N_\mathrm{a} h_\mathrm{a}^2 - N_\mathrm{b} h_\mathrm{b}^2$$

[77] $$= (N_\mathrm{a} h_\mathrm{a} + N_\mathrm{b} h_\mathrm{b}) - (N_\mathrm{a} h_\mathrm{a}^2 + N_\mathrm{b} h_\mathrm{b}^2)$$

and since the derivative of the first parenthesis is zero at hitmax, the derivative b′(N_b) cannot vanish there unless $(N_\mathrm{a} h_\mathrm{a}^2 + N_\mathrm{b} h_\mathrm{b}^2)'$ vanishes at hitmax, which, in general, it will not do (exactly). Further consideration of this line of thought is deferred to Section 4f below, but we note here that when b(N_b) is written in terms of the two differences between hit-rates and their squares,

[78] $$\mathrm{b}N_\mathrm{b} = N_\mathrm{a}(h_\mathrm{a} - h_\mathrm{a}^2) + N_\mathrm{b}(h_\mathrm{b} - h_\mathrm{b}^2)$$

we get an intuitive appreciation of why the hitmax cut almost minimizes (SS_a + SS_b), since the difference between a hit-rate and its square is a decreasing monotone function of the hit-rate, so that if the two hit-rates are not too disparate, these ($h - h^2$) differences are changing at nearly the same rate as we slide the *x*-cut along in the intermediate region.

f. ***A theorem concerning sums of sums of cross-products above and below a cut in terms of the latent mean differences on two output indicators and the latent hit-rates on the input indicator***

A relationship analogous to that shown in the just preceding subsection (4-e) of course obtains for the (frequency-weighted) covariance of a pair of output variables as it does for the (frequency-weighted) variance of a single output variable. Consider an input variable *x* on which we are locating a sliding cut, each value of which determines two covariance statistics on the output-variable pair (*y, z*), taken about the manifest means $(\overline{y}_a, \overline{z}_a)$ and $(\overline{y}_b, \overline{z}_b)$ calculated within the N_{ax} cases above and the N_{bx} cases below the *x*-cut. Then from the general formula for the covariance of a mixed population, analogous to Equation [59] for the variance, the (*yz*)-covariance within the N_a above-cut cases is expressible in latent terms thus:

$$[79] \quad \mathrm{cov}_a(yz) = p_a \mathrm{cov}_s(yz) + q_a \mathrm{cov}_n(yz) + p_a(\overline{y}_s - \overline{y}_a)(\overline{z}_s - \overline{z}_a) + q_a(\overline{y}_n - \overline{y}_a)(\overline{z}_n - \overline{z}_a)$$

and similarly the covariance below the cut can be written in latent terms, analogously to Equation [60], as

$$[80] \quad \mathrm{cov}_b(yz) = p_b \mathrm{cov}_s(yz) + q_b \mathrm{cov}_n(yz) + p_b(\overline{y}_s - \overline{y}_b)(\overline{z}_s - \overline{z}_b) + q_b(\overline{y}_n - \overline{y}_b)(\overline{z}_n - \overline{z}_b)$$

From here on the derivation proceeds, mutatis mutandis, as in the preceding subsection Equation [61]-[75], terminating in the covariance analogue of [75] which is identical with [75] as to its variable bracket, differing from [75] in the constant terms only, that is,

$$[81] \quad N_a \mathrm{cov}_a(yz) + N_b \mathrm{cov}_b(yz) = N_s \mathrm{cov}_s(yz) + N_n \mathrm{cov}_n(yz) + \Delta\overline{y}\Delta\overline{z}\left[N_a h_a(1-h_a) + N_b h_b(1-h_b)\right]$$

the first two terms being constant, and the product of the latent mean differences $\Delta\overline{y}\,\Delta\overline{z}$ being constant, analogous to the squared latent mean difference $\Delta\overline{y}^2$ in Equation [75]. Hence the sum of sums of cross-products is minimized by minimizing the bracket $\left[N_a h_a(1-h_a) + N_b h_b(1-h_b)\right]$.

g. ***Two consistency tests based upon median cut of one manifest distribution and maximizing sign-concordance with another***

We show that when an indicator *y* is cut at its manifest median, and a sliding cut on *x* is chosen so as to maximize *xy*-sign-concordance in a fourfold table, then (1) the *y*-mean for cases

lying at that x-cut must equal the grand mean, and (2) this x-cut is the cut that maximizes the numerator of a phi-coefficient.

Proof

Locate the median of the manifest y-distribution and fix this value as a y-cut, above which cases are called y^+ and below which they are called y^-. Then in latent terms (base-rates, valid and false positive and negative rates) we can write the equality of manifest frequencies above and below as

[82] $$Pp_{sy} + Qp_{ny} = Pq_{sy} + Qq_{ny}$$

that is,

[83] $$\frac{q_{ny} - p_{ny}}{p_{sy} - q_{sy}} = \frac{P}{Q}$$

i.e., the latent valid-minus-invalid rate differences are inversely proportional to the base-rates.

We now locate an x-cut such that the sum of frequencies in the concordant cells of the (xy)-fourfold table is a maximum. In PR-65-2 (pp. 17-18, Equation [10]) we showed that the maximum of concordance C(x, y) for any fixed y-cut satisfies the latent condition

[84] $$\frac{f_s(x)}{f_n(x)} = \frac{q_{ny} - p_{ny}}{p_{sy} - q_{sy}}$$

Combining [83]-[84] we have

[85] $$\frac{f_s(x)}{f_n(x)} = \frac{P}{Q}$$

That is, the (unrelativized) ordinates of the latent frequency functions $f_s(x)$ and $f_n(x)$ at this x-cut are directly proportional to the base-rates. Hence the ratio of schizotypes to nonschizotypes among cases lying within a small interval containing these ordinates is P : Q, and the manifest y-mean of those cases should be equal to the grand mean of y, in accordance with the weighted composite

[86] $$\overline{y}_i = P\overline{y}_s + Q\overline{y}_n = \overline{y}_t$$

If this is not approximately true, we infer that the postulated latent situation is somewhere incorrect, most likely as to the intra-taxon independence assumption, the $\overline{y}$ -mean in this x-interval being displaced from $\overline{y}$ because of an intra-taxon (xy)-correlation.

Corollary: In PR-65-2 we showed (Section 8, Equations [18]-[19]) that the "phimax" cuts on *x* and *y*, those cuts which maximize the numerator of the phi-coefficient $\left[= p^{+}_{xy} - p^{+}_{x} p^{+}_{y}\right]$, are the cuts which equalize the relativized frequency functions $\left[\phi_{\mathrm{s}}(x) = \phi_{\mathrm{n}}(x) \text{ and } \phi_{\mathrm{s}}(y) = \phi_{\mathrm{n}}(y)\right]$ on each indicator. But the *x*-cut where $\phi_{\mathrm{s}}(x) = \phi_{\mathrm{n}}(x)$ is of course the cut where

[87] $$\frac{f_{\mathrm{s}}(x)}{f_{\mathrm{n}}(x)} = \frac{P}{Q}$$

$\left(\text{since } f_{\mathrm{s}}(x) = P\phi_{\mathrm{s}}(x) \text{ and } f_{\mathrm{n}}(x) = Q\phi_{\mathrm{n}}(x)\right)$, and that is the *x*-cut located above. Hence this correspondence is a further consistency test.

h. ***A 3-indicator system of equations based upon hitmax interval statistics and the grand means***

Once the hitmax interval is located and one or more consistency tests reassure us that the latent situation is well approximated by the model, there is available a very short and direct path to estimating the latent means and base-rates, relying solely upon hitmax interval statistics and grand means. Its defect is in its reliance on a system of five equations (three of which are in the second degree), and hence presumably subject to considerable sampling instability unless the *N* is very large. We require a set of three indicators, say, *w* as an "input" indicator with *x* and *y* as "output" indicators. We first locate *w*-hitmax by one or more of the hitmax-locating methods described in PR-65-2 or the present report, and test consistency insofar as possible with this limited determination. Then in the hitmax interval we infer, for the output indicators *x* and *y*, the relations between w_h-interval manifest means and the latent taxon means,

[88] $$\overline{y}_{hw} = \frac{1}{2}\overline{y}_{\mathrm{s}} + \frac{1}{2}\overline{y}_{\mathrm{n}}$$

[89] $$\overline{x}_{hw} = \frac{1}{2}\overline{x}_{\mathrm{s}} + \frac{1}{2}\overline{x}_{\mathrm{n}}$$

For the observed (*xy*)-covariance of cases falling within this *w*-interval we can write

[90] $$\mathrm{cov}(xy)_{hw} = \frac{1}{4}\left(\overline{x}_{\mathrm{s}} - \overline{x}_{\mathrm{n}}\right)\left(\overline{y}_{\mathrm{s}} - \overline{y}_{\mathrm{n}}\right)$$

Finally, we know that the two grand (manifest) means of the entire mixed population are expressible in terms of the latent means and base-rates,

[91] $$P\overline{x}_{\mathrm{s}} + Q\overline{x}_{\mathrm{n}} = \overline{x}_{\mathrm{t}}$$

[92] $$P\overline{y}_{\mathrm{s}} + Q\overline{y}_{\mathrm{n}} = \overline{y}_{\mathrm{t}}$$

Solving the system [88]-[92] for the 5 latent variables $\overline{x}_{\mathrm{s}}$, $\overline{x}_{\mathrm{n}}$, $\overline{y}_{\mathrm{s}}$, $\overline{y}_{\mathrm{n}}$, and P we have our main parameters directly.

Consistency check: Over the mixed population the observed grand covariance should be calculable from the five inferred latent values by the relation

[93] $$\mathrm{cov}(xy)_{\mathrm{t}} = PQ\left(\overline{x}_{\mathrm{s}} - \overline{x}_{\mathrm{n}}\right)\left(\overline{y}_{\mathrm{s}} - \overline{y}_{\mathrm{n}}\right)$$

i. ***A 5-indicator system of equations based solely upon hitmax interval covariances***

The general expression for the covariance of an output indicator-pair x, y among cases falling in an input-interval w_i is

[94] $$\mathrm{cov}_{w}(xy) = p_i q_i \left(\overline{x}_{\mathrm{s}} - \overline{x}_{\mathrm{n}}\right)\left(\overline{y}_{\mathrm{s}} - \overline{y}_{\mathrm{n}}\right)$$

where p_i and q_i are the proportions of schizotypes and nonschizotypes in that interval. In the hitmax interval this becomes

[95] $$\mathrm{cov}(xy)_{hw} = \frac{1}{4}\left(\overline{x}_{\mathrm{s}} - \overline{x}_{\mathrm{n}}\right)\left(\overline{y}_{\mathrm{s}} - \overline{y}_{\mathrm{n}}\right)$$

as previously shown. Equation [95] is of second degree (cross-products) in four unknowns. Each indicator added to an indicator-family presents two more unknowns, i.e., the means of the two latent taxa. (The covariance of a given indicator-pair in the hitmax interval of w should, of course, be the same as it is at hitmax of a different input indicator, say, v. Hence this does not boost the number of available equations, but merely provides a consistency test.)

So we have $2k$ unknowns for a k-indicator family. The number of such equations being the same as the number of indicator-pairs, for solubility we must have

[96] $$2k = \frac{k(k-1)}{5}$$

which is satisfied by $k = 5$. If $k < 5$ the system is indeterminate; if $k > 5$ it is (generally) inconsistent, appropriately handled as an overdetermined system containing error. When $k = 5$, as with an indicator-family (x, y, z, u, v), we have a system of 10 equations

[97]
$$\begin{cases} \overline{x}_{\mathrm{s}}\overline{y}_{\mathrm{s}} - \overline{x}_{\mathrm{s}}\overline{y}_{\mathrm{n}} - \overline{x}_{\mathrm{n}}\overline{y}_{\mathrm{s}} + \overline{x}_{\mathrm{n}}\overline{y}_{\mathrm{n}} = \mathrm{C}_1 \\ \overline{x}_{\mathrm{s}}\overline{z}_{\mathrm{s}} - \overline{x}_{\mathrm{s}}\overline{z}_{\mathrm{n}} - \overline{x}_{\mathrm{n}}\overline{z}_{\mathrm{s}} + \overline{x}_{\mathrm{n}}\overline{z}_{\mathrm{n}} = \mathrm{C}_2 \\ \vdots \\ \vdots \\ \overline{u}_{\mathrm{s}}\overline{v}_{\mathrm{s}} - \overline{u}_{\mathrm{s}}\overline{v}_{\mathrm{n}} - \overline{u}_{\mathrm{n}}\overline{v}_{\mathrm{s}} + \overline{u}_{\mathrm{n}}\overline{v}_{\mathrm{n}} = \mathrm{C}_{10} \end{cases}$$

where the right-hand constants are obtained from hitmax-interval observed covariances by expressions of the form

[98]
$$\mathrm{C}_m = 4\,\mathrm{cov}(xy)_{hz}$$

3. A single-indicator method, relying on the (testable) hypothesis of intra-taxon normality

a. The main methods presented in PR-65-2 and the present report deliberately avoid the familiar assumptions of normality and homogeneity of variance, assumptions which are not likely to obtain (and have frequently been shown not to obtain) when the domain of investigation is a clinical taxon such as schizotypy. However, we have hypothesized that some of the intra-taxon covariances are negligible, and that others are at least approximately equal; and as has been repeatedly emphasized in the text, these are approximations, whose effect upon the accuracy of the methods proposed remains for more thorough investigation. At present it is not apparent whether mild departures from the assumptions of zero (or equal) intra-taxon covariances are more, or less, damaging than the normality assumption which we have thus far avoided. It would seem appropriate therefore to investigate, in both Monte Carlo and empirical tests of the method, the robustness of a taxon-identification technique relying upon intra-taxon normality.

Secondly, to the extent that this assumption is directly or indirectly testable within the data — and therefore, more properly, to be spoken of as a statistical hypothesis rather than an assumption — estimates of the latent parameters made upon that assumption (and not wholly redundant with the other methods) can provide additional consistency tests and increase our confidence in the estimates.

Thirdly, there may arise situations, especially in the very early stages of the research in a domain, where only two single indicators are available for study on a sufficiently large sample, or even where there is only a single indicator for which clinical experience, previous research,

and theoretical considerations jointly lead to a legitimately high prior belief that the indicator discriminates powerfully between the latent taxa of interest.

Further, just as it was pointed out that one can improve the approximation to zero or equal intra-taxon covariances by transformations (or, in the case of psychometric devices, by item-analysis) so here it may be possible, through a preliminary item-analytic procedure employing crude clinical criteria, to reduce extreme departures from the normality assumption. While I do not here consider the details, it seems plausible to suppose that even if intra-taxon normality is a poor approximation, and the variables are refractory to score-transformations or item-analysis aimed at improving this approximation sufficiently, it would still be possible to employ the method to be described in this section. What we would need then to do is to retreat from postulated normality to a somewhat more general form of Pearson's generalized frequency function, e.g., one requiring, say, only the weaker assumption of unimodality within taxa.

b. The basic idea is extremely simple and relies upon the fact that the manifest (= mixed-taxa) frequency distribution is the sum of the latent distributions. It is obvious that one cannot assign all six latent parameters independently. Considering the three distribution parameters required on the normality assumption (i.e., the base-rate, mean and sigma within a taxon), as soon as we choose a triplet of values (N_s, $\bar{x}_s$, σ_s) for the schizotypes, then the manifest distribution's values determine the corresponding parameters (N_n, $\bar{x}_n$, σ_n) for the nonschizotypes. The "searching procedure" therefore consists of assigning arbitrary (sliding) values to the base-rate frequencies N_s, N_n [$N_n = N - N_s$], then to the latent means, and finally to the latent sigmas. This logical tree terminates in predicted resultant values for the observed (mixed-taxa) frequency distribution. We then compute a chi-square on the discrepancy between the predicted and observed frequencies, and record it. It serves first as a significance test (testing departure from the postulated latent model-*cum*-parameter values) but also, more importantly, as a rough measure of the poorness of our approximation. For each value of $\bar{x}_s$, we can generate a family of curves of these chi-squares, each curve showing the chi-square values as a function of the arbitrarily assigned σ_s. A super-family of such curve-families is generated by each assigned N_s. The N_s-super-family which contains the curve-family which in turn contains the curve whose minimum chi-square is smallest is then the best approximation we can get. Ideally it would be a nonsignificant chi-square and would corroborate the intra-taxon normality assumption. However, since the null hypothesis is always false (Lykken 1966, Meehl 1967, Lykken 1968) and all such

assumptions are approximations (easily refutable by large samples such as the present method demands) it is sufficient for practical purposes if the minimum achievable chi-square is "small," although statistically significant.

c. The basic equations for the latent values of the six parameters ($\overline{x}_s$, $\overline{x}_n$, σ_s, σ_n, N_s, N_n) which completely characterize the latent situation on the intra-taxon normality hypothesis are:

[99] $$N_n = N - N_S$$

[100] $$\overline{x}_n = \frac{N\overline{x}_t - N_s\overline{x}_s}{N_n}$$

$$N_n\overline{x}_n = N\overline{x}_t - N_s\overline{x}_s$$

$$N_n\overline{x}_n + N_s\overline{x}_s = N\overline{x}_t$$

$$\sum x_n + \sum x_s = \sum x_t \qquad \text{An identity}$$

The 3-line expansion on Eq [100] was added by Meehl in 1989 for clarification.

[101] $$\sum^{N_s} x_s^2 = N_s\left(\sigma_s^2 + \overline{x}_s^2\right)$$

[102] $$\sum^{N_n} x_n^2 = \sum^{N} x_t^2 - \sum^{N_s} x_s^2$$

[103] $$\sigma_n^2 = \frac{1}{N_n}\sum^{N_n} x_n^2 - \overline{x}_n^2$$

Thus an arbitrary assignment of schizotypic base-frequency N_s determines the nonschizotype base-frequency N_n as in [99]. Given those frequencies, an arbitrary assignment of the schizotype mean $\overline{x}_s$ then determines the nonschizotype mean $\overline{x}_n$ as in [100]. Given these frequencies and means, an arbitrary assignment of the schizotypic standard deviation σ_s determines the nonschizotypic standard deviation σ_n by solving serially [101] - [102] - [103]. Each of the latent normal distributions being completely characterized by its three parameters (N, $\overline{x}$, σ), we enter a table of normal-curve integrals to get the frequencies n_s and n_n with which the two taxa should occur in each class-interval, and their sum $n_{ci} = (n_s + n_n)$ is the calculated mixed-taxon frequency for that interval. We then compute the chi-square over all intervals on the discrepancies between the calculated values n_{ci} and the observed frequencies n_{oi}.

There being three independently assignable parameters, what we deal with computationally is a super-family of curve-families, each curve within a family showing the relation between goodness-of-fit and the arbitrary values of the third parameter [= σ]. Thus:

(1) First curve-family: Parameters N_s (and hence $N_n = N - N_s$) fixed.

(a) First curve of first family: Parameters N_s, N_n, $\bar{x}_s$ (and hence $\bar{x}_n$) fixed.

((1)) First point on first curve of first family: Parameters N_s, N_n, $\bar{x}_s$, $\bar{x}_n$, σ_s (and hence σ_n) fixed. Chi-square of theoretical – observed discrepancy plotted as ordinate against values of σ_s as abscissa.

((2)) Second point on first curve of first family: Parameters N_s, N_n, $\bar{x}_s$, $\bar{x}_n$, as in ((1)), but a new arbitrary value of σ_s (and hence σ_n).

(b) Second curve of first family: Parameters N_s, N_n fixed as in (a), but a new arbitrary value of $\bar{x}_s$ (hence $\bar{x}_n$).

(c) And so on for set of curves of first family.

(2) Second curve-family: New parameters N_s (and hence $N_n = N - N_s$) assigned.

Then repeat the entire process as under (1), and so forth. We thereby produce a super-family of curve-families of curves (of Chi-squares as a function of σ_s, for fixed N_s and $\bar{x}_s$). Each curve will have a minimum (best fit) value for an arbitrary σ_s-assignment. Each curve-family will have a minimum of these minima. We want the minimum of these minima of minima. That minimum-of-minima-of-minima is then the best fit attainable by optimal arbitrary assignments of N_s, $\bar{x}_s$, σ_s, given the normality hypothesis. Statistical significance of this chi-square refutes the exact normality hypothesis. But a "small" numerical value of borderline significance would suggest that the approximation is good.

Intuitively, it seems that the orderliness exhibited by the changes in minima is a rough test of the basic assumptions. That is, if the latent model is essentially valid, each curve should pass through a clear, inspectional minimum, showing a definite trend down to this unique low value and up again. The minima of the several curves in each family should also exhibit an orderly progression with respect to the arbitrary $\bar{x}_s$-values determining these curves; and these minima-of-minima should display an orderly passage through a minimum as we move along the abscissa N_s within the superfamily.

As of this writing, a preliminary trial of the method utilizing "real data," the taxa being sex and the single indicator-variable being MMPI Scale 5, is not completely analyzed. The graphs show a minimum very close to the true parameter values, and their steepness is apparently much

greater in the vicinity of the true values. But one "maverick" curve has a minimum uncomfortably close to the best one, and arises from parameter assignments that are grossly erroneous. This curve could, hopefully, have been spotted as aberrant by virtue of its being clearly unlike the others in its curve-family (e.g., they are parallel and orderly in progression, it is not); but its mere occurrence raises questions about the method. When thoroughly analyzed, these data will be discussed in a forthcoming Report of the Psychiatric Research Laboratories (Meehl, Lykken, Burdick and Schoener, 1969)

4. Suggested further developments, and some queries

a. Iterative procedure

In PR-65-2 it was suggested, on the basis of some numerical examples, that slight-to-moderate departures from the assumption of zero (or at least equal) intra-taxon covariance might nevertheless permit reasonably accurate estimates of the latent parameters. The various consistency tests (PR-65-2, Section 9) will, hopefully, provide some empirical check on the extent to which this conclusion is fulfilled by a set of real data. If it is not, the general method may perhaps still be useable by "bootstrapsing" through a series of successive approximations. Suppose that the intra-taxon covariances are not sufficiently close to zero as between the "output" indicators x and y, but that they are negligible as between an "input" indicator w and each of the former. Then the latent means $\overline{x}_w, \overline{y}_w$ of cases lying within successive w-intervals will vary solely as a function of the schizotype-proportions p_i of these w-intervals; but the (xy)-covariance will not have its maximum at $p_i = q_i = 1/2$ (i.e., the hitmax cut on w will be mis-located by the covariance-maximizing search of PR-65-2, Section 3). Therefore the inferred $p_i q_i = 1/4$ for the pseudo-hitmax interval will be slightly in error (see, e.g., PR-65-2, pp. 51-53) and, hence, the constant $\mathrm{K} = \Delta\overline{x}\Delta\overline{y}$ will be in systematic error. When we proceed by the simple method presented in Section 1 of the present report to compute the schizotype-rates p_i for other w-intervals, those estimates will then be doubly biased, first by the systematic error in K, and also by the fact that $\mathrm{cov}_w(xy)$ is not adequately approximated by the expression

$$\mathrm{cov}_w(xy) = \mathrm{K}\, p_i\, q_i$$

but requires instead the more general expression (for possibly unequal intra-taxon covariances)

[104] $$\mathrm{cov}_w(xy) = \mathrm{K}\, p_i\, \mathrm{cov}_s(xy) + q_i\, \mathrm{cov}_n(xy) + \mathrm{K}\, p_i\, q_i$$

But perhaps we can employ the parameter-estimates obtained in Section 1 above as first approximations to reach improved estimates via the hitmax cut, and re-cycle? Such an approach seems intuitively plausible, but I have put it in this "suggestion" section, lacking an analytic proof that such an iterative procedure will converge to successively better approximations. At least it seems plausible to argue that, so long as no direct circularity is involved, an improvement in the consistency-tests produced by such an iterated method would tend to corroborate the hypothesis that it was converging successfully.

By the simplified method of Section 1 we have moved sequentially through hitmax-location ⟶ $\mathrm{cov}_h(xy)$-estimate ⟶ K-estimate ⟶ p_i-estimates for all w-intervals ⟶ N_s-estimates for all w-intervals ⟶ estimates of base-rate P, and thence of latent intra-taxon means $\overline{x}_\mathrm{s}, \overline{x}_\mathrm{n}, \overline{y}_\mathrm{s}, \overline{y}_\mathrm{n}, \overline{w}_\mathrm{s}, \overline{w}_\mathrm{n}$ to first approximations. We now draw an arbitrary w-cut on the manifest w-distribution (say, at the w-median). We note that this cut does not rely on any (x, y)-statistics, and in particular, does not depend upon (xy)-covariance data of any sub-sample or of the whole sample. Now the (xy)-covariances of the subsets of patients lying above, and below, this w-median are observable quantities, and "experimentally independent." Our first cycle has, however, drawn us the inferred latent distribution, so we "know" (to a first approximation) the latent hit-rates p_sa and q_nb for these two sub-samples defined by the arbitrary w-cut. For the cases lying above the w-median, we write the observed (mixed-taxon) (xy)-covariance in terms of these latent quantities, from generic Equation [104] above,

$$[105] \qquad \begin{aligned} \mathrm{cov}_{aw}(xy) &= p_\mathrm{sa}\,\mathrm{cov}_\mathrm{s}(xy) + p_\mathrm{na}\,\mathrm{cov}_\mathrm{n}(xy) + p_\mathrm{sa}p_\mathrm{na}\left(\overline{x}_\mathrm{s} - \overline{x}_\mathrm{n}\right)\left(\overline{y}_\mathrm{s} - \overline{y}_\mathrm{n}\right) \\ &= p_\mathrm{sa}\,\mathrm{cov}_\mathrm{s}(xy) + q_\mathrm{sa}\,\mathrm{cov}_\mathrm{n}(xy) + \mathrm{K}'p_\mathrm{sa}q_\mathrm{sa} \end{aligned}$$

indicating by K′ the revised K-estimate calculated from the approximate latent means. Similarly for the cases below the w-median we have

$$[106] \qquad \mathrm{cov}_{bw}(xy) = p_\mathrm{sb}\,\mathrm{cov}_\mathrm{s}(xy) + q_\mathrm{sb}\,\mathrm{cov}_\mathrm{n}(xy) + \mathrm{K}'p_\mathrm{sb}q_\mathrm{sb}$$

The quantities p_sa, q_sa, p_sb, q_sb and K′ are "knowns" (first approximation) and the left-hand covariances $\mathrm{cov}_\mathrm{aw}(xy)$, $\mathrm{cov}_\mathrm{bw}(xy)$ are manifest statistics. So we have a pair of simultaneous equations linear in the intra-taxon "unknowns" $\mathrm{cov}_\mathrm{s}(xy)$ and $\mathrm{cov}_\mathrm{n}(xy)$, which we solve to get a second approximation to these latter. We can now drop the idealized initial assumption of zero-

or-equal intra-taxon covariance, and rewrite the general *w*-interval expression for manifest $\text{cov}_i(xy)$ as

[107] $$\text{cov}_i(xy) = p_i \text{cov}_\text{s}(xy) + q_i \text{cov}_\text{n}(xy) + \text{K}' p_i q_i$$

This equation is then solved as in Section 1 for every *w*-interval, yielding a revised set of schizotype-proportions p_i' for all the *w*-intervals, in turn yielding new absolute schizotype-frequencies N_s', a new base-rate P', and new latent means $\bar{x}_\text{s}'$, $\bar{x}_\text{n}'$, $\bar{y}_\text{s}'$, $\bar{y}_\text{n}'$, $\bar{w}_\text{s}'$, and $\bar{w}_\text{n}'$. The whole process is then repeated, and we continue iterating until (hopefully) the estimated latent values cease to change and the consistency tests are reasonably satisfied.

The presently unanswered questions are, of course, (1) Does the iteration rely upon some subtle "circularity" leading to a misleading convergence?, and (2) Does the latent model insure that convergence will occur? Obviously the first question is the more important one, since if the second question is answered negatively, we will readily discover that fact in any empirical situation of non-convergence.

b. Preliminary tests of the manifest distribution to warrant method's use

(1) If there are available, as will usually be the case, any "crude, gross" taxonomic criteria (e.g., medical chart diagnosis, previously well-validated MMPI patterns) it would presumably be desirable, before investing time and money in application of these techniques, to examine some preliminary questions, for which such coarse criteria probably suffice. There are two main preliminary questions:

(a) Does each of the contemplated indicator-variables show a moderate-to-large discrimination against the crude criteria? It is difficult, pending adequate Monte Carlo study, to set any minimum value on the coarse concurrent validity level as a condition precedent to "bootstrapsing" construct validity by procedures like the present one (Cronbach and Meehl, 1955; Meehl, 1959). But contemplation of graphical frequency-distributions (and a little numerical juggling) will perhaps convince most readers that mean differences of less than around one sigma are not encouraging. I do not myself believe that this unduly restricts the applicability of the method, considering the fact that mean concurrent-validity differences in the range 2-3 sigma are often attainable with MMPI scales, when reasonable care is exercised in conducting the investigation.

(b) The second consideration is, of course, the intrataxon covariance assumption. Ideally all the intrataxon covariances should be zero. As has been shown, however, it will suffice if the intrataxon covariances are approximately equal for those indicator-pairs employed as output indicators; although we must still require that the intrataxon covariance of each of these with an input indicator should be near zero, otherwise the estimate of latent mean values from values in the hitmax interval will be invalid.

(c) If preliminary study suggests that the intrataxon correlations are too large and unequal, but there are theoretical reasons (e.g., qualitative nature of a potential indicator suggested by a theory of the schizotaxic defect) or previous empirical data supporting the indicator's high validity, either or both arguing for retention of the indicator in the trial-indicator family; then score transformations should be attempted and, if the indicator is a psychometric one (e.g., an MMPI scale or sub-scale) item-analysis should be carried out in which the basis for retaining an item in the modified indicator is a conjunctive condition, namely, low item-scale correlation between the item and other indicators in the provisional family, and moderate-to-high item-validity against the crude criterion.

(2) It may happen (although very rarely) that no such crude criteria are available, or that their intrinsic validity (or net validity as affected by unreliability) is not trusted sufficiently to use them even for preliminary assessment of the situation. This will hardly be the case in any clinical population with respect to the establishment of each indicator's respectable discriminating power. But it might be true with respect to the intrataxon covariance assumptions, since the diagnostic habits of clinicians will "force" some degree of spurious intrataxon covariance between indicators that are already clinically recognized as clustered and hence presumably contaminative. There also arise situations, such as psychometric studies of family members of schizophrenic probands, where we have no theoretical or empirical assurance that an indicator behaves the same way with regard to compensated schizotypy as it does with regard to diagnosable decompensated schizotypy, and adequate (intensive-study) diagnostic assessment here will often be absent if the *N* is sufficiently large to justify use of the method. In such situations I know of no way to check on the intrataxon covariance assumption except by applying the method followed by the consistency tests. If they indicate that something is badly wrong with the assumptions, one may run a direct empirical test of the intrataxon covariance assumption upon cases provisionally identified by the first iteration of the technique.

However, the requirement of respectable single-indicator validity can be refuted, with some confidence, without using any external criteria. If a latent taxonomy exists, and if it is productive of sizable mean differences between the two latent frequency distributions on a single indicator, the manifest (mixed-taxa) distribution of that indicator should reflect the latent situation by showing a considerable departure from normality. Specifically, the effect of two latent distributions being superimposed to generate the manifest distribution will be platykurtosis. Here again, thorough Monte Carlo study will probably be necessary in order to set a suitable condition on the minimum amount of platykurtosis; although, if the sample size is sufficient to justify using the method at all, at least a statistically significant departure from normality, in the platykurtic direction, should surely exist. A rough idea of the quantitative relationships involved here may be gleaned from the engineer Hald (1952) who discusses the present problem in an interesting section — little known among psychologists — on "heterogeneous distributions." Hald shows, for the special case of two normal populations with equal sigma, that when the distance between the two means is three times the standard deviation (of the single latent distribution) the manifest distribution is clearly bimodal. With a distance of two sigma between the latent means, the joint distribution is strikingly platykurtic. Whereas for a mean difference of only one sigma (i.e., suppose the mean T-score of schizotypes were only T = 60 on the MMPI scale Sc), the manifest distribution when merely "eyeballed" does not differ conspicuously from a normal curve. These rough illustrative values are reassuring with regard to the kinds of empirical situations one is likely to meet. Hald also discusses our problem of "dissection of a heterogeneous distribution" in terms of plotting the cumulative manifest distribution frequencies on probability paper, where the latent heterogeneity is reflected graphically by marked departures from linearity in the plot. (See his pp. 152-158, and the prior explanation of the "fractile" concept, pp. 66-67, 102-103, 127-140).

c. Estimating latent means from tails of manifest distribution

In PR-65-2 (Section 14d) it was suggested that the latent means $\bar{x}_s$ and $\bar{x}_n$ might be estimated from the statistics of extreme high and low intervals of the manifest *w*-distribution. It is obvious that, just as the hitmax interval on *w* provides maximal "mixture" of the latent taxa ($p_s = q_s = 1/2$), so the very high or very low regions of *w* will be occupied by sub-populations which are very preponderantly schizotypic, or nonschizotypic, respectively. Whether this relative

taxonomic "purity" at the two tail-ends of the manifest *w*-distribution is sufficient to provide useful estimates of $\bar{x}_s, \bar{x}_n, \bar{y}_s, \bar{y}_n$, ... will depend upon *w*-overlap (i.e., validity of the input indicator) in relation to sample size. If the overlap on *w* is excessive, only very extreme *w*-regions will provide "unmixed" populations, these regions being perforce chosen so far out that the absolute frequencies in these intervals are too small to yield stable statistics. And, as always, the assumption of negligible intra-taxon covariance may give trouble, especially since rather slight departures from $cov_s(xw) = cov_n(xw)$, harmless in intermediate regions, might produce disturbingly large systematic error in estimating $\bar{x}_s$ and $\bar{x}_n$ when working with extremely deviant *w*-values on the input side. That is, the value of $\bar{x}_s$ estimated upon cases lying within the very high *w*-intervals will reflect not only the $p_s \gg p_n$ "taxon impurity" of those intervals, but will also reflect any within-taxon dependence of *x* upon *w*.

However, one is not confined to mere hoping that the approximations are close enough for reliance upon the tail-statistics. Before utilizing *w*-tail values of $\bar{x}_w$ as estimators of the latent means $\bar{x}_s$, $\bar{x}_n$ we can apply some consistency tests to the data that should forestall unjustified reliance upon the adequacy of the latent model's approximation in these respects, as follows:

(1) Behavior of $cov_w(xy)$ at *w*-tails:

In locating the hitmax interval of *w* we plotted $cov_w(xy)$ to find its maximum. Considering parameter values, to the extent that the postulated latent model is a good approximation, the sole source of within-interval departures from $cov_w(xy) = 0$ is the taxon mixture of cases lying therein. Hence, neglecting sampling errors, the values of $cov_w(xy)$ should fall off steadily as we consider *w*-intervals farther removed from *w*-hitmax in either direction, and should approach zero as $p_s(w) \rightarrow 1$ at upper tail and $p_s(w) \rightarrow 0$ at lower tail, respectively. The first observational condition for reliance upon tail-statistics as latent mean estimators is, therefore, a finding that the graph of $cov_w(xy)$ becomes and remains flat (at $\simeq$ zero) for several *w*-intervals at the high and low extremes. In practice, we would look for random-sampling fluctuations around a central value close to zero. (I am not clear whether an appropriate significance test exists for this purpose.) Actually the required cov(*xy*) behavior combines two conditions, (a) The covariances should cease to change systematically over these intervals, and (b) Their representative value over these intervals should be close to zero. If either of these conditions is appreciably violated, either the postulated latent situation is not an adequate

approximation, or the *w*-indicator has too much overlap to generate intervals for which $p_s(w) \simeq 1$ [or $p_s(w) \simeq 0$] with the sample size under study. It is evident that the two conditions on tail-behavior of $\text{cov}_w(xy)$ reflect distinguishable aspects of the numerical approximation, only the first of which must be adequate before one can rely upon *w*-tails as estimators of the latent intra-taxon values. Failure of cov(*xy*) to become constant over successive *w*-intervals reflects primarily further changes in $p_s(w)$ i.e., the latent distributions overlap excessively, so that extreme *w*-intervals are "mixed" rather than "pure." Failure of $\text{cov}_w(xy)$ to stabilize at zero (but settling down at some other value > 0) reflects mainly intra-taxon correlation between *x* and *y*, and while this departure from the model impairs the validity of certain other procedures (e.g., most of the consistency tests) it should not, so far as I can see, invalidate the use of tail-means as estimators. But it is worth noting that a stabilization of $\text{cov}_w(xy) \simeq C > 0$ for extreme *w*-intervals warns us that the tail method will not work when the (*x*, *y*) pair of indicators is treated as input and output variable. And, contrariwise, if $\text{cov}_w(xy) \simeq 0$ over a set of extreme *w*-intervals, we may infer the legitimacy of employing *x*-tails for estimating latent $\bar{y}_s$, $\bar{y}_n$ (and *y*-tails for estimating $\bar{x}_s$, $\bar{x}_n$) insofar as the (*xy*)-covariance assumption is fulfilled; but of course the taxon-"purity" of the extreme *x*-intervals (or *y*-intervals) for that purpose remains to be tested.

(2) Behavior of $\bar{y}_w$ (and of $\bar{x}_w$) at *w*-tails:

Whereas satisfactory behavior of $\text{cov}_w(xy)$ gives us information about intra-taxon correlation between *x* and *y*, and also about non-mixture of taxa within sufficiently extreme *w*-intervals, the possibility of an unwanted dependence of *y* (and *x*) upon the input variable *w* is not reflected therein, but instead by systematic change in the *y*-means (and *x*-means) with change in *w*. If *y* [all the subsequent remarks in this sub-section apply, mutatis mutandis, to *x* as an alternative output variable] continues to increase with increases in *w*, even when we are working in extreme *w*-intervals for which $\text{cov}_w(xy) \simeq K$, such a constellation of findings tells us that the assumption of zero intra-taxon covariance for *y* and *w* is false. Because if we can find sufficiently high (and sufficiently low) *w*-intervals to stabilize $\text{cov}_w(xy)$, whether at zero or some other value, we can infer that taxon mixture has become and remained negligible over these extreme intervals (i.e., we are "out far enough" on *w* so that $p_s(w) \simeq 1$ and $p_s(w) \simeq 0$ in the high-*w* and low-*w* regions respectively.) But if taxon-mixture is negligible over a *w*-region, any systematic dependence of $\bar{y}_w$ on *w* shows that *y* and *w* are correlated within taxa. If we find that

$\overline{y}_w$ continues to change with *w* in the extreme high region but not in the extreme low region, we infer that the assumption of zero intra-taxon correlation holds for nonschizotypes but is violated by schizotypes. And of course if *y* covaries with *w* within either taxon, tail-values of $\overline{y}_w$ are biased estimators of the latent mean for that taxon.

If $\text{cov}_w(xy)$ behaves in a satisfactory manner at both *w*-tails, and if $\overline{y}_w$ and $\overline{x}_w$ show zero slope within the same region, we have strong corroboration of the latent model, and trustworthy estimators of the latent means. Several consistency tests are immediately available by virtue of this situation. Example: The *w*-interval with y_w-mean = $\frac{1}{2}\overline{x}_s + \frac{1}{2}\overline{y}_n$ should be the same as that with x_w-mean = $\frac{1}{2}\overline{x}_s + \frac{1}{2}\overline{x}_n$, since each is the hitmax interval of *w*. Example: The *w*-interval for which $\overline{y}_w = \frac{1}{2}\overline{y}_s + \frac{1}{2}\overline{y}_n$ should correspond to that found by the other hitmax-locators. Example: The numerical value of $\text{cov}_w(xy)$ at its maximum should agree with that calculable from the relation

[108] $$\text{cov}_w(xy)_{\text{Max}} = \frac{1}{4}(\overline{x}_s - \overline{x}_n)(\overline{y}_s - \overline{y}_n)$$

Example: By substituting the tail-estimated latent means $\overline{x}_s$, $\overline{x}_n$, $\overline{y}_s$, $\overline{y}_n$ in the formula

[109] $$\text{cov}_t(xy) = PQ(\overline{x}_s - \overline{x}_n)(\overline{y}_s - \overline{y}_n)$$

we should get the observed value of the grand covariance $\text{cov}_t(xy)$.

As always, some of the consistency tests might instead be employed as primary estimating procedures. E.g., in the last example above, once having concluded that our *w*-tail regions do provide good estimates of the latent means, we could use the grand covariance expression itself as a basis for getting the base-rate, solving the quadratic

[110] $$p^2 - p - \frac{\text{cov}_t(xy)}{(\overline{x}_s - \overline{x}_n)(\overline{y}_s - \overline{y}_n)} = 0$$

for *P* — assuming we have some basis for choosing between the two roots, e.g., which inequality $P > Q$ or $Q > P$ is compatible with the manifest-latent relations

$$\overline{y}_t = P\overline{y}_s + Q\overline{y}_n$$
$$\overline{x}_t = P\overline{x}_s + Q\overline{x}_n$$

That the values involved in using the *w*-tails are not impracticable, so as, e.g., to demand an outlandish sample size, is suggested by the following simple numerical example. Suppose a 2-sigma difference exists between the latent *w*-means of schizotypes and nonschizotypes, $P = Q$, and intra-taxon normality obtains. Then the hitmax cut is at +1σ below $\overline{w}_{\mathrm{n}}$. At +1 P.E. of the schizotype distribution [= 75%ile] we are at +2.67σ on the nonschizotype distribution, above which point lie only .004 of that taxon's cases. Hence the proportion of nonschizotypes in this "upper region" of *w* is, given the equal base-rates, only .004/.250 = .016 < 2 percent. Therefore the observed $\overline{y}$-mean in that region is a weighted function of the latent means with weight = .98 applied to the schizotypic mean, i.e., a quite accurate estimate of the desired latent mean $\overline{y}_{\mathrm{s}}$. If our sample size were as small as $N = 200$, which is probably somewhat low for use of the whole approach, the $\overline{y}_{w}$-statistic is still fairly stable from the sampling standpoint, being based on approximately 25 schizotypes in the upper tail region of *w*.

As to change in the taxon-purity over successive intervals in this region, a rough idea can be gleaned by considering the ordinates. At +2.67σ of the nonschizotypes' curve, which corresponds to + .67σ [= 1 P.E.] on the schizotype curve, the ordinates are .011 and .319 respectively, a ratio

$$\frac{.011}{.319} = .034$$

for taxon-mixture at that point (or, in an interval taken sufficiently small around that point). Moving up to 90th percentile on the schizotype curve, that is, to +1.29σ and +3.29σ on the two latent frequency functions, the corresponding ordinates are .1736 and .0018, a ratio of

$$\frac{.0018}{.1736} = .010$$

so the proportion of nonschizotypes has declined interval-by-interval from 3% to 1% over the region taken. These values seem reassuring as to the tail-method's practicability with moderate-to-large samples and decent luck with the latent distribution shapes.

d. Problem of a general proof that hitmax cut quasi-maximizes sum of hit-rates above and below cut

There must, I think, be a less clumsy way to reach the result of Section 2b above, but I have been unsuccessful. A more mathematically adept reader may wish to try his hand at the problem in its general form, thus: Define

[111] $$\mathrm{U}_\mathrm{a} = h_\mathrm{a} - \frac{1}{2} = \frac{H_\mathrm{s}}{N_\mathrm{a}} - \frac{1}{2}$$

[112] $$\mathrm{U}_\mathrm{b} = h_\mathrm{b} - \frac{1}{2} = \frac{H_\mathrm{n}}{N_\mathrm{b}} - \frac{1}{2}$$

Then the conditions are that

[113] $$-\frac{1}{2} \le \mathrm{U}_\mathrm{a} \le \frac{1}{2} \qquad \text{everywhere}$$

[114] $$-\frac{1}{2} \le \mathrm{U}_\mathrm{b} \le \frac{1}{2} \qquad \text{everywhere}$$

[115] $$\mathrm{U}_\mathrm{a} > 0 \qquad \text{if } N_\mathrm{b} \ge k$$

[116] $$\mathrm{U}_\mathrm{b} > 0 \qquad \text{if } N_\mathrm{b} \le k$$

if $N_\mathrm{b} = k$ is the cumulative manifest frequency up to the hitmax cut. Also we have

[117] $$\frac{d\mathrm{U}_\mathrm{a}}{dN_\mathrm{b}} > 0 \qquad \text{everywhere}$$

[118] $$\frac{d\mathrm{U}_\mathrm{b}}{dN_\mathrm{b}} < 0 \qquad \text{everywhere}$$

[110] $$\frac{dN_\mathrm{a}}{dN_\mathrm{b}} = -1 \qquad \text{everywhere}$$

[120] $$\frac{\dfrac{d\mathrm{U}_\mathrm{a}}{dN_\mathrm{a}}}{\dfrac{d\mathrm{U}_\mathrm{b}}{dN_\mathrm{b}}} = -\frac{N_\mathrm{b}}{N_\mathrm{a}}\frac{\mathrm{U}_\mathrm{a}}{\mathrm{U}_\mathrm{b}} \qquad \text{at } N_\mathrm{b} = k$$

[121] $$\frac{d}{dN_\mathrm{b}}\left(\mathrm{U}_\mathrm{a}N_\mathrm{a} + \mathrm{U}_\mathrm{b}N_\mathrm{b}\right) = 0 \qquad \text{at } N_\mathrm{b} = k$$

[122] $$\frac{d^2}{dN_\mathrm{b}^2}\left(\mathrm{U}_\mathrm{a}N_\mathrm{a} + \mathrm{U}_\mathrm{b}N_\mathrm{b}\right) < 0 \qquad \text{at } N_\mathrm{b} = k$$

and, finally,

[123] $N_a \gg U_a$

[124] $N_b \gg U_b$

in vicinity of N_b = k, by factors of 100 or more. (It may be that my difficulty arises from failure to include all of the actual conditions of the physical situation in the above statement of hypothesis).

Then, to prove, that ($U_a + U_b$) has a maximum at $N_b = k$, or very close thereto.

e. Possible use of sums of squares above and below cut as a good approximate hitmax locator

In Section 2e above we raised the question whether an additional hitmax cut locator might be developed from the intuitive notion that when a cut on an "input" indicator minimizes the latent taxonomic misclassifications, the dispersions on an output indicator within groups above and below, should, in a general way, tend to be smaller than when the sub-sets above and below the cut are more "mixed" (i.e., when the cut is poorly chosen, so that many cases are misclassified). In Section 2e it was shown that the sum of sums of squares of an output indicator (about its means taken above and below a sliding input-indicator cut) is a minimum when the frequency-weighted sum of hit-rates times their complements is a minimum; and that the input cut thus located is not, in general, exactly at hitmax. The present Section will make it plausible that the error in locating the hitmax cut in this way is likely to be small, at least for intra-taxon distributions departing not too grossly from normal. But since I have not tried out a wide range of distribution types as to numerical values, the proposed hitmax-locator is presented here among mere "Suggested further developments and queries."

We want to show that if sums of squares of deviations on an output-indicator *y* are computed about the *y*-means of cases falling above and below a sliding *x*-cut, the sum of these two sums of squared deviations will be a minimum when the *x*-cut is close to hitmax. (Re-read concluding portion of Section 2e before continuing.) "Close to hitmax" I take to mean "within the hitmax interval or, at worst, one interval displaced."

We define a function S(N_b) which is the cut-variable portion of the sum of sums of squared deviations of *y* about the *y*-means of cases falling above, and below, a sliding *x*-cut:

[125] $$SS_a + SS_b = H_s\sigma_s^2 + N_n\sigma_n^2 + \Delta\bar{y}^2\left[N_a h_a\left(1-h_a\right) + N_b h_b\left(1-h_b\right)\right]$$

[126] $$S\left(N_b\right) = \Delta\bar{y}^2\left[N_a h_a\left(1-h_a\right) + N_b h_b\left(1-h_b\right)\right]$$

In what follows through Equaton [171], all derivatives are taken with respect to the cumulative frequency [= N_b] below the *x*-cut. The bracketed term in [126] expands as

[127] $$b(N_b) = N_a h_a + N_b h_b - N_a h_a^2 - N_b h_b^2$$

and as pointed out in Section 2e above, the first two terms of this expression (= total hits) have zero derivative at hitmax, hence the derivative db/dN_b will not, in general, vanish precisely at hitmax, since the subtracted portion $\left(N_a h_a^2 + N_b h_b^2\right)$ will not be minimized exactly where $\left(N_a h_a + N_b h_b\right)$ is maximized. The practical question is, then, "How far off is it?" It will be convenient to estimate this error in hitmax-location in terms of the cumulative frequency N_b, and an error $\Delta N_b < 10$ is quite acceptable, since this would usually locate the correct class-interval, and would never err by more than one class-interval. The combined effect of errors due to (a) Unreliability, (b) Sampling, and (c) Grouping will very likely be greater than the error due to the present approximation.

The first derivatives of the terms in [127] with respect to N_b are

[128] $$\frac{d}{dN_b}\left(N_a h_a\right) = N_a h_a' + N_a' h_a = N_a h_a' - h_a$$

[129] $$\frac{d}{dN_b}\left(N_b h_b\right) = N_b h_b' + N_b' h_b = N_b h_b' + h_b$$

[130] $$\frac{d}{dN_b}\left(N_a h_a^2\right) = N_a \left(h_a^2\right)' + N_a' h_a^2 = 2N_a h_a h_a' - h_a^2$$

[131] $$\frac{d}{dN_b}\left(N_b h_b^2\right) = N_b \left(h_b^2\right)' + N_b' h_b^2 = 2N_b h_b h_b' + h_b^2$$

The hit-rate derivatives which occur in these expressions are

[132] $$h_a' = \frac{d}{dN_b}\left(\frac{H_s}{N_a}\right) = \frac{N_a H_s' - H_s N_a'}{N_a^2} = \frac{1}{N_a}\left(\frac{N_a H_s' + H_s}{N_a}\right) = \frac{1}{N_a}\left(H_s' + h_a\right)$$

[133] $$h_b' = \frac{d}{dN_b}\left(\frac{H_n}{N_b}\right) = \frac{N_b H_n' - H_n N_b'}{N_b^2} = \frac{1}{N_b}\left(\frac{N_b H_n' - H_n}{N_b}\right) = \frac{1}{N_b}\left(H_n' - h_b\right)$$

Substituting these hit-derivative expressions in [128]-[131] we obtain

[134] $$N_a h_a' - h_a = H_s'$$

[135] $$N_b h_b' + h_b = H_n'$$

[136] $$2N_{\mathrm{a}}h_{\mathrm{a}}h'_{\mathrm{a}} - h_{\mathrm{a}}^2 = h_{\mathrm{a}}\left(2H'_{\mathrm{s}} + h_{\mathrm{a}}\right)$$

[137] $$2N_{\mathrm{b}}h_{\mathrm{b}}h'_{\mathrm{b}} - h_{\mathrm{b}}^2 = h_{\mathrm{b}}\left(2H'_{\mathrm{n}} - h_{\mathrm{b}}\right)$$

To minimize the sum of sums of squares we want the zero of its derivative, that is,

[138] $$\frac{d}{dN_{\mathrm{b}}}\left(SS_{\mathrm{a}} + SS_{\mathrm{b}}\right) = 0$$

which zeros the drivative of its cut-dependent portion S(N_{b}), so we want

[139] $$\mathrm{S}'\left(N_{\mathrm{b}}\right) = \Delta\bar{y}^2\left[N_{\mathrm{a}}h_{\mathrm{a}} + N_{\mathrm{b}}h_{\mathrm{b}} - N_{\mathrm{a}}h_{\mathrm{a}}^2 - N_{\mathrm{b}}h_{\mathrm{b}}^2\right]' = 0$$

Substituting the expressions [134]-[137] into the bracket, we obtain

[140] $$\mathrm{S}'\left(N_{\mathrm{b}}\right) = \Delta\bar{y}^2\left[H'_{\mathrm{s}} + H'_{\mathrm{n}} - h_{\mathrm{a}}\left(2H'_{\mathrm{s}} + h_{\mathrm{a}}\right) - h_{\mathrm{b}}\left(2H'_{\mathrm{n}} - h_{\mathrm{b}}\right)\right] = 0$$

as the condition for minimizing (SS_{a} + SS_{b}). That this is a minimum rather than a maximum or flex-point is physically apparent, but can also be shown from the fact that S″(N_{b}) > 0 in this vicinity.

What is the value of this derivative at the hitmax cut? We would like it to be zero, so that (SS_{a} + SS_{b}) = Min could be employed as a hitmax cut locator. That this is not <u>exactly</u> true we see as follows: Re-arranging the bracketed portion of S′(N_{b})

[141] $$\begin{aligned}\mathrm{b}'\left(N_{\mathrm{b}}\right) &= H'_{\mathrm{s}} + H'_{\mathrm{n}} - h_{\mathrm{a}}\left(2H'_{\mathrm{s}} + h_{\mathrm{a}}\right) - h_{\mathrm{b}}\left(2H'_{\mathrm{n}} - h_{\mathrm{b}}\right)\\ &= H'_{\mathrm{s}}\left(1 - 2h_{\mathrm{a}}\right) + H'_{\mathrm{n}}\left(1 - 2h_{\mathrm{b}}\right) + h_{\mathrm{b}}^2 - h_{\mathrm{a}}^2\end{aligned}$$

What is this at hitmax? Recalling that the frequency-functions intersect at hitmax (i.e., the ordinates are equal, and half of the <u>area</u> in a small interval surrounding hitmax cut is contributed by each taxon), we know that, at hitmax,

[142] $$H'_{\mathrm{s}} = \frac{dH_{\mathrm{s}}}{dN_{\mathrm{b}}} = -\left(1/2\right)$$

[143] $$H'_{\mathrm{n}} = \frac{dH_{\mathrm{n}}}{dN_{\mathrm{b}}} = +\left(1/2\right)$$

and substituting these into [141] we have, at hitmax,

[144] $$\begin{aligned}\mathrm{b}'\left(N_{\mathrm{b}}\right) &= -\left(1/2\right)\left(1 - 2h_{\mathrm{a}}\right) + \left(1/2\right)\left(1 - 2h_{\mathrm{b}}\right) + h_{\mathrm{b}}^2 - h_{\mathrm{a}}^2\\ &= \left(h_{\mathrm{b}}^2 - h_{\mathrm{a}}^2\right) + \left(h_{\mathrm{a}} - h_{\mathrm{b}}\right)\\ &= \left(h_{\mathrm{b}}^2 - h_{\mathrm{a}}^2\right) - \left(h_{\mathrm{b}} - h_{\mathrm{a}}\right)\end{aligned}$$

$$= \left(h_b - h_a\right)\left(h_b + h_a\right) - \left(h_b - h_a\right)$$

[145] $$= \left(h_b - h_a\right)\left(h_b + h_a - 1\right)$$

and putting this in terms of

[146] $$U_a = h_a - 1/2$$

[147] $$U_b = h_b - 1/2$$

we have, at hitmax,

[148] $$b'(N_b) = \left(U_b - U_a\right)\left(U_b + U_a\right)$$

[149] $$= \left(U_b^2 - U_a^2\right)$$

So the first derivative of $(SS_a + SS_b)$ at hitmax is

[150] $$\left(SS_a + SS_b\right)' = \Delta\overline{y}^2\left(U_b^2 - U_a^2\right) \neq 0$$

which will be close to zero but not exactly zero in general, since it can be exactly zero only if there is perfect symmetry between the hit-rates above and below the cut,

[151] $$U_a^2 = U_b^2$$

[152] $$\text{so} \quad U_a = U_b$$

[153] $$\text{so} \quad h_a = h_b$$

which is, in general, false. In what follows we shall assume its falsity, permitting division by $(U_a - U_b)$ at one step. If it should happen to be true in a particular empirical situation, the following evaluation of error ΔN_b is not necessary, since the hitmax cut exactly minimizes $(SS_a + SS_b)$ in that special case. Otherwise the value of $(SS_a + SS_b)'$ at hitmax is, of course, also the error in $(SS_a + SS_b)'$ at hitmax.

We wish to evaluate the error made when one attempts to locate the hitmax cut by minimizing $(SS_a + SS_b)$. That is, if a cut is located by empirically minimizing this statistic, how far will such a cut deviate from the hitmax cut? It is easier to look at this error from the other direction, i.e., if we were at hitmax and there calculated $(SS_a + SS_b)$, how far "off" (in terms of cumulative frequency N_b) would we be from the cut that would exactly minimize $(SS_a + SS_b)$? That is, if $(SS_a + SS_b)'$ is not exactly zero at hitmax, how many more (or fewer) cases ΔN_b would have to be cumulated to bring this derivative to zero? We approximate this by the differential,

[154] $$\text{Error} = \Delta N_b \simeq \frac{d^2\left(SS_a + SS_b\right)}{dN_b^2}\Delta\left(SS_a + SS_b\right)'$$

and of course for present purposes the algebraic sign of this error is of no interest.

Taking the second derivative of ($SS + SS_b$) with respect to N_b,

[155] $$\left(SS_a + SS_b\right)'' = S''\left(N_b\right) = \Delta\overline{y}^2\left[N_a h_a\left(1-h_a\right) + N_b h_b\left(1-h_b\right)\right]''$$

the bracketed (cut-variable) portion of this being, differentiating [141],

[156] $$b''\left(N_b\right) = \frac{d}{dN_b}\left[H_s'\left(1-2h_a\right) + H_n'\left(1-2h_b\right) + h_b^2 - h_a^2\right]$$

$$= H_s'\left(1-2h_a\right)' + H_s''\left(1-2h_a\right) - 2h_a h_a' + H_n'\left(1-2h_b\right)' + H_n''\left(1-2h_b\right) + 2h_b h_b'$$

$$= H_s'\left(-2h_a'\right) + H_s''\left(1-2h_a\right) - 2h_a h_a' + H_n'\left(-2h_b'\right) + H_n''\left(1-2h_b\right) + 2h_b h_b'$$

$$= H_s'\left(-\frac{2}{N_a}\left(H_s' + h_a\right)\right) + H_s'\left(1-2h_a\right) - 2h_a\left(\frac{1}{N_a}\left(H_s' + h_a\right)\right)$$

$$+ H_n'\left(-\frac{2}{N_b}\left(H_n' - h_b\right)\right) + H_n'\left(1-2h_b\right) + 2h_b\left(\frac{1}{N_b}\left(H_n' + h_b\right)\right)$$

At hitmax, putting in the hitmax values

$$H_s' = -\left(1/2\right)$$

$$H_n' = +\left(1/2\right)$$

we obtain

[157] $$b''\left(N_b\right) = \frac{1}{N_a}\left(h_a - 1/2\right) + H_s''\left(1-2h_a\right) - 2\frac{h_a}{N_a}\left(h_a - 1/2\right)$$

$$+ \frac{1}{N_b}\left(h_b - 1/2\right) + H_n''\left(1-2h_b\right) - 2\frac{h_b}{N_b}\left(h_b - 1/2\right)$$

$$= \frac{1}{N_a}\left(h_a - 1/2\right)\left(1-2h_a\right) + \frac{1}{N_b}\left(h_b - 1/2\right)\left(1-2h_b\right)$$

$$+ H_s''\left(1-2h_a\right) + H_n''\left(1-2h_b\right)$$

[158] $$= \left(1-2h_a\right)\left[\frac{1}{N_a}\left(h_a - 1/2\right) + H_s''\right] + \left(1-2h_b\right)\left[\frac{1}{N_b}\left(h_b - 1/2\right) + H_n''\right]$$

which, reversing signs, dividing and multiplying by 2, and putting (U + 1/2) for the hit-rates, yields

[159] $$b''(N_b) = -2\left[U_a\left(\frac{U_a}{N_a} + H_s''\right) + U_b\left(\frac{U_b}{N_b} + H_n''\right)\right]$$

Eq [159] ... + H_s'' was originally mis-typed as ...+ h_s''

so at hitmax the second derivative of $(SS_a + SS_b)$ with respect to the cumulative frequency is

[160] $$\frac{d^2}{dN_b^2}(SS_a + SS_b) = -2\Delta\bar{y}^2\left[U_a\left(\frac{U_a}{N_a} + H_s''\right) + U_b\left(\frac{U_b}{N_b} + H_n''\right)\right]$$

Lemma: The second derivatives H_s'' and H_n'' of the hit-frequencies H_s and H_n with respect to cumulative frequency N_b are equal (everywhere, not only at hitmax), thus:

[161] $$\begin{aligned} H_s &= H_s - M_s \\ &= N_s - (N_b - H_n) \\ &= H_n - N_b + N_s \end{aligned}$$

Differentiating with respect to N_b,

[162] $$\frac{dH_s}{dN_b} = \frac{dH_n}{dN_b} - 1$$

Differentiating again with respect to N_b,

[163] $$\frac{d^2H_s}{dN_b^2} = \frac{d}{dN_b}\left(\frac{dH_n}{dN_b}\right) = \frac{d^2H_n}{dN_b^2}$$

that is, we can everywhere make a substitution based upon the equality

[164] $$H_s'' = H_n''$$

Returning to the problem of estimating our error ΔN_b in locating hitmax, the error in finding the zero of the first derivative is closely approximated by the differential

[165] $$\Delta(SS_a + SS_b)' \simeq (SS_a + SS_b)''\,\Delta N_b$$

[166] $$\Delta N_b \simeq \frac{\Delta(SS_a + SS_b)'}{(SS_a + SS_b)''}$$

and substituting the error in the first derivative $\neq 0$ from [150] and the hitmax value of the second derivative from [160] we have

[167] $$\Delta N_b \simeq \frac{\Delta\bar{y}^2(U_b^2 - U_a^2)}{-2\Delta\bar{y}^2\left[U_a\left(\frac{U_a}{N_a} + H_s''\right) + U_b\left(\frac{U_b}{N_b} + H_n''\right)\right]}$$

which, substituting H_n'' for H_s'' and cancelling $\Delta\bar{y}^2$ above and below, yields

[168] $$\Delta N_b \simeq -\frac{1}{2}\frac{U_b^2 - U_a^2}{\frac{U_a^2}{N_a} + U_a H_n'' + \frac{U_b^2}{N_b} + U_b H_n''}$$

[169] $$\simeq -\frac{1}{2}\frac{U_b^2 - U_a^2}{\left(\frac{U_a^2}{N_a} + \frac{U_b^2}{N_b}\right) + H_n''(U_a + U_b)}$$

First bracket of denominator is always positive, so neglecting this term will therefore lead to an overestimate of the absolute error $|\Delta N_b|$. We then have

[170] $$|\Delta N_b| < \frac{1}{2}\left|\frac{(U_b + U_a)(U_b - U_a)}{H_n''(U_a + U_b)}\right|$$

and cancelling ($U_a + U_b$) we get

[171] $$|\Delta N_b| < \frac{1}{2}\left|\frac{(U_b - U_a)}{H_n''}\right|$$

Thus, the error, in cumulative frequency below, in locating hitmax cut by minimizing ($SS_a + SS_b$) on an output variable, does not exceed half the difference between latent hit-rates below and above, divided by the second derivative of nonschizotypic hits with respect to the cumulative frequency N_b below the cut.

To evaluate this error numerically we must express the second derivative $H_n'' = \frac{d^2 H_n}{dN_b^2}$ terms of the latent frequency functions $f_s(x)$, $f_n(x)$, and $f_t(x) = f_s(x) + f_n(x)$. It will be convenient to consider the derivatives of these unrelativized frequency functions taken with respect to the abscissa variable x instead of the cumulative frequency N_b. But of course what we are evaluating on the left is the second derivative of H_n taken with respect to N_b. The reader must remember that the prime-sign on the right-hand side of all equations following designates differentiations w.r.t. the input variable x itself. We have

[172] $$\frac{d^2 H_n}{dN_b^2} = \frac{d}{dN_b}\left(\frac{dH_n}{dN_b}\right) = \left[\frac{d}{dx}\left(\frac{dH_n}{dN_b}\right)\right]\frac{dx}{dN_b} \qquad \text{Chain rule}$$

Expressing the inner parenthesis in terms of derivatives w.r.t. x, by Chain Rule we have

[173] $$\frac{dH_{\mathrm{n}}}{dN_{\mathrm{b}}} = \frac{dH_{\mathrm{n}}}{dx}\,\frac{dx}{dN_{\mathrm{b}}} = f_{\mathrm{n}}(x)\frac{dx}{dN_{\mathrm{b}}}$$

[174] $$= f_{\mathrm{n}}(x)\frac{1}{\frac{dN_{\mathrm{b}}}{dx}} = \frac{f_{\mathrm{n}}(x)}{f_{\mathrm{t}}(x)}$$

Substituting [174] in [172] we obtain

[175] $$\frac{d^2H_{\mathrm{n}}}{dN_{\mathrm{b}}^2} = \left[\frac{d}{dx}\left(\frac{f_{\mathrm{n}}(x)}{f_{\mathrm{t}}(x)}\right)\right]\frac{dx}{dN_{\mathrm{b}}} = \left[\frac{d}{dx}\left(\frac{f_{\mathrm{n}}(x)}{f_{\mathrm{t}}(x)}\right)\right]\frac{1}{f_{\mathrm{t}}(x)}$$

Since, everywhere,

[176] $$\frac{dN_{\mathrm{b}}}{dx} = f_{\mathrm{t}}(x)$$

Differentiating [175] within the bracket,

[177] $$\frac{d}{dx}\left(\frac{f_{\mathrm{n}}(x)}{f_{\mathrm{t}}(x)}\right) = \frac{f_{\mathrm{t}}(x)f_{\mathrm{n}}'(x) - f_{\mathrm{n}}(x)f_{\mathrm{t}}'(x)}{\left[f_{\mathrm{t}}(x)\right]^2}$$

and substituting [177] in [175] we obtain, for the second derivative of H_{n} with respect to N_{b}, everywhere,

[178] $$\frac{d^2H_{\mathrm{n}}}{dN_{\mathrm{b}}^2} = \frac{f_{\mathrm{t}}(x)f_{\mathrm{n}}'(x) - f_{\mathrm{n}}(x)f_{\mathrm{t}}'(x)}{\left[f_{\mathrm{t}}(x)\right]^3}$$

remembering that the *f*'s on the right hand side are ordinates, and derivatives of ordinates, of the unrelativized latent frequency functions, taken with respect to *x*. Since

[179] $$f_{\mathrm{t}}(x) = f_{\mathrm{s}}(x) + f_{\mathrm{n}}(x)$$

everywhere, then

[180] $$f_{\mathrm{t}}'(x) = f_{\mathrm{s}}'(x) + f_{\mathrm{n}}'(x)$$

everywhere, and substituting [179]-[180] in [178] we obtain

[181] $$\frac{d^2H_{\mathrm{n}}}{dN_{\mathrm{b}}^2} = \frac{f_{\mathrm{s}}(x)f_{\mathrm{n}}'(x) + f_{\mathrm{n}}(x)f_{\mathrm{n}}'(x) - f_{\mathrm{n}}(x)f_{\mathrm{s}}'(x) - f_{\mathrm{n}}(x)f_{\mathrm{n}}'(x)}{\left[f_{\mathrm{t}}(x)\right]^3}$$

and at hitmax cut, where $f_{\mathrm{s}}(x) = f_{\mathrm{n}}(x)$, we substitute $f_{\mathrm{n}}(x)$ for $f_{\mathrm{s}}(x)$ in [181] to get

[182] $$\frac{d^2H_{\mathrm{n}}}{dN_{\mathrm{b}}^2} = \frac{f_{\mathrm{n}}(x)\left[f_{\mathrm{n}}'(x) - f_{\mathrm{s}}'(x)\right]}{\left[f_{\mathrm{t}}(x)\right]^3}$$

which, since $f_{\mathrm{t}}(x) = 2f_{\mathrm{n}}(x)$ at hitmax, yields there the value

[183] $$\frac{d^2H_{\mathrm{n}}}{dN_{\mathrm{b}}^2}=\frac{f_{\mathrm{n}}(x)\left[f_{\mathrm{n}}'(x)-f_{\mathrm{s}}'(x)\right]}{8\left[f_{\mathrm{n}}(x)\right]^3}$$

[184] $$=\frac{f_{\mathrm{n}}'(x)-f_{\mathrm{s}}'(x)}{8\left[f_{\mathrm{n}}(x)\right]^2}$$

Re-writing this as two fractions and substituting $f_{\mathrm{s}}(x)$ for $f_{\mathrm{n}}(x)$ in the second one,

[185] $$\frac{d^2H_{\mathrm{n}}}{dN_{\mathrm{b}}^2}=\frac{1}{8}\left[\frac{f_{\mathrm{n}}'(x)}{\left[f_{\mathrm{n}}(x)\right]^2}-\frac{f_{\mathrm{s}}'(x)}{\left[f_{\mathrm{s}}(x)\right]^2}\right]$$

Expressing the f's and their derivatives in terms of relativized frequency functions in accordance with the relations

[186] $$\begin{cases} f_{\mathrm{s}}(x)=N_{\mathrm{s}}\phi_{\mathrm{s}}(x) \\ f_{\mathrm{n}}(x)=N_{\mathrm{n}}\phi_{\mathrm{n}}(x) \\ f_{\mathrm{s}}'(x)=N_{\mathrm{s}}\phi_{\mathrm{s}}'(x) \\ f_{\mathrm{n}}'(x)=N_{\mathrm{n}}\phi_{\mathrm{n}}'(x) \end{cases}$$

we have, at hitmax,

[187] $$\frac{d^2H_{\mathrm{n}}}{dN_{\mathrm{b}}^2}=\frac{1}{8}\left[\frac{1}{N_{\mathrm{n}}}\frac{\phi_{\mathrm{n}}'(x)}{\left[\phi_{\mathrm{n}}(x)\right]^2}-\frac{1}{N_{\mathrm{s}}}\frac{\phi_{\mathrm{s}}'(x)}{\left[\phi_{\mathrm{s}}(x)\right]^2}\right]$$

which is the denominator of [171] as far as it can be taken without plugging in numerical values. We write the unsigned error in cumulative frequency as

[188] $$\left|\Delta N_{\mathrm{b}}\right|<\frac{1}{2}\frac{\left(\mathrm{U}_{\mathrm{b}}-\mathrm{U}_{\mathrm{a}}\right)}{\frac{1}{8}\left[\frac{1}{N_{\mathrm{n}}}\frac{\phi_{\mathrm{n}}'(x)}{\left[\phi_{\mathrm{n}}(x)\right]^2}-\frac{1}{N_{\mathrm{s}}}\frac{\phi_{\mathrm{s}}'(x)}{\left[\phi_{\mathrm{s}}(x)\right]^2}\right]}$$

Eq [188] subscript mistyping: denominator term $1/N_{\mathrm{s}}$ was originally mis-typed as $1/N_{\mathrm{n}}$

[189] $$<\frac{4\left(h_{\mathrm{b}}-h_{\mathrm{a}}\right)}{\frac{1}{N_{\mathrm{n}}}\frac{\phi_{\mathrm{n}}'(x)}{\left[\phi_{\mathrm{n}}(x)\right]^2}-\frac{1}{N_{\mathrm{s}}}\frac{\phi_{\mathrm{s}}'(x)}{\left[\phi_{\mathrm{s}}(x)\right]^2}}$$

at hitmax.

Note that the denominator is not actually a difference, but a sum, of numerical values, because

[190] $\phi_{\mathrm{n}}'(x)<0$ to right of its mode

[191] $\phi_{\mathrm{s}}'(x)<0$ to left of its mode

so the first term of denominator is negative at hitmax, the second term positive (but to be subtracted), hence the numerical value of the denominator is a sum,

[192] $$\text{Denom} = \frac{1}{N_{\text{n}}}\left|\frac{\phi'_{\text{n}}(x)}{\left[\phi_{\text{n}}(x)\right]^2}\right| + \frac{1}{N_{\text{s}}}\frac{\phi'_{\text{s}}(x)}{\left[\phi_{\text{s}}(x)\right]^2}$$

I have not been able to simplify this further, or to prove any general statements about it useful for present purposes. To set a constraint on the error ΔN_{b} one must plug in extreme numerical values of the hit-rates above and below, and evaluate the ratios of derivatives to squares of frequency functions. The behavior of this fraction when the intra-taxon distributions are normal is very reassuring. Thus, for example, at +1σ the ratio of the derivative $\phi'(x)$ of the ordinate to the square $\left[\phi(x)\right]^2$ of the ordinate is ≃ 4.00, which with hit-rates as asymmetrical as $h_{\text{a}} = .80$ and $h_{\text{b}} = .60$ yields an error, if $N_{\text{s}} = N_{\text{n}} = 100$, of less than 10 cases, i.e., the cut would usually fall in the correct interval, and would never deviate by more than one interval from the correct one. What rough numerical trials I have done on very skewed distributions (e.g., on a chi-square distribution for 1 d.f.) are similarly reassuring. Pending thorough canvass of distribution forms in relation to base-rates and overlap, I opine with some confidence that minimizing ($SS_{\text{a}} + SS_{\text{b}}$) can serve as a pretty accurate hitmax locator.

f. Possible use of covariances above and below cut as a good approximate hitmax cut locator.

In the first method for hitmax cut location (PR-65-2, Section 3, and Section 1 of this report) we relied upon the fact that, given the assumption of zero (or equal) latent intra-taxon covariances between two indicators, their covariance within a sub-population is an increasing function of the latter's "mixture" (heterogeneity, taxon impurity), i.e., the largest (*xy*)-covariance should be observed among cases lying in the *w*-interval where *p* (schizotypy) = 1/2 = *p* (nonschizotypy), therefore in the hitmax interval on *w*. This same line of reasoning should apply for sets of *w*-intervals, including the whole set lying above a *w*-cut and the set lying below. The slopes of the (*xy*)-regression lines within the two sub-populations defined by a sliding cut on input-variable *w* should have a tendency to increase when these sub-populations are more "mixed," and to decline when they are relatively "pure." Thus, for example, if the intra-taxon (*xy*)-correlation were strictly zero, and a *w*-cut could be found such that no latent taxon-

misclassifications resulted from that cut, then the regression line of y on x would be horizontal for the N_{aw} cases above w_h (all being schizotypes), and the same would be true for the y-on-x regression line below the cut (all N_{bw} of these being nonschizotypes). And intuition generalizes from this idealized case of a hitmax-infallible "input" indicator to the general dependence of these two slopes upon the goodness-of-classification achieved by various input-indicator cuts. There should be some method of weighting-and-combining yx-slopes above and below a sliding w-cut, such that this function has a minimum at, or close to, the hitmax cut w_h.

In Section 2f above, as a corollary to the theorem concerning a latent expression for sums of squares on a single output indicator, we found further that when a pair of "output" indicators is studied, their frequency-weighted (xy)-covariances sum to a minimum when the latent condition

$$[193] \qquad N_{aw}\,\mathrm{cov}(xy)_{aw} + N_{bw}\,\mathrm{cov}(xy)_{bw} = N_s\,\mathrm{cov}_s(xy) + N_n\,\mathrm{cov}_n(xy) + \Delta\bar{x}\Delta\bar{y}\left(N_a p_a q_a + N_b p_b q_b\right) = \mathrm{MIN}$$

is satisfied. By reasoning strictly analogous to that of the just preceding sub-section 4f, where the variable quantity to be minimized, in minimizing ($SS_a + SS_b$) on a single output variable, was identical with the parenthetic quantity ($N_a p_a q_a + N_b p_b q_b$) in [193], we conclude, as a corollary to the theorem of Section 4e, that one can closely approximate the hitmax cut on one indicator by finding the cut such that the sum of the two cross-product deviation sums above and below the cut, i.e., the left-hand side of Equation [193], has the minimum value empirically attainable.

g. Some important unsettled questions

I suspect that the reader who has stuck with it to this point is feeling rather like the little boy who returned a library book, saying, "Ma'am, this book tells more about penguins than I wanted to know." Perhaps more time has been spent on these derivations than is warranted, pending a full-scale Monte Carlo study of the approach. All we have to go on is the Seth thesis (1965) and preliminary results on the single-indicator method. My justification lies in the fact that illness and financial problems slowed up a doctoral candidate's progress on the full Monte Carlo job; but in the meantime scores of requests for PR-65-2, elicited by the cites in Dawes and Meehl (1966) and Hays (1968), have been responded to. I therefore felt it imperative to put out the present report, lest other workers waste time through ignorance of these more recent developments. Furthermore, trials of the method on diverse real-data problems, both psychological and biological, are perhaps more valuable at this stage than Monte Carlo runs.

Before expending further effort on mathematical features of the idealized latent model, we need to know four kinds of things about the overall approach, as I see it:

(1) Size of random sampling errors in the estimators, as a function of base-rates, indicator validities, distribution shapes, ratio of their variances, and *N*.

(2) Size and direction of systematic errors due to empirical departures from the idealization (e.g., discontinuity) and especially the unrealistic hypothesis of zero intra-taxon correlation (required in many of the derivations) or, at least, equal covariances between taxa (required in almost all).

(3) Numerical values to be imposed upon the results of the various consistency tests, i.e., "how far off" can they be without impugning the idealization's adequacy?

(4) When the safe numerical limits on (3) are exceeded, and the iterative method is relied upon to "bootstraps" a closer approximation, does convergence (and satisfaction of consistency tests) assure us about the final estimates, or does the bootstrapsing sometimes converge, through circular reasoning or subtle dependencies, to a biased estimate showing pseudo-consistency?

Answering these four kinds of questions would seem to have highest priority, but analytic solutions are far beyond my powers (and, I suspect, cannot be provided without a stronger model). If these issues can be settled in a satisfactory direction, it will then be time to investigate some less urgent but still important matters, such as the following:

(5) Obviously a kind of symmetry exists between main estimators and what I have called "consistency tests," in that the roles can usually (not always) be reversed. Thus, for example, we have in this report elevated a consistency test presented in PR-65-2 (Section 9c, the "Hitmax interval covariance test") to a prime role as estimator of the constant $K = \Delta x \Delta y$, which is then employed to draw the latent frequency functions (Section 1 of this report). There should be some rational or empirical basis for assigning one or the other role to a manifest-latent equation, especially when we have several multiple estimates available for the same parameter, and also more consistency tests than we need.

(6) There are, presumably, differential sensitivities, among the consistency tests, to different departures from the idealization; and the same must be true of the alternative estimators. It seems plausible to suppose that the pattern of consistency-test "danger signals" might inform us as to which estimators should be relied upon in preference to others, in a particular real-data problem.

(7) Aside from the pattern of consistency tests (e.g., assuming they are all reasonably satisfied by a set of data) the several methods for estimating the same latent value are surely not equally informative. Thus we have, all told, some half-dozen hitmax cut locators (see PR-65-2, Sections 3 and 4; and in the present report, Sections 2a, 2b, 4e, and 4f). They differ considerably in how much of the total data each relies on, as well as probable intrinsic instability or "crudeness," e.g., the equating of instantaneous slope and an interval statistic in the method of Section 2a is likely to be a rather touchy operation in comparison with the method of Section 2b. Which methods are best? Should the poorer ones be ignored, or relegated to the role of consistency tests, or what?

(8) Should some, or all, of the latent values be estimated by pooling numerical results of the several methods? I take it that this is a question of the "sufficiency," in Fisher's sense, of the estimator, and therefore should be answerable analytically, if one has the mathematical competence.

(9) Allied to the preceding query is the question of pooling the several numerical values arrived at via the same sub-method but employing different indicators (in the "input" and "output" roles). For example, the simplified sequence set out in Section 1 yields, among other things, a base-rate estimate. For any triad of three indicators (x, y, z) the method of Section 1 can be applied three ways, i.e., with x as "input" and the (y, z)-pair as output, with y as input and the (x, z)-pair as output, and with z as input and the (x, y)-pair as output. In a four-indicator family (x, y, z, w) there arise $3 \times 4 = 12$ such indicator-patterns for use with the simplified procedure, and hence we can obtain 12 numerical estimates of P, partially but not wholly overlapping as to data, and some of the sets being almost completely independent, e.g., (x, y)-pair against z as input, and (z, w)-pair against x as input. In a 5-indicator family there are entirely non-overlapping sets of data yielding estimates of P. Should all available P-estimates obtainable from an indicator-family be averaged, weighted or unweighted, before re-cycling within each triad, or how should these values be optimally used?

(10) Alternatively, can a more general system of latent-manifest equations be written including the whole set of k available indicators at once?

(11) Can an "appearance of taxonomy" arise spuriously? If so, how? Can the consistency tests also be satisfied by such an unlucky pseudo-taxonomic situation?

(12) Since so many of the sub-routines involve maximum (or minimum) values, which in practice will be somewhat fuzzy due to errors of measurement, sampling, and grouping, some defensible graph-smoothing operation is in order. Thus, for example, location of the hitmax cut on *w* by finding the numerical maximum of $\text{cov}(xy)_w$ over the *w*-intervals can easily be distorted through a bad sampling fluctuation in one near-maximum class-interval, which might be avoided by considering the behavior of the cov(*xy*) function in the adjacent regions. It was apparent in Mr. Seth's M.A. thesis research (Seth, 1965) that these problems would have to be dealt with somehow, even if nothing more rationally defensible than one of the well-known crude smoothing operations could be found to do the job.

REFERENCES

Cronbach, L. J., & Meehl, P. E. (1955) Construct validity in psychological tests. *Psychol. Bull*., *52*, 281-302.

Dawes, R. M., & Meehl, P. E. (1966) Mixed group validation: A method for determining the validity of diagnostic signs without using criterion groups. *Psychol. Bull*., *66*, 63-67.

Hald, A. (1952) *Statistical theory with engineering applications*. New York: Wiley and Sons.

Hays, W. C. (1958) Statistical theory. In P. R. Farnsworth, M. R. Rosenzweig & J. T. Polefka, (Eds.) *Annual Review of Psychology, Vol. 19* (pp. 417-436).

Lykken, D. T. (1966, December 30) *Statistical significance in psychiatric research*. Reports from the Research Laboratories of the Department of Psychiatry, University of Minnesota, Report Number PR-66-9.

Lykken, D. T. (1968) Statistical significance in psychological research, *Psychol. Bull*., *70*, 151-159.

Meehl, P. E. (1959) Some ruminations on the validation of clinical procedures. *Canad. J. Psychol*., *13*, 102-128.

Meehl, P. E. (1965, May 25) *Detecting latent clinical taxa by fallible quantitative indicators lacking an accepted criterion*. Reports from the Research Laboratories of the Department of Psychiatry, University of Minnesota, Report Number PR-65-2.

Meehl, P. E. (1967) Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*, 103-115.

Meehl, P. E. (1973). MAXCOV-HITMAX: A taxonomic search method for loose genetic syndromes. In Meehl, *Psychodiagnosis: Selected papers* (pp. 200-224). Minneapolis: University of Minnesota Press.

Meehl, P. E., & Golden, R. (1982). Taxometric methods. In P. Kendall & J. Butcher (Eds.), *Handbook of research methods in clinical psychology* (pp. 127-181). New York: Wiley.

Meehl, P. E., & Rosen, A. (1955) Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychol. Bull*., *52*, 194-216.

Meehl, P. E., & Yonce, L. J. (1994) Taxometric analysis: I. Detecting taxonicity with two quantitative indicators using means above and below a sliding cut (MAMBAC procedure). *Psychological Reports, 74*, 1059-1274.

Meehl, P. E., & Yonce, L. J. (1996) Taxometric analysis: II. Detecting taxonicity using covariance of two quantitative indicators in successive intervals of a third indicator (MAXCOV procedure). *Psychological Reports, 78*, 1091-1227.

Meehl, P. E., Lykken, D. T., Burdick, M. R., & Schoener, G. R. (1969) *Identifying latent clinical taxa, III. An empirical trial of the normal single-indicator method, using MMPI Scale 5 to identify the sexes*. Reports from the Research Laboratories of the Department of Psychiatry, University of Minnesota, Report No. PR-69-1.

Seth, R. E. (1965) *Hitmax cut accuracy of a latent taxa model as a function of sample size*. Unpublished M.A. Thesis, University of Minnesota.

pdf by LJY Nov 2014