**Optimum ways of testing for equality of thresholds as a null hypothesis, and deriving a likelihood function for magnitude of threshold difference between conditions**

Three approaches to consider, all based on likelihood considerations:

Psyfit delivers, for each condition analysed, the probability of the data as a function of threshold, t, and psychometric width parameter w. For each condition, the marglh element of the statstruct output structure is the absolute probability of the data D1 or D2 as a function of t or w, numerically averaged over values of the other psychometric parameter; thus prob(D | t) is the average of prob(D| t,w), the probability of the data for given values of t and w, which is also proportional to the posterior joint probability density of t and w (given the data D1 or D2 respectively and a flat prior), for each width w and threshold t. The marginals prob(D1 | t) = p1(t) for condition 1 and p(D2 | t) = p2(t) for condition 2 form the first row of the marglh element of the relevant statstruct, i.e. statstruct(1) for condition 1 and statstruct(2) for condition 2. These differ from the relative likelihoods generated as marglh in previous versions of psyfit.m in that they are simply estimated probabilities for the entire data set, with no normalization of marglh to a peak of 1 or to a sum of 1 over the t values.

Averaging over values of the free parameter (psychometric width) to form the expected probability for given t value is tricky, and an ‘empirical Bayes’ weighting is adopted in which each value of w is given a weight proportional to its own (posterior estimated) likelihood. This amounts to averaging the squares of the probabilities, see comments in psyfit.m and (2) below. This needs further rumination and justification, as per emails with Ed, or Wikipedia on empirical Bayes. But it is obvious that with typical not too sparse data sets, there is little possibility of underestimation of uncertainty by adopting the sample as a post hoc basis for assessing the uncertainty of parameter values.

(1) A simple approach avoids these difficulties: note the **mode** of p1 and p2; these are the likelihoods at the **maximum a posteriori (MAP)** estimates of threshold for each condition. Multiply these to obtain a number proportional to the relative probability of all data for both conditions together given the MAP estimates of threshold for each condition. Compare this with the mode of the likelihood function (relative pdf) obtained when the combined data of both conditions are analysed, which is the likelihood at the single MAP threshold estimate obtained under the constraint that the threshold is the same for both conditions. For the data of alantest.m the estimated ratio of the independent to the constrained likelihoods is 2.2 x 10^-4, as compared with .2 if the data for both conditions are the same (should be 1, but the rough approximations made could explain the deviation?)[[1]](#footnote-1).

(2) But as usual, use of the likelihood at the MAP value of the threshold parameter is a poorly justified substitute for a comparison of the **probability of the data averaged for all free parameter values**, for the constrained and unconstrained analyses. The **unweighted** integral is the **volume under the probability distribution** for all of the data. The **volume under the probability distribution** is the posterior probabilities of the data when the parameters are not selected for optimal values but allowed to range freely. This was the procedure apparently advocated by Laplace and it defines the ‘Bayes Factor’ (Kass and Raftery), a quantity much more difficult to compute than the likelihood ratio based on the MAP points. But as K &R describe (their Equation 4), if the likelihood distributions can be approximated by bivariate normals (which is a reasonable approximation for reasonably complete psychophysical data sets), a fairly simple correction factor generates the volume ratio from the MAP ratio: multiply each MAP by the square root of the determinant of the covariance matrix of the likelihood (and the covariance matrix is the inverse (with a sign change) of the Hessian matrix, that is made up of the second derivatives of the likelihood function; so the determinant is the reciprocal of the determinant of the Hessian). The difference this makes as compared with method (1) can be assessed by comparing the standard deviations, sigmalatticemarg. If the threshold-width correlation implicit in the two likelihood surfaces (for independent threshold values, vs equal threshold values) is comparable, the likelihood ratio will change by the geometric mean of the ratios of sigma values for parameters 1 and 2. For the data of alantest.m, the change is less than a factor of 2.

Numerical integration to estimate the volumes under l(t,w) for the constrained and unconstrained analyses is an alternative to the Laplace approximate method. This would be easy within psyfit.m as it is approximated by just comparing the totals or means of the probability values in the errlh matrices in statstruct. This is what Laplace’s procedure approximates, but it doesn’t weight the different parameter value choices by their posterior likelihood. Instead there is a case for weighting the probabilities associated with each parameter value vector by the likelihood of that parameter vector as indicated by the data themselves (along with any prior): ‘empirical Bayes’, see Wikipedia.

In this approach the probability of the data under a model would be assessed with a likelihood-weighted average over both model parameter values--not the unweighted average that defines the volume under the likelihood function in parameter space. This probability is available in statstruct.probdata. For the unconstrained model, where both threshold parameter values can differ for the two conditions, the probability of the combined data appropriately averaged over *both* the threshold parameter values (t and w) for the two conditions, is easily obtained by multiplication as statstruct1.probdata\*statstruct2.probdata, where statstruct1 and statstruct2 are from the separate psyfit outputs for data sets D1 and D2. For the constrained model, a run fitting the merged data from both sets would yield, say, statstruct1\_2.probdata. The ratio of this to the product statstruct1.probdata\*statstruct2.probdata is the relative probability of the data, under the constraint of identical threshold parameter values, to the probability of the data without that constraint (but using instead the ‘empirical Bayes’ estimate of the joint posterior pdf of the parameter values for each data set independently).

(2b) Often we want to **constrain one of the threshold parameters, say w, to equality** and derive some kind of fiducial interval for the difference in the other, unconstrained parameter, say t. In principle, the likelihood of threshold inequality should be assessed taking account of uncertainty in each estimate of t by averaging over the plausible range of t in each fit. Psyfit provides, in marglh for t, the probability of the data as a function of threshold, t, for each condition separately, prob(D1 | t) for condition 1 and prob(D2 | t) for condition 2. The probability of the combined data, as a function of *both* the threshold estimates t1 and t2 for the two conditions, can be obtained by multiplication,

prob(D1,D2 | t1, t2) = prob(D1 | t1)\*prob(D2|t2);

this product defines a single surface if the two threshold values t1 and t2 are adopted as Cartesian coordinates. Along the diagonal, the null hypothesis holds—the thresholds are equal. (The MAP point for the merged data is the peak and is generally off the diagonal). If the null hypothesis has 50% prior probability, the Bayesian posterior probability for the null hypothesis is the ratio of the average (likelihood-weighted) diagonal probability value to the weighted average probability over the entire surface.

This procedure for accepting or rejecting the null hypothesis of threshold equality follows Wetzels? Rouder? Others? One can alternatively avoid the binary accept/reject procedure (and the associated allocation of substantial prior probability to the point or line representing the null hypothesis) by deriving the full likelihood function for the difference in threshold, obtained by averaging prob(D1,D2 | t1, t2) along the parallel diagonal lines t1 = t2 + delta to yield the marginal likelihood as a function of delta. A problem for this method is that psyfit doesn’t generally use exactly the same parameter values to analyse different conditions. So interpolation is called for, see [threshdifflkhd.m](file:///C:\toolbox\stair\threshdifflkhd.m).

As always, Monte Carlo runs can be tried to check this. Also a simpler approximate approach can be based on mean and sigma of the threshold estimates for the two conditions (code for that on Rouder’s web site). The script [threshdifflkhd.m](file:///C:\toolbox\stair\threshdifflkhd.m) does the full analysis above, calling psyfit.m, but the results seem generally to support the simpler procedure, so the more elaborate analysis is arguably a waste of time to consider.

1. Kass and Raftery, Well known basic ref on Bayes Factors and BIC: JAMSA,

<http://www.jstor.org/stable/10.2307/2291091>

2.Wetzels et al: http://pps.sagepub.com/content/6/3/291

DOI: 10.1177/1745691611406923

Perspectives on Psychological Science 2011 6: 291

Ruud Wetzels, Dora Matzke, Michael D. Lee, Jeffrey N. Rouder, Geoffrey J. Iverson and Eric-Jan Wagenmakers, Statistical Evidence in Experimental Psychology : An Empirical Comparison Using 855 t Tests

3. Psychonomic Bulletin & Review 2009, 16 (2), 225-237

doi:10.3758/PBR.16.2.225

Bayesian t tests for accepting and rejecting the null hypothesis

JEFFREY N. ROUDER, PAUL L. SPECKMAN, DONGCHU SUN, AND RICHARD D. MOREY University of Missouri, Columbia, Missouri GEOFFREY IVERSON University of California, Irvine, California

1. The meaningfulness of the ratio is compromised by the loss of one degree of freedom in choosing the thresholds independently. Correspondingly for non-identical data, the ratio would always be < 1. Can this be handled analogously to Bessel’s correction of (N-1/N) for sample size in standard errors, by simply doubling the constrained likelihood, effectively a conservative discounting of one binary response’s contribution to the decline of the likelihood of the data? The rationale is that independent choice of threshold would allow a single trial and response to be accommodated… but this is vague reasoning, and the analogy between analysis of binary vs continuous data is far from perfect. The Bayesian Information Criterion (BIC: Jeffreys; Kass and Raftery tutorial) might suggest a much larger correction for the added free parameter (and there is an added free parameter, even though no data are being fit with a more complex model). But this is a separate, and less important, issue. All values of t and w are considered in both the constrained and the unconstrained analysis. Yet, in the extreme case where the data are divided into blocks with only 2 trials per block, values could be found for the 2 free parameters that predict the 2 responses in each block.So it seems appropriate to scale the log probability by the ratio of degrees of freedom to total number of trials ((N-2)/N), since the log probability is expected to accumulate linearly over trials. Check, and try with identical data…but in any case, this correction would be quite inconsequential for typical data sets. [↑](#footnote-ref-1)