Appendix 1: Statistical Techniques

Logit or "Log-odds" Transform

If subject i of N makes r_i errors in a sample of n_i responses, giving a basic proportional score of $p_i=r_i/n_i$, then the log-odds score is defined as:

$$L_i = ln\left(\frac{p_i}{1-p_i}\right) = ln\left(\frac{r_i}{n_i - r_i}\right) \qquad \text{Equation 9-1}$$

This logarithmic transform "stretches" proportional scores just above zero and just below one, making some allowance for floor and ceiling effects. However, the transform is not defined at these extremina exactly, i.e., when $r_i=0$ or $r_i=n_i$.

Empirical Log-odds

To handle situations when $r_i=0$ or $r_i=n_i$ $(n_i>0)$, an empirical log-odds score can be defined (Cox & Snell, 1989):

$$L_i = ln\left(\frac{r_i + 0.5}{n_i - r_i + 0.5}\right) \qquad \text{Equation 9-2}$$

This caters for measured proportions of zero or one. However, it makes no allowance for the fact that one proportional measurement may be based on a large number of observations (when n_i is large), and hence likely to be more accurate than one based on only a small number of observations (when n_i is small).

Weighted Log-odds

The empirical log-odds score defined above has an associated variance:

$$U_i = \frac{(n_i + 1)\ (n_i + 2)}{n_i\,(r_i + 1)\ (n_i - r_i + 1)} \qquad \text{Equation 9-3}$$

Empirical log-odds scores can be weighted by the inverse of their variances:

$$W_i = \frac{1}{U_i} \qquad \text{Equation 9-4}$$

to give a weighted mean across subjects:

$$\bar{L} = \frac{\sum W_i L_i}{\sum W_i} \qquad \text{Equation 9-5}$$

where the summand is from subject $i=1$ to $i=N$.

Transforming back into the original coordinates, the estimated mean proportion over subjects, $\bar{p}$, has variance, $\bar{V}$, given by:

$$\bar{p} = \frac{e^{\bar{L}}}{1+e^{\bar{L}}} \qquad \bar{V} = \bar{U}\bar{p}^2 (1-\bar{p})^2 \qquad \text{Equation 9-6}$$

Testing Related, Weighted, Log-odds

To test a difference in means of two, related log-odds, L_i and L'_i, let:

$$d_i = L_i - L'_i \qquad \text{Equation 9-7}$$

A combined weight, w_i, can be determined from:

$$w_i = \left(\frac{1}{W_i} + \frac{1}{W'_i}\right)^{-1} \qquad \text{Equation 9-8}$$

The weighted mean difference score is then:

$$\bar{d} = \frac{\sum w_i d_i}{\sum w_i} \qquad \text{Equation 9-9}$$

and the standard error of the difference scores is:

$$\sqrt{\frac{1}{\sum w_i}} \qquad \text{Equation 9-10}$$

which enables testing of the standardised score:

$$Z(N) = \frac{\sum w_i d_i}{\sqrt{\sum w_i}} \qquad \text{Equation 9-11}$$

To get a measure of the weighted mean difference in terms of proportions, $\bar{q}$:

$$\bar{q} = \frac{\bar{p}(1-\bar{p})(e^{\bar{d}}-1)}{1+\bar{p}(e^{\bar{d}}-1)} \qquad \text{Equation 9-12}$$

given a particular baseline proportion $\bar{p}$.

Testing Unrelated, Weighted, Log-odds

To test the difference between log-odds of two, unrelated groups (with N subjects):

$$Z(N) = \frac{\bar{L} - \bar{L'}}{\sqrt{\frac{1}{\sum W_i} + \frac{1}{\sum W_i'}}} \qquad \text{Equation 9-13}$$

Testing 2x2 Contingency Tables

Given r_i errors and $(n_i\text{-}r_i)$ correct responses on one measure, and s_i errors and $(n_i\text{-}s_i)$ correct responses on another, such that the 2x2 contingency table for each subjects is:

x_i	..	r_i
..	..	$n_i\text{-}r_i$
s_i	$n_i\text{-}s_i$	n_i

Then, under the null hypothesis that the two measures are uncorrelated, the expected number of cases where they are in agreement, *E(x),* is:

$$x = \sum x_i \qquad E(x) = \sum \frac{r_i s_i}{n_i} \qquad \text{Equation 9-14}$$

and the variance of *x* is:

$$V(x) = \sum \frac{r_i s_i (n_i - r_i)(n_i - s_i)}{n_i n_i (n_i - 1)} \qquad \text{Equation 9-15}$$

which allows a combined test of significance of individual subjects' two-by-two contingency tables by a Z-score:

$$Z(N) = \frac{x - E(x)}{\sqrt{V(x)}} \qquad \text{Equation 9-16}$$

giving a measure of the association or correlation between the two measures. Note that this assumes homogeneity across subjects, such that they all show a similar association in their individual contingency tables.

Testing Conditional Error Probabilities

Given m_j reports which have no errors on positions *1...j-1* and assuming r_j errors are made on position *j*, then:

$$m_{j+1} = m_j - r_j \qquad \text{Equation 9-17}$$

Let the conditional probability of an error on position *j* be q_j. The maximum likelihood estimate of q_j is:

$$\hat{q}_j = \frac{r_j}{m_j} \qquad \text{Equation 9-18}$$

If there is no change in conditional probability of an error across positions *j* and *j+1*, then the common maximum likelihood estimator is:

$$\hat{q} = \frac{r_{j+1} + r_j}{m_{j+1} + m_j} \quad \text{Equation 9-19}$$

To test the hypothesis that $q_j = q_{j+1}$, the goodness of fit of the following model is tested:

$$E[r_j] = m_j\hat{q} \qquad E[r_{j+1}] = m_j(1-\hat{q})\,\hat{q}$$

$$E[m_{j+2}] = m_j(1-\hat{q})\,(1-\hat{q}) \quad \text{Equation 9-20}$$

The corresponding X^2 statistic has (approximately) a Chi-squared distribution on one degree of freedom under the null hypothesis. A combined χ^2 across subjects can be obtained by summing individual, signed Z scores, squaring and dividing by *N*.

Significance of Multiple Pairwise Comparisons

Given *N* pairwise, a priori comparisons with individual significance levels of α, the appropriate familywise significance level, α_F, according to a Bonferroni correction is:

$$\alpha_F = \frac{\alpha}{N} \quad \text{Equation 9-21}$$

This correction is very conservative. A more powerful approach is Holm’s method (Howell, 1992), for which the Bonferroni correction is applied iteratively to each individual comparison, testing the largest absolute difference against α_F above, and testing *i*th next largest difference ($i=1..N-1$) against α_i, where:

$$\alpha_i = \frac{\alpha}{N-i} \quad \text{Equation 9-22}$$

The iteration continues until the *i*th comparison is nonsignificant, whence all remaining comparisons are also deemed nonsignificant.

Hotelling's T-squared Test

Hotelling's one-sample T^2-test can be used to test p means of samples taken from n subjects against hypothesised values (Mardia, Kent & Bibby, 1979). Let **h** be the vector of hypothesised means, **d** be the vector of actual means, and **S** be the matrix of the sums of squares of data values. Then the vector **t** of the differences of hypothesised and actual means:

$$\boldsymbol{t} = \boldsymbol{d} - \boldsymbol{h} \qquad \text{Equation 9-23}$$

gives the T^2 statistic:

$$T^2 = n\,(\boldsymbol{t}^T \boldsymbol{S}^{-1} \boldsymbol{t}) \qquad \text{Equation 9-24}$$

which can be tested by the F-ratio:

$$F(p, n-p) = \frac{(n-p)}{p\,(n-1)} T^2 \qquad \text{Equation 9-25}$$

Note that T^2 statistic, by taking into account the variances and covariances of the p data samples, does not have to assume independence of the p means.