412TW-PA-16437

Ridit Analysis for Cooper-Harper & other Ordinal Ratings for Sparse Data –

A Distance-based Approach

ARNON HURWITZ, PhD

AIR FORCE TEST CENTER
EDWARDS AFB, CA

SEPTEMBER 2016

4
1
2
T
W

412TH TEST WING
EDWARDS AIR FORCE BASE, CALIFORNIA
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)* 09-19-2016	**2. REPORT TYPE** TECHNICAL PAPER	**3. DATES COVERED** *(From - To)* 09-19-2016

4. TITLE AND SUBTITLE

RIDIT ANALYSIS FOR COOPER-HARPER & OTHER ORDINAL RATINGS FOR SPARSE DATA – A DISTANCE-BASED APPROACH

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

ARNON HURWITZ, PH.D.

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) AND ADDRESS(ES)

412th Test Wing
307 East Popsun Ave
Edwards AFB CA 93524-6630

8. PERFORMING ORGANIZATION REPORT NUMBER

412TW-PA-16437

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

N/A

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release A: distribution is unlimited.

13. SUPPLEMENTARY NOTES

CA: Air Force Test Center Edwards AFB CA CC: 012100

14. ABSTRACT

Ordinal categorical data (OCD), such as opinion rankings, are common in many areas of application. In the Air Force, Cooper-Harper ratings are used extensively for the assessment of Flying Qualities. OCD is not, however, a ratio-scale measurement and cannot be treated as ordinary numbers. Notwithstanding this, the ordinal scores are often regarded as ratio-scale and analyzed incorrectly using means and variances. A method of correct analysis of OCD leading to statistically valid hypothesis tests and based on a method of probability scoring or 'Ridits,' has found wide applicability for other large-data-set applications such as Epidemiology. This paper explains the use of Ridits and examines how we might effect a Ridit analysis on the often sparse data sets in many Flying Qualities applications. All flying qualities data in this paper is synthetic, and has been simulated to illustrate Ridit analysis. The method of this paper is to fit empirical Beta distributions to observed data, and then to use a randomization approach to make inferences on the difference between distributions based on a distance metric.

15. SUBJECT TERMS *Borg scale rating; Cooper-Harper; flying qualities; Hellinger distance; human factors; ordinal categorical data; Ridit; sparse data; statistical defensibility.*

16. SECURITY CLASSIFICATION OF: **Unclassified**			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** 412 TENG/EN (Tech Pubs)
a. REPORT Unclassified	**b. ABSTRACT** Unclassified	**c. THIS PAGE** Unclassified	None	19	**19b. TELEPHONE NUMBER** *(include area code)* 661-277-8615

Ridit Analysis for Cooper-Harper & other Ordinal Ratings for Sparse Data – A Distance-based Approach

Arnon Hurwitz, PhD

Statistical Methods Office
Edwards Air Force Base
Building 1440, #115
E. Popson Avenue
Edwards, CA 93524

Approved for public release; distribution unlimited. 412TW-PA-16437

Abstract

Ordinal categorical data (OCD), such as opinion rankings, are common in many areas of application. In the Air Force, Cooper-Harper ratings are used extensively for the assessment of Flying Qualities. OCD is not, however, a ratio-scale measurement and cannot be treated as ordinary numbers. Notwithstanding this, the ordinal scores are often regarded as ratio-scale and analyzed incorrectly using means and variances. A method of correct analysis of OCD leading to statistically valid hypothesis tests and based on a method of probability scoring or 'Ridits,' has found wide applicability for other large-data-set applications such as Epidemiology. This paper explains the use of Ridits and examines how we might effect a Ridit analysis on the often sparse data sets in many Flying Qualities applications[i]. The method of this paper is to fit empirical Beta distributions to observed data, and then to use a randomization approach to make inferences on the difference between distributions based on a distance metric.

Key words: *Borg scale rating; Cooper-Harper; flying qualities; Hellinger distance; human factors; ordinal categorical data; Ridit; sparse data; statistical defensibility.*

1 Introduction

Ordinal categorical data[ii], such as opinion rankings for categories of products or other items, are common in many fields where ratio-scale measurements are unavailable[iii]. In the Air Force, Cooper-Harper ratings (Cooper & Harper, 1969; Harper & Cooper, 1986) are used extensively for the assessment of Flying Qualities (Wilson & Riley, 1989, 1990). An often-made assumption (Agresti, 1984, pg. 2) is that there is a latent but unobserved continuous ratio scale[iv] underlying the observed ordinal choices. However, such ordinal scores are often treated as ratio-scale measurements, which they are not, and analyzed incorrectly using means and variances. A method of correct analysis leading to statistically valid hypothesis tests and based on a method of probability scoring or ‘Ridits,’ was first proposed by Bross (1958). Bross named the Ridit after *‘with **R**eference to an **I**dentified **D**istribution.’* Ridit analysis was later formalized by Brockett and Levine (1977), grounding the concept of a Ridit on the basis of intuitively reasonable postulates.

The following exposition explains the use of Ridit analysis and examines how we might analyze sparse data, common in many Cooper-Harper Flying Qualities applications, using Ridit and related analysis techniques. Ridit analysis is simple to compute, and permits statistics such as hypothesis test power, necessary to determine if a proposed test plan is statistically defensible, to be estimated as well. We base our presentation on two examples:

1. A college course evaluation by students. This introduces basic concepts and notation.[v]
2. Fatigue scores by pilots flying several sorties at increasing levels of G-stress.

In many applications where one is asked to compare OCD results taken under different conditions—for example, different flight configurations—one is faced with the problem of small sample sizes. The standard Ridit analysis, as found in the literature (for example, Selvin, 1977, 2004) applies correctly to large sample sizes and it is thus necessary to discover a way to better treat the analysis in the case of small samples. It is the contribution of this paper to suggest an approach that does not depend on the large-sample Normal-distribution approximation.

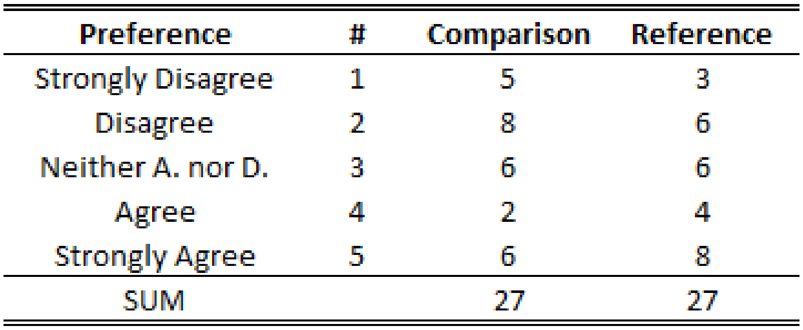
We introduce standard Ridit analysis in Example 1 below, and then apply it to a small sample case in Example 2. This second example will show how erroneous confidence intervals arise using the standard approach. In the last section of this paper, we derive an alternative analysis that does not require Normal-distribution assumptions, and apply it to the data of Example 2.

2 Examples

2.1 Example 1: Evaluation of a course by students

Consider the following data analysis (Croushore & Schmidt, 2010): Students were asked to enter scores to the question: ‘This course fulfilled my expectation’ on a questionnaire with their answers chosen from 5 ordered rankings from ‘Strongly Disagree’ through ’Strongly Agree.’ There were 5 in the #1 ‘Comparison’ category, 8 in the #2 category, etc. These cumulated scores were compared to the previous year’s score (the ‘Reference’ column) where there were 3 in the #1 category, 6 in the #2 category, etc.

Table 1 Student preference score frequencies in each of two years



Preference	#	Comparison	Reference
Strongly Disagree	1	5	3
Disagree	2	8	6
Neither A. nor D.	3	6	6
Agree	4	2	4
Strongly Agree	5	6	8
SUM		27	27

The important aspects of this type of data are:

1. The score categories are ranked in some ascending (or descending) order
2. The rankings are recognized as possibly quite different in ‘distance’ apart

Let V and X denote independent discrete random variables taking values in the set {1, …, K}, drawn respectively from a Reference population V and a Comparison population X. In this example, K=5. Let $q_k = P(V = k), and\ p_k = P(X = k), with\ k\ in\ \{1, \ldots, K\},$ and denote the column vectors $(q_1, \ldots, q_K)'$ and $(p_1, \ldots, p_K)'$ by **q** and **p** respectively. Vectors such as **q** and **p** form probability distributions over {1, …, K}.

p is estimated by dividing the observed frequency, or count, in the comparison population’s cell entries by the total number of X counts (m) for that population; for example, $\hat{p}_1 = \frac{5}{27} = 0.185,$ where ‘^’ indicates an estimated value. Thus $\hat{\boldsymbol{p}}' = \{\hat{p}_1, \ldots, \hat{p}_K\}$. In what follows, we dispense with the ‘^’ notation and refer simply to **p**, mentioning the difference as a need arises. $\hat{\boldsymbol{q}}$ is defined similarly, with the total number of V counts being n. In the current example, both m and n equal 27.

Definition:

The k-th Ridit for the ***reference*** *population V is*

$$r_k = \begin{cases} \dfrac{q_1}{2} \quad for\ k = 1, \\ q_1 + \cdots + q_{k-1} + \dfrac{1}{2} q_k \quad for\ k > 1 \end{cases} \tag{1}$$

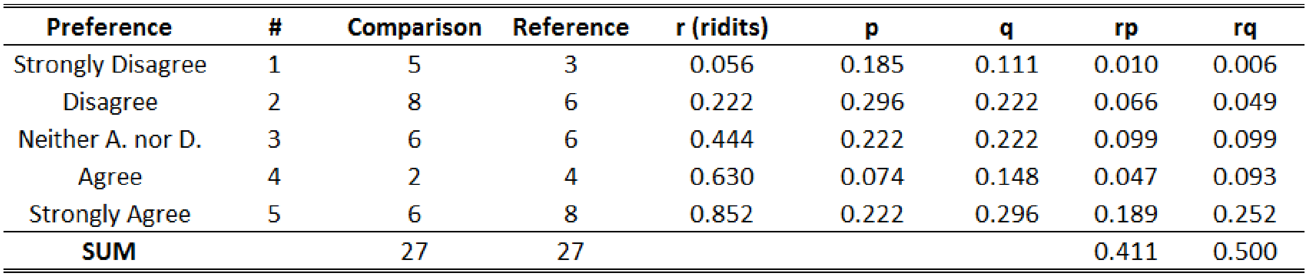
The k-th Ridit for the ***comparison*** *population X is*

$$t_k = \begin{cases} \dfrac{p_1}{2} \quad for\ k = 1, \\ p_1 + \cdots + p_{k-1} + \dfrac{1}{2} p_k \quad for\ k > 1 \end{cases}$$

(2)

Intuitively, a Ridit is akin to the cumulated probability density function of its given population, with a splitting of the k-th category in half. All the quantities necessary for a Ridit analysis are easily computed in a spreadsheet program, as we see in Table 2, which shows Table 1 expanded by estimated values for **p, q, rp** and **rq.**

Table 2 Original data plus Ridit calculations for student scores



Preference	#	Comparison	Reference	r (ridits)	p	q	rp	rq
Strongly Disagree	1	5	3	0.056	0.185	0.111	0.010	0.006
Disagree	2	8	6	0.222	0.296	0.222	0.066	0.049
Neither A. nor D.	3	6	6	0.444	0.222	0.222	0.099	0.099
Agree	4	2	4	0.630	0.074	0.148	0.047	0.093
Strongly Agree	5	6	8	0.852	0.222	0.296	0.189	0.252
SUM		27	27				0.411	0.500

The sum of column **rp**, which is the inner product of vectors **r** and **p** (that is, **r**'**p** in vector notation), equals 0.411. This sum is denoted R(**p**|**q**) and is called *'the mean Ridit of the reference population with respect to the comparison population.*'[vi] That is

$$R(\boldsymbol{p}|\boldsymbol{q}) = \sum_{k=1}^{K} r_k p_k = 0.411$$

(3)

This quantity is an estimate of the expectation of the Ridits of V (that is, the **r**) under the distribution (that is, the **p**) of the Comparison population X, or $E_X(\boldsymbol{r})$.

For any k ≤ K, r_k is the cumulated sum of the known or observed probabilities of the Reference population V (i.e. the $q_{j<k}$) up to and including $\frac{q_k}{2}$; it is thus, intuitively, the probability that a response from the Reference population V is less than the 'middle' of the k-th category.[vii] It can be shown[viii] that

$$R(\boldsymbol{p}|\boldsymbol{q}) = P(V < X) + \frac{1}{2} P(V = X)$$

(4)

Consider an interpretation of the mean Ridit R(**p**|**q**): By (4) it is clear that R(**p**|**q**) is the probability that the reference distribution V lies to the left of the comparison distribution X, with the 'break-even' situation being R(**p**|**q**) = ½ . If R(**p**|**q**) > ½, (i.e. if the probability R(**p**|**q**) is higher than ½), then $\{q_k\}$, the probability mass of V will lie mostly to the left of $\{p_k\}$, the probability mass of X, as equation (1) implies. In Student Evaluation terms, this implies V is closer to the ‘Strongly Disagree’ end of the scale than X. The reverse is true of R(**p**|**q**) < ½, in which case X will be to the ‘left’ of V or, on average, closer to the ‘Strongly Disagree’ end of the scale than V.

$R(\boldsymbol{p}|\boldsymbol{q})$ is thus a proxy for the probability that, on average, an individual drawn at random from the Reference population V is ‘to the left’ of a random individual from the Comparison population; in other words: $R(\boldsymbol{p}|\boldsymbol{q}) \approx$ P($\bar{V} < \bar{X}$). The higher the value of $R(\boldsymbol{p}|\boldsymbol{q})$, the more likely a V-individual will be ‘to the left’ (that is, in this example, to score ‘#1=Strongly Disagree’) compared to an X-individual, and vice-versa.

Since $R(\boldsymbol{q}|\boldsymbol{q}) = \sum_{k=1}^{K} r_k q_k = \boldsymbol{r}'\boldsymbol{q} = \frac{1}{2}\sum_{k=1}^{K} q_k^2 + \sum_{i \neq j} q_i q_j = \frac{1}{2}(q_1 + q_2 + \cdots + q_K)^2$, it follows that $R(\boldsymbol{q}|\boldsymbol{q})=$ 0.5 , and this holds for any **q**. That is, the mean Ridit of the Reference population V with respect to itself is the inner product $\boldsymbol{r}'\boldsymbol{q}$, and always equals $\frac{1}{2}$.

A one-sided null hypothesis of interest in comparing the mean Ridits of the two populations to test if {V is ‘to the left’ of X} is

$$H_0: R(\boldsymbol{p}|\boldsymbol{q}) > \frac{1}{2}$$

(5)

If we reject this hypothesis, by observing that $R(\boldsymbol{p}|\boldsymbol{q})$ is significantly less than ½, we may conclude that the probability that {V is to the left of X} is low, and therefore it is rather the X scores that are ‘to the left’ of V scores on average. A test of the null hypothesis can be executed by forming the statistic Z, where

$$Z = \frac{\bar{\bar{r}} - \frac{1}{2}}{\sqrt{Var(\bar{\bar{r}})}}$$

(6)

and where $\bar{\bar{\boldsymbol{r}}} = R(\hat{\boldsymbol{p}}|\hat{\boldsymbol{q}})$, the estimate of **r** given by entering the observed values of **p** and **q** into R(**p**|**q**) . Z has an approximately Normal (0, 1) distribution for large enough m and n.[ix]

The variance of $\bar{\bar{r}}$ is sometimes[x] taken as 1/(12m); this assumes that q is known, which it seldom is and this variance estimate is better replaced by the more conservative estimate

$$Var(\bar{\bar{r}}) = \frac{1}{12m} + \frac{1}{12n}$$

(7)

which is a variance formula given by Selvin (1977). Doing the Z-test for the Student Scores example with these formulas gives

$$Z = \frac{0.411 - 0.5}{\sqrt{\frac{1}{12x27} + \frac{1}{12x27}}}$$

(8)

So Z = - 0.089/0.0785 = -1.132. The critical Normal (0, 1) left-tail Z value for a one-sided hypothesis test at the 95% level, $Z_{0.05} = -1.645$, so the test result is: ‘No significant difference detected’ between this year’s and last year’s student scores.[xi] That is, there is no evidence to suspect that X is ‘to the left’ of V.

The example above served to introduce Ridit definitions and the usual mean-Ridit test for the factor ‘Student Year’ being presented at two levels, namely: Current year and Previous year. We now we examine an example where the input factor is given over several levels. In addition, we examine how we might construct confidence intervals for the Ridit means, and how the fact of small data samples may affect the confidence intervals for these means.

2.2 Example 2: Analysis of Borg-scale Fatigue levels over several stages of G

Five pilots were assigned to fly several repeated sorties at increasing G (gravitational stress) levels and their ‘Fatigue’ was measured by responses scored by an adjusted Borg-scale measure. The standard Borg scale measures physiological exertion expressed on a range of 6 to 20, with 6 being ‘no exertion at all.’ The adjusted scale used in the present example went from 0 through 10, with 0 being ‘no exertion at all.’ This adjusted scale is seen to be very similar to the Cooper-Harper scale. These scores are plotted in Figure 1. Note that G1.5 was set at slightly above stationary, ground-level G. G85 and G95 refer to repeated maneuvers at G8 and G9 respectively. No observed score exceeded level 5.

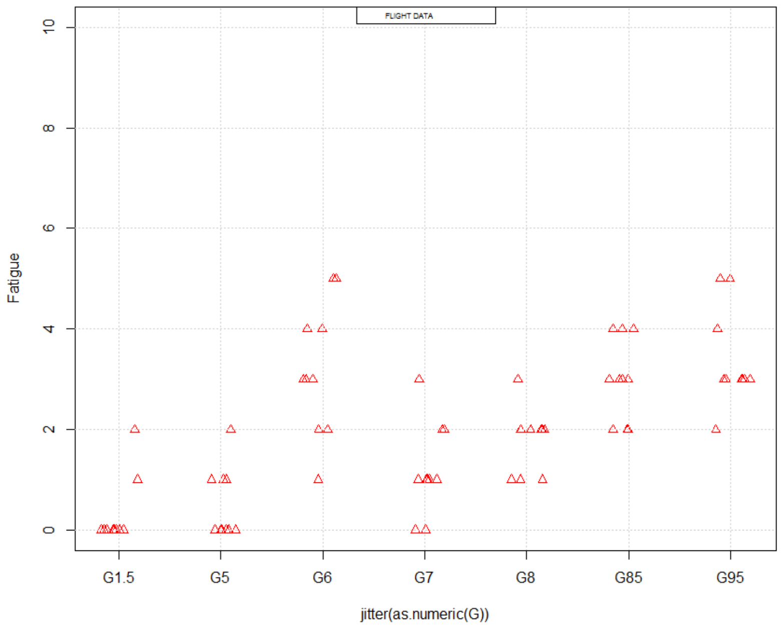


Figure 1 Adjusted Borg scores for 'Fatigue' vs. increasing G levels

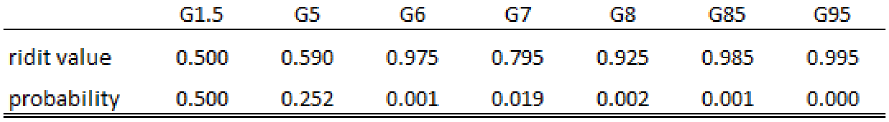
The counts of recorded Borg scores for each G level were entered into Table 3. For example, in column G1.5, there were 8 scores of '0', one score of '1', etc. This layout enables the Ridit calculations to be easily done.

Table 3 Adjusted Borg scores given by pilots flying at increasing G levels

Score	G1.5	G5	G6	G7	G8	G85	G95	Reference
0	8	6	0	2	0	0	0	8
1	1	3	1	5	3	0	0	1
2	1	1	2	2	6	3	1	1
3	0	0	3	1	1	4	6	0
4	0	0	2	0	0	3	1	0
5	0	0	2	0	0	0	2	0
SUM	10	10	10	10	10	10	10	10

Taking G1.5 as a fixed baseline, and comparing the fatigue scores for increasing G levels, the 95% Bonferroni-adjusted t-value (lower-tail) percentile for k=6 comparisons against G1.5 as the reference distribution (d.f. =18, and using 1-α/2k) is 0.004; thus we see that by these t-tests that G6, G8, G85 and G95 are significantly different to G1.5. The complete results are given in Table 4.

Table 4 Ridits for the Borg Fatigue scores vs. increasing G levels, and the probability that each value is greater than the G1.5 reference level by Bonferroni-adjusted t-value



	G1.5	G5	G6	G7	G8	G85	G95
ridit value	0.500	0.590	0.975	0.795	0.925	0.985	0.995
probability	0.500	0.252	0.001	0.019	0.002	0.001	0.000

95% confidence intervals at each G-value might be constructed using these Ridit values and the adjusted t-value, as shown in Figure 2.

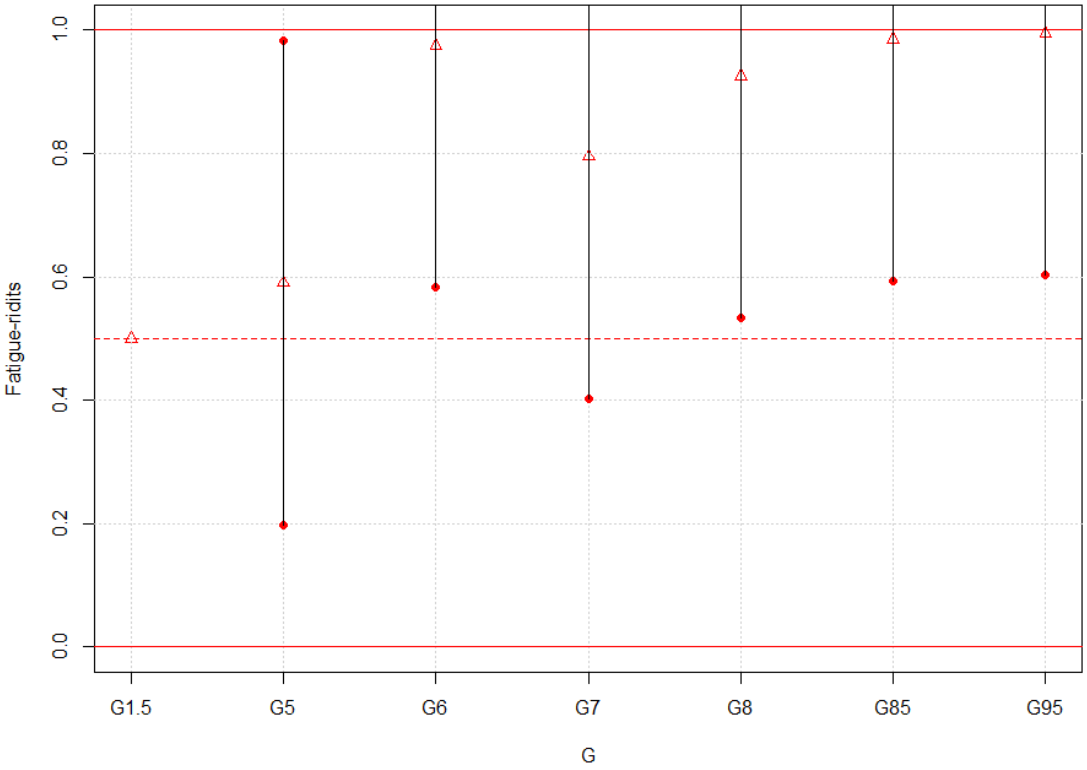


Figure 2 Plot of Ridits and their 95% Bonferroni-adjusted confidence intervals for the Borg Fatigue scores vs. G

The figure bears out the conclusions of Table 4 except for an overlap of the 0.5 line at G7. As will also be noted, some of the confidence intervals overlap the endpoints of the (0, 1) interval to which Ridits are constrained, and the graph of this analysis should be taken as an approximate indication of significant fatigue-level differences.

3.1 A Distance-Based Approach

So far we have shown that standard Ridit analysis applies quite well to inferences on the differences of means even in the case of small samples. However, the problem of incorrect confidence intervals is a problem that needs be addressed. One solution to this problem is to take a transform of the Ridits that might better approximate a Normal distribution for the small-sample case. Such an approach is discussed in Hurwitz (2015); the transform that was taken was the logistic transform, and an assumption was made that this gave a Normally-distributed situation that could be used in making inferences about Ridit means and their confidence intervals. This solved the problem of inappropriate confidence bounds as evinced in Figure 2 above. However, it is not always certain that the logit transform—or any transform—will give an adequate approximation to Normality. In the following discussion we will take a different approach to the problem.

Consider the problem of comparing the 'distance' between any two discrete probability distributions $\{p_i\}$ and $\{q_i\}$ defined over a common domain. One such measure is the discrete-probability-distribution version of the (squared) 'Hellinger Distance' (Yang & Le Cam, 2000)

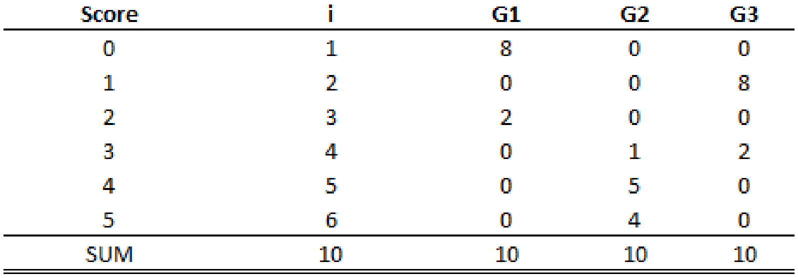
$$H^2(p,q) = 1 - BC(p,q) \tag{9}$$

where BC(p, q) is the 'Bhattacharyya Coefficient' (Bhattacharyya, 1943)

$$\text{BC(p, q)} = \sum_{all\ i} \sqrt{p_i q_i}\ . \tag{10}$$

The maximum Hellinger distance 1 is achieved when p_i assigns probability zero to every set to which q_i assigns a positive probability, and vice versa. In Table 5 we show the consequence of using H^2 to gauge the distance between discrete distributions of fatigue ratings.

Table 5 Three Hypothetical Discrete OCD Fatigue Distributions

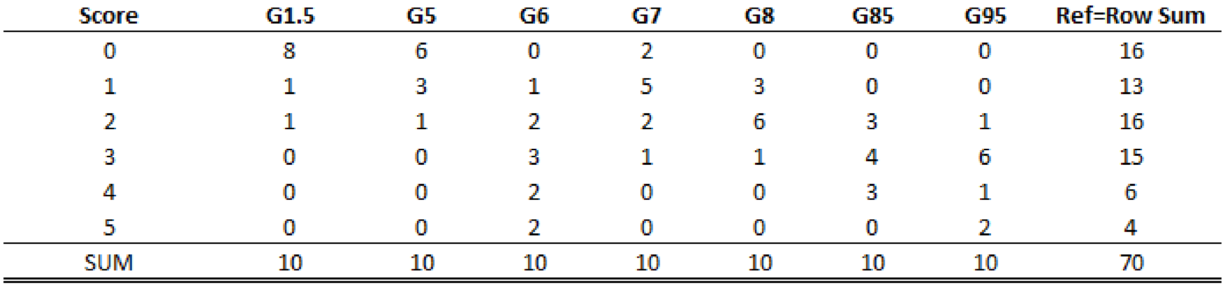


Score	i	G1	G2	G3
0	1	8	0	0
1	2	0	0	8
2	3	2	0	0
3	4	0	1	2
4	5	0	5	0
5	6	0	4	0
SUM	10	10	10	10

It is clear what will happen to H^2 when the ratings are turned into their corresponding probability values – column G1 vs. G2 will have H^2 = 0, as we'd expect (as the distributions are 'far apart,') but columns G1 vs. G3 will also have H^2 = 0, as we do not expect between two distributions that are 'quite close together.'

A solution here is to fit continuous distributions over the discrete ones we observe, and then to apply the continuous-distribution version of H^2 to these instead. First, however, we need to decide on how to construct the 'common domain' of Ridits (r_i) on which the definition of H^2 will depend. In our previous 'standard' Ridit treatment, as can be seen in Table 3 above, we took one column—namely G1.5—as the 'reference distribution' and compared the other columns to it. This has the advantage of carrying the idea of independent means through to our inferences. However, the 'domain' that we implicitly use is then restricted to those three rows of Table 3 where the q_i corresponding to G1.5 are non-zero, namely rows 1, 2, and 3. The remaining three rows then have Ridits equal to 1.0, as those are the cumulated value of the probabilities for the G1.5 observations. A more satisfactory construct for a domain would be to have Ridit values distributed more evenly across the [0, 1] range, and this can be achieved by making the marginal *row sums* the new reference distribution, as shown in Table 6.

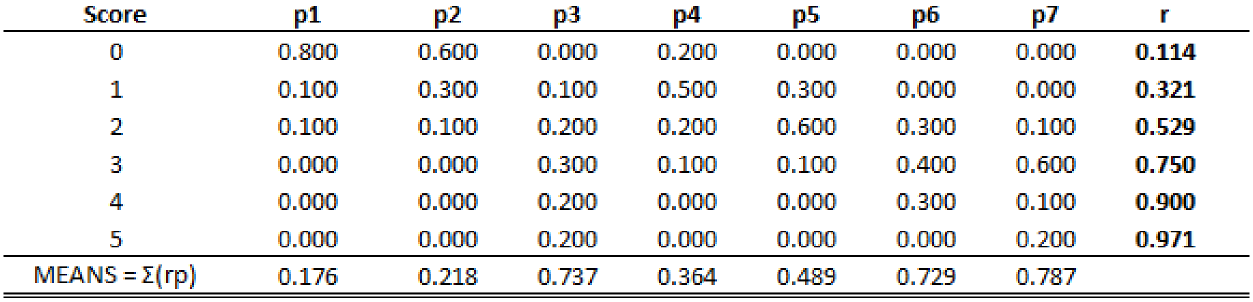
Table 6 Ridit Reference Distribution based on Row Sums



Score	G1.5	G5	G6	G7	G8	G85	G95	Ref=Row Sum
0	8	6	0	2	0	0	0	16
1	1	3	1	5	3	0	0	13
2	1	1	2	2	6	3	1	16
3	0	0	3	1	1	4	6	15
4	0	0	2	0	0	3	1	6
5	0	0	2	0	0	0	2	4
SUM	10	10	10	10	10	10	10	70

Table 7 shows the corresponding probability computations for the seven columns of observed Fatigue scores. These will be the same as those computed for Table 3. The difference here is that the Ridit column '**r**' is now based on the marginal row sums rather than just on the G1.5 column. The domain of our probabilities is now more evenly spread across [0, 1]. The mean Ridits are also shown; formula is the usual $\bar{\boldsymbol{x}} = \mathbf{rp}$ = $\sum_i r_i p_i$, one mean for each column.

Table 7 Prob. Distributions of Fatigue Scores, with Ridits (r) based on Row Sums. Probability means over 'r' are shown on last line.



Score	p1	p2	p3	p4	p5	p6	p7	r
0	0.800	0.600	0.000	0.200	0.000	0.000	0.000	**0.114**
1	0.100	0.300	0.100	0.500	0.300	0.000	0.000	**0.321**
2	0.100	0.100	0.200	0.200	0.600	0.300	0.100	**0.529**
3	0.000	0.000	0.300	0.100	0.100	0.400	0.600	**0.750**
4	0.000	0.000	0.200	0.000	0.000	0.300	0.100	**0.900**
5	0.000	0.000	0.200	0.000	0.000	0.000	0.200	**0.971**
MEANS = Σ(rp)	0.176	0.218	0.737	0.364	0.489	0.729	0.787	

Figure 3 illustrates the shapes of the seven observed probability distributions, given as vertical lines across the domain of the **r**'s, with a line connecting the tops of each vertical line.

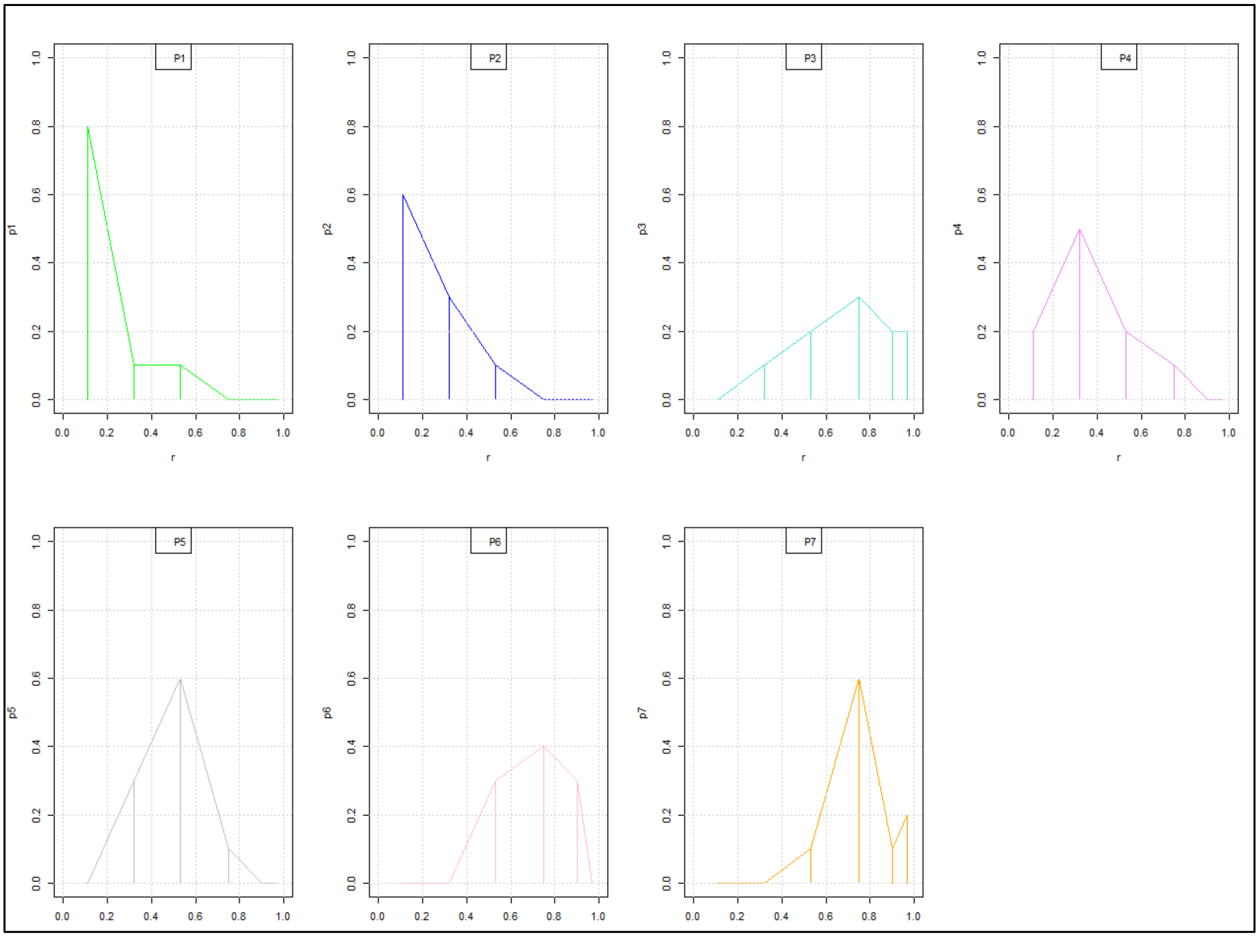


Figure 3   Observed Probability Distributions of Fatigue Scores

The next step is to fit appropriate continuous distributions to the observed probabilities. A flexible choice for this is the Beta(α, β) family of distributions which have support (i.e. domain) over (0, 1) and shapes similar to those of the observed (discrete) distributions. We obtain estimates of the required seven (α, β) pairs via the method of moments formulas for the Beta distribution:

$$\hat{\propto} \;= \bar{x}\left(\frac{\bar{x}(1-\bar{x})}{\bar{v}} - 1\right), \text{ if } \;\bar{v} < \bar{x}(1-\bar{x}) \quad \text{where the } \bar{x}'s \text{ are given by the } \Sigma rp\ 's \text{ in Table 7} \tag{11}$$

$$\hat{\beta} \;= (1-\bar{x})\left(\frac{\bar{x}(1-\bar{x})}{\bar{v}} - 1\right), \text{ if } \;\bar{v} < \bar{x}(1-\bar{x}), \text{ and } \bar{v} \text{ is an estimated variance.} \tag{12}$$

We already have the estimates for the $\bar{x}$'s. The estimates for the $\bar{v}'s$ are derived by following the formulas for a weighted variance (our observations are given in weighted form, the weights being the probabilities p_i). The weighted variance formula, with weights $w_i, \sum w_i = 1$,

$$\bar{v} = \sum w_i (x_i - \bar{x})^2 \tag{13}$$

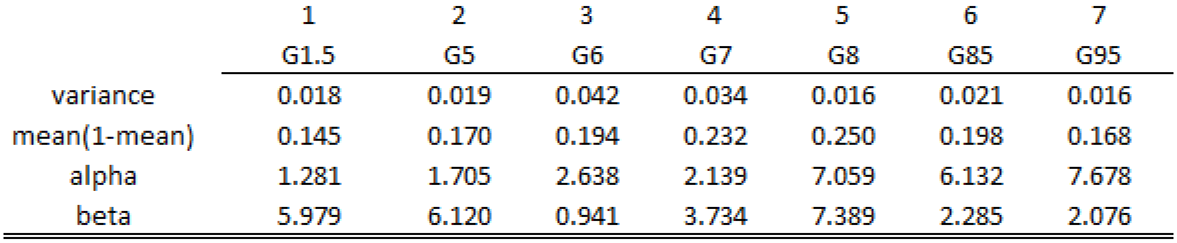
This translates, in our case, to

$$\bar{v} = \sum p_i (r_i - \bar{x})^2. \tag{14}$$

$\bar{x}$ is the 'rp' mean for that column as given in Table 7.`

One variance is computed for each column of $p_i's$. The results are shown in Table 8. The first two rows are for the check that $\bar{v} < \bar{x}(1-\bar{x})$; all instances pass this check.

Table 8 Variance, check, and Beta Distribution parameters by Method of Moments



	1	2	3	4	5	6	7
	G1.5	G5	G6	G7	G8	G85	G95
variance	0.018	0.019	0.042	0.034	0.016	0.021	0.016
mean(1-mean)	0.145	0.170	0.194	0.232	0.250	0.198	0.168
alpha	1.281	1.705	2.638	2.139	7.059	6.132	7.678
beta	5.979	6.120	0.941	3.734	7.389	2.285	2.076

Now we are in a position to construct fitted Beta distributions—one for each column of the Fatigue scores—against the observed probability histogram. Figure 4 shows cumulated distribution fits for all seven distributions.

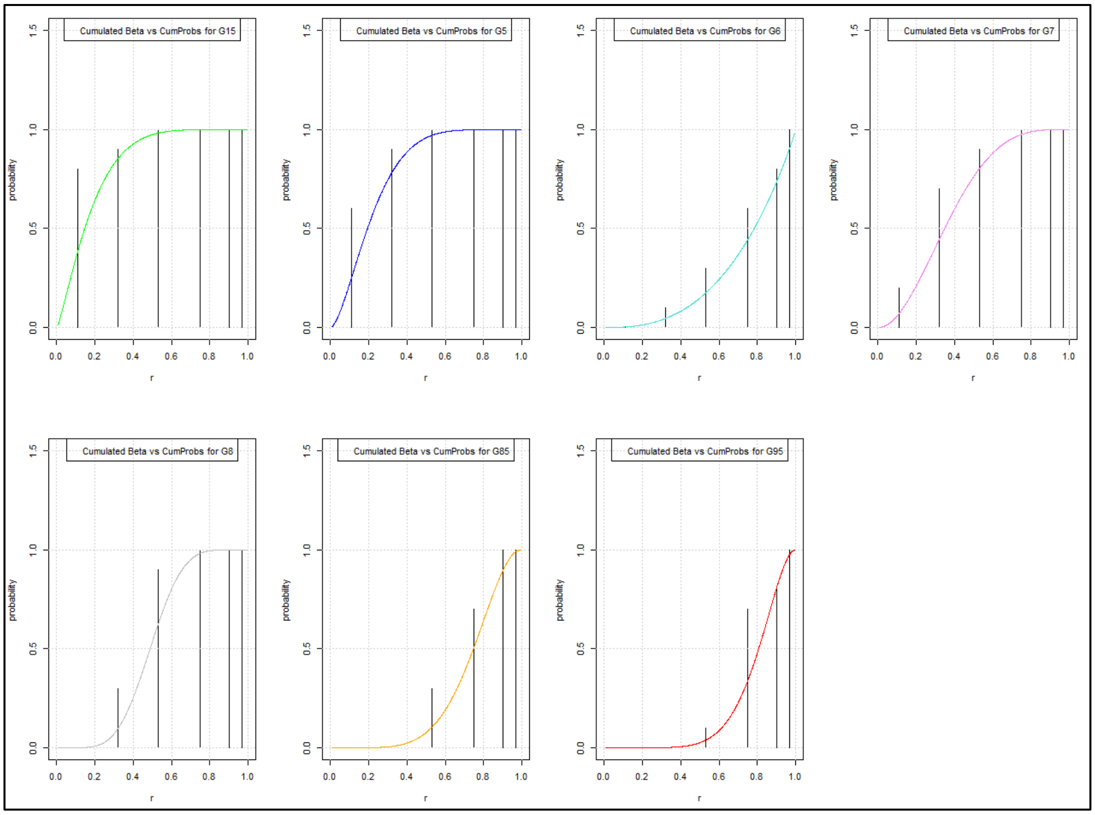


Figure 4 Cumulated Beta (fitted) vs (observed) Cumulated Probabilities

As can be seen, the fitted continuous Beta distributions are reasonable approximations to the observed discrete probability distributions. This will form the basis of our inferences.

We now compute the squared Hellinger distance between the continuous G1.5 and G8 distributions, and use it as our metric for 'distance apart.' The formula for H^2 given for continuous distributions is

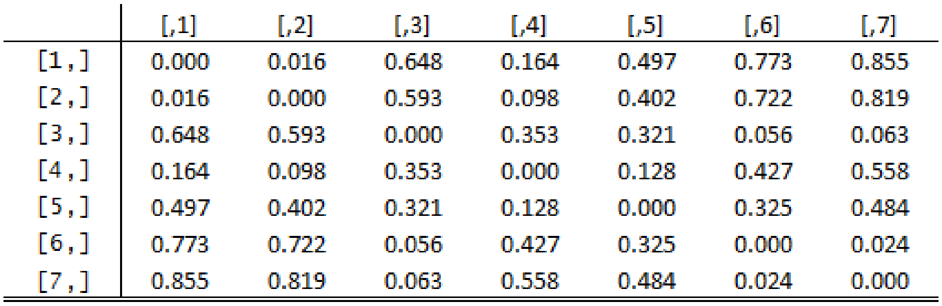
$$H^2 = 1 - \frac{B(\frac{\propto_1 + \propto_2}{2}, \frac{\beta_1 + \beta_2}{2})}{\sqrt{B(\alpha_1, \beta_1) B(\alpha_2, \beta_2)}}$$

(15)

where 'B' is the Beta function.

Computing all the H^2 distances gives a symmetric matrix with 0's on the diagonal; the 0's show the distance of a distribution from itself, and the matrix is symmetric since the distance is symmetric either way. This is shown on Table 9.

Table 9 Squared Hellinger Distances between all seven Fatigue distributions

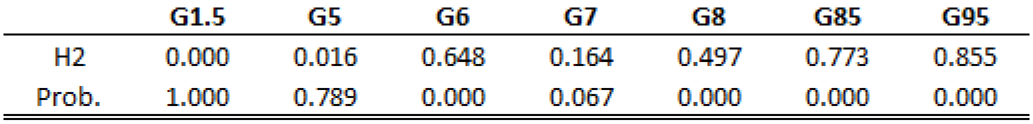


	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]
[1,]	0.000	0.016	0.648	0.164	0.497	0.773	0.855
[2,]	0.016	0.000	0.593	0.098	0.402	0.722	0.819
[3,]	0.648	0.593	0.000	0.353	0.321	0.056	0.063
[4,]	0.164	0.098	0.353	0.000	0.128	0.427	0.558
[5,]	0.497	0.402	0.321	0.128	0.000	0.325	0.484
[6,]	0.773	0.722	0.056	0.427	0.325	0.000	0.024
[7,]	0.855	0.819	0.063	0.558	0.484	0.024	0.000

The distance between distribution '1' (G1.5), and distribution '2' (G5) is shown as 0.16. Could this distance have happened by chance? If one were dealing with independent Normal distributions, one could use a t-test to answer this question. In our present case, we can answer the question using a randomization trial as follows:

Set the null hypothesis as Ho: G1.5 = G5. The alternative is 'G1.5 ≠ G5'. Assume the null to be true. Draw a random sample size 10 (recall, 10 was our original Fatigue sample size for G1.5). Draw a second random sample from G1.5 as well. Compute H^2 for this sample pair. Repeat (say) 10,000 times and collect the 10,000 values of H^2. Now ask: 'What proportion of the 10,000 distances are greater than or equal to 0.16?' This can be computed as a simple ratio from the H^2 data, and gives the probability 'p' that the null hypothesis is true. Doing this, we obtain: p = 0.789. This is a high probability, so we conclude that Ho, the null hypothesis, is true. G1.5 and G5 are too close to tell apart, so we have no evidence that they are different distributions. Taking the first row of H^2 distances and doing the same using two random samples, *both from G1.5,* gives the probabilities of Table 10.

Table 10 H^2 distances, and the probability that each distribution = G1.5



	G1.5	G5	G6	G7	G8	G85	G95
H2	0.000	0.016	0.648	0.164	0.497	0.773	0.855
Prob.	1.000	0.789	0.000	0.067	0.000	0.000	0.000

The results in Table 10 are in general agreement with the standard Ridit results given in Table 4. G1.5 is can be taken as equivalent to G5 and G7, but the other distributions are different to G1.5.

3.2 Confidence intervals for the differences in means

We have, so far in Section 3, examined the probabilities that two observed *distributions* of OCD data are the same or different and we have done so using the squared Hellinger distance. We could continue and develop confidence intervals around the squared Hellinger distances, but the results would not be easily interpretable in Engineering terms. We will, instead, take a more intuitive approach and compute confidence intervals around the Ridit (***rp***) means, and we do this by using the fitted Beta distributions. Our (randomization-based) method is:

1. Draw two samples, each size n=10, at random from each of two Beta distributions

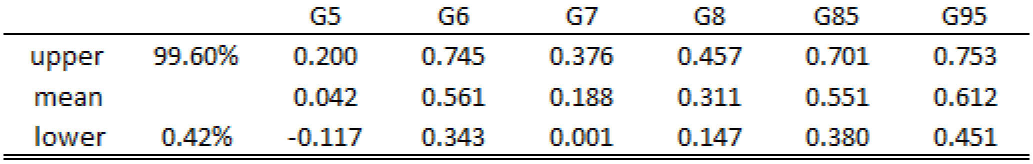
2. Compute the difference between the means

3. Do this 10,000 times

4. Compute a 100(1-alpha/2KK) CI based on proportions where KK=number of comparisons we will make (and gives the Bonferroni correction).

This method, comparing G1.5 mean to all other means, gives the Confidence intervals in Table 11.

Table 11 Bonferroni-adjusted CI's on mean differences: G1.5 vs other means



		G5	G6	G7	G8	G85	G95
upper	99.60%	0.200	0.745	0.376	0.457	0.701	0.753
mean		0.042	0.561	0.188	0.311	0.551	0.612
lower	0.42%	-0.117	0.343	0.001	0.147	0.380	0.451

Figure 5 give a graphical example of what we have done for the case of the difference between the means for G1.5 and G95.

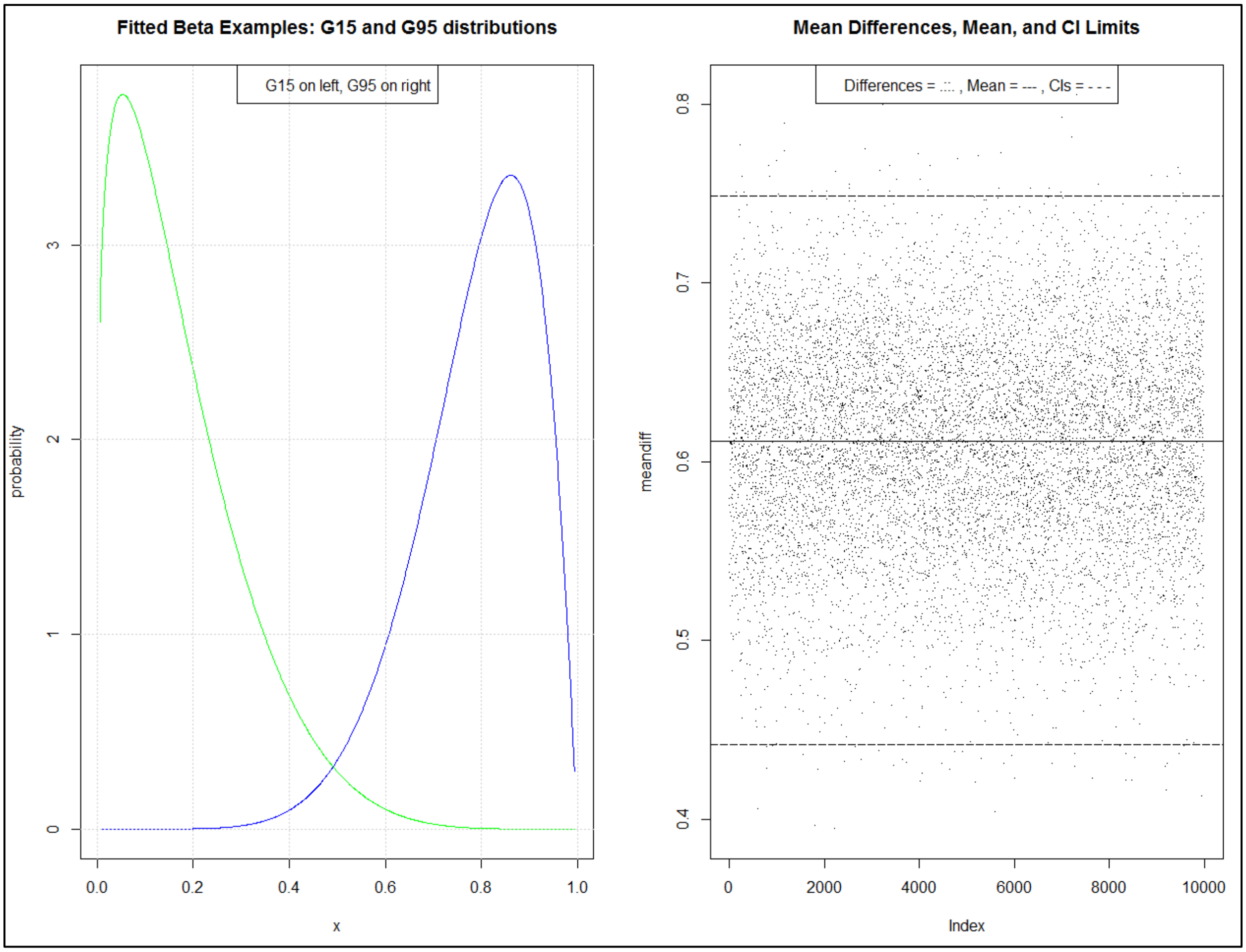


Figure 5 Fitted Beta distributions for G1.5 and G95; Mean, Differences & C.I.'s

Table 10 shows that, with reference to G1.5, the distributions of G5 and G7 are either no different or close to no different. Table 11 bears this out, with the Ridit confidence intervals for G5 and G7 either including zero or close to including zero. For the other Ridit means difference to G1.5, Table 11 gives, for example, a CI for the G6 difference of [0.343, 0.745] around a ridit mean of 0.561. The distribution of OCD scores (and fitted Beta distribution) for G6 lies to the right of that for G1.5, and these results indicate that the mean Ridit value for G6 is above that for G1.5 by, on average, 0.561. This says that we can state, at a 95% level of confidence, that the probability distribution based on the G6 OCD results has an average that is 0.561 higher than that for the OCD results taken at G1.5: G6 gives significantly greater Fatigue scores than G1.5.

Conclusion

We have developed a new method for comparing the results of OCD data distributions that is not based on the standard large-sample Ridit analysis methods. We have used a distance-based metric and randomization tests to give inferences on distributions, and for confidence intervals on mean differences. The method used in our construction induces some dependence between the means, and this item needs to be further investigated. However, the results presented here are in line with independent-mean results derived earlier and our new method, we believe, gives a path to a better analysis of small sample OCD especially as no Normal-distribution assumptions need be made, and the confidence intervals so derived do not violate the bounds of the probability limits.

Appendix A

A1:

$\boldsymbol{R}(\boldsymbol{p}|\boldsymbol{q}) = P(V < X) + \frac{1}{2}P(V = X).$

Proof of A1:

If drawings from V and X are independent, and for any k ≤ K, then the proof follows from:

$r_k p_k = (q_1 + \cdots + q_{k-1} + \frac{1}{2}q_k\,)p_k$, that is

$r_k p_k$ = P ({V=1}$\cap\{X = k\}) + \cdots + P(\{V = k - 1\} \cap \{X = k\}) + \frac{1}{2}\,P(\{V = k\} \cap \{X = k\})$.

Acknowledgements

The authors would like to thank the Statistical Methods Group at Edwards AFB for giving us the time to pursue this research, to the Engineering community at EAFB for providing feedback on earlier versions of this paper, to the Edwards Test Pilots' School for giving permission to use some of their Fatigue Study data, to Technical Publications for their timely authorizations, and to the Technical Library at EAFB for always giving their best support in finding the several publications we found necessary to pursue this investigation.

References

1. Agresti, A. (1984). *Analysis of Ordinal Categorical Data*. New York: Wiley & Sons

2. Agresti, A. (1996). *Categorical Data Analysis*. New York: Wiley & Sons

3. Bhattacharyya, A. (1943). *On a measure of divergence between two statistical populations defined by their probability distributions*. Bulletin of the Calcutta Mathematical Society: 99–109.

4. Bauer, D. F. (1972). Constructing Confidence Sets Using Rank Statistics. *Journal of the American statistical Association*, vol. 67, No. 339. 687-690.

5. Beder J.H. and Heim, R.C. (1990). On the use of Ridit Analysis. *Psychometrika*, vol. 55, No. 4. 603-616.

6. Borg, G. (1970). Perceived exertion as an indicator of somatic stress. *Scandinavian journal of rehabilitation medicine* **2** (2): 92–98.

7. Box, G.E.P., Hunter, W.G., and Hunter, J.S. (1978). *Statistics for Experimenters*. Wiley, New York.

8. Bross, I.D.J. (1958). How to use Ridit analysis. *Biometrics*, 14. 18-38.

9. Cooper, G. and Harper, R. (1969). The use of pilot rating in the evaluation of aircraft handling qualities. *NASA Technical Report* TN D-5153

10. Croushore, D. & Schmidt, R.M. (2010). Ridit analysis of student score evaluations. *Robins School of Business, Univ. of Richmond*, VA 23173 (USA).

11. Dunn, Olive Jean. (1961). Multiple Comparison Among Means. *J. of the American Statistical Association*, vol. 56, No. 293. 52-64

12. Harper, R. and Cooper, G. (1986). Handling Qualities and Pilot Evaluation. *AIAA, J. of Guidance Control and Dynamics*, vol. 9, No. 5, pp. 515-529

13. Hollander, M. and Woolf, D.A. (1973). *Nonparametric Statistical Methods*. Wiley & Sons. New York

14. Hurwitz, A. M. (2015). *Ridit Analysis & Confidence Intervals for Borg Scale, Cooper-Harper & Other Ordinal Ratings*. Proceedings of the Society of Flight Test Engineers. https:// www.dropbox.com/sh/vx1sa54rw8cw4ga/AAD-ZGPKR0JVbmODG7KIZIRa?dl=0

15. Jansen, M.E. (1984). "Ridit analysis, a review." *Statistica Neerlandica*, 38, pp. 141-158

16. R Core Team. (2013). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria.

17. Selvin, S. (1977). A further note on the interpretation of Ridit analysis. *American Journal of Epidemiology*, 105. 16-20.

18. Selvin, S. (2004) *Statistical Analysis of Epidemiologic Data*, 3rd ed. OUP. 176-177

19. Wilson, D.J. & Riley, D.R. (1989). Cooper-Harper Pilot Rating Variability. American Institute of Aeronautics & Astronautics. (© McDonnell Douglas Aircraft Corp., St. Louis, MO 63166). *AIAA, 16th Atmospheric Flight Mechanics Conference, Boston, MA, USA.*

20. Wilson, D.J. & Riley, D.R. (1990). More on Cooper-Harper Pilot Rating Variability. (© McDonnell Douglas Aircraft Corp., St. Louis, MO). *AIAA, Atmospheric Flight Mechanics Conference, Portland, OR*. August 20-22, 1990.

21. Yang, Grace Lo & Le Cam, Lucien M. (2000). *Asymptotics in Statistics: Some Basic Concepts*. Berlin: Springer.

Endnotes

[i] All flying qualities data in this paper is synthetic, and has been simulated to illustrate Ridit analysis

[ii] A general reference is Agresti, 1996

[iii] For a discussion on measurement taxonomy, see http://en.wikipedia.org/wiki/Level_of_measurement

[iv] Measurements which are comparable to each other in terms of size or distance apart.

[v] The terminology and notation given in Beder & Heim (1990) is followed closely

[vi] Beder & Heim, 1990, reverse this wording and give: '*the mean Ridit of the comparison population with respect to the reference population'*; our wording, however, is more in line with the actual construct of $R(\boldsymbol{p}|\boldsymbol{q})$.

[vii] The 'middle' or the 'median' of a category is not an exact term as a category is ordinal, not ratio-scale

[viii] See Appendix A, result A1

[ix] Beder & Heim, 1990, formula (17).

[x] Bross (1958)

[xi] Note that in Croushore & Schmidt (2010), the variance was taken as 1/(12m), so estimated standard error in that paper is 0.056

Ridit Analysis for Borg-Scale & Cooper-Harper Ratings: A Distance-Based Approach for Small Samples

Arnon Hurwitz, PhD.

STATISTICAL METHODS OFFICE

EDWARDS AFB, EDWARDS, CA

arnon.hurwitz@us.af.mil

661-527-4809

Approved for public release; distribution is unlimited.

412TW-PA-16437

Overview

- Ridit method/example – Course Rating by Students
  - Basic Method & Notation

- Ridit example – Borg Scores for Fatigue Levels
  - Means & Confidence Intervals using standard method
  - Distribution comparisons using a distance-based method
  - Confidence Intervals using randomization

Ridit Analysis

- Consider a simple example: 27 students are asked to answer ‘Course was good?’ from #1 (Strongly Disagree) to #5 (Strongly Agree)

This year’s scores

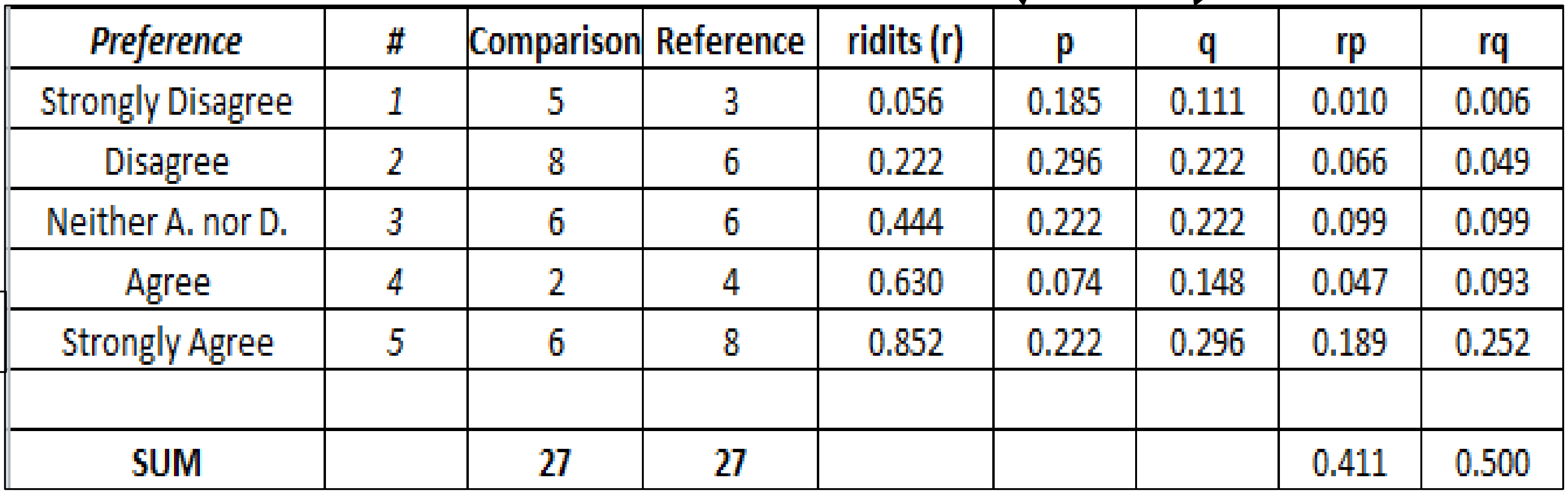
Last year’s scores

This year’s proportions

Last year’s proportions

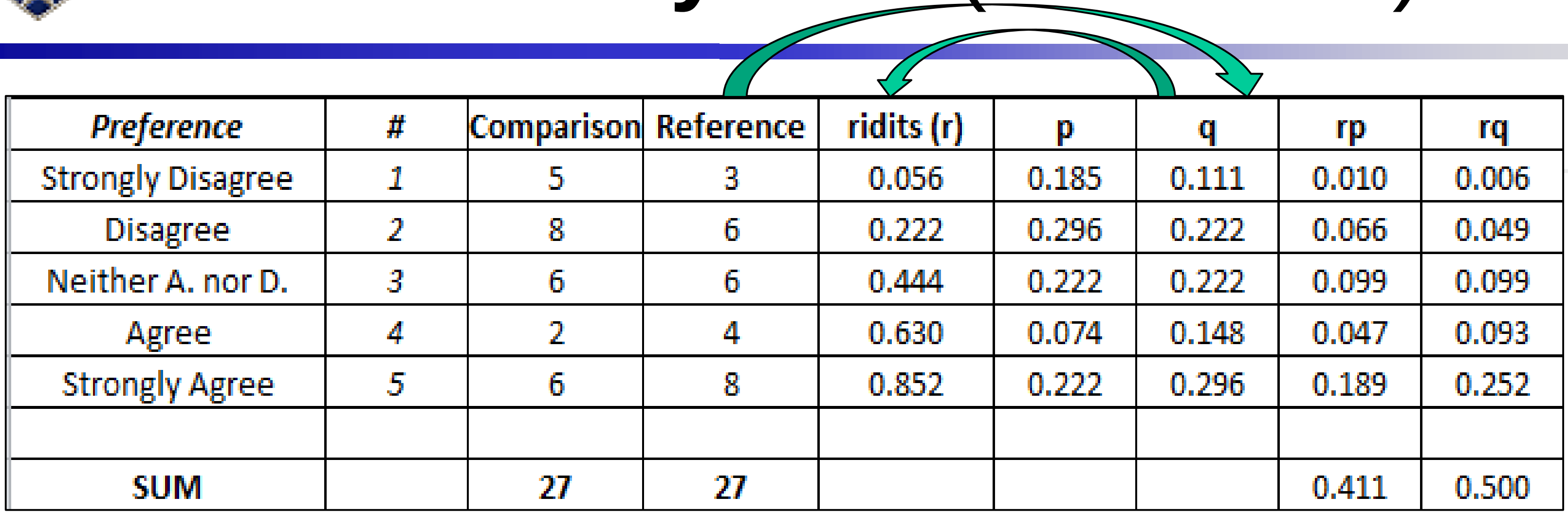
Bad

Good



Preference	#	Comparison	Reference	ridits (r)	p	q	rp	rq
Strongly Disagree	1	5	3	0.056	0.185	0.111	0.010	0.006
Disagree	2	8	6	0.222	0.296	0.222	0.066	0.049
Neither A. nor D.	3	6	6	0.444	0.222	0.222	0.099	0.099
Agree	4	2	4	0.630	0.074	0.148	0.047	0.093
Strongly Agree	5	6	8	0.852	0.222	0.296	0.189	0.252
SUM		**27**	**27**				0.411	0.500

Ridit Analysis – (continued)

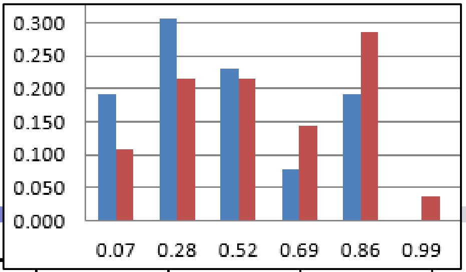


Preference	#	Comparison	Reference	ridits (r)	p	q	rp	rq
Strongly Disagree	*1*	5	3	0.056	0.185	0.111	0.010	0.006
Disagree	*2*	8	6	0.222	0.296	0.222	0.066	0.049
Neither A. nor D.	*3*	6	6	0.444	0.222	0.222	0.099	0.099
Agree	*4*	2	4	0.630	0.074	0.148	0.047	0.093
Strongly Agree	*5*	6	8	0.852	0.222	0.296	0.189	0.252
SUM		**27**	**27**				0.411	0.500

- Proportions **p** and **q** (i.e. estimated probabilities) are computed from the data. E.g. 0.185 = 5/27, etc.
- A population (Last year's) is set as the 'Reference'
- The *k-th ridit of the Ref. population* is defined as:

$$r_k = \begin{cases} \dfrac{q_1}{2} \quad for\ k = 1, \\ q_1 + \cdots + q_{k-1} + \frac{1}{2} q_k \quad for\ k > 1 \end{cases}$$

Ridit Analysis



To the left ↑

To the rt. ↓

Preference	#	Comparison	Reference	ridits (r)	p	q	rp	rq
Strongly Disagree	*1*	5	3	0.056	0.185	0.111	0.010	0.006
Disagree	*2*	8	6	0.222	0.296	0.222	0.066	0.049
Neither A. nor D.	*3*	6	6	0.444	0.222	0.222	0.099	0.099
Agree	*4*	2	4	0.630	0.074	0.148	0.047	0.093
Strongly Agree	*5*	6	8	0.852	0.222	0.296	0.189	0.252
SUM		**27**	**27**				0.411	0.500

- Form columns **rp** and **rq**, and sum (∑) each one
- ∑ **rp** = 0.411 is the probability that the **Reference** pop. will be '**to the left**' of the Comparison pop.
  - If the p's are 'bunched' to the right versus the q's, then ∑ **rq** < ∑ **rp**
  - that is, high ∑ **rp** ⟹ **Reference** pop. (q's) is bunched '**to the left**' of p's
  - that is, high ∑ **rp** ⟹ **Reference** pop. (last year) was **worse** than this year
- Our HYPOTHESIS is that ∑ **rp** ≥ 0.5 What does this mean?
  - If true, then last year's (**Reference**) scores are **worse** than this year's
  - However, it's obvious that ∑ **rp** = 0.411 ≤ 0.5 - So was last year better?
  - Can only say this *if experimental error = 0* → We need a statistical test!

Ridit Analysis – Hypothesis Test

- ‘Experimental error’ means that, if the underlying situation stays the same, but we draw a new sample, the numbers (p’s and q’s) we see will be somewhat different. So conclusions might change

- To test Ho: ∑ **rp** ≥ ∑ **rq = 0.5** , form **t** = (∑ **rp** – **0.5**) / $\sqrt{\left[\frac{1}{12m}+\frac{1}{12n}+\frac{1}{12mn}\right]}$

  m = n = 27. So t = (0.411- 0.5)/sqrt(0.0063)= -1.12, with d.f.= m+n-2 = 52

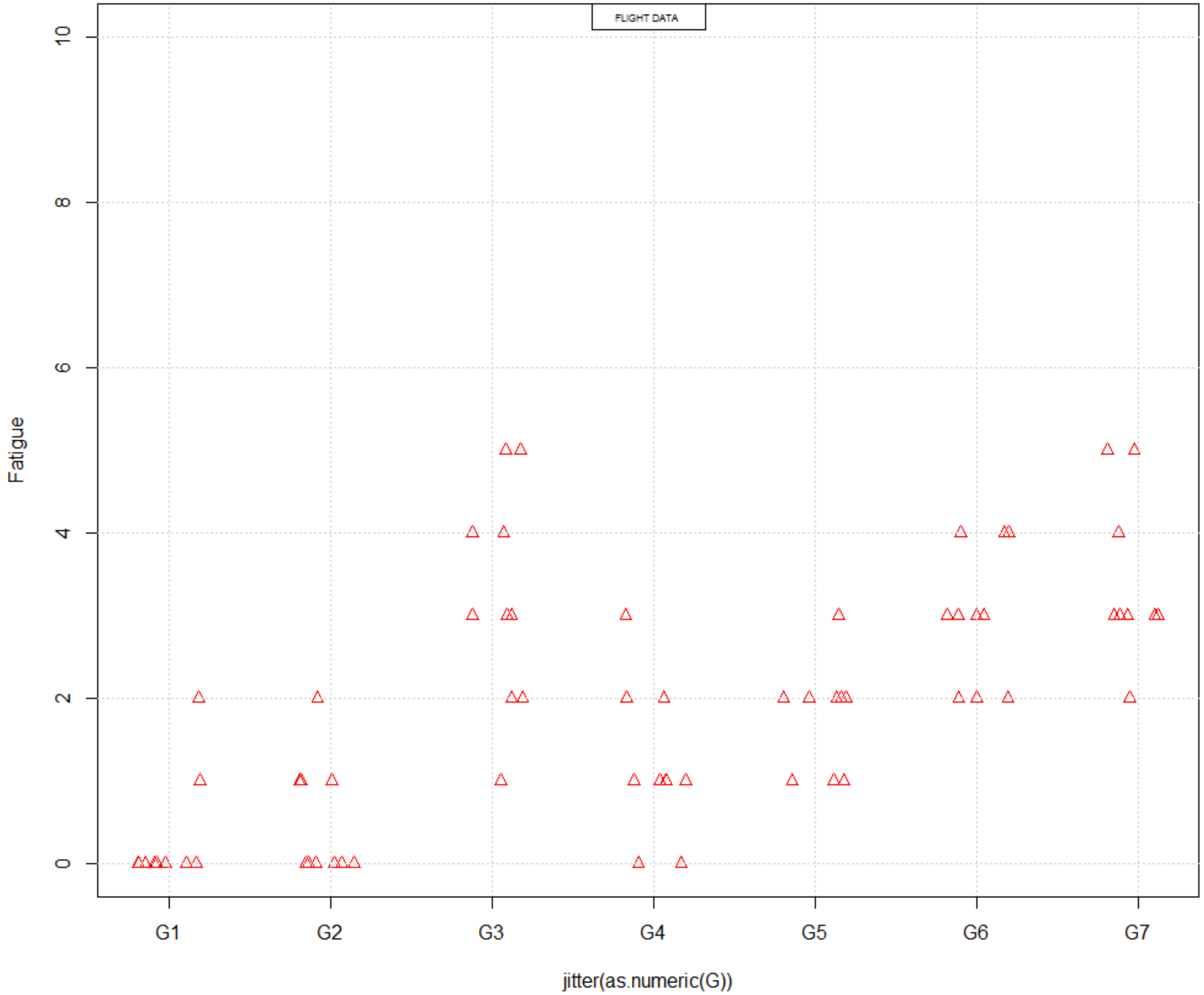
- Left-tail, critical t (at 95% confidence, d.f.=52) = -1.675, so do not reject Ho → We cannot say that this year’s scores are any better than last year’s

- NOTE: If we had another distribution (e.g. **p**-scores from another school, $\mathbf{p^o}$), we could test Ho: ∑ **rp** ≠∑ $\mathbf{rp^o}$ using **q** as ref., and var = $\sqrt{\left[\frac{1}{12m}+\frac{1}{12n}\right]}$

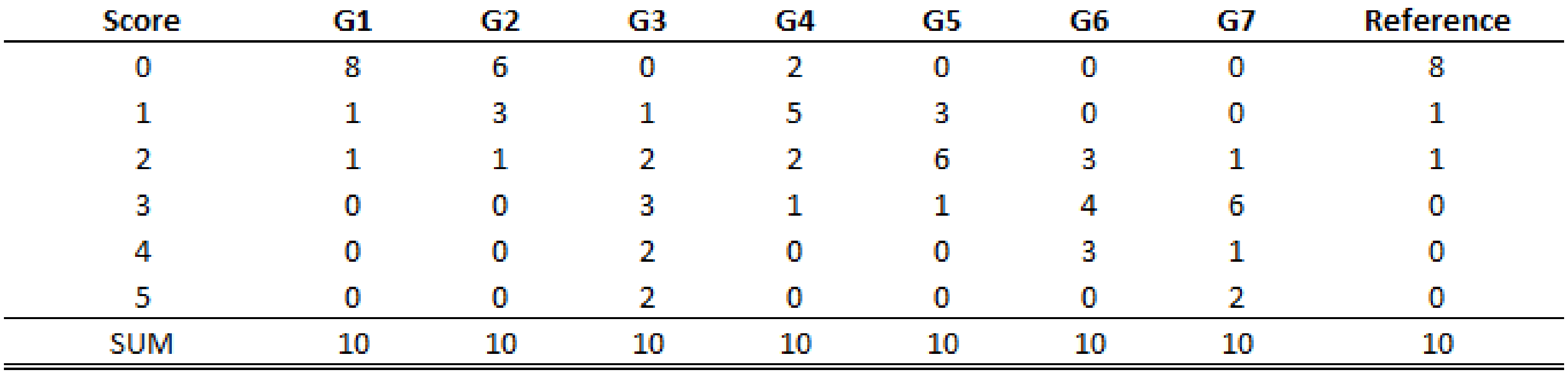
Borg-scale Fatigue Levels vs. G

- The Borg Scale measures physiological exertion and is given over a range of 6 through 20, with 6 being ‘No exertion at all’
- Five pilots flew several repeat sorties at different G levels and recorded ‘Fatigue’ on a modified Borg scale of 0 through 10 – (very similar to a Cooper-Harper scale)
- The G levels were: G1, G2, … ,G6, G7 with G1 slightly above ground-level zero G as a ‘baseline,’ and G6 and G7 being repeated maneuvers at 8G and 9G respectively
- Is Fatigue at higher G levels significantly greater than Fatigue at G1 ? No observed Fatigue rating was > 5

Adjusted Borg scores for ‘Fatigue’ vs. increasing G levels

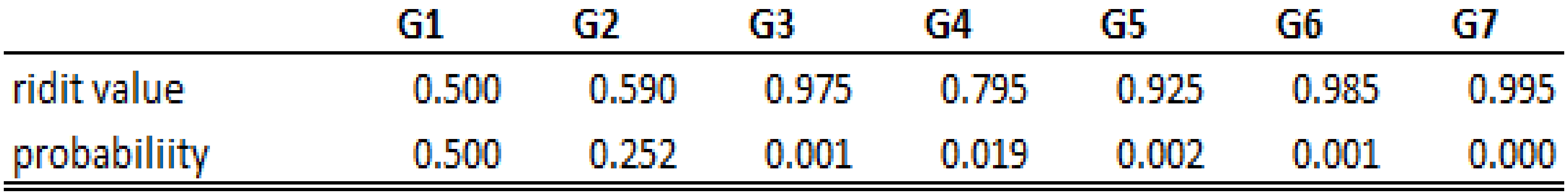


Adjusted Borg scores given by pilots flying at increasing G levels



Score	G1	G2	G3	G4	G5	G6	G7	Reference
0	8	6	0	2	0	0	0	8
1	1	3	1	5	3	0	0	1
2	1	1	2	2	6	3	1	1
3	0	0	3	1	1	4	6	0
4	0	0	2	0	0	3	1	0
5	0	0	2	0	0	0	2	0
SUM	10	10	10	10	10	10	10	10

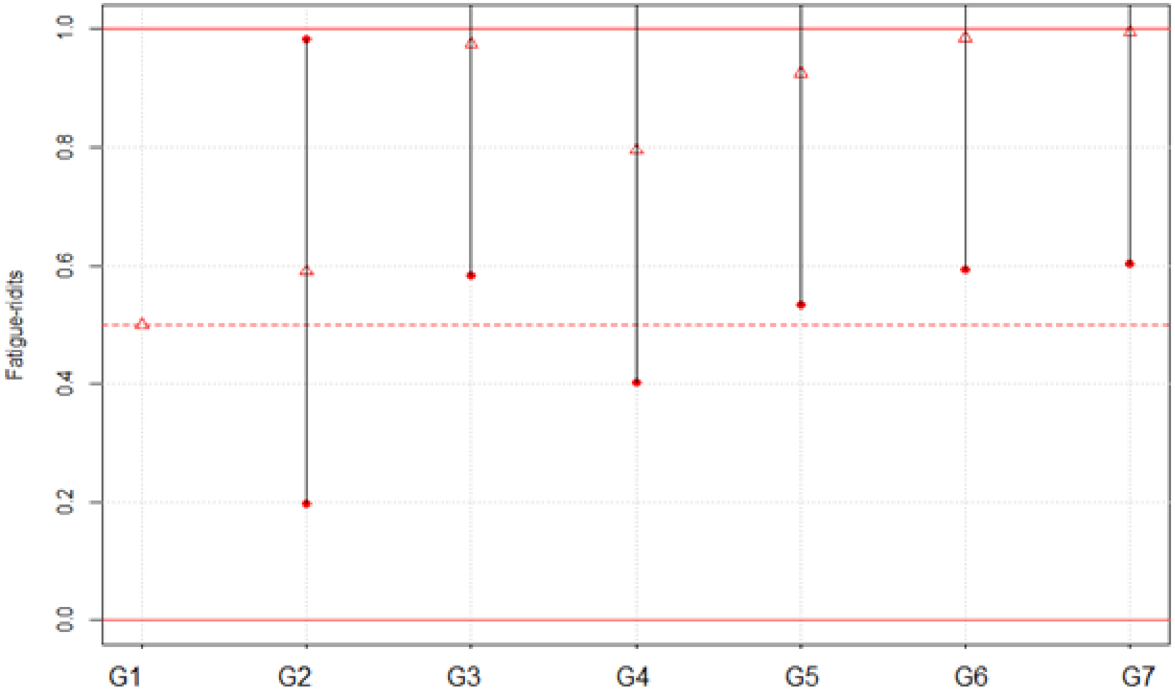
Ridit Analysis of the Borg scores



	G1	G2	G3	G4	G5	G6	G7
ridit value	0.500	0.590	0.975	0.795	0.925	0.985	0.995
probabiliity	0.500	0.252	0.001	0.019	0.002	0.001	0.000

- **Ridits for the Borg Fatigue scores vs. increasing G levels, and the probability that each value is less than the G1 reference level (by Bonferroni-adjusted t-value)**

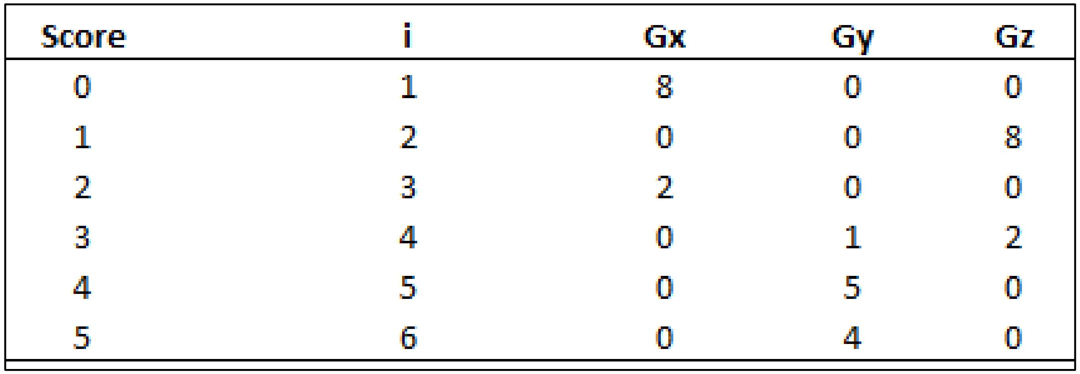
Ridit Plot + Confidence Intervals



- Plot of mean ridits and their 95% Bonferroni-adjusted confidence intervals for the Borg Fatigue scores vs. G
- Problem: Several CI's show overlap of (0, 1).
  - This implies that our distribution theory is only approximate

COMPARING TWO DISTRIBUTIONS

- In OCD analysis, we really want to test if two score distributions—for example, G1 and G7 scores—differ
- Consider how we’d compare two distributions that we’ve turned into probabilities, like: G1→ $\{p_i\}$, G7→ $\{q_i\}$
- A ‘distance’ measure is $H^2 = 1 - \sum \sqrt{p_i q_i}$
- H^2 is called the ‘Squared Hellinger Distance’, and is 1 if the $\{p_i\}, \{q_i\}$, do not overlap, and in [0, 1) if they do. This seems OK as a ‘distance’, but there’s a problem: In the table below, Gx vs Gy has $H^2 = 0$, which we’d expect, but so do Gx and Gz



Score	i	Gx	Gy	Gz
0	1	8	0	0
1	2	0	0	8
2	3	2	0	0
3	4	0	1	2
4	5	0	5	0
5	6	0	4	0

FITTING BETA DISTRIBTIONS

- A solution to this problem is to fit continuous distributions to the observed discrete probability data.
- A flexible distribution to fit, in the case of a discrete distribution lying in [0, 1], Is the Beta(α, β) distribution
- For any given discrete probability distribution, we need an estimate of α and β. These are given by the ‘Method of Moments’ formulas:

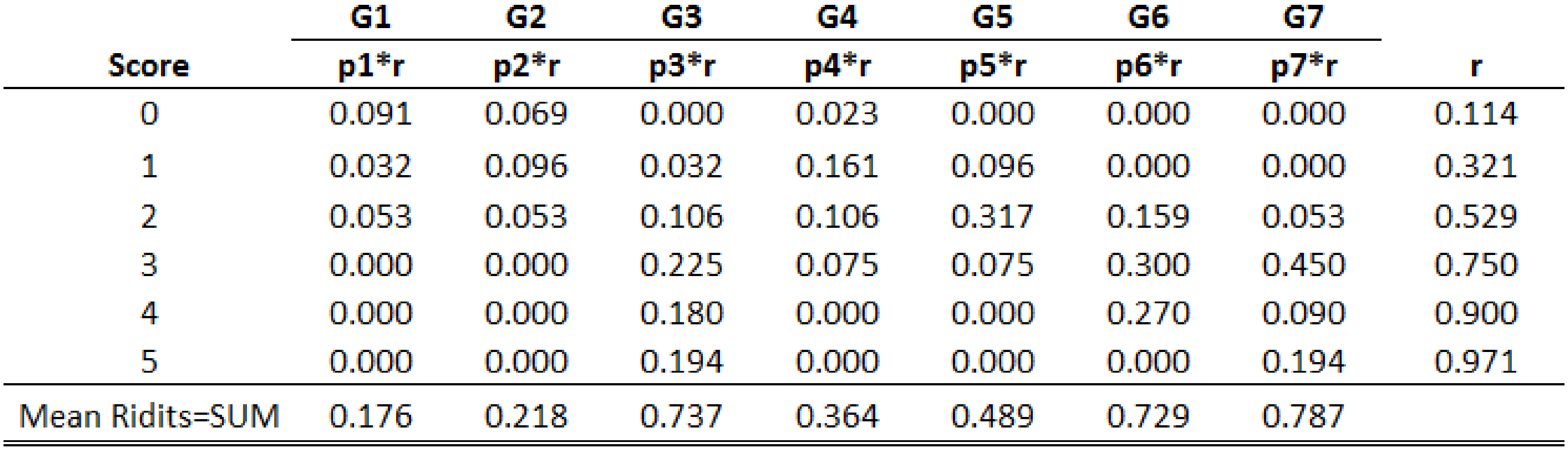
$\widehat{\propto} = \bar{x}\,(\frac{\bar{x}(1-\bar{x})}{\bar{v}} - 1)$, if $\bar{v} < \bar{x}(1-\bar{x})$, $\bar{x}$ is a mean

$\widehat{\beta} = (1-\bar{x})\,(\frac{\bar{x}(1-\bar{x})}{\bar{v}} - 1)$, if $\bar{v} < \bar{x}(1-\bar{x})$, $\bar{v}$ a variance

$\bar{x} = \sum r_i\, p_i$ for each distribution, $\bar{v} = \sum p_i(r_i - \bar{x})^2$.

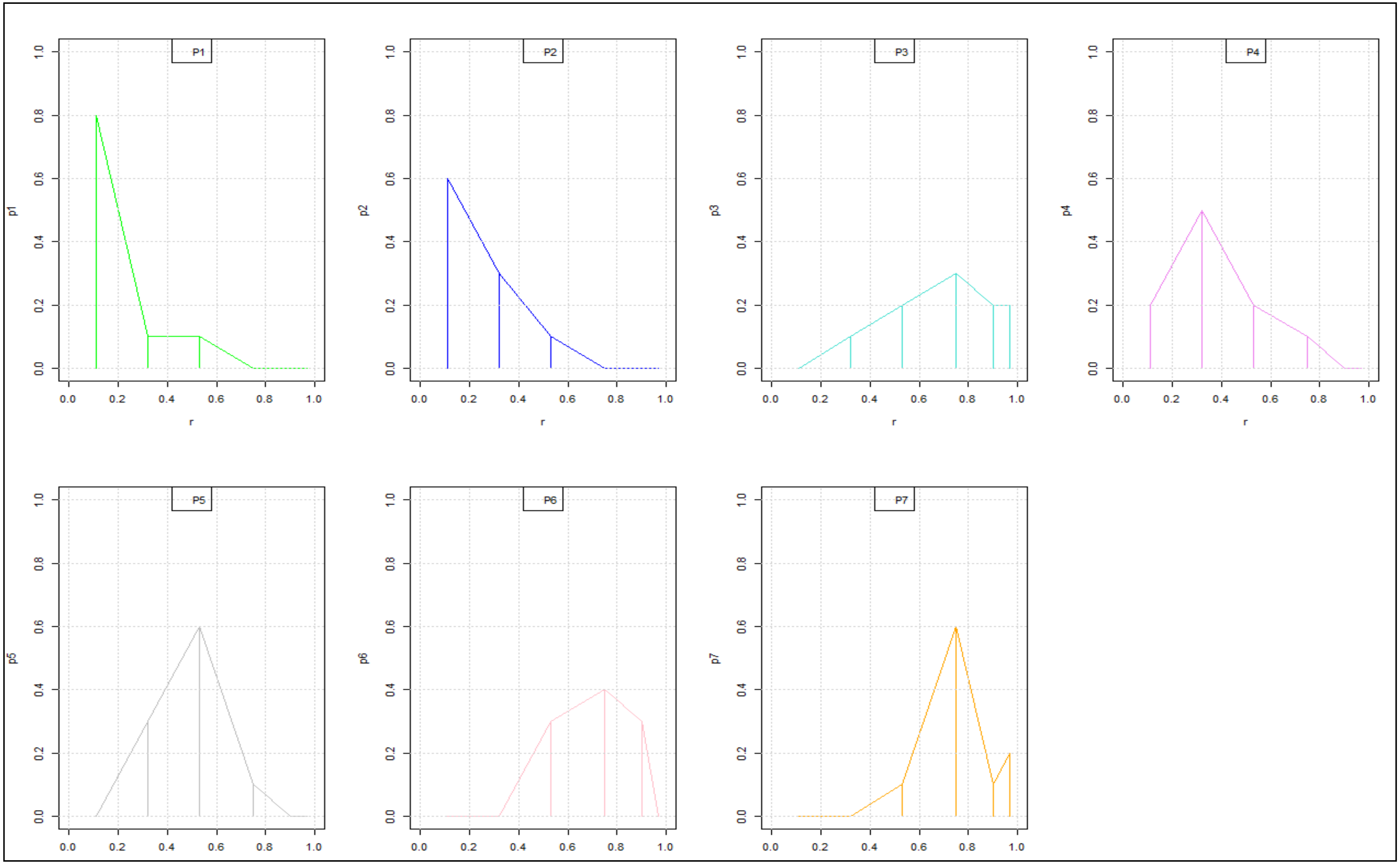
A Common Domain is Required

- Hellinger's Distance requires that the distributions we are comparing be defined over a common domain
- Our reference-derived ridits will serve as the domain over [0,1] but, using G1 as the ref. basis, pushes us to the left
- A solution is to use the row sums of G1…G7 as reference:
- This spreads the domain out better, and gives new means

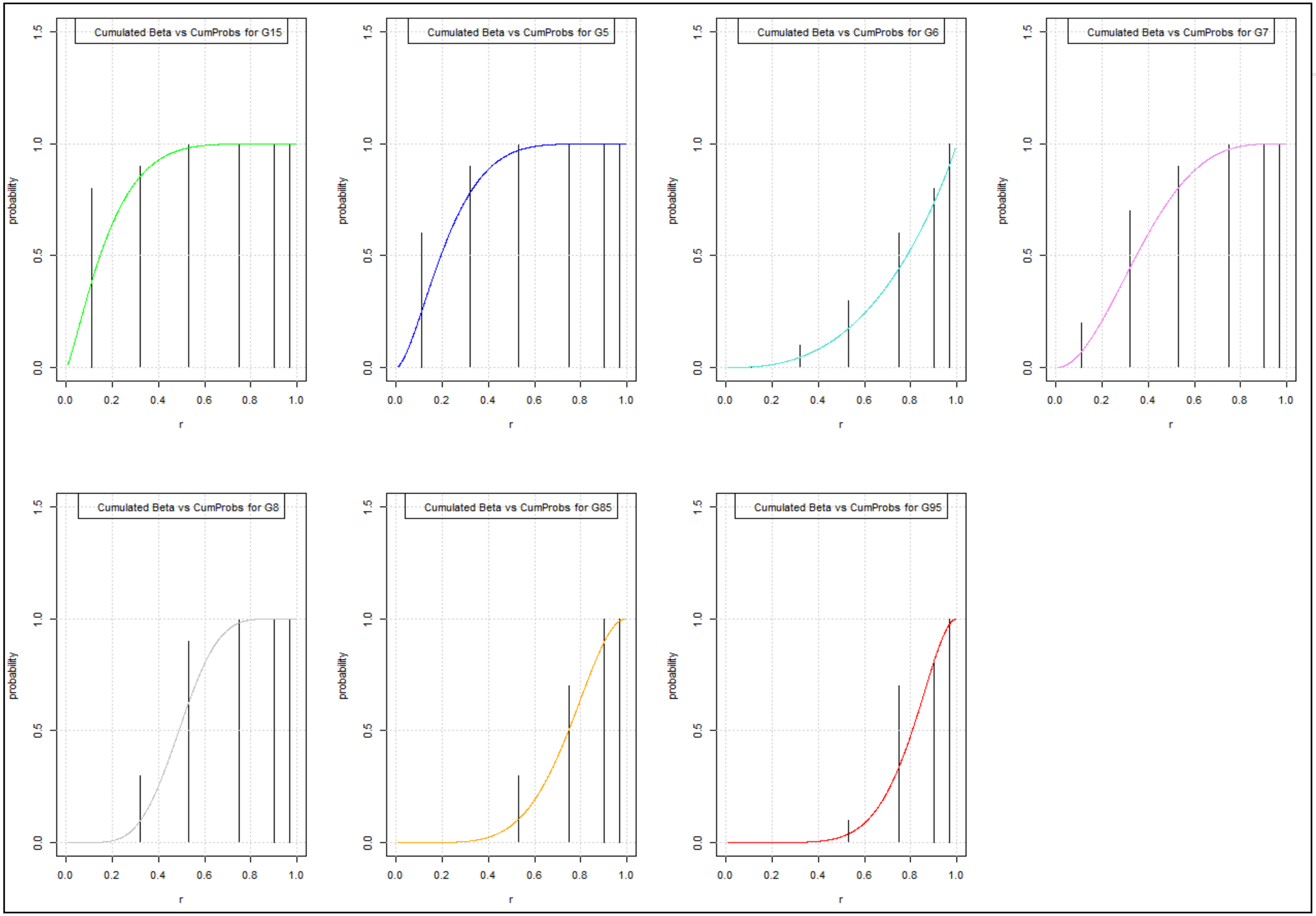


	G1	G2	G3	G4	G5	G6	G7	
Score	p1*r	p2*r	p3*r	p4*r	p5*r	p6*r	p7*r	r
0	0.091	0.069	0.000	0.023	0.000	0.000	0.000	0.114
1	0.032	0.096	0.032	0.161	0.096	0.000	0.000	0.321
2	0.053	0.053	0.106	0.106	0.317	0.159	0.053	0.529
3	0.000	0.000	0.225	0.075	0.075	0.300	0.450	0.750
4	0.000	0.000	0.180	0.000	0.000	0.270	0.090	0.900
5	0.000	0.000	0.194	0.000	0.000	0.000	0.194	0.971
Mean Ridits=SUM	0.176	0.218	0.737	0.364	0.489	0.729	0.787	

Observed Probability ~'s for G1…G7



Fitted Beta ~'s for G1…G7

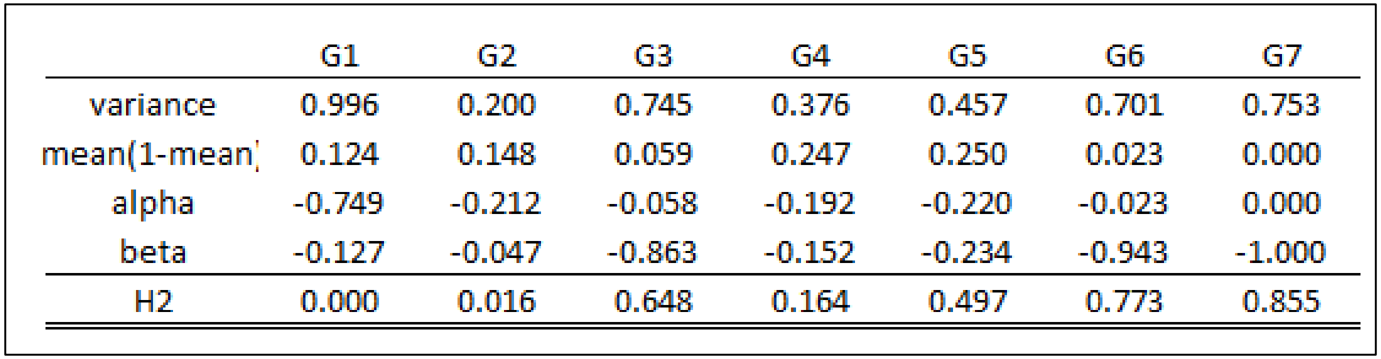


The H^2 Distances Computed

- Hellinger distances for the continuous Beta distributions can now be computed (B is the beta-function) as:

$$H^2 = 1 - \frac{B(\frac{\propto_1+\propto_2}{2},\frac{\beta_1+\beta_2}{2})}{\sqrt{B(\alpha_1,\beta_1)B(\alpha_2,\beta_2)}}$$

- Applying the above formulas, obtain {α, β} for all seven distributions and their H² distances from G1:

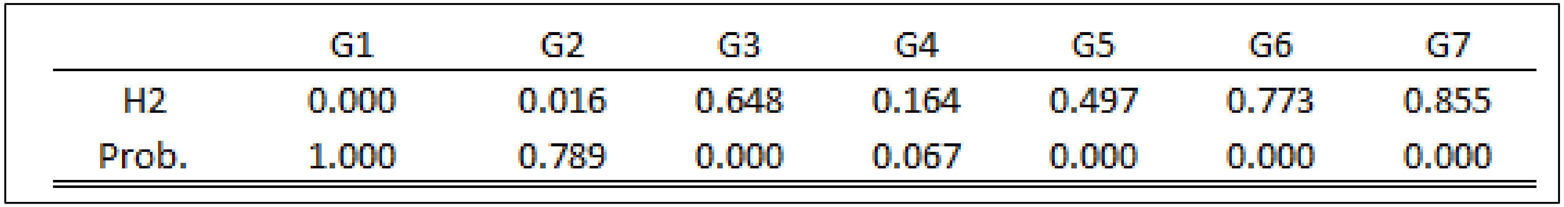


	G1	G2	G3	G4	G5	G6	G7
variance	0.996	0.200	0.745	0.376	0.457	0.701	0.753
mean(1-mean)	0.124	0.148	0.059	0.247	0.250	0.023	0.000
alpha	-0.749	-0.212	-0.058	-0.192	-0.220	-0.023	0.000
beta	-0.127	-0.047	-0.863	-0.152	-0.234	-0.943	-1.000
H2	0.000	0.016	0.648	0.164	0.497	0.773	0.855

Use H^2 to test distribution differences

- Now we have the distance from G1 (as a reference) to all the other six distributions, we can use the H²'s to run randomization tests to find the probability that a null hypothesis of the type Ho: G1=G2 is false:

1. Given: H^2(G1, G2) = 0.016
2. Take a random sample of size n=10 from G1, and another from G1 again. Compute H^2 between these two samples. Do this 10,000 times
3. Compute {number of times $H^2 \geq 0.016$} / 10000. This is the estimated probability P that H^2 =0.016 will occur given Ho is true
4. P = 0.789, so Ho is likely to be true. All prob.'s shown below:

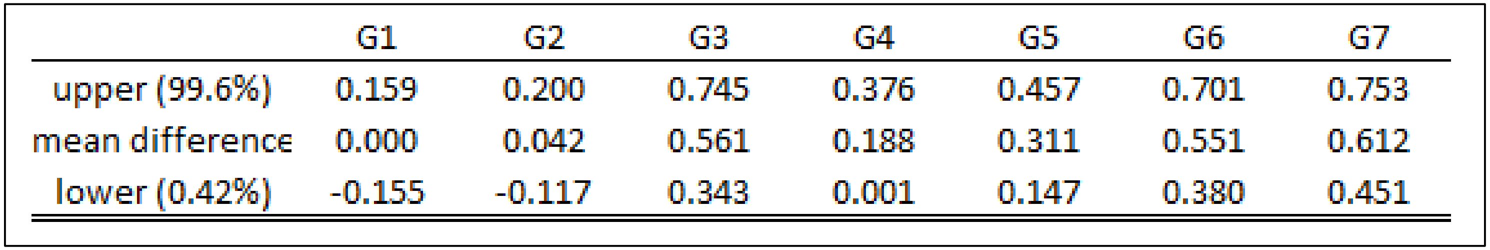


	G1	G2	G3	G4	G5	G6	G7
H2	0.000	0.016	0.648	0.164	0.497	0.773	0.855
Prob.	1.000	0.789	0.000	0.067	0.000	0.000	0.000

Use $\boldsymbol{Beta \sim' s}$ for confidence intervals on differences between means

1. Draw two samples, each size n=10, at random from each of two Beta distributions
2. Compute the difference between the means.
3. Do 1 & 2 10,000 times
4. Compute 100(1-alpha/2k) C.I. quantiles based on proportions where k=number of comparisons we will make (and gives the Bonferroni correction for a 95% overall confidence level; k=6; upper / lower quantile = 99.6% / 0.42%).

This method, comparing G1's ridit mean to all other means, gives C.I.'s: So G3, (G5), G6 & G7 are all > G1: The probability that their OCD distributions are to the right of G1 is confirmed



	G1	G2	G3	G4	G5	G6	G7
upper (99.6%)	0.159	0.200	0.745	0.376	0.457	0.701	0.753
mean difference	0.000	0.042	0.561	0.188	0.311	0.551	0.612
lower (0.42%)	-0.155	-0.117	0.343	0.001	0.147	0.380	0.451

Summary & Conclusions

- It is important to use ridits, or some other nonparametric method, when comparing different flight-test situations with ordinal categorical data (OCD) ratings such as Cooper-Harper, or the Borg scale

- RIDIT ANALYSIS is recommended as a simple technique to replace the incorrect use of ordinal categorical data as ratio-scale numbers.

- We have demonstrated the use of ridit analysis in its standard form, and examined it for the case of student scores and Borg-scale ratings

- We have shown that ridit analysis applies to these cases, and that a new method –using fitted Beta distributions, a Distance-based method, along with randomization trials produce comparisons of mean values, and C.I.'s, that give valid probability results.

References

1. Agresti, A. (1984). Analysis of Ordinal Categorical Data. New York: Wiley & Sons
2. Beder and Heim (1990). On the use of Ridit Analysis. Psychometrika, vol. 55, No. 4. 603-616.
3. Borg, G. (1970). Perceived exertion as an indicator of somatic stress. *Scandinavian journal of rehabilitation medicine* 2 (2): 92–98.
4. Box, G.E.P., Hunter, W.G., and Hunter, J.S. (1978). Statistics for Experimenters. Wiley, New York.
5. Bross, I.D.J. (1958). How to use ridit analysis. Biometrics, 14. 18-38.
6. Cooper, G. and Harper, R. (1969). The use of pilot rating in the evaluation of aircraft handling qualities. Technical Report TN D-5153, NASA
7. Croushore, D. & Schmidt, R.M. (2010). Ridit analysis of student score evaluations. Robins School of Business, Univ. of Richmond, VA 23173 (USA).
8. R Core Team. (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.
9. Selvin, S. (1977). A further note on the interpretation of ridit analysis. American Journal of Epidemiology, 105. 16-20.
10. Selvin, S. (2004) Statistical Analysis of Epidemiologic Data, 3rd ed. OUP. 176-177
11. Wikipedia, 2014. Cooper-Harper. http://en.wikipedia.org/wiki/Cooper-Harper
12. Wilson, D.J. & Riley, D.R. (1989). Cooper-Harper Pilot Rating Variability. American Institute of Aeronautics & Astronautics. (© McDonnell Douglas Aircraft Corp., St. Louis, MO 63166)
13. Wilson, D.J. & Riley, D.R. (1990). More on Cooper-Harper Pilot Rating Variability. American Institute of Aeronautics & Astronautics. (© McDonnell Douglas Aircraft Corp., St. Louis, MO)
14. Wu, C. 2007). On the application of Grey relational analysis and Ridit analysis to Likert Scale surveys. International Mathematical Forum, 2, No. 14. 675-687.