War-Winning Capabilities … On Time, On Cost

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

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 → Comparison

Last year’s scores → Reference

This year’s proportions → p

Last year’s proportions → q

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:

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

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 **z** = (∑ **rp** – **0.5**) / $\sqrt{\left[\frac{1}{12m}+\frac{1}{12n}+\frac{1}{12mn}\right]}$
- For large samples, z is a Normal variate. However, in many tests we deal with small samples, so we shall rather use a t-distribution approximation, and with:

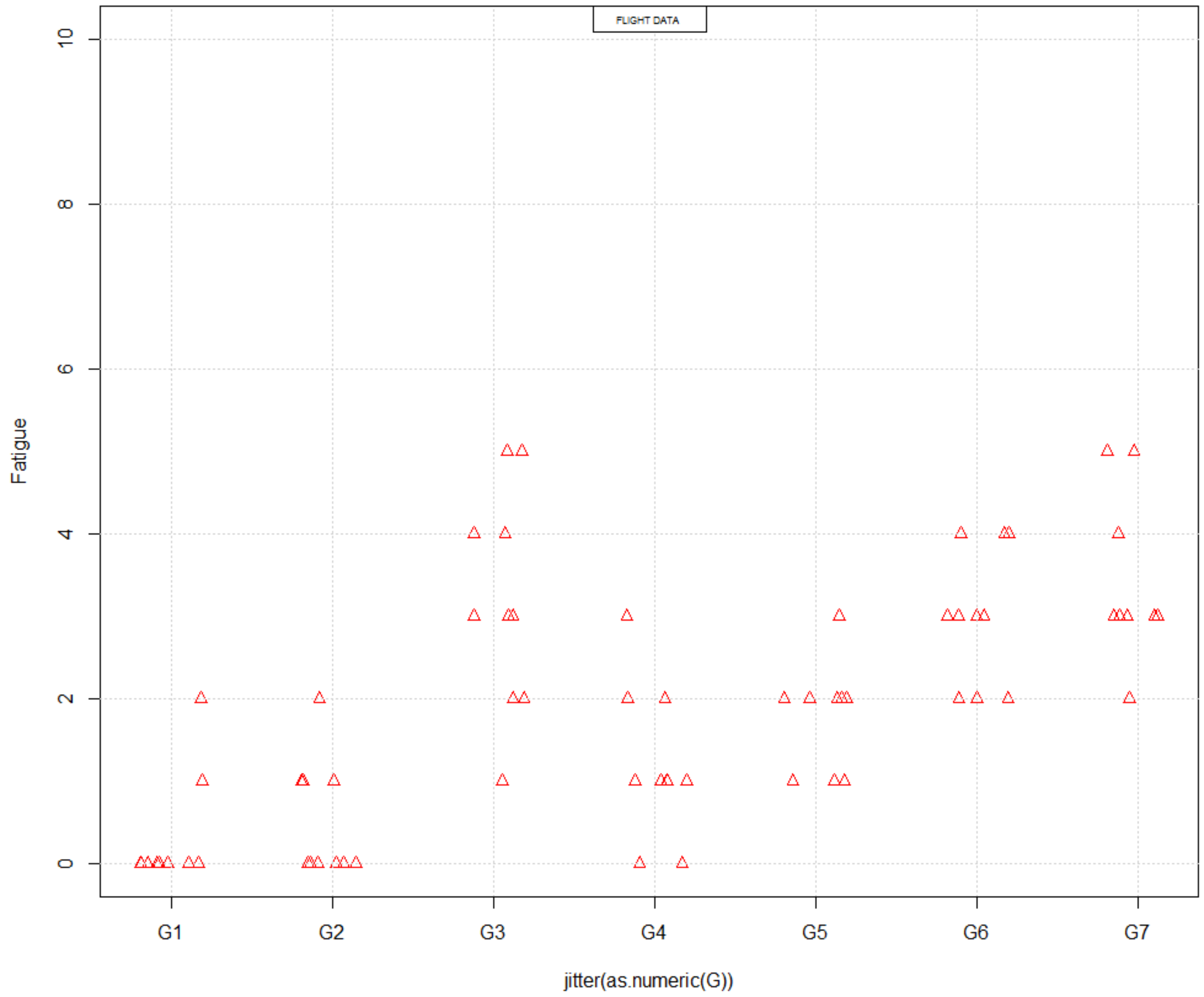
  m = n = 27. So t = (0.411- 0.5)/sqrt(0.0063)= -1.12, with d.f.= m+n-2 = 52

- 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: We can also test Ho: ∑ rp = ∑ rq in the usual way
- 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

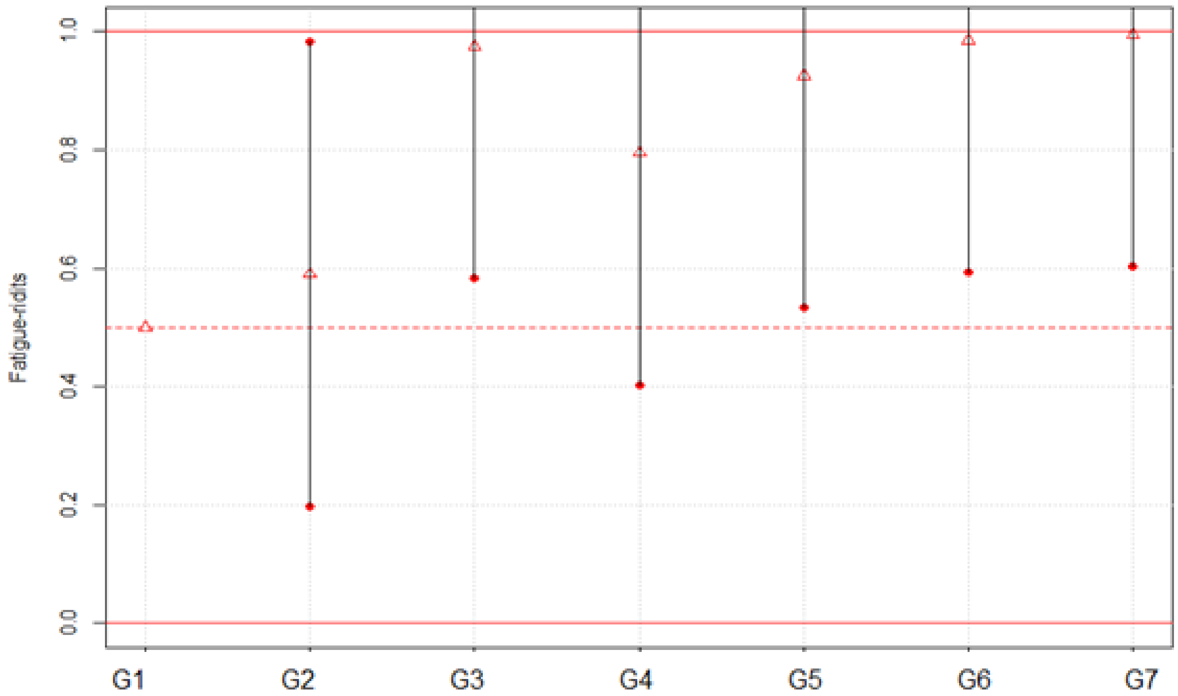
Score	G1	G2	G3	G4	G5	G6	G7	Reference
0	8	6	0	2	0	0	0	8
1	1	3	1	5	3	0	0	1
2	1	1	2	2	6	3	1	1
3	0	0	3	1	1	4	6	0
4	0	0	2	0	0	3	1	0
5	0	0	2	0	0	0	2	0
SUM	10	10	10	10	10	10	10	10

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 $H^2 = 1$, 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:

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

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

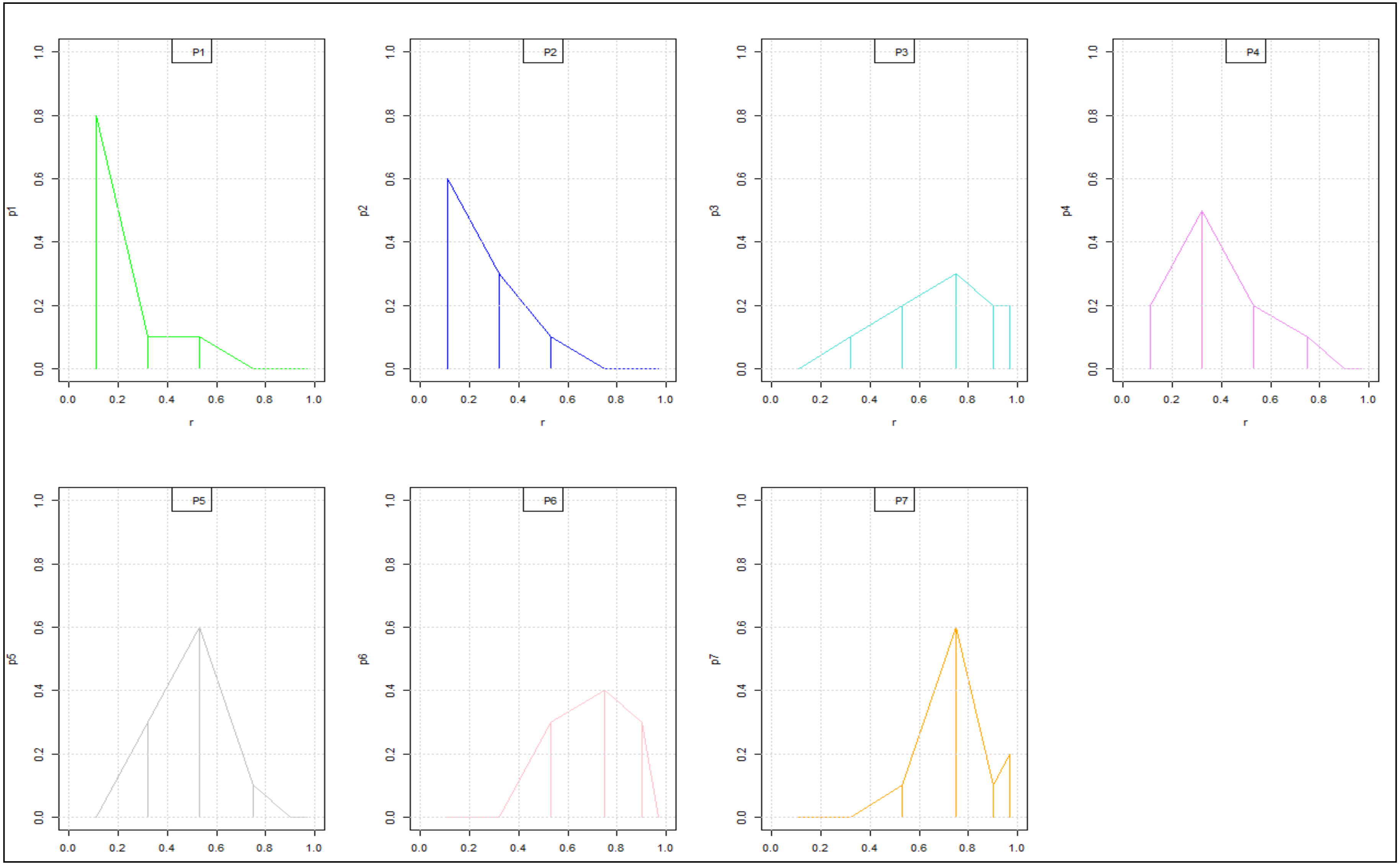
$\bar{x} = \sum r_i p_i$ for each distribution, $\bar{v} = \sum p_i (r_i - \bar{x})^2$.

A Common Domain is Required

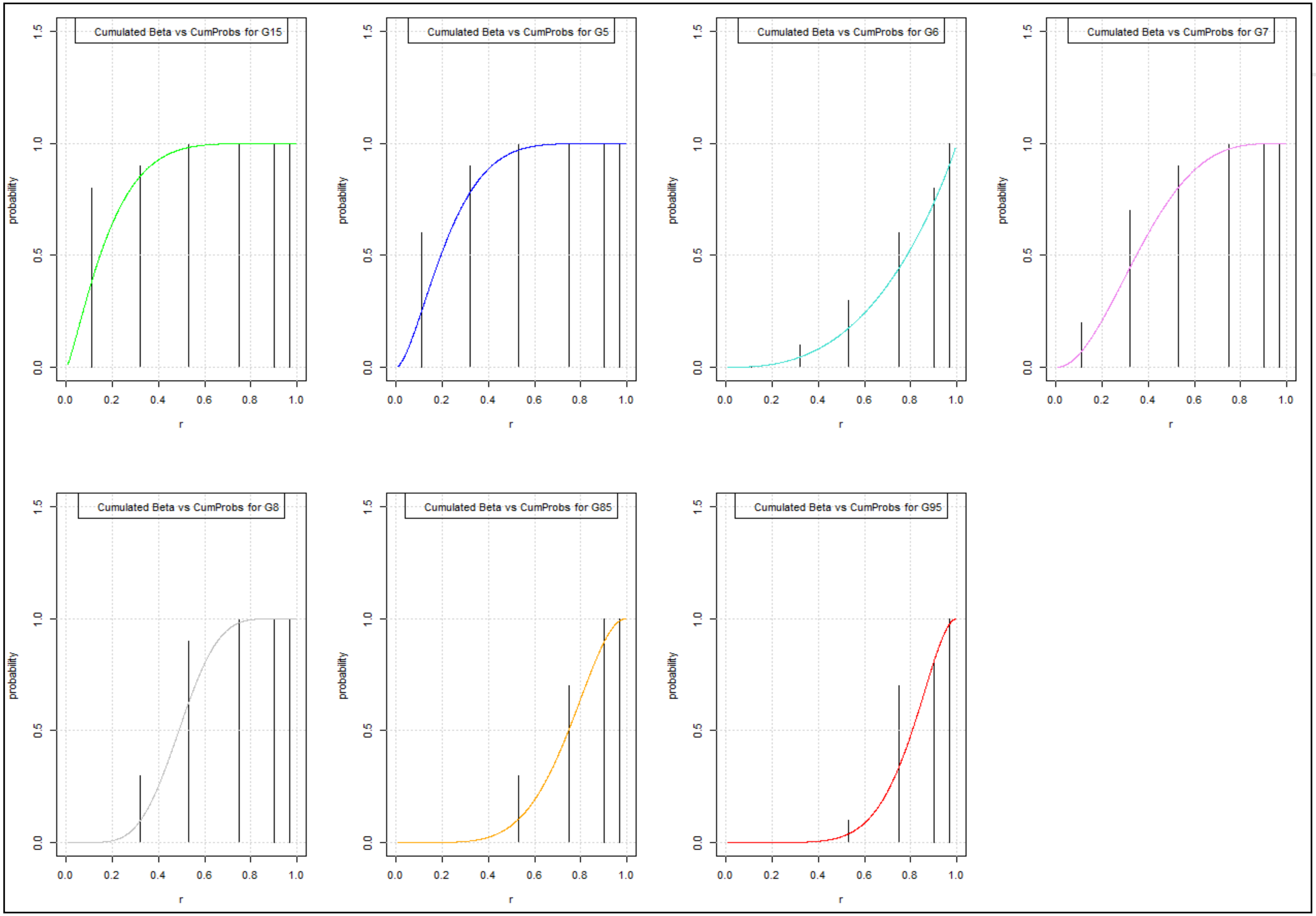
- Hellinger's Distance requires that the distributions we are comparing be defined over a common domain
- Our reference-derived ridits will serve as the domain over [0,1] but, using G1 as the ref. basis, pushes us to the left
- A solution is to use the row sums of G1…G7 as reference:
- This spreads the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

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

- Applying the above formulas, obtain {α, β} for all seven distributions and their H^2 distances from G1:

	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

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 H^2'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 $Beta \sim' s$ 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, …,G7 are all ≠ G1: The probability that their OCD distributions are different to G1 is confirmed. (G4 is 'borderline').

	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.