War-Winning Capabilities … On Time, On Cost

Ridit Analysis for Cooper-Harper and other Ordinal Ratings given for Sparse Data

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 No.: 15357

Overview

- Ridit Analysis Method – introductory example
- Ridit Analysis – Copper-Harper Ratings example
- Ordinal Categorical Data Analysis – some methods compared
- Ridit analysis – example of Borg-scale Fatigue Levels over increasing stages of G-stress

Ridit Analysis

412TW

- 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)

412TW

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:

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

Ridit Analysis – (continued)

412TW

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 ∑ **rp** > ∑ **rq**
  - that is, high ∑ **rp** ⟹ **Reference** pop. (the q's) is bunched '**to the left**'

- 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: $\sum$ **rp** ≥ $\sum$ **rq = 0.5** , form **t** = ($\sum$ **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

Ridit Analysis – FQ by CHR's

412TW

- Two pilots fly 4 runs each, for factor A (*for example, let* A=speed)
  - ❑ **-1** → fly 'slow' configuration
  - ❑ **1** → fly 'fast' configuration
- If 'slow' is the Reference, is 'fast' better?
- Rearranging so as to facilitate a ridit analysis
  - For example, the slow setting for A had one 4 and three 5's
- Note: Now 'to the left' ➔ 'better'

Factor A	PILOT 1 -CH	PILOT 2 - CH
1	2	3
-1	4	5
1	4	3
-1	5	5

CHR	A (-1)	A (+1)	SUM
1	0	0	0
2	0	1	1
3	0	2	2
4	1	1	2
5	3	0	3
6	0	0	0
7	0	0	0
8	0	0	0
9	0	0	0
10	0	0	0
SUMS	4	4	8

This places the 2 medians near the CH boundaries: '*Satisfactory w/o improvement*,' & '*Deficiencies*'

Ridit Analysis – FQ by CHR's

412TW

Converting the above CHR data to ridit analysis, R(p|q) = 0.0313

	REFERENCE	COMPARISON
CHR	Slow	Fast
1	0.000	0.000
2	0.000	1.000
3	0.000	2.000
4	1.000	1.000
5	3.000	0.000
6	0.000	0.000
7	0.000	0.000
8	0.000	0.000
9	0.000	0.000
10	0.000	0.000

	p	q	r	rp	rq
	0.0000	0.0000	0.0000	0.0000	0.0000
	0.2500	0.0000	0.0000	0.0000	0.0000
	0.5000	0.0000	0.0000	0.0000	0.0000
	0.2500	0.2500	0.1250	0.0313	0.0313
	0.0000	0.7500	0.6250	0.0000	0.4688
	0.0000	0.0000	1.0000	0.0000	0.0000
	0.0000	0.0000	1.0000	0.0000	0.0000
	0.0000	0.0000	1.0000	0.0000	0.0000
	0.0000	0.0000	1.0000	0.0000	0.0000
	0.0000	0.0000	1.0000	0.0000	0.0000
SUM	1.0000	1.0000	5.7500	0.0313	0.5000

$$t = \frac{0.0313 - 0.5}{\sqrt{\frac{1}{12*4} + \frac{1}{12*4} + \frac{1}{12*4*4}}} = \frac{-0.469}{0.216} = -2.17$$

The 'critical' one-sided t-statistic at 95% confidence and 4+4-2=6 degrees of freedom is -1.94, so we reject the hypothesis: Ho: 'slow' is better (i.e. closer to '1) than 'fast'.

- Thus we conclude that 'fast' is, in fact, better than 'slow'

Other OC Data Test Methods

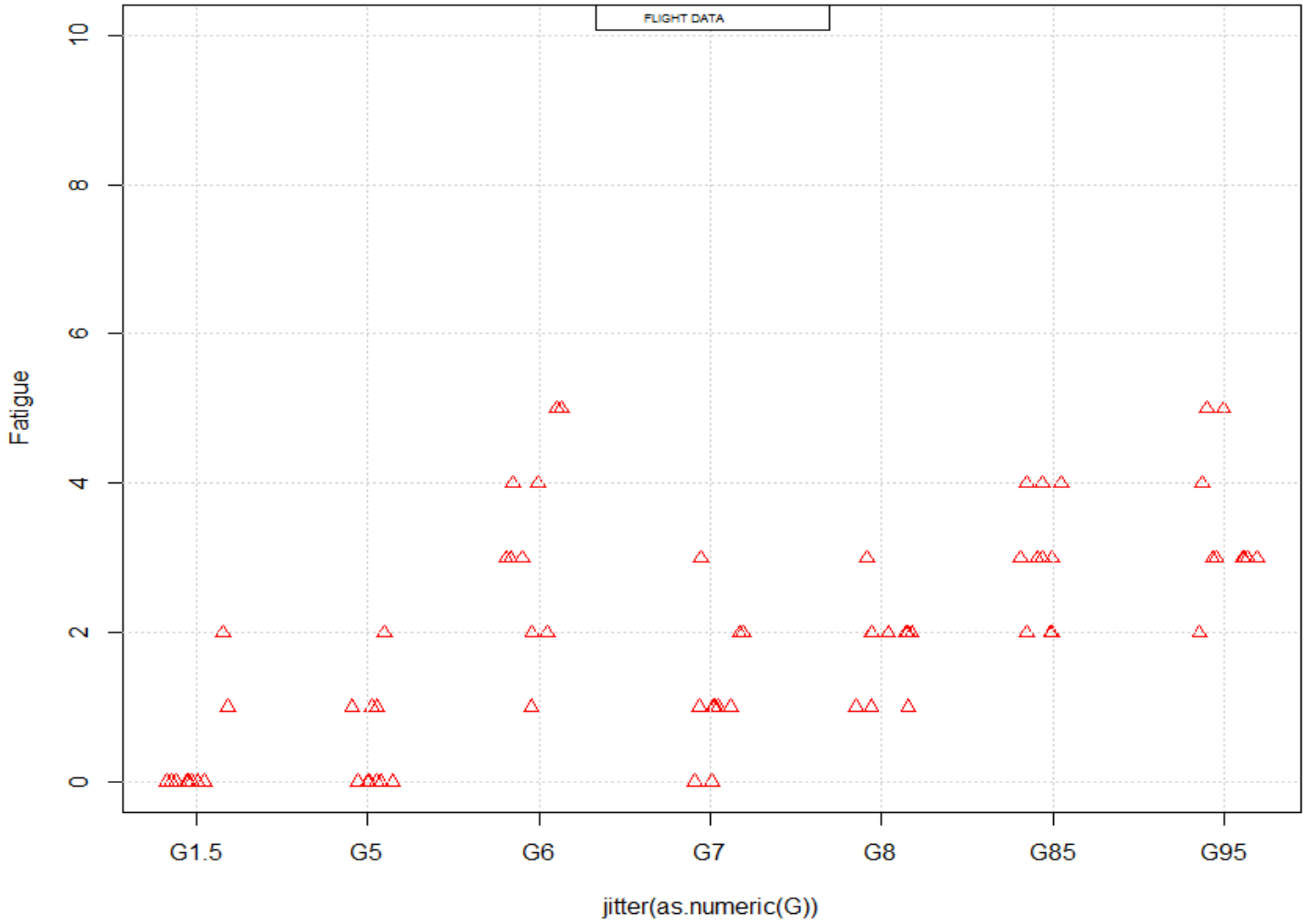
- The t-statistic we used in the above hypothesis test computed out as -2.17 and with 6 degrees of freedom
- The probability of t = -2.17 under the null hypothesis of ‘no difference’ is 0.036
- If we regarded the -2.17 as being from a ‘Z’ (standard Normal distribution), Probability(-2.17) = 0.015. This is a less conservative value than for t
- Another test we could do on the ‘Slow’ vs. ‘Fast’ airspeed case is the Wilcoxon Exact Rank-Sum test, which gives a probability of 0.028
- An ‘approximate’ Wilcoxon test gives 0.018
- The above shows that the t-test is most conservative, and this is usually the case; so we prefer the t-test

Borg-scale Fatigue Levels vs. G

412TW

- 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.
- The G levels were: G1.5, G5, G6, G7, G8, G85, G95, with G1.5 slightly above ground-level zero G as a ‘baseline,’ and G85 and G95 being repeated maneuvers at G8 and G9 respectively
- Is Fatigue at higher G levels significantly greater than Fatigue at G1.5?

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



Adjusted Borg scores given by pilots flying at increasing G levels

412TW

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

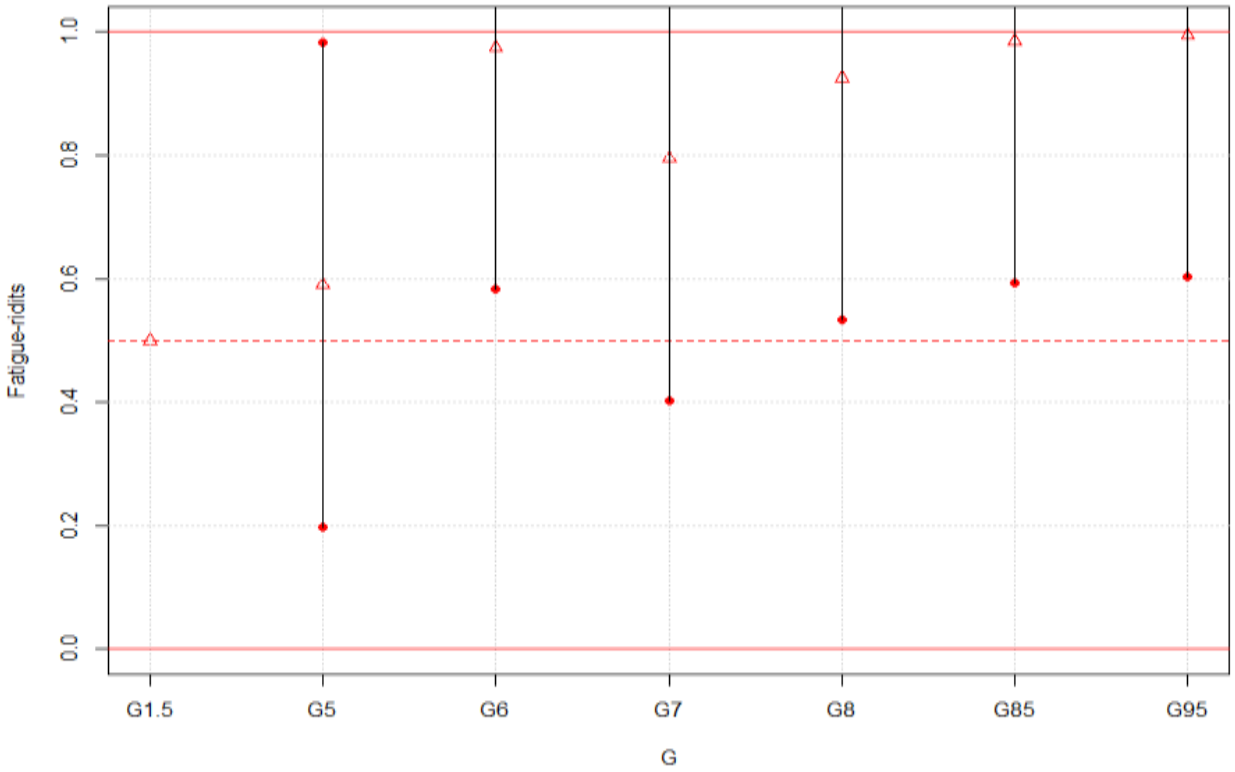
Ridit Analysis of the Borg scores

412TW

	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

- **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)**

Ridit Plot + Confidence Intervals



- **Plot of mean ridits and their 95% Bonferroni-adjusted confidence intervals for the Borg Fatigue scores vs. G**

Summary & Conclusions

412TW

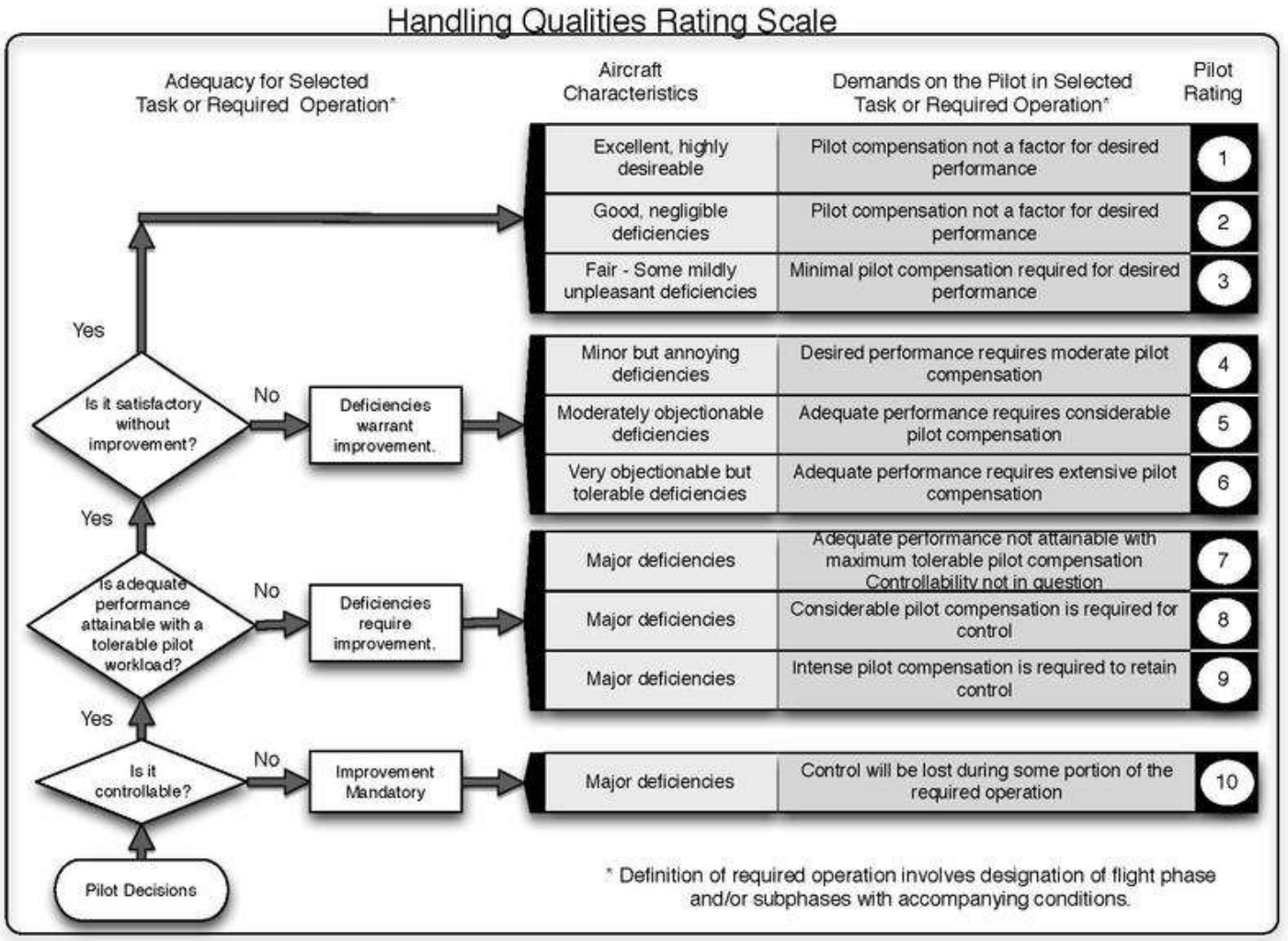
- We have demonstrated the use of ridit analysis in its standard form, and examined it for the case of Cooper-Harper and Borg-scale ratings where observations are often sparse. This latter situation is common in flight test situations

- We have shown that ridit analysis applies well to the sparse-data case. Ridit analysis of sparse ordinal categorical data, and based on the t-statistic, gives statistically defensible tests of hypotheses

- This is important for the use of ridits when comparing different flight-test situations with ordinal categorical ratings such as Cooper-Harper, and is recommended as a technique to replace the incorrect use of ordinal categorical data as ratio-scale numbers.

Cooper-Harper flowchart

412TW



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.