Ordinal Logistic Regression as an alternative analysis strategy for the comparison of two independent samples.

Bilski B[1], Derrick B[1], Toher D[1] and White P[1]*

[1]School of Data Science and Mathematics, University of the West of England, Bristol, BS16 1QY

* Corresponding author paul.white@uwe.ac.uk

Abstract

The two group between subjects design is pervasive with analyses often performed using the Mann Whitney Rank Sum test or using the Welch variant of the t-test. Using simulation it is shown that a dummy variable ordinal logistic regression (OLR) model provides an alternative analysis strategy for n >= 16 per group retaining Type I error robustness for both continuous and tied data. OLR is demonstrated to have comparable power to the Mann Whitney test under non-normal alternatives and with comparable power to the Welch t-test for normal distributed data. This opens the possibility for the ordinal logistic model to be a general analysis technique for higher order designs.

Keywords: Ordinal logistic regression; Mann Whitney test; Welch t-test

1. Introduction

The comparison of two independent samples comprising scale outcome data is long established. To detect a location shift, common long-established approaches include using either the Mann Whitney Wilcoxon Rank sum test referred to as the Mann Whitney test (Mann and Whitney, 1947) or the independent samples t-test assuming homogeneity of variances (Fisher, 1925), or the separate variances version of the t-test known as the Welch test (Welch, 1947). It is acknowledged that other approaches exist, such as the Yuen Welch t-test (Yuen, 1974), or the Brunner and Munzel test (2000) test (Brunner and Munzel, 2000), or using permutation tests or the bootstrap, but these approaches generally have less uptake.

The independent samples t-test is the uniformly most powerful test for normally distributed data providing distribution variances are identical (Zimmerman, 1987). In application, the population or distribution may be approximately Normal, and the assumption of precisely equal population variances might be too restrictive. In these cases, the Welch (separate variances) t-test is often championed as the preferred analytical approach. When the assumption of normality has not been grossly violated, Ruxton (2006), Derrick et al (2016), Delacre et al. (2017), amongst others, have recommend using the Welch t-test as the default t-test particularly when sample sizes differ between the two groups irrespective of whether variances are equal or not. For non-normal data it is well established that the non-parametric Mann Whitney test will retain Type I error rates and may confer power advantages relative to the assumed equal variances t-test or Welch t-test (Fagerland and Sandvik, 2009a).

An analyst may be tempted to assess the data for normality and to assess other assumptions to help decide between applying a form of the t-test or the Mann Whitney test. Formal tests of normality will have low power when sample sizes are small, negating their use. With large sample sizes, formal tests of normality will have increased power to detect lack of normality and this is when an assumption of normality is typically less critical, again negating their use. In alignment with the famous quotations "*Normality is a myth; there never has, and never will be, a normal distribution*" (Geary, 1947) and "*All models are wrong, but some are useful*", the real question is whether deviation from normality is sufficiently great to such an extent which would invalidate the use of a t-test. However, no such formal test exists or can exist with point null hypothesis testing. In the absence of such a test an analyst might place reliance on rules of thumb (e.g. with respect to the degree of skew, or kurtosis of residuals) and/or through a graphical assessment of distributional properties (e.g. histograms, Quantile-Quantile plots, box-and-whisker plots) in deciding upon a valid analytical approach. However, allowing the data to select an analysis technique might impact on conclusions obtained (see Pearce and Derrick (2019), Derrick et al, 2018).

The independent samples t-test is a special case of the one-way between subjects ANOVA; the one-way ANOVA defaults to the independent samples t-test when the number of factor levels is two. The Welch t-test may similarly be generalised to, or be considered a special instance of, the robust one-way Welch ANOVA (Delacre et al, 2019). Likewise, the non-parametric equivalent of the one-way ANOVA, the Kruskal-Wallis test, is logically equivalent to the Mann Whitney test when there are only two factor levels. The one-way ANOVA may be readily extended to a one-way analysis of covariance (ANCOVA) with or without corrections for homogeneity of variances. For non-parametric analyses, the Quade test (Quade, 1967) may be used to undertake a rank-based analysis including a covariate.

ANOVA and ANCOVA may be used with higher order designs (e.g., two-way between subjects factorial designs with or without covariates). Aligned Rank Transformation ANOVA (ART-ANOVA) permits higher order non-parametric analyses (Leys and Schumann, 2010) and the Scheirer-Ray-Hare test (SRH-ANOVA, Scheirer, Ray, and Hare, 1976). However, there has been little uptake of ART-ANOVA, or SRH-ANOVA and little uptake of the Quade test in the applied literature.

Brought to prominence by McCullagh (1980), Ordinal Logistic Regression (OLR) is a regression model for ordinal dependent variables. It is typically championed for analysing ordinal categorical data and usually used when there are a small number of ordered categorical response options (see e.g. Elamir and Sadeq. 2010; Menard 2010). However, there is no limit to the number of options or ordered categories a response variable can take in ordinal logistic regression and categories may consist of a single observation. Essentially, data which may be ranked may form ordinal categories with tied data occupying the same category. As such OLR with a two-level factor dummy variable coded may be seen as being similar to the Mann Whitney test and similar to the Mann Whitney test, OLR is not reliant on an assumption of normality Inference under OLR relies on asymptotic approximations being acceptable and this may be a limiting problem in applying OLR when sample sizes are small.

The aim of this paper is to determine whether ordinal logistic regression provides an alternative analytical strategy for the analysis of ordinal and scale outcome data for the two groups situation. In this situation the two-level factor is dummy variable coded and included in the regression model as a single predictor. If such an approach is reasonable, particularly with small sample sizes, then this would lead to providing flexible models potentially incorporating covariates, multiple factor levels, and multiple factors without an express assumption of normality. This paper is designed to assess how each of OLR, the Welch t-test, and the Mann-Whitney test provide control of Type I error rates when a Null Hypothesis is true. Comparison is also made in a non-null situations to ascertain relative power advantages.

2. Methodology

The basic aim is to compare statistical inferences from applying the OLR model, the Mann Whitney test and the Welch t-test to the two group between-subjects design (i.e. data stochastically independent both within and between two groups). Robustness of statistical techniques is examined via Monte-Carlo simulation techniques (see e.g. Derrick et al 2015, Mirtagioğlu 2017, Lyhagen 2021). An assessment of the *p*-values under the tests will be undertaken for all three methods and two-sided statistical inferences will be considered at the nominal significance level of α = 0.10, 0.05, and 0.01.

The normal approximation often used in the Mann Whitney test is known to work well when sample sizes are 16 or larger (Fagerland and Sandvik, 2009b). For this reason, the simulation design will be restricted to equal sample sizes between groups (i.e. $n_1 = n_2 = n$) with n = 16, 32, 64, 128.

The focus is on a location shift and therefore the simulation will be restricted to simple location shift models preserving equality of variance (i.e. $\sigma_1^2 = \sigma_2^2 = \sigma^2$). Both symmetric and skewed distributions will be considered.

Simulation using sampling from inherently continuous distributions uses full machine precision. However, real world data is not perfectly normal and has finite precision (e.g., see Stahl, 2006). We will therefore additionally consider rounded data as the dependent variable. For instance, data sampled from the standard normal distribution when rounded to the nearest integer (0 decimal

places) will have a reduced support typically -3, -2, -1, 0, +1, +2, +3 (i.e., the majority of the time within +/- 3.5 standard deviations). Rounding will increase the probability of tied rank positions in the data.

For each cell of the design, a total of 10,000 instances is considered. A summary of the simulation design parameters under both H_0 (to investigate type I error rates) and under H_1(to investigate power for Type I error robust procedures) is given in Table 1.

Table 1. Simulation design parameters

Distribution	Under H_0	Under H_1
Normal	μ_1=μ_2=0, σ_1=σ_2=1	μ_1= 0.5, μ_2= 0, σ_1=σ_2=1
Chi-square χ_2^2	μ_1=μ_2=2, σ_1=σ_2=2	μ_1=3, μ_2=2, σ_1=σ_2=2
Rounding	None, 2 d.p., 1 d.p., 0 d.p.	
Sample size	16, 32, 64, 128	
α	0.10, 0.05, 0.01	
Number of iterations	10,000 per cell	
Programming language	R version 4.1 (seed=2642)	

Using R version 4.1, the *polr* command from the 'MASS' package is used to estimate the ordinal logistic regression model. The *t.test* command and the *wilcox.test* command from the 'STATS' package are used to calculate the Welch t-test and the Mann Whitney test corrected for ties.

A valid statistical test with a true null hypothesis will generate *p*-values which are uniformly distributed over the interval [0,1] (Hung et al. 1997). Bradley (1978) suggested a stringent, moderate and liberal criterion fort test robustness. Specifically, for a given nominal α, stringent robustness criteria are met if the empirical Type I error rate is between $\alpha \pm 0.1\ \alpha$, moderate robustness criteria are met if the empirical Type I error rate is between $\alpha \pm 0.2\ \alpha$, and liberal robustness criteria are met if the empirical Type I error rate is between $\alpha \pm 0.5\ \alpha$. For instance, if the nominal α = 0.05 is considered and the empirical Type I error rate is below 0.025 then the test is said to be conservative, and if above 0.075 the test is considered to be liberal (Bradley, 1978).

When the alternative hypothesis is true, the distribution of the p-values is a function of sample size and effect size and is typically positively skewed reflecting the power of the test (Hung et al, 1997). Power of the OLR approach is compared to that of the Mann-Whitney test and the Werlch t-test under identical sample size and effect size conditions. Under H_1, parameters are chosen to represent a simple location shift.

3. Results

Results under the null hypothesis are considered to obtain Type I error rates for each of Ordinal Logistic Regression (OLR), the Mann-Whitney test (M-W) and Welch's form of the independent samples t-test (t-test). This is followed by results under the alternative hypothesis to obtain a power comparison between the three methods.

Type I error rates

When the null hypothesis is true, *p*-values would be expected to observe characteristics of the uniform distribution with domain [0, 1]. When Ho is true and the outcome data is normally distributed then it is well known that both the Mann-Whitney test and the Welch t-test are valid tests with uniformity of p-values. Figure 1 graphically depicts the distribution of the p-values under the simulation for the OLR model for normal deviates under a true null hypothesis. Visually, the simulated derived p-values

appear compatible with a claim of uniformly distributed data although for n = 16 a critical reviewer may argue that there is some unevenness across some histogram bins.

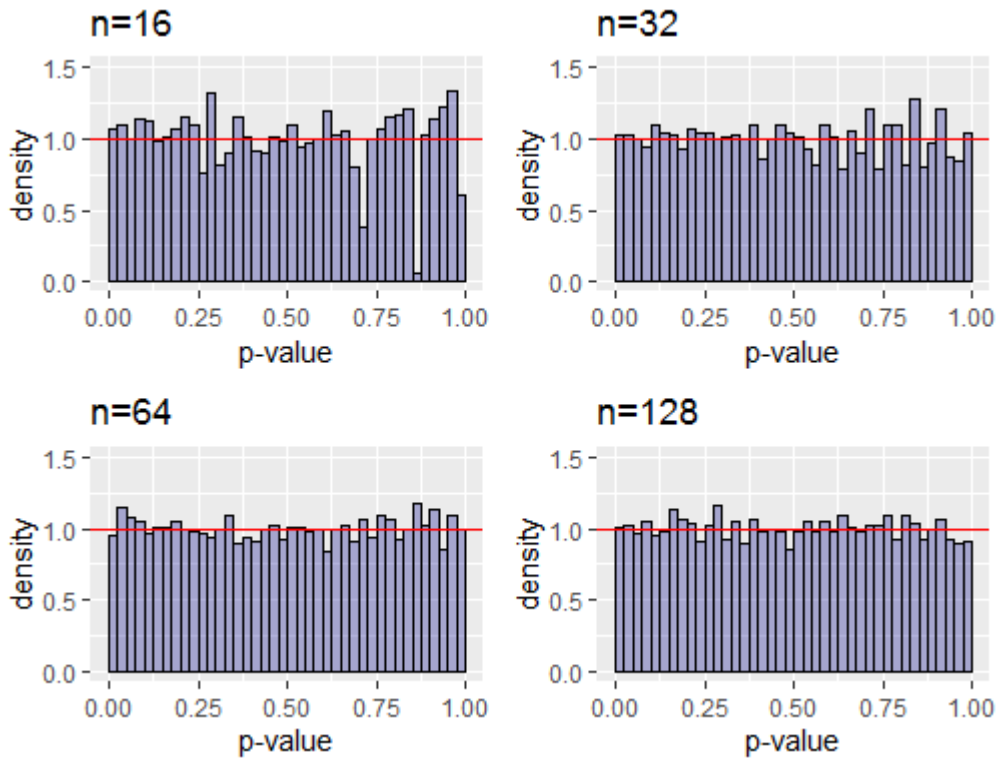

Figure 1. OLS p-values, Normal distribution (no rounding).

The scatterplot of p-values for the three methods when the null hypothesis is true (independent normal random variables, no rounding) given in Figure 2 (OLR and Welch's test) and in Figure 3 (OLR and the Mann-Whitney test) shows a strong linear correlation between OLR and the two other methods with respect to observed p-values These data relate to the normal distribution without tied values, and it is noticeable that in these situations the OLR procedure and the Mann Whitney test become practically indistinguishable with respect to the p-values with increasing sample size.

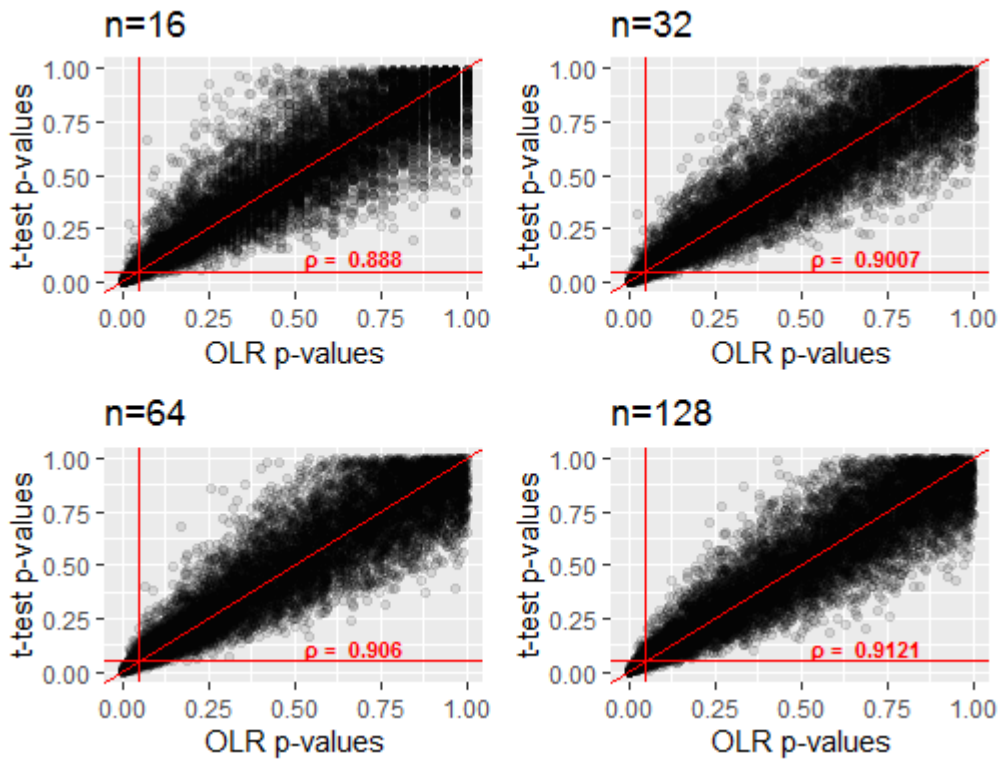

Figure 2. OLR and t-test p-values, Normal distribution (no rounding). Reference lines at the 5% significance level. Pearson's correlation coefficient calculated.

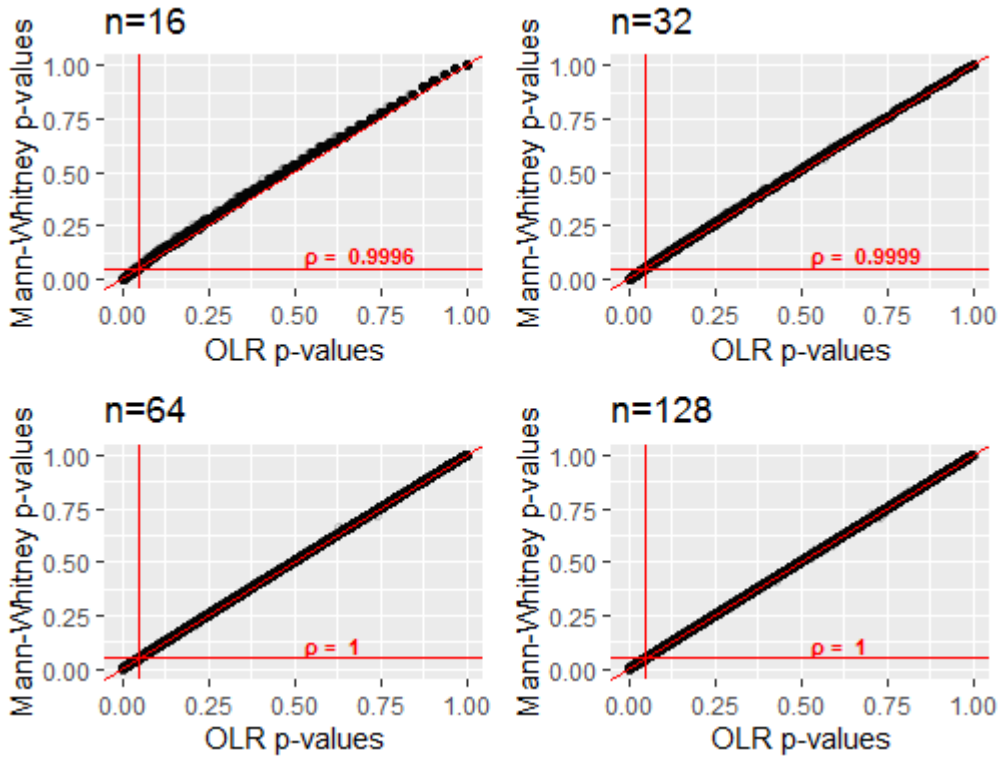

Figure 3. OLR and Mann-Whitney test p-values, Normal distribution (no rounding). Reference lines at the 5% significance level. Pearson's correlation coefficient calculated.

Table 2 provides simulation observed Type I error rates for the three methods (OLR, Welch t-test, Mann Whitney test) when the outcome variable has the standard normal distribution without rounding and with rounding to the nearest integer. Across all significance levels the null hypothesis rejection rate is approximately equal to the nominal α value, indicating Type I error robustness under the normality condition for n >= 16 per group. This Type I error robustness is observed across the simulation design including the extreme most case of rounding to 0 d.p.

Table 2. Type I error rates, Normal distribution (mean zero, standard deviation 1)

	No rounding			Rounded to 0 decimal places		
	OLR	t-test	M-W	OLR	t-test	M-W
α=0.10, n=16	0.1077	0.0994	0.0907	0.1025	0.0986	0.0947
α=0.10, n=32	0.0998	0.0976	0.0934	0.0960	0.0958	0.0928
α=0.10, n=64	0.1057	0.1023	0.1027	0.1042	0.1004	0.1028
α=0.10, n=128	0.1009	0.1000	0.0993	0.1006	0.0993	0.1000
α=0.05, n=16	0.0543	0.0481	0.0468	0.0482	0.0469	0.0432
α=0.05, n=32	0.0514	0.0490	0.0470	0.0486	0.0485	0.0466
α=0.05, n=64	0.0526	0.0515	0.0502	0.0507	0.0500	0.0500
α=0.05, n=128	0.0507	0.0496	0.0497	0.0482	0.0500	0.0479
α=0.01, n=16	0.0095	0.0082	0.0088	0.0073	0.0091	0.0075
α=0.01, n=32	0.0104	0.0091	0.0100	0.0097	0.0102	0.0091
α=0.01, n=64	0.0087	0.0077	0.0081	0.0069	0.0075	0.0068
α=0.01, n=128	0.0100	0.0100	0.0099	0.0081	0.0085	0.0078

Relatedly, Figure 4 shows the proportion of times where a correctly specified null hypothesis is rejected at the 5% significance level under four scenarios (standard normal, standard normal rounded to 2 decimal places, 1 decimal place and 0 decimal places). It can be seen that there is negligible difference in the robustness of OLR going from no rounding to 2 d.p. or 1 d.p. This is true for the other two methods too. However, all three methods under consideration reject the null hypothesis less frequently when the normally distributed values are rounded to 0 d.p. Nevertheless, across all conditions simulated under normality, all three methods satisfy stringent Type I error robustness criteria stipulated by Bradley (1978), apart from the Mann Whitney test for n = 16 per group with rounding to 0 decimal places but which satisfies moderate Type I error robustness.

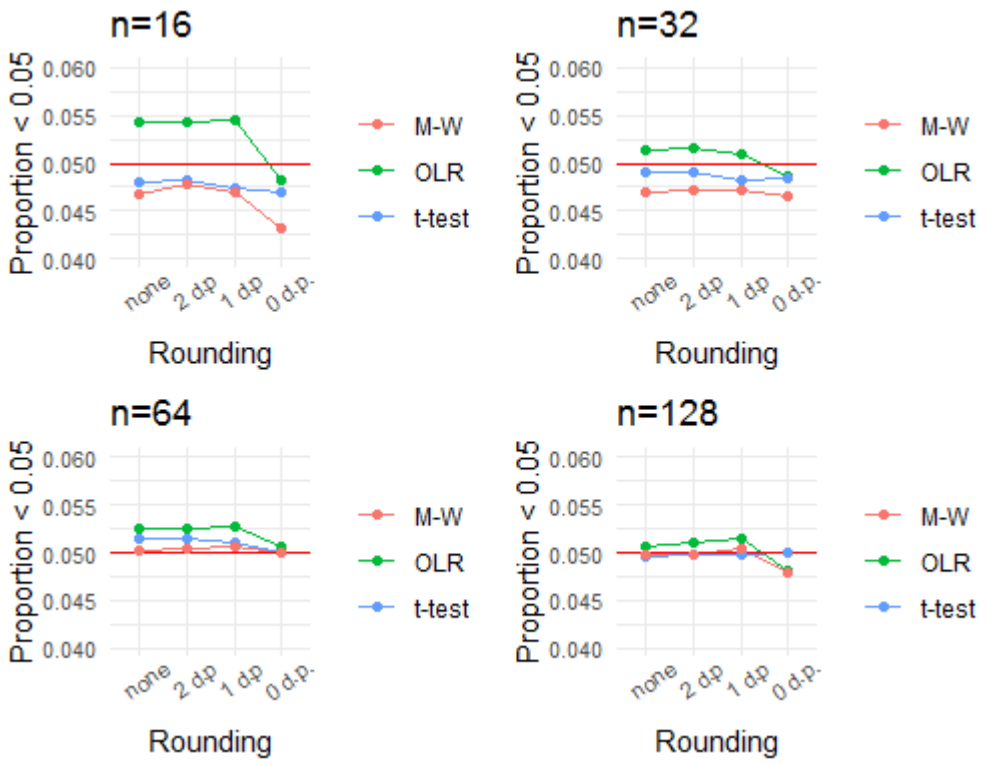

Figure 4. Type I error rates, Normal distribution with various levels of rounding.

Having confirmed Type I error robustness for OLR under the condition of normality, the Type I error robustness for skewed data is assessed. Figure 5 displays the distribution of p-values when performing OLR on two samples taken from the Chi-square distribution χ^2_2.

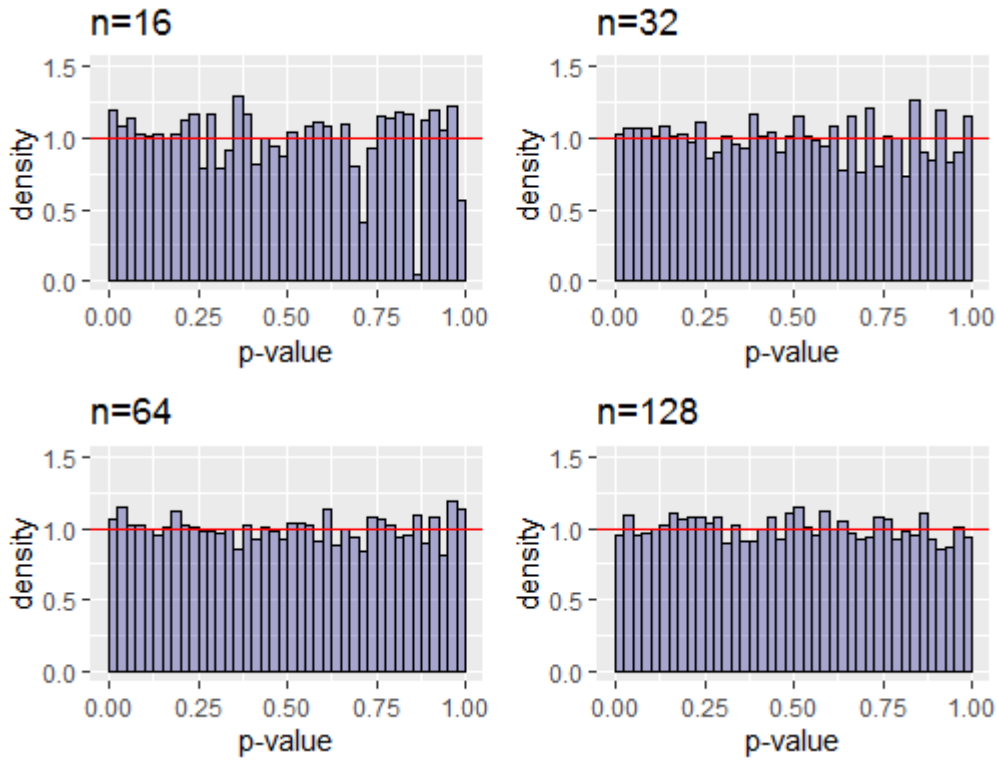

Figure 5. OLR p-values, Chi-square distribution (no rounding).

Figure 5 suggests that the proposed OLR approach can be extended to this non-normal condition. For *n*=16 there is perhaps some apparent deviation from uniformity in the upper region of the distribution but otherwise approximate uniformity is observed. Further detail for typical levels of α are given in Table 3. Across all significance levels the null hypothesis rejection rate is approximately equal to the nominal α value, indicating Type I error robustness is retained for skewed distributions.

Table 3. Type I error rates, outcome sampled from the chi-square distribution on 2 degrees of freedom.

	No rounding			0 decimal places		
	OLR	t-test	M-W	OLR	t-test	M-W
α=0.10, n=16	0.1114	0.0988	0.0951	0.1072	0.0985	0.0993
α=0.10, n=32	0.1063	0.0939	0.0996	0.1041	0.0935	0.0990
α=0.10, n=64	0.1064	0.0990	0.1033	0.1045	0.1007	0.1022
α=0.10, n=128	0.0988	0.1028	0.0977	0.0988	0.1021	0.0978
α=0.05, n=16	0.0570	0.0465	0.0493	0.0558	0.0472	0.0502
α=0.05, n=32	0.0525	0.0476	0.0500	0.0494	0.0461	0.0478
α=0.05, n=64	0.0554	0.0486	0.0537	0.0527	0.0491	0.0520
α=0.05, n=128	0.0511	0.0472	0.0509	0.0501	0.0499	0.0499
α=0.01, n=16	0.0123	0.0077	0.0119	0.0114	0.0083	0.0102
α=0.01, n=32	0.0102	0.0077	0.0097	0.0091	0.0071	0.0086
α=0.01, n=64	0.0116	0.0104	0.0111	0.0103	0.0102	0.0101
α=0.01, n=128	0.0090	0.0086	0.0088	0.0084	0.0084	0.0084

Figure 6 shows the proportion of iterations where the null hypothesis is rejected at the 5% significance level for sample data from the Chi-square distribution discretised to the specified number of decimal places. There is an apparent trend for OLR to result in the rejection of the null hypothesis more frequently than the other two methods, which is less apparent with increasing sample size. Nevertheless, across all conditions simulated, all three methods satisfy liberal Type I error robustness criteria stipulated by Bradley (1978).

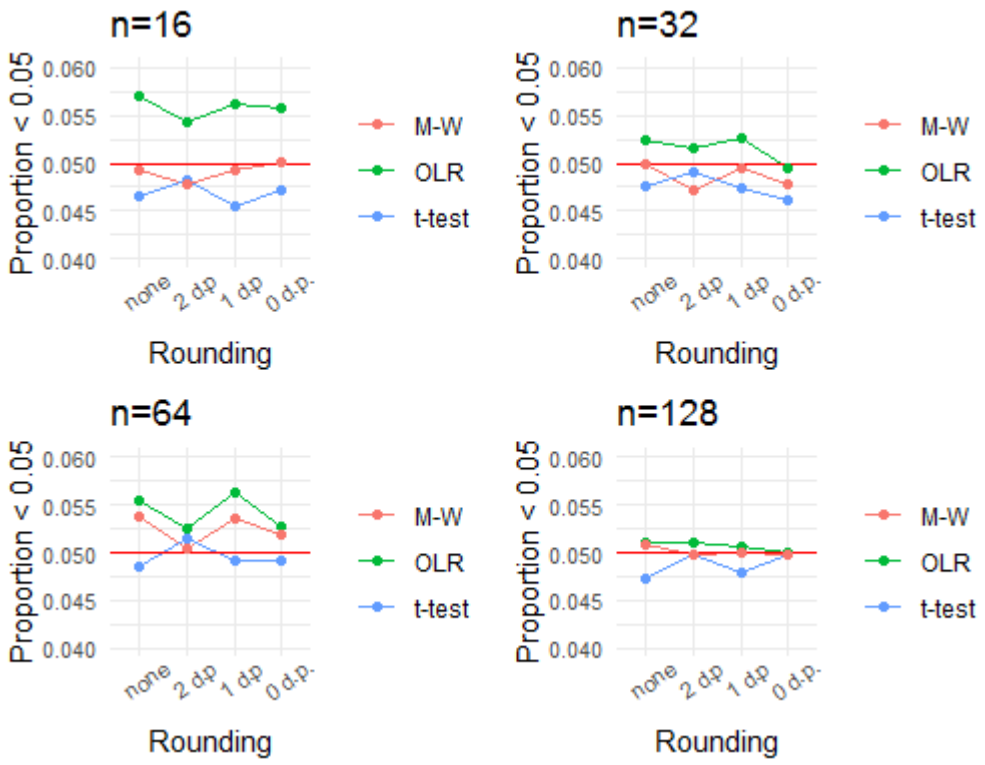

Figure 6. Type I error rates, Chi-square distribution with various levels of rounding.

Power

The Type I error robustness of all three methods determines that all are suitable providing n >= 16 per group and hence decisions on the best approach can be made in terms of power.

Figure 7 shows the proportions of occasions where the null hypothesis is rejected at the α=0.05 significance level with effect size *d*=0.5, thus representing the power of the competing tests. Under conditions of normality, there is little to separate the three competing methods and all three exhibit similar power properties. When the normality assumption is violated, OLR and M-W have superior power to the t-test, with OLR marginally outperforming M-W in the case where samples are taken from χ^2_2. Further detail can be seen in Table 4 and Table 5.

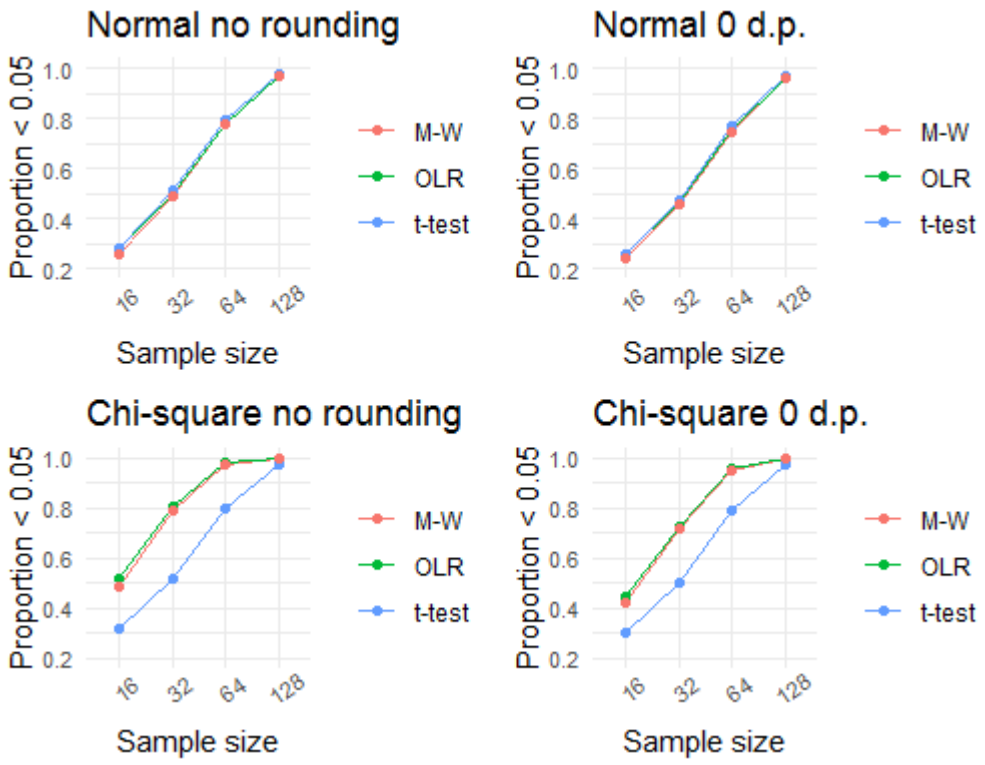

Figure 7. Power of the alternative methods for Normal distribution and Chi-square distribution.

Table 4. Null hypothesis rejection rates for the Normal distribution under H_1

	No rounding			0 decimal places		
	OLR	t-test	M-W	OLR	t-test	M-W
α=0.10, n=16	0.4011	0.3950	0.3740	0.3807	0.3766	0.3642
α=0.10, n=32	0.6248	0.6387	0.6122	0.5979	0.6065	0.5910
α=0.10, n=64	0.8618	0.8747	0.8593	0.8405	0.8497	0.8385
α=0.10, n=128	0.9886	0.9899	0.9886	0.9837	0.9861	0.9834
α=0.05, n=16	0.2810	0.2789	0.2579	0.2571	0.2590	0.2449
α=0.05, n=32	0.5007	0.5092	0.4865	0.4672	0.4768	0.4606
α=0.05, n=64	0.7763	0.7947	0.7731	0.7510	0.7652	0.7484
α=0.05, n=128	0.9726	0.9780	0.9725	0.9647	0.9691	0.9643
α=0.01, n=16	0.0996	0.1047	0.0948	0.0823	0.0958	0.0816
α=0.01, n=32	0.2567	0.2661	0.2527	0.2317	0.2439	0.2280
α=0.01, n=64	0.5536	0.5758	0.5464	0.5105	0.5289	0.5088
α=0.01, n=128	0.9024	0.9161	0.9010	0.8781	0.8925	0.8772

Table 5. Null hypothesis rejection rates for the Chi-square distribution under H_1

	No rounding			0 decimal places		
	OLR	t-test	M-W	OLR	t-test	M-W
α=0.10, n=16	0.6440	0.4287	0.6092	0.5721	0.4185	0.5529
α=0.10, n=32	0.8791	0.6428	0.8692	0.8279	0.6273	0.8199
α=0.10, n=64	0.9883	0.8766	0.9876	0.9770	0.8664	0.9761
α=0.10, n=128	0.9999	0.9881	0.9999	0.9997	0.9849	0.9997
α=0.05, n=16	0.5132	0.3149	0.4817	0.4415	0.3006	0.4184
α=0.05, n=32	0.8061	0.5199	0.7917	0.7257	0.5018	0.7161
α=0.05, n=64	0.9770	0.8008	0.9757	0.9542	0.7862	0.9517
α=0.05, n=128	0.9999	0.9746	0.9999	0.9993	0.9694	0.9992
α=0.01, n=16	0.2628	0.1283	0.2516	0.2014	0.1196	0.1883
α=0.01, n=32	0.5790	0.2892	0.5654	0.4790	0.2766	0.4653
α=0.01, n=64	0.9179	0.5942	0.9099	0.8549	0.5754	0.8505
α=0.01, n=128	0.9991	0.9123	0.9990	0.9957	0.9011	0.9955

Table 4 and Table 5 indicate that the relative power advantage is irrespective of the significance level reported.

Conclusion

The OLR approach maintains Type I error robustness for all of the scenarios within the simulation design, and hence it may be considered to be a valid a competitor to existing approaches particularly where the normality assumption is violated. Moreover, using the OLR approach as default means that there may be no requirement for researchers to test the normality assumption.

When the alternative hypothesis is true, and data is skewed, OLR is not inferior in power compared to the Mann Whitney test and superior in power compared to the Welch t-test. When data is normally distributed, any relative loss in power is minor when comparing OLR to the Welch t-test (subject to equal sample sizes >= 16 per group).

In general, OLR is capable of incorporating covariates. In general, OLR is capable of including more than one factor in modelling and can be extended to include interaction terms. This research paves the way to further research comparing the three methods in higher order designs and therefore opening the possibility of researchers having an extended toolkit.

References

Bradley, J.V., (1978) Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*(2), 144-152.

Brunner, E., & Munzel, U. (2000) The nonparametric Behrens-Fisher problem: Asymptotic theory and a small-sample approximation. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, Vol *42, No* 1, 17-25.

Delacre, M., Lakens, D. & Leys, C. (2017) Why psychologists should by default use Welch's t-test instead of Student's t-test, *International Review of Social Psychology*, 30 (1), 92- 101.

Delacre, M., Leys, C., Mora, Y. L., & Lakens, D. (2019) Taking parametric assumptions seriously: arguments for the use of Welch's *F*-test instead of the classical *F*-test in one-ay ANOVA. *International Review of Social Psychology*, *32*(1), 13.

Derrick, B., Dobson-Mckittrick, A., Toher, D., & White, P. (2015). Test statistics for comparing two proportions with partially overlapping samples. *Journal of Applied Quantitative Methods*, *10*(3).

Derrick, B., Ruck, A., Toher, D., & White, P. (2018). Tests for equality of variances between two samples which contain both paired observations and independent observations. *Journal of Applied Quantitative Methods*, *13*(2), 36-47.

Derrick, B., Toher, D., & White, P. (2016). Why Welch's test is Type I error robust. *The Quantitative Methods in Psychology*, *12*(1).

Elamir, E., & Sadeq, H. (2010). Ordinal regression to analyze employees' attitudes towards the application of total quality management. *Journal of Applied Quantitative Methods*, *5*(4).

Fagerland, M., and Sandvik, L. (2009a) Performance of five two-sample location tests for skewed distributions with unequal variances. *Contemporary clinical trials*, 30, 490-496.

Fagerland. M., and Sandvik, L. (2009b) The Wilcoxon –Mann- Whitney test under scrutiny. *Statistics in Medicine*, 28, 1487-1497.

Fagerland, M.W. (2012) M.W. T-tests, Non-parametric tests, and large studies—a paradox of statistical practice?, *BMC Medical Research Methodology*, 12(1), 78.

Fisher R A (1925). Applications of 'Student's' distribution. *Metron* 5, 90-104.

Hung, J.H.M., O'Neill, R.T., Bauer, P. & Kohne, K. (1997) The behavior of the p-value when the alternative hypothesis is true. *Biometrics*, Vol 53(1), 11-22.

Leys C & Schumann S (2010) A nonparametric method to analyse interactions: The adjusted rank transformation test, *Journal of Experimental Social Psychology*, 46(4), 684 – 688.

Lyhagen, J. (2021). Size and power of tests for dependency in regressions with ordinal variables. *Journal of Applied Quantitative Methods*, *16*(1).

Mann, H.B. & Whitney, D.R. (1947) On a test of whether one of two random variables Is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18, 50- 60.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *42*(2), 109-127.

Mirtagioğlu, H., Yiğit, S., Mendeş, E., & Mendeş, M. (2017). A Monte Carlo simulation study for comparing performances of some homogeneity of variances tests. *Journal of Applied Quantitative Methods*, *12*(1), 1-11.

Menard, S. (2010) Ordinal Logistic Regression. *Logistic Regression: From Introductory to Advanced Concepts and Applications*, Thousand Oaks, CA: SAGE Publications, 193-221.

Pearce, J., & Derrick, B. (2019). Preliminary Testing: The devil of statistics? *Reinvention: An International Journal of Undergraduate Research*, *12*(2)

Ruxton, G.D. (2006) The unequal variance t-test Is an underused alternative to Student's t-test and the Mann–Whitney U Test, *Behavioral Ecology*, 17, 28-35.

Scheirer CJ, Ray WS and Hare N (1976) The analysis of ranked data derived from completely randomized factorial designs, *Biometrics*, Vol 32, 429 – 434.

Stahl, S. (2006). The evolution of the normal distribution. *Mathematics magazine*, *79*(2), 96-113.

Quade D (1967) Rank analysis of covariance, *Journal of the American Statistical Association*, 62(320), 1187 – 1200.

Welch, B.L. (1947) The generalization of Student's problem when several different population variances are involved, *Biometrika*, 34, 28-35.

Wilcoxon, F. (1945) Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6), 80-83.

Yuen, K. K. (1974). The two-sample trimmed t for unequal population variances, *Biometrika* 61(1), 165–170.

Zimmerman, D. W. (1987). Comparative power of Student t test and Mann--Whitney U test for unequal sample sizes and variances. *The Journal of Experimental Education*, 55, 171-174.