Title:

A new statistical test for trends: establishing the properties of a test for repeated binomial observations on a set of items

Introduction

Many studies of therapies with single subjects involve testing with a set of items each of which can be right or wrong on two or more occasions. In some studies, performance is probed frequently, perhaps in each therapy session. These studies typically (though not invariably) emanate from North America and eschew statistical analysis, relying instead on visual inspection of the data to detect significant trends (Franklin, Gorman, Beasley, & Allison, 1997). The reason for this suspicion of statistics is that it is well-known that analysis of such data is severely compromised by autocorrelation: performance on the test on one day may be related to performance on the previous test occasion. There are methods that allow the analyst to estimate and adjust for such autocorrelation effects, but these require at least 100 test occasions to be feasible, but these data are in practice never available (Franklin et al., 1997). In a few studies visual analysis methods are supplemented by use of statistical tests known to have very serious shortcomings (e.g. Tryon's S see e.g. (Gorman & Allison, 1997))or have no clear statistical basis or interpretation (e.g. measures of effect size).

The other group of studies (from, mainly, Europe and Australia) tend to use much larger number of items but probe performance on these much less often (e.g. just before and after each therapy phase) and use statistical tests for significance usually from the Kendall family (e.g. McNemar's test).

This paper aims to show that a test derived from this family (i) keeps the type I error rate to the desired level irrespective of any serial dependency in the data, and (ii) provides a powerful test of ordinal trend. Simulations are used to explore how the power of the test How the test is affected by (i) the number of test items, (ii) the number of test occasions, and (iii) the degree of serial dependence in the data is investigated. Simulations explore (i) comparison of changes for two sets of items tested on the same occasions, and (ii) comparisons for one set tested over two phases (e.g. pre-therapy and during therapy).

Motivating the test.

Mann's test for trend calculates a statistic *S*: for each observation calculate the number of later observations that have a higher value minus the number with the lower value, and sum this across trials. Note that this requires only that trials are ordered, and makes no assumptions about spacing or linearity. The expected value of *S* across a series of trials under a null hypothesis of no trend is zero. Marascuilo & MacSweeney (1977)and Meddis (1984) extend this general concept to *t* tests with *n* items with a binomial response, but base their analyses on the permutational probabilities across trials assuming no serial dependence. Where there is serial dependence their analyses do not apply. However, whatever the degree of dependence, under the null hypothesis the mean value of *S* is zero with a symmetrical distribution, although its specific shape – and hence the standard deviation - depends on the degree of serial dependence. The calculation of *S* for an example dataset is illustrated in Box 1. Given that the expected value of *S* is zero with an unknown but

symmetrical distribution, its significance could be assessed with either a Wilcoxon one-sample test (that assumes only ordinality in the responses) or a t test (that assumes a near-normal distribution of mean S_i when the null hypothesis is true). Which of these tests performs better in the evaluation of mean S_i is tested in a simulation.

---------- Insert Box 1 about here ----------

Simulations: general method

The performance of this statistic in assessing a trend was investigated in a series of simulations. These varied the number of tests *t* (with values of 2, 5 and 8), the number of items (10, 25, 50, and 100) and the degree of consistency. Consistency was expressed as a variable *k* for the odds ration of correct performance comparing accuracy on trial *n* for items that were correct on trial *n-1* with accuracy for items that were incorrect on trial *n-1*. *k*=1 corresponds to no serial dependence (accuracy on trial *n* is independent from accuracy on the previous trial); larger values represent greater accuracy on trials where the previous item was correct than where it was incorrect. Varying the odds ratio was chosen because it guarantees that the probability correct for each item is always bounded by 0 and 1. *k* was normally varied from 1 to 10, 50 and 100.

Responses for individual items were generated randomly on the basis of these parameters. For the evaluation of the null hypothesis, i.e. there was no improvement, 100,000 pseudo-experiments were performed. Evaluation of power was based on 10,000 pseudo-experiments.

Whether the t statistic or the Wilcoxon one sample test, which with just two tests was implemented as McNemar's test was treated as an empirical issue. The Wilcoxon tests were evaluated with and without a continuity correction; an exact version of McNemar's test was used but the equivalent of a version without a continuity correction was generated by using a mid-*p* correction.

Note that the mean value of S_i is equal to the mean trend across sessions (i.e. the gradient of the best-fitting regression line) if mean S_i is divided by $(t^3 - t)/6$. The consequence is that (i) it is possible to calculate the mean rate of improvement over testing sessions, and (ii) confidence intervals on this mean rate of improvement, making no assumptions about linearity of improvement. This provides an ideal measure for the effectiveness of therapy suitable for meta-analysis. It has the advantage that (i) it can take into account the number of items involved in therapy (a 50% improvement with 10 items is less than a 50% improvement with 50 items) (ii) it is not affected by item-consistency among items during a baseline period, and (iii) it makes no assumptions about the linearity of effects.

Results

(i) Evaluation of a single trend

This deals with *n* items over *t* tests for detecting if there is significant improvement. Consistency is varied as described above.

When the null hypothesis is true, i.e. there is no improvement, the t test consistently outperforms the Wilcoxon one-sample test. In the remainder of this paper results from a t test on mean S_i will therefore be reported. As can be seen in Table 1, the type I error is

consistently maintained at around or below the 5% level (the 95% CI from the binomial theorem is .0516 to .0493). With small numbers of items, and especially with high consistency, however, the probability of a type I error falls well below the .05 level suggesting a real loss of power.

---------- Insert Table 1 about here ----------

Investigations of power show that while power increases sharply with increasing numbers of items (*n*) there is a much less substantial effect of the number of trials (*t*). When *k* – intertrial consistency - is greater than around 3, power tends to decrease with more items (see Figure 1 illustrating just extreme values of consistency).

(ii) Comparisons between two trends over the same period.

As should be obvious from the previous simulations, when performance with two sets of items tested over the same period are compared, using a two sample t test, the test performs well across all variations of the parameters. Power is, as with a single test, reduced with small *n*.

(iii) Comparison between trends over two periods.

When experiments seek to compare the rate of improvement over two periods (say pre-therapy vs therapy), any comparison needs to assume some kind of 'shape' for the rate of improvement over time. The obvious 'shape' to assume is a linear rate of improvement, but this is very difficult to motivate as a null hypothesis as it assumes that the test items are equally graded in difficulty for the subject under investigation. It is hard to see how this assumption can be motivated given that the factors affecting item difficulty may vary between individual subjects, and the gradient of difficulty of items in a test are not known (and how that is to be established is not clear).

Assuming that improvement is linear across the two phases, in simulations, the test performs reasonably well. Assuming other 'shapes' of improvement, however, - for example, a sigmoid function – results in a very substantially increased rate of type I error.

Conclusion

This development of Mann's and Marascuilo and MacSweeney's test offers a method that can statistically evaluate and place confidence intervals on the rate of improvement during therapy. It does not, unlike other tests, depend on stochastic independence between trials. The simulations show that there is only benefit from multiple probes during therapy when inter-item consistency is low. Power is increased a great deal by more items. Therapy studies will have much greater statistical power if they employ large numbers of items tested on few occasions.

References

Franklin, R. D., Gorman, B. S., Beasley, T. M., & Allison, D. B. (1997). Graphical display and visual analysis. In R. D. Franklin & D. B. Allison & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 119-158). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Gorman, B. S., & Allison, D. B. (1997). Statistical alternatives for single-case designs. In R. D. Franklin & D. B. Allison & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 159-214). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Marascuilo, L. A., & MacSweeney, M. (1977). *Nonparametric and distribution-free methods for the social sciences*. Monterey, Calif.. Brooks Cole.

Meddis, R. (1984). *Statistics using ranks*. Oxford: Oxford University Press.

Box 1. Calculation of the statistic.

There are n items tested on t tests. If item i on test j is correct x_{ij}=1 otherwise x_{ij}=0. The trials are weighted by λ coefficients with a spacing of 2 and summing to zero (more technically $\lambda_j = 2j - t - 1$). S for item i is the sum of the individual scores multiplied by the λ coefficients: $S_i = \sum_{i=1}^{t} x_{ij}\lambda_j$.

Worked example:

Trial j	1	2	3	4	
Weighting λ_j	-3	-1	1	3	S_i
Item i					
1	1	0	1	1	1
2	0	0	1	1	4
3	0	0	0	0	0
4	0	1	0	1	2
5	0	0	0	1	3
6	1	0	1	0	-2
7	0	0	1	0	1
8	1	1	1	1	0
9	0	1	1	0	0
10	1	0	1	1	1
Mean	0.4	0.3	0.7	0.6	
		Mean S_i			1
		Standard deviation S_i			1.70
		Standard error S_i			0.54
		t			1.86
		Degrees of freedom			9
		p (one tailed)			0.048
		Mean rate of improvement (items per session)			1.0

Table 1. The probability of type I error using a t test, varying item consistency (*k*), the number of tests and the number of items.

	Number of tests											
t	2				5				8			
k	1	10	50	100	1	10	50	100	1	10	50	100
Number of items												
10	0.0397	0.0167	0.0062	0.0039	0.0324	0.0204	0.0115	0.0075	0.0480	0.0385	0.0206	0.0135
25	0.0520	0.0471	0.0289	0.0186	0.0518	0.0485	0.0427	0.0373	0.0516	0.0505	0.0509	0.0470
50	0.0511	0.0486	0.0506	0.0450	0.0509	0.0504	0.0505	0.0493	0.0506	0.0517	0.0508	0.0515
100	0.0486	0.0519	0.0496	0.0495	0.0516	0.0508	0.0506	0.0504	0.0509	0.0501	0.0506	0.0498

Figure 1.Illustrating that with high item-consistency power decreases with more tests.