ABSTRACT

Title of Document: **BASELINE ADJUSTMENT FOR ORDINAL COVARIATES BY INDUCING A PARTIAL ORDERING IN RANDOMIZED CLINICAL TRIALS.**

Na An, Master of Arts, 2007

Directed By: Professor Paul J. Smith
Department of Mathematics

In two-armed randomized clinical trials (RCTs) designed to compare a new treatment with a control, a key endpoint is often measured and analyzed both at baseline and after treatment for two groups. More powerful and precise statistical inferences are possible once the between-group comparisons have been adjusted for covariates. In this thesis we propose a new method for ordered categorical outcomes which adjusts for baseline without relying on any specific assumptions on the data generating process. Based on baseline and post-treatment values, data are composed of counts of patients who have improved from one category to another, stayed the same or deteriorated. Not all patterns are comparable. Hence, the ordering is only partial. We develop an approach to test the treatment effects based on comparing each observation in one group to each observation in the other group to which it is comparable. The power comparisons of this test with four common approaches are conducted in our simulation study.

BASELINE ADJUSTMENT FOR ORDINAL COVARIATES BY INDUCING A PARTIAL ORDERING IN RANDOMIZED CLINICAL TRIALS.

By

Na An

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Master of Arts
2007

Advisory Committee:
Professor Paul J. Smith, Chair
Professor Benjamin Kedem
Professor Abram M. Kagan

Acknowledgements

First, I am grateful to Dr. Vance W. Berger at NIH for suggesting the topic and providing the necessary resources for the thesis.

I also would like to thank Professor Benjamin Kedem and Professor Abram M. Kagan for agreeing to serve on my committee on a very short notice.

I would like to thank my parents and my husband for their extended support and encouragement throughout my entire graduate program.

Foremost, I am indebted to my advisor Professor Paul J. Smith for helping me in many fundamental ways and providing some valuable comments. I truly appreciate his guidance, suggestions and patience during my numerous mishaps. His knowledge and encouragement helped me all the time in research for and writing of this thesis.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Many randomized trials involve measuring an ordinal outcome at baseline and after treatment to determine the effectiveness of treatment. For example, in the simplest pretest-posttest designs (only one measurement is made after treatment), consider the evaluation of an endovascular approach relative to standard procedure for the treatment of abdominal aortic aneurysm. Each patient condition may be classified as good (G), fair (F), serious (S) or critical (C). After treating the patient for a period of time, their health conditions are again rated on same scale from good to critical. The purpose of such clinical trials is to assess the effectiveness of a new treatment relative to a standard control approach in improving the state of patients, or in reducing the magnitude of deterioration.

Adjusting between-group comparisons for covariates often improves the analysis (Senn, 1989). The most common approaches to adjust for an ordinal covariate seem to be treating it as binary, nominal, or continuous.

When the covariate is binary or nominal, the adjustment generally consists of comparing outcomes across treatment groups, within each level of the covariate. One typical nonparametric test is Fisher's exact test, which combines categories to create a 2 × 2 table to test homogeneity of each outcome probability among the rows. Moses, Emerson, and Hosseini (1984) and Zimmermann (1993) cited this common practice as inefficient because ignoring the ordering among the categories or collapsing categories will result in a loss of power.

To exploit the ordering, numerical scores may be assigned to the ordered categories, and simply subtract baseline values from post-treatment values. The primary response variable is then the change on the pain scale from baseline. Thus, we have a single vector-valued endpoint which captures both baseline and subsequent pain measurements. When the choice of scores is not apparent, integer (equally spaced) scores are often assigned. Berger and Ivanova (2001) showed this practice generally leads to unnecessarily conservative tests.

By treating the ordinal response variable as continuous, we can use the analysis of covariance (ANCOVA) with the post-treatment value as the response variable and the baseline values as the covariate (Maurer and Commences, 1988; Laird and Wang, 1990). ANCOVA is a merger of ANOVA and regression for continuous variables. ANCOVA tests whether certain factors have an effect after removing the variance for which quantitative predictors (covariates) account. The inclusion of covariates can increase statistical power because it accounts for some of the variability.

Another method to adjust for baseline is to resort to ordinal regression models which utilize the ordinal nature of the data by describing various models of stochastic ordering and thus eliminating the need of assigning scores. The most widely used model in ordinal regression is the cumulative logit model which models cumulative logits by combining the probability of the event and all events that are ordered before it. This model has a complete set of parameter estimates for each cumulative logit (that is, multiple intercepts and multiple estimates for each predictor). A popular submodel of the cumulative logit models is the proportional odds model (see Agresti, 1990). The model assumes that the odds of responses below a given response level are constant regardless of the level you pick. The proportional odds model plays an immensely important role in the practical application of analysis of categorical data. Readers interested in further details are referred to McCullagh, 1980 and Agresti, 1990. However, compared with design-based non-parametric tests, regression based tests are less transparent in terms of interpretation and inference. Also, regression based methods may not be appropriate when the model does not fit the data.

In this thesis, we explore a new nonparametric method to adjust for baseline which does not rely on any assumptions. Specifically, we consider the information-preserving composite endpoint (Berger, 2002), which consists of the pair of values for each patient, one at baseline and one after treatment, and determine which of these patterns indicate the most improvement. It will turn out that some pairs cannot be ranked above, equivalent to or below others, resulting in only a partial ordering. To the extent that pairs of categories, and therefore pairs of observations, are comparable, the experiment is still informative. We exploit the information that is present to compute a modified U-statistic (Serfling, 1988).

In Chapter 2 we illustrate, through a series of examples, some situations in which partial ordering arise in RCTs. In Chapter 3 we present several methods for adjusting for an ordinal baseline variable, and explore the partial ordering on the outcome levels induced by each. In Chapter 4 we develop an exact approach to between-group analysis adjusting for ordinal baseline covariates (Berger, 2004) based on the partial ordering discussed in Chapter 3. Three traditional methods for categorical data analysis (Fisher’s exact test, ANCOVA, proportional odds regression) are introduced in Chapter 5, and we conduct a series of simulations to compare these conventional tests with our proposed procedure in term of unconditional power. The results are summarized and discussed in Chapter 6.

Chapter 2: Partial ordering in clinical trials

In this section we define partial orderings and illustrate, through a series of examples, how they may arise in RCTs. The partial ordering is defined mathematically as a mapping on the product space of the elements of a set into the space {>, <, =, ≠}, where a ≠ b indicates that a and b are not comparable, or that none of a < b, a = b, or a > b would be accurate. For example, if the set is 1, 2, A, B, then there are six pairs of elements, and one partial ordering on this set might be 1 < 2, 1 ≠ A, 1 ≠ B, 2 ≠ A, 2 ≠ B, and A < B. Any partial ordering satisfies reflexivity (a = a), and antisymmetry (if a > b, then b < a; if a = b, then b = a; if a < b, then b > a; if a ≠ b, then b ≠ a). In addition, a proper partial ordering will satisfy the property of transitivity, so that a > b > c implies that a > c (Kolmogorov and Fomin, 1970). Partial orderings can arise naturally in a variety of settings within the general guise of RCTs. In the remainder of this section we illustrate the diversity of RCT situations which result in partial orderings.

Example 2.1 (Partially Ordered Sample Space with a Completely Ordered Endpoint)

Suppose that two patients are randomized to each of the experimental treatment E and the standard of care control S, and suppose further that the primary efficacy endpoint is trichotomous, with three completely ordered outcome levels. For example, these outcome levels may be cure (C), improvement (I), or failure (F) in the evaluation of pneumonia, or other disease. Even though these three outcome levels are completely ordered (C > I > F), the permutation sample space is only partially ordered because the endpoint is ordinal but not interval. To see this, suppose that the 2 × 3 contingency table (by convention, we list the S row first, then the E row, separated by a semi-colon, with columns separated by a comma and listed in order of increasing benefit, or F, I, C) is observed to be (1, 1, 0; 0, 1, 1), indicating that in the S group there was one F and one I, while in the E group there was one I and one C. For simplicity, we may also write this as (F, I; I, C). The permutation sample space is the set of 2 × 3 contingency tables that preserve the row margin (2, 2) and the column margin (1, 2, 1). With these fixed margins, there are two degrees of freedom, so may denote a 2 × 3 contingency table (viewed as a point in the permutation sample space) by only the first two elements. The observed data are then considered as (1, 1). The other points of the sample space are (0, 1) = (I, C; F, I), (0, 2) = (I, I; F, C), and (1, 0) = (F, C; I, I). Clearly, (F, I; I, C) provides the most evidence that E is superior to S, and (I, C; F, I) provides the least. But it sis not clear how (I, I; F, C) and (F, C; I, I) compare to each other without making judgments concerning the relative spacing among C, I, and F. That means it is hard to compare two improved patients to one cured patient and another patient with no improvement.

Example 2.2 (Multivariate response with ordinal margin)

Stevens (1951) distinguishes the classification of scale types as nominal, ordinal, interval and ratio scales. However this list is incomplete since only a partial order may exist among the categories. More complex order structure arises when a bivariate or a multivariate response is observed even though the categories for each margin are ordinal. For instance, consider two binary endpoints, y1 and y2, each scored as 0 and 1, with 1 corresponding to the better outcomes. We may consider the pair (y1, y2) as a single vector-valued endpoint, and each patient may be classified as (0, 0), (0, 1), (1, 0), or (1, 1). It is clear that (1, 1) > (1, 0) > (0, 0) and (1, 1) > (0, 1) > (0, 0), but (1, 0) ≠ (0, 1), which results in a partial order.

Example 2.3 (Censored Data)

Consider survival data with right-censoring. The usual complete ordering on uncensored observations still holds. That is, death at nine months is better than death at six months (9 > 6). It remains to compare censored observations to censored and uncensored observations. Obviously, equality holds if and only if both the time and the censoring indicator are common to the two observations. It seems reasonable to define the censored observation to be greater than the uncensored one if and only if its time is equal to or greater than the time of the uncensored one (11+ > 11, 6+ > 1). If the time of the censored observation is less than the time of the uncensored one, there is no way to compare these quantities. For example, if we were to try to compare 6+ to 8, then without assuming some sort of model which enable us to estimate the actual time of death of the patient whose survival time was right-censored at six months, we would only conclude that 6+ ≠ 8. Two observations with different censoring times may or may not be considered comparable, e.g., 6+ < 10+ or 6+ ≠ 10+.

Example 2.4 (Missing Data)

Consider a phase III clinical trial with missing data, where each patient might be classified on their final result as missing, failure or success. We can consider missing as better than failure but worse than success, or we can just consider that it is non-comparable to either one.

Example 2.5 (Adjustment for Ordinal Baseline)

Consider the evaluation of a new therapy for functional gastro-intestinal disorder. Each patient may be classified based on pain as disabling (D), severe (S), moderate (M), mild (L), slight (T), or none (N). Obviously, these six outcome levels are completely ordered, but they are different from the outcomes in Example 2.1. These outcomes represent a point in time, and not change, so the baseline value needs to be considered. Suppose that to enter the study a patient would need to be in one of the four categories D, S, M, or L. At the end of the study, the patient can be in any of the six states. Then we have a single vector-valued endpoint which captures both baseline and subsequent pain measurements (Berger, 2002), with $4 \times 6 = 24$ partially ordered outcome levels, as we will study in detail in Chapter 3. This study is precisely the kind of problem that motivated this research.

Chapter 3: Adjusting for an ordinal baseline variable by inducing partial ordering

Based on the study described in Example 2.5, the development of partial orderings on the 24 categories is informative. If we ignore the comparison of a given category to itself, then there are *IJ (IJ* – 1)/2 pairs of distinct categories for an *I* × *J* contingency table, or, with 24 categories, 24(23)/2 = 276 pairs of distinct categories. In this section we present several methods for adjusting for an ordinal baseline variable, and we can actually linearly order these partial orderings by how many of the pairs of categories they treat as comparable. This is important, because comparative information derives from comparisons of categories. Hence, a partial ordering that compares more pairs of categories will provide a more informative analysis. However, as we will see, there is a danger in pretending that certain categories can be compared when in fact they cannot. We first present the partial ordering for the specific case of the 4 × 6 contingency table (Table 1), then generalize to an *I* × *J* contingency table. We remark that the orderings are based not on the perspective of the patient, who would regard as best starting at L and ending at N, but rather from the perspective of the evaluation of the medical intervention. This being the case, the most clinical benefit derives then for the (D, N) pattern.

Table 1: The 4 × 6 contingency table of Example 2.5

Baseline Pain	Post Treatment Pain Assessment					
	D	S	M	L	T	N
D	(D, D)	(D, S)	(D, M)	(D, L)	(D, T)	(D, N)
S	(S, D)	(S, S)	(S, M)	(S, L)	(S, T)	(S, N)
M	(M, D)	(M, S)	(M, M)	(M, L)	(M, T)	(M, N)
L	(L, D)	(L, S)	(L, M)	(L, L)	(L, T)	(L, N)

3.1 Stratify by Baseline

The idea behind stratifying for baseline is that two categories are comparable only if they have the same first component (baseline), or are in the same row of Table 1. Now each category is comparable to five other categories, resulting in (4 × 6 × 5)/2 = 60 comparable pairs of categories out of 24(23)/2 = 276 pairs of categories. In general, with an $I \times J$ contingency table, each row category would be comparable to $J - 1$ categories, and the number of comparable pairs of categories would be $IJ\ (J - 1)/2$, out of $IJ\ (IJ - 1)/2$ pairs of categories. Obviously, this is a sparse partial ordering, which is tantamount to treating baseline as a nominal variable (when in fact it is ordinal), and does not treat as comparable the categories (D, N) and (L, D), even though the former represents improvement from disabling pain to no pain and the latter represents degradation from mild pain to disabling pain.

3.2 Forward and Backward Stratification

With forward or backward stratification, two outcome levels are comparable only if they have the same first (baseline) or second (post-treatment) component, or are in either the same row or the same column of Table 1. Now each category is comparable to 5 + 3 = 8 other categories, resulting in (4 × 6 × 8)/2 = 96 comparable pairs of categories out of 276 pairs of categories. In general, with an $I \times J$ contingency table, each category would be comparable to $(J -1) + (I - 1)$ categories, and the number of comparable pairs of categories would be $IJ\ (J + I - 2)/2$, out of *(IJ*

– 1)/2 pairs of categories. This partial ordering is still sparse, and still considers (D, N) ≠ (L, D).

3.3 Enrichment

One can enrich the partial ordering of Section 3.2 by making it transitive. Thus, if (D, N) > (D, D), which is meaningful because going from disabling pain to no pain reflects better on the treatment than starting with disabling pain and remaining with disabling pain, and if (D, D) > (L, D), which is also meaningful because starting with disabling pain and remaining with disabling pain reflects better on the treatment than going from mild pain to disabling pain, then it is only reasonable that (D, N) > (L, D). Define two categories as comparable if one dominates a category that dominates the other. Any category is then comparable to any other category Northeast or Southwest of it (Table 1). To find the total number of comparable pairs of categories, consider the four cells (categories), at which a pair of rows and a pair of columns intersect (Diaconis and Sturmfels, 1998). This gives 4!/[(2!)(2!)] = 6 pairs of cells, of which five (all but the upper-left vs. the lower-right) are comparable. As there are 4!/[(2!)(2!)] = 6 pairs of rows, and 6!/[(2!)(4!)] = 15 pairs of columns (Table 1), there are 6 × 15 = 90 pairs of non-comparable categories, and 276 – 90 = 186 pairs of comparable categories. In general, with an $I \times J$ contingency table, there would be $I!/[(2!)(I - 2)!]$ pairs of rows and $J!/[(2!)(J - 2)!]$ pairs of columns, or $I!J!/[(2!)(I – 2)!(2!)(J – 2)!]$ pairs of non-comparable categories.

An equivalent derivation is to start with the $IJ\ (I + J - 2)/2$ from Section 3.2, and then recognized that symmetry half of the remaining $[IJ\ (IJ – 1) – IJ\ (I + J -2)]/2$ pairs of categories are comparable, and the other half are not. Yet a third derivation, which is also instructive, comes from using the five comparable pairs of categories from each of the $\mathrm{I}!J!/[(2!)(I – 2)!(2!)(J – 2)!]$ pairs of rows and columns and then subtracting away the over count, which is $IJ\ [(I – 2)(J – 1) + (J -2)(I – 1)]/2$. This is evident because each categories is compared to each of the other $(J – 1)$ categories in its row $(I – 1)$ times instead of once, and each category is compared to each of the other $(I – 1)$ categories in its column $(J – 1)$ times instead of once.

3.4 Direction of Effect

The aforementioned partial orderings do not compare come improvement categories, such as (L, N), to some worsening categories, such as (M, D). If both dimensions are measured on the same scale, then one can enrich the partial ordering by considering as comparable pairs of categories which differ in the direction of effect. For instance, (M, D) < (M, M) < (L, N). To find the number of comparable categories, consider rows r_1 and $\mathrm{r}_2 > \mathrm{r}_1$, columns c_1 and $\mathrm{c}_2 > \mathrm{c}_1$, such that they are not interweaving, i.e., the two columns are either within the interval of the two rows or outside the interval. Mathematically, $\mathrm{r}_1 \leq \mathrm{c}_1 < \mathrm{c}_2 \leq \mathrm{r}_2$ or $\mathrm{c}_1 \leq \mathrm{r}_1 < \mathrm{r}_2 \leq \mathrm{c}_2$, and both equalities cannot hold at the same time. These two pairs will intersect at four cells, which give six pairs of cells. All of these are comparable (the upper left vs. lower right is also comparable since one is above or on the diagonal and the other is below or one the diagonal, but they are not on the diagonal at the same time). Obviously, the number of ways to choose the columns from outside the interval (r$_1$, r$_2$) is:

$$(r_1 - 1)(J - r_2).$$

The number of ways to choose columns inside the interval (r$_1$, r$_2$), where $I_{(r_2 - r_1 - 1)}$ is an index function is:

$$(r_2 - r_1 - 1)(r_2 - r_1 - 2)I_{(r_2 - r_1 - 1)} / 2.$$

The number of ways to choose one at the endpoint and the other outside the interval is:

$$(J - r_2) + (r_1 - 1).$$

The number of ways to choose one at the endpoint and the other inside the interval is:

$$2(r_2 - r_1 - 1).$$

So for a fixed pair of rows, there are:

$$K(r_1, r_2) = (r_1 - 1)(J - r_2) + (r_2 - r_1 - 1)(r_2 - r_1 - 2)I_{(r_2 - r_1 - 1)} / 2 + (J - r_2) + (r_1 - 1) + 2(r_2 - r_1 - 1)$$

ways to choose a pair of columns such that they intersect at four cells, of which a total of six pairs are comparable. Hence, the total number of non-comparable pairs is:

$$\sum_{(r_1, r_2)} (J(J-1)/2 - K(r_1, r_2)),$$

where the sum is over all possible pairs of rows (r$_2$ > r$_1$).

In our example of pain, I = 4, J = 6, and there are six possible pairs of rows. We find that:

$$\begin{aligned} K(1,2) &= (1-1)(6-2) + (2-1-1)(2-1-2)I_{(2-1-1)} / 2 + (6-2) + (1-1) + 2(2-1-1) \\ &= 0 + 0 + 4 + 0 + 0 = 4, \end{aligned}$$

$K(1,3) = 0 + 0 + (6 - 3) + 0 + 2(3 - 1 - 1) = 5$,

$K(1,4) = 0 + (4 - 1 - 1)(4 - 1 - 2)/2 + (6 - 4) + 0 + 2(4 - 1 - 1) = 7$,

$K(2,3) = (2 - 1)(6 - 3) + 0 + (6 - 3) + (2 - 1) + 0 = 7$,

$K(2,4) = (2 - 1)(6 - 4) + 0 + (6 - 4) + (2 - 1) + 2(4 - 2 - 1) = 7$,

$K(3,4) = (3 - 1)(6 - 4) + 0 + (6 - 4) + (3 - 1) + 0 = 8$.

So the total number of non-comparable pairs is (15)(6) – (4 + 5 + 7 + 7 + 7 + 8) = 52, and the number of comparable pairs is 276 – 52 = 224.

3.5 Compare Non-change

One additional modification is to consider the non-change categories as comparable. None of the previously discusses partial orderings would consider (D, D) comparable to (S, S), for example. It is not entirely clear how these categories are to be compared. One might argue that all of these categories represent no change, so they are all equivalent. However, one could also argue that more baseline pain means more room (and need) for improvement, so that (L, L) > (M, M) > (S, S) > (D, D). The opposite ranking would result if one were to take the view that the healthier the patient is to start with, the easier it is to improve. It is not our intention to resolve this issue, but rather to point out that these categories may or may not be considered comparable. If they are, then there are $k!/[(2!)(k - 2)!]$ fewer pairs of non-comparable cells than in Section 3.4, where $k = \min(I, J)$. When $I = 4$ and $J = 6$, $k = 4$, and there are 46 pairs of non-comparable categories, and 230 comparable pairs of categories.

3.6 Sort by Change

In Section 3.4 the main diagonal (representing no change) was used as a line of demarcation to separate improvement from deterioration. Other diagonals can be used the same way. Using all diagonals parallel to the main diagonal in this way, and equating all cells within a given diagonal, is tantamount to assigning equally-spaced scores assigned to the six pain evaluations (say D = 1, S = 2, M = 3, L = 4, T = 5, and N = 6), and then ranking the categories by the change from baseline (delta). This would be a complete (and obviously transitive) ordering which would consider all 276 pairs of categories as comparable. However, the relative spacings among categories measured on an ordinal but not an interval scale are unknown (Bajorski and Petkau, 1999), so it is artificial to compare overlapping changes unless one set contains the other. The comparison of pairs of categories not considered comparable by the partial ordering in Section 3.5, e.g., (D, S) and (S, M), provides only pseudo-information.

Chapter 4: An exact approach based on partial ordering

In this section we will develop our methodology for constructing an exact permutation test based on partial ordering of the baseline outcome pairs of categories. We will give the definition of the test statistic and technical details of how to compute p-values and power of our test. Theoretically, similar to Fisher's exact test, this approach can be explained as follows. Enumerate all possible tables consistent with the given margins, and calculate the statistic value of each. The significance value (p-value) of the observed table is then the percentage of those test statistics which are no less than the observed one. One real life sample is presented for better illustration. Furthermore, in order to extend the bounds of feasibility of our exact procedure for practical use, we explore an efficient algorithm which finds the approximate significance level of an $I \times J$ contingency table without enumerating all possible tables.

4.1 Methodology

The idea of our approach is that although baseline-outcome pairs cannot be ordered completely, a partial ordering can still be obtained based on the relationships defined in Chapter 3. Our analysis is based on the partial ordering presented in Section 3.5, equating all no-change categories. Then an exact permutation test statistic can be defined and the null reference distribution function can be derived.

The statistical analysis will be dictated by the design to be a two-armed parallel RCT with 1:1 randomization to each of the experimental arm E and the standard of care control arm S, and the partial ordering. However, a word of caution is required here that a philosophical decision needs to be reached prior to performing the analysis. It is desirable to settle whether to treat one category as better than another if it does not necessarily reflect superiority of E to S. An extreme example is given below to clarify this issue.

Suppose that there are 100 patients randomized to each of E and S. Consider that each patient randomized to E has outcome (D, N), and each patient randomized to S has outcome (L, N). Then every patient on each arm leaves the study pain-free. The difference in outcomes is actually a difference only in the baseline component of the outcomes, one of which is 100% disabling pain and the other is 100% slight pain. Obviously, going from D to N is better than going from L to N, as discussed in Section 3.2. But in the evaluation of one treatment relative to another, can this superiority be explained by the difference in treatments? Unless there is selection bias (Berger and Exner, 1999), randomization ensures that the baseline distribution within each arm is necessarily the same, so the observed difference must be a random occurrence (Senn, 1994), and the apparent superiority may not be attributable to the treatments. This situation can be avoided by stratifying for the baseline pain score in the design. So we consider this to be a philosophical issue and not a statistical one.

Once the issues of partial ordering have been settled, then under the hypothesis that active treatment is about the same or superior to control treatment, we want to test the hypotheses:

$$H_0 : Active = Control\,,$$

$$H_A : Active > Control\,.$$

Suppose n_1 patients have been given the active treatment and n_2 patients have been given the control treatment. The outcomes are from an *I* × *J* contingency table. In a two-armed RCT, the data structure is a 2 × *IJ* contingency table, with two rows (one for each of the active and control treatments) and *IJ* columns, with some partial ordering on these *IJ* columns. The row margins are n_1 and n_2 respectively. Given the two samples and in absence of any further assumption about the samples, the modified U-statistic is the ratio of pairs favorable to the active group to the total number of informative pairs (pairs that are favorable to one of the two groups), i.e.

$$T = \frac{\#\ of\ \ pairs\ \ favorable\ to\ \ the\ active\ group}{Total\ \#\ of\ \ comparable\ pairs\ \ favorable\ to\ the\ active\ or\ control\ group}\,.$$

This is an estimator of $\theta = \dfrac{P(A > C)}{P(A > C) + P(C > A)}$, where $P(A > C)$ is the probability that the active treatment will produce the better outcome and the control will produce the worse outcome. While we deal with the ties differently from Munzel and Tamhane (2002), we retain the expression “tendentiously larger” for the active group if $\theta > 0.5$, or for the control group if $\theta < 0.5$. Under the null hypothesis the cell probabilities are common to both groups, so $\theta = 0.5$.

In order to efficiently compute the test statistic based on partial ordering, one needs to keep track of the set of comparable cells for each cell in contingency table. To this end, a comparison matrix M can be defined in order to calculate the newly defined test statistic. For a 2 × 9 contingency table with 3 baseline and 3 outcome categories ranging from 1 = best to 3 = worst, partial ordering has a comparison pattern of Table 2:

Table 2: Comparison pattern

	Active								
Control	(1,1)	(1,2)	(1,3)	(2,1)	(2,2)	(2,3)	(3,1)	(3,2)	(3,3)
(1,1)	=	C	C	A	=	C	A	A	=
(1,2)	A	=	C	A	A	≠	A	A	A
(1,3)	A	A	=	A	A	A	A	A	A
(2,1)	C	C	C	=	C	C	A	≠	C
(2,2)	=	C	C	A	=	C	A	A	=
(2,3)	A	≠	C	A	A	=	A	A	A
(3,1)	C	C	C	C	C	C	=	C	C
(3,2)	C	C	C	≠	C	C	A	=	C
(3,3)	=	C	C	A	=	C	A	A	=

In Table 2, A or C means active or control treatment is favored by this comparison, “=” means equal treatment effect and “≠” represents non-comparable pairs. Based on Table 2, the comparison matrix M (33 by 33) is defined as follows:

$$\underline{M}_{33\times33}=\begin{pmatrix} \underset{\sim}{\underline{0}_{10\times10}}, & \underline{0}_{3\times3}, & \underline{0}_{7\times7}, & \underline{0}_{3\times3}, & \underline{0}_{7\times7}, & \underline{0}_{3\times3} \\ \underline{0}_{3\times10}, & \underline{A_1}_{3\times3}, & \underline{0}_{3\times7}, & \underline{A_2}_{3\times3}, & \underline{0}_{3\times7}, & \underline{A_3}_{3\times3} \\ \underline{0}_{7\times10}, & \underline{0}_{7\times3}, & \underline{0}_{7\times7}, & \underline{0}_{7\times3}, & \underline{0}_{7\times7}, & \underline{0}_{7\times3} \\ \underline{0}_{3\times10}, & \underline{A_4}_{3\times3}, & \underline{0}_{3\times7}, & \underline{A_1}_{3\times3}, & \underline{0}_{3\times7}, & \underline{A_2}_{3\times3} \\ \underline{0}_{7\times10}, & \underline{0}_{7\times3}, & \underline{0}_{7\times7}, & \underline{0}_{7\times3}, & \underline{0}_{7\times7}, & \underline{0}_{7\times3} \\ \underline{0}_{3\times10}, & \underline{A_5}_{3\times3}, & \underline{0}_{3\times7}, & \underline{A_4}_{3\times3}, & \underline{0}_{3\times7}, & \underline{A_1}_{3\times3} \end{pmatrix},$$

where

$$\underline{A_1}_{3\times3} = \begin{pmatrix} 0 & 1 & 1 \\ -1 & 0 & 1 \\ -1 & -1 & 0 \end{pmatrix}, \qquad \underline{A_2}_{3\times3} = \begin{pmatrix} -1 & 0 & 1 \\ -1 & -1 & 0 \\ -1 & -1 & -1 \end{pmatrix}, \qquad \underline{A_3}_{3\times3} = \begin{pmatrix} -1 & -1 & 0 \\ -1 & -1 & -1 \\ -1 & -1 & -1 \end{pmatrix},$$

$$\underline{A_4}_{3\times3} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ -1 & 0 & 1 \end{pmatrix}, \qquad \underline{A_5}_{3\times3} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix}.$$

As we mentioned above, if the active group has n$_1$ patients with the observed pairs of values $\{(a_1^0,\, a_1^1),\, (a_2^0,\, a_2^1),\, \ldots,\, (a_{n_1}^0,\, a_{n_1}^1)\}$ and the control group has n$_2$ patients with the observed pairs $\{(c_1^0,\, c_1^1),\, (c_2^0,\, c_2^1),\, \ldots,\, (c_{n_2}^0,\, c_{n_2}^1)\}$, and the category levels are less than 9 (this is normal in the practical research), we can rewrite these pairs as $\{a_1^0 a_1^1,\, a_2^0 a_2^1,\, \ldots,\, a_{n_1}^0 a_{n_1}^1\}$ and $\{c_1^0 c_1^1,\, c_2^0 c_2^1,\, \ldots,\, c_{n_2}^0 c_{n_2}^1\}$ based on the formula:

Baseline × 10 + Post-treatment.

To obtain all possible pair combinations between two treatment groups, the observed combination matrix X has the following form:

$$\underline{X}_{n_1 n_2\times2} = \begin{pmatrix} \underline{B_1}_{n_2\times2} \\ \underline{B_2}_{n_2\times2} \\ \cdots \\ \cdots \\ \underline{B_{n_1}}_{n_2\times2} \end{pmatrix},$$

where

$$\underline{B_i}_{n_2 \times 2} = \begin{pmatrix} a_i^0 a_i^1 \ , & c_1^0 c_1^1 \\ a_i^0 a_i^1 \ , & c_2^0 c_2^1 \\ \cdots \ , & \cdots \\ \cdots \ , & \cdots \\ a_i^0 a_i^1 \ , & c_{n_2}^0 c_{n_2}^1 \end{pmatrix}_{n_2 \times 2} .$$

The two components of each row, for example, $[a_1^0 a_1^1,\, c_2^0 c_2^1]$ decides the position in comparison matrix M , and the corresponding values (1, -1 or 0) in matrix M give us the comparison result of this pair of observed values based on the partial orderings. Therefore, we can easily calculate the total number of comparable pairs which are favorable to the active or control treatment by using matrix M and X together, i.e., the counts of -1 and 1 represent the number of pairs favorable to active and control group respectively. Thus, the test statistic can be easily calculated using the observed data. We have developed the S-Plus code, which includes building the comparison matrix M and the observed combination matrix X , and providing the value of the test statistic T .

4.2 Conditional P-value

In this section we discuss regarding computations of p-values of the tests we have proposed. Exact calculation of the conditional p-value requires enumerating all possible tables under fixed row and column margins. The immediate difficulty in exact calculation is that the required computation can very easily grow beyond the capacity of even modern computers. The sample space can very quickly grow to be something that limits implementation of any exhaustive procedures. Next we provide technical details for computation of the p-values.

4.2.1 Exact conditional p-value

One usually looks at a conditional sample space where the entries are conditioned on the margins of the contingency table. Because the marginals are sufficient statistics the conditional inference is optimal. Under the null hypothesis of no association between row and column categories, the probability of the sampled $r \times c$ table with total sample size N is:

$$P = \frac{\prod_{i=1}^{r} x_{i.}! \prod_{j=1}^{c} x_{.j}!}{N!\,(\prod_{i=1}^{r} \prod_{j=1}^{c} x_{ij}!)} .$$

Recall that a p-value is the probability of the observed data or more extreme data occurring under the null hypothesis, thus, for $2 \times k$ contingency table in two-armed CRTs, the conditional probability of obtaining a test statistic that is same as or more extreme than the observed one under the null hypothesis (i.e. conditional p-value) is:

$$P(T \geq t \mid c) = \sum_{x \in \Gamma'_{c(t)}} P(X = x \mid H_0, c)$$

$$= \sum_{x \in \Gamma'_{c(t)}} \frac{\dfrac{1}{\prod_{j=1}^{k} x_{1j}!\,(t_j - x_{1j})!}}{\sum_{\underline{X} \in \Gamma_c} \dfrac{1}{\prod_{j=1}^{k} x_{1j}!\,(t_j - x_{1j})!}},$$

where the rejection region is the set $\Gamma'_{c(t)} = \{x \in \Gamma_c : T \geq t_{obs}\}$, Γ_c is the set of all tables $\underline{X}$ with the marginal fixed at c. $c = [c_1, c_2, \cdots, c_k]$ is the vector of column marginal counts. For size α level, the critical region is

$$t_\alpha(c) = \min[\, t : P(T \geq t \mid H_0, c) \leq \alpha\,].$$

4.2.2 Monte Carlo estimate of conditional p-value

Substantial research has been done on exact inference for contingency tables over the past decade, in terms of developing both new analysis and efficient algorithms for computation. The main problem of applying the exact test is that for moderate sized tables, the number of table probability to be enumerated can easily reach into the billions. Thus, in order to make our exact procedure feasible for practical use, an appropriate algorithm needs to be explored.

It has been shown that the number of possible table grows factorially fast as the number of baseline categories, number of outcome categories, or the total sample size increases. Thus, the number of operations to enumerate $\Gamma(c)$ grows faster than any polynomial in the minimum margin count. To extend the bounds of feasibility, much research has been done both in exploring new methods for complete enumeration and enhancing Monte Carlo approximation accuracy.

Pagano and Halvorsen (1981) came up with an efficient algorithm which finds the exact significance level of $I \times J$ contingency table without enumerating all possible tables. Later in 1983, Pagano and Trichler gave another algorithm which reduced the computing time to polynomial time as opposed to exponential time otherwise. However, because it involves inverting the characteristic function of the statistic, this algorithm is only good for statistics which are linear combinations of either the original observations or the ranks, such as the Wilcoxon test. At the same time, Mehta and Patel (1983) gave a network algorithm by recursively summing the probability in the required contingency tables, which eventually lead to creation of StatXact. Morgan and Blumenstein (1991) gave another algorithm for exact conditional tests for hierarchical models in multidimensional contingency tables. Both network and Morgan's algorithms depend on complete enumeration, and will thus give the exact p-value.

In this thesis, a Monte Carlo procedure given by Patefield (1981) was developed as a function Permu( ) in S-Plus to approximate significance levels of the proposed exact test on $r \times c$ table. It efficiently generates random tables under fixed row and column margins. The idea is as follows:

Let a_{ij} denote cell counts in a $r \times c$ contingency table with the row and column totals $(a_{i.}, 1 \le i \le r; a_{.j}, 1 \le j \le c)$. The conditional probability distribution of entry a_{lm} given the entries in the previous rows, i.e. $(a_{ij}, i=1,\ldots,l-1,\ j=1,\ldots,c)$ and the previous entries in row l, i.e. $(a_{ij},\ j=1,\ldots,m-1)$ is found to be

$$\begin{aligned}P_{lm} &= (a_{l.} - \sum_{j=1}^{m-1} a_{ij})!\,(N - \sum_{i=1}^{l} a_{i.} - \sum_{j=1}^{m-1} a_{.j} + \sum_{i=1}^{l}\sum_{j=1}^{m-1} a_{ij})! \\ &\times (a_{.m} - \sum_{i=1}^{l-1} a_{im})!\ [\sum_{j=m+1}^{c} (a_{.j} - \sum_{i=1}^{l-1} a_{ij})]! \\ &\times \{a_{lm}!\,(a_{.m} - \sum_{i=1}^{l} a_{im})!\,(a_{l.} - \sum_{j=1}^{m} a_{lj})! \\ &\times (N - \sum_{i=1}^{l} a_{i.} - \sum_{j=1}^{m} a_{.j} + \sum_{i=1}^{l}\sum_{j=1}^{m} a_{ij})!\,[\sum_{j=m}^{c} (a_{.j} - \sum_{i=1}^{l-1} a_{ij})]!\,\}^{-1},\end{aligned}$$

where $N = \sum_{i=1}^{r} a_{i.}$ and $(1 \le l \le r-1, 1 \le m \le c-1)$. The above conditional probability is valid when either $l=1$ or $m=1$ if the convention $\sum_{i=1}^{0}(\,.\,) = \sum_{j=1}^{0}(\,.\,) = 0$ is employed.

To ensure the rest cell counts are at least zero, the range of a_{lm} is such that all the factorial terms are non-negative. The conditional expected value of a_{lm} given previous entries and the row and column totals is

$$E_{lm} = \frac{(a_{.m} - \sum_{i=1}^{l-1} a_{im})\ (a_{l.} - \sum_{j=1}^{m-1} a_{lj})}{\sum_{j=m}^{c} (a_{.j} - \sum_{i=1}^{l-1} a_{ij})}.$$

When the denominator in this expression is zero, $E_{lm} = 0$.

Although formula for the conditional probability distribution of a_{lm} appears rather complicated, each of the terms is evaluated sequentially as $l = 1, \ldots, r-1$; $m = 1, \ldots, c-1$ in only a few lines of code. For each (l, m) the algorithm generates a random number, *RAND*, between 0 and 1. The probability distribution of a_{lm} is then accumulated, starting with a_{lm} equal to the nearest integer to E_{lm}. The value of a_{lm} is chosen with the required conditional probability when the cumulative probability exceeds *RAND*. Given the fixed marginals, the function Permu( ) can provide one sampled table as the output. Thus, the procedure for using our proposed exact test is to sample table by a large number of calls of Permu( ) and to estimate the significance level by the proportion of samples with a value of the test statistic T (as we defined in Section 4.1) that provides at least as much evidence against the null hypothesis as the value of the observed statistic t_{obs}, that is,

$$P(T \geq t \mid c) = \frac{\#\ of\ \ sampled\ \ \ tables\ \ \ with\ \ T \geq t_{obs}}{Total\ \ \#\ of\ \ sampled\ \ \ tables}.$$

To evaluate the performance of the Monte Carlo procedure, we compare the estimated p-values with the exact p-values produced by a complete enumeration of tables using a recursive method.

Consider a 2×6 contingency table with 2 baseline (2 and 3) and 3 outcome categories (1, 2, and 3) where 1 = best and 3 = worst. Two sets of RCT samples with different sizes are given in Table 3.

Table 3: 2 Sets of RCT 2 × 6 Contingency Tables

Sample	Size	Treatment	(2,1)	(2,2)	(2,3)	(3,1)	(3,2)	(3,3)
1	10	Active	1	1	0	4	2	2
	10	Control	2	2	3	1	1	1
2	20	Active	7	3	2	5	2	1
	20	Control	6	3	2	4	3	2

The comparisons of the exact and approximate p-values for these two samples are listed in Table 4.

Table 4: P-value comparisons

Sample	t_{obs}	Type	# of Permutations	P-value
1	0.8076923	Complete Enumeration	782	0.032
		Monte Carlo	500	0.030
2	0.5490909	Complete Enumeration	7,532	0.349
		Monte Carlo	1,000	0.346

As can be seen from the tables, the approximate p-values are reliable whereas the time required to generate random tables is much less dependent on sample size. The difference between complete enumeration and the Monte Carlo algorithm is quite substantial when dealing with large tables, which enables us to easily handle the large tables that previously would have been impractical to calculate.

4.3 Power

The performance of our proposed exact test needs to be evaluated compared to other widely used tests in term of power. As a preamble to that investigation we start off by discussing the formula involved in the power calculation.

4.3.1 Exact conditional power

The exact size α conditional power is computed by integrating the point probabilities of each table in sample space under an alternative hypothesis H_A over the rejection region of the null hypothesis H_0:

$$P_{\theta_A}(T(x) \geq t_{obs} \mid c) = \sum_{x \in \Gamma_c} P_{\theta_A}(X = x \mid c)\ I_{T(x) \geq t_\alpha} \ ,$$

where $t_\alpha = \min[t : P(T \geq t_{obs} \mid c, H_0) \leq \alpha]$ and Γ_c is the set of all possible tables given row and column margins. Due to the discrete nature of the data, the significance level α would not be exhausted fully.

Because of the nice exponential family form, the conditional point probability $P_\theta(X = x \mid c)$ has nice form. Let $x_{11}, \cdots, x_{1k}$ and $x_{21}, \cdots, x_{2k}$ represent the samples from population F_a (active group) and F_c (control group), which arise from multinomial $(n_{1.}, \underline{\pi_1})$ and $(n_{2.}, \underline{\pi_2})$ respectively, where the $\pi's$ are the associated probabilities. Let $c_i = x_{1j} + x_{2j}$, the column totals. Thus, for each element in the sample space, the point probabilities can be calculated. Under H_0 it has a simple form as follows:

$$P(\underline{X} = x \mid c, H_0) = \frac{\dfrac{1}{\prod_{j=1}^{k} x_{1j}!\, x_{2j}!}}{\sum_{\underline{Y} \in \Gamma_c} \dfrac{1}{\prod_{j=1}^{k} y_{1j}!\, y_{2j}!}} .$$

Under H_A, it has the form:

$$P(\underline{X} = x \mid c, H_A) = \frac{\prod_{j=1}^{k} \frac{\pi_{ij}^{x_{ij}}}{x_{ij}!} \prod_{j=1}^{k} \frac{\pi_{2j}^{x_{2j}}}{x_{2j}!}}{\sum_{\underline{Y}\in\Gamma_c} \prod_{j=1}^{k} \frac{\pi_{1j}^{y_{1j}} \, \pi_{2j}^{y_{2j}}}{y_{1j}! \, y_{2j}!}} \quad .$$

4.3.2 Exact unconditional power

It is obvious that when conducting a conditional test we compute the exact p-value with the marginal responses fixed at their observed values. When designing a study, however, the marginal responses that will arise once the data are gathered are unknown; a priori we can specify only the distributions of the responses, $\underline{\pi_1}$ and $\underline{\pi_2}$. Consequently, we must compute power unconditionally with respect to all possible margins. We can then obtain exact unconditional power as the expected value of these terms,

$$P_{\theta_\alpha}(T(x) \geq t) = \sum_{c\in\Omega} P_{\theta_\alpha}(T(x) \geq t \mid c) \; P(C = c) \, ,$$

where $\Omega = \{c : \sum c_j = n_{1.} + n_{2.}\}$, $P_{\theta_\alpha}(C = c) = \sum_{x\in\Gamma_c} P_{\theta_\alpha}(\underline{X_1} = \underline{x_1}) \; P_{\theta_\alpha}(\underline{X_2} = \underline{x_2})$.

The computation is practically infeasible since even for a moderate sample size, Ω can be quite large. For example, for $K = 5$ and $n_{1.} + n_{2.} = 50$, Ω contains 316,251 distinct vectors c.

Alternatively, to reduce the computational burden we can instead estimate exact power from a sample of Ω, given $n_{1.}$ and $n_{2.}$. In simulation studies power is usually estimated by the crude Monte Carlo estimator, $\hat{\alpha}$, which is given by:

$$\hat{\alpha} = \frac{\sum_{i=1}^{N} I\{P(T_i \geq T_{obs}) \leq \alpha\}}{N},$$

where N is the number of Monte Carlo samples.

We are now in a position to put our methodology to the litmus test. We will compare it with Fisher's exact test, ANCOVA, and proportional odds model via extensive simulations in Chapter 5.

Chapter 5: Comparison with other tests

In Chapter 4 we have developed an exact approach to between-group analysis adjusting for ordinal baseline covariates. In order to evaluate the performance of our methodology, we compare it with Fisher's exact test, ANCOVA and the proportional odds model, three widely used tests for categorical data as well to decide if our test is really desirable. The choice of which method to use can be determined by analysis of the statistical properties of each. An important criterion for a good statistical method is that it should reduce the rate of false negative (β). The β of a statistical test is usually expressed in terms of statistical power ($1-\beta$). A method that requires relatively fewer data to provide a certain level of statistical power is described as efficient.

5.1 Other tests for categorical data

5.1.1 Fisher's exact test

Fisher's exact test is a statistical significance test used in the analysis of categorical data where sample sizes are small. It is named after its inventor, R. A. Fisher, and is one of a class of exact tests. The test is used to examine the significance of the association between two variables in 2 × 2 contingency table. With large samples, a chi-squared test can be used in this situation. However, this test is not suitable when sample sizes are small or when the "expected value" in any of the cells of the table is below 10, that is, when the data are very unequally distributed among the cells of the table. The Fisher test is, as its name states, exact, and it can therefore be used regardless of the sample characteristics. It is also very useful for highly imbalanced tables.

Fisher's exact test is based on exact probabilities from the hypergeometric distribution. Under H_0, the exact probability of observing one particular sampled table, given fixed row and column margins, has been given in Section 4.1.2. The one-sided probability for the Fisher’s exact test is calculated by generating all tables that are more extreme than the table given by the user, in the direction specified by the one-sided alternative. The p-values of these tables are added up, including the p-value of the table itself. Because the calculation of Fisher’s exact test involves permuting the observed cell frequencies it is also referred to as a permutation test, like our proposed exact test in Chapter 4.

In two-armed RCTs, the data often include small and zero cell counts. If the response is ordinal, we can combine categories to create a 2 × 2 table by treating the ordinal covariate as binary (improved or non-improved). Obviously, this converting is inefficient because ignoring the ordering among the categories or collapsing categories will result in a loss of power. This can be verified by the simulations in Section 5.2.

5.1.2 Analysis of covariance (ANCOVA)

In most experiments the scores on the covariates are collected before and after the experimental treatment. By treating the ordinal response variable as continuous, we can use ANCOVA with the post-treatment value as the response variable and the baseline value as the covariate.

ANCOVA, or analysis of covariance is a general linear model with one continuous explanatory variable and one or more factors. ANCOVA is a merger of ANOVA (analysis of variance) and regression for continuous variables. ANCOVA tests whether certain factors have an effect after removing the variance of which quantitative predictors (covariates) account. The inclusion of covariates can increase statistical power because it accounts for some of the variability. The analysis of covariance includes the same assumption as the analysis of variance: independent sampling, equal corresponding population variances and normally distributed corresponding populations. In addition it includes two other assumptions related to the relationship between the covariate and the dependent variable. It is assumed that the covariance between the covariate and the dependent variable, within each sample or column, do not differ significantly from each other. In other words, if we were to compute prediction equations within each sample or column, the slopes of the lines would not differ significantly from each other. It is also assumed that the relationship between the covariate and the dependent variable is linear--that the relationship is best described by a straight line.

In this thesis, the ANCOVA model is written as:

$$Y_{ij} = \mu + \alpha_i + \beta x_{ij} + \varepsilon_{ij}, \quad i = 1, 2; \quad j = 1, \cdots, n_i ,$$

where Y_{ij} is the response of the j-th unit, receiving treatment i with associated baseline covariates x_{ij}. In this model, the effect of the i-th treatment is modeled via the parameter α_i. The i values 1, 2 represent the active and control treatment respectively.

Now, our interest of examining whether or not there is any treatment effect becomes to test the null hypothesis $H_0 : \alpha_1 = \alpha_2$ against the alternative hypothesis $H_A : \alpha_1 < \alpha_2$ (lower level is better). The p-value can be derived by using the F distribution. Readers interesting in further details about the computation of p-value are referred to Rencher, 2000.

5.1.3 Proportional odds model

The most well-known cumulative logit model for ordinal response is the proportional odds model. Under this model, we assume that the log of the cumulative odds ratio is proportional to the distance between the values of the explanatory variables.

Let *Y* be an ordinal variable with *k* levels, and let $P(Y_i \leq j)$ be the cumulative probability of responding of categories j from group i (active or control group).

$$P(Y_i \le j) = \pi_{i1} + \pi_{i2} + \cdots + \pi_{ij},\ j = 1, 2, \cdots, k; i = 1, 2.$$

Then we have $k-1$ cumulative probabilities to look at, since $P(Y_i \le k) = 1$. Let

$$\begin{aligned}\text{logit } P(Y \le j) &= \log \frac{P(Y \le j)}{P(Y > j)} \\ &= \log \frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_k},\quad j = 1, 2, \cdots, k-1\end{aligned}$$

be the cumulative logits. We can incorporate all $k-1$ cumulative logits into the following model (Agresti, 1990):

$$\text{logit}[P(Y \le j)] = \alpha_j + \beta' X,\quad j = 1, \cdots, k-1.$$

where β denotes the effects of explanatory variables X. The covariate X can be continuous or categorical. In our case, there are two explanatory variables: binary treatment factor (active or control) and ordinal baseline variable.

The above cumulative logit model satisfies

$$\begin{aligned}\text{logit}[P(Y \le j \mid X_a)] - \text{logit}[P(Y \le j \mid X_c)] &= \log\left[\frac{P(Y \le j \mid X_a)/P(Y > j \mid X_a)}{P(Y \le j \mid X_c)/P(Y > j \mid X_c)}\right] \\ &= \alpha_j + \beta' X_a - \alpha_j - \beta' X_c \\ &= \beta'(X_a - X_c).\end{aligned}$$

The odds of making response $\le j$ at $X = X_a$ is $\exp[\beta'(X_a - X_c)]$ times the odds at $X = X_c$. The log cumulative odds ratio is proportional to the distance between X_a and X_c. The same proportionality constant applies to each logit. Cumulative logit models simultaneously fit all $k-1$ logit models for the k categories of the response. There are different methods to fit the proportional odds model. In this thesis, we use GEE (Generalized Estimating Equations) by using the GENMOD procedure in SAS to estimate the parameters. See Liang, Zeger and Qaqish (1992) and Lipsitz, Kim and Zhao (1994) for a description of GEE methods for ordinal responses.

<u>5.2 Simulations: Comparison of performance</u>

5.2.1 Methods

As comparison we employ the three common tests in the previous section and our proposed exact permutation test based on partial ordering. We compare more generally the unconditional power of these four one-sided tests (active is better than control) in moderate and small samples.

Let us consider the 2 × 3 contingency tables with 2 baseline and 3 outcome categories ranging from 1 = best to 3 = worst. In a two-armed RCT, the data structure is 2 × 6 contingency tables, with the active samples from multinomial $(n_{1.}, \underline{\pi_1})$ and the control samples form multinomial $(n_{2.}, \underline{\pi_2})$, where the π's are the associated transition probability vectors from baseline to post-treatment. The initial probabilities to generate the baseline data is (0, 0.55, 0.45). Nine sets of transition probability vectors are given in Table 5.

Table 5: Transition Probabilities

No.	Transition Probabilities
$\underline{\pi}_1$	(0.63, 0.23, 0.14; 0.56, 0.26, 0.18)
$\underline{\pi}_2$	(0.54, 0.27, 0.19; 0.49, 0.29, 0.22)
$\underline{\pi}_3$	(0.45, 0.30, 0.25; 0.42, 0.31, 0.27)
$\underline{\pi}_4$	(0.36, 0.33, 0.31; 0.35, 0.33, 0.32)
$\underline{\pi}_5$	(0.30, 0.33, 0.37; 0.29, 0.32, 0.39)
$\underline{\pi}_6$	(0.27, 0.32, 0.41; 0.25, 0.30, 0.45)
$\underline{\pi}_7$	(0.18, 0.30, 0.52; 0.14, 0.23, 0.63)
$\underline{\pi}_8$	(0.12, 0.25, 0.63; 0.07, 0.16, 0.77)
$\underline{\pi}_9$	(0.03, 0.13, 0.84; 0.01, 0.04, 0.95)

For each $\underline{\pi}_i$, the first three components are the conditional probabilities of falling in category 1, 2, 3 after treatment given that the initial level 2 (sum = 1); the last three components represent the conditional probabilities of falling in category 1, 2, 3 category given that the baseline level is 3 (sum = 1). From Table 4, it can be seen that with the index number increases, the transition probabilities more and more shift to the higher level, which means given the same baseline values, the sample generated from $\underline{\pi}_i$ tends to be better than the one generated from $\underline{\pi}_j$ if $i < j$. The initial probabilities are always same for two treatment groups; however, due to sampling error, with random assignment of subjects to the groups, the baseline status may differ between treatment groups, that is, they may have same distributions with different realizations. By choosing different combinations of set of transition probabilities and sample sizes, we generate a variety of datasets for different cases. For equal treatment effects (H_0: Active = Control is true), the same transition probabilities are used; for unequal treatment effects (explicitly H_A: Active > Control is true), the index number of the transition probabilities $\underline{\pi}_a$ for active group should be smaller than $\underline{\pi}_c$ for control group. For each combination of set of transition probabilities and sample sizes, we generate 500 pseudo-datasets by sampling from the baseline distribution and the distribution created by the given set of the transition probabilities and find out how many of the 500 p-values are under significance level $\alpha = 0.1$ or 0.05, i.e. the unconditional power.

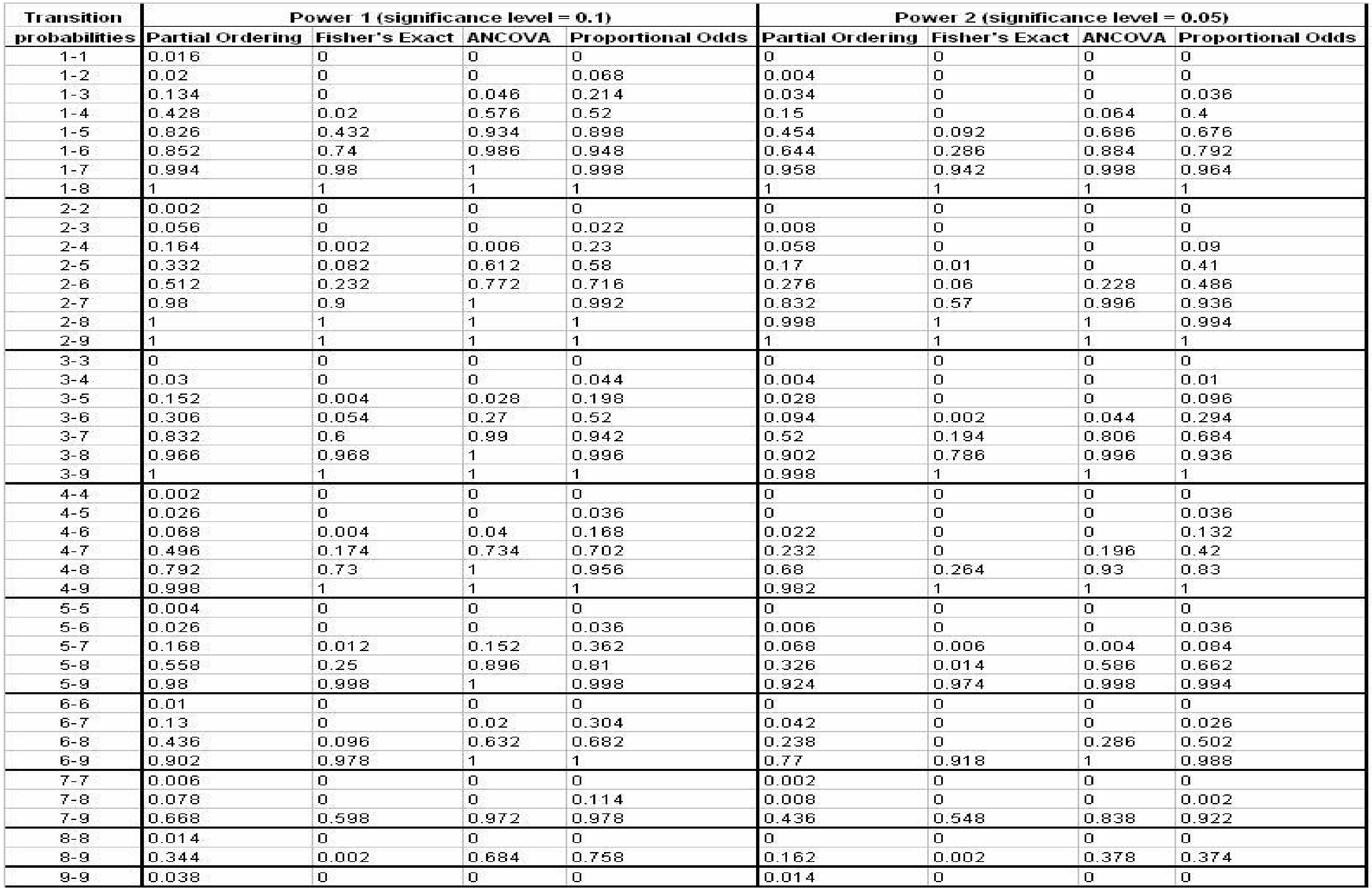
5.2.2 Results

<u>Simulation 1:</u>

Table 6 displays the power of the four tests for all possible combinations of set of transition probabilities, when $n_{1.} + n_{2.} = 30$, in balanced ($n_{1.} = n_{2.}$=15) samples. To compute the p-value for our proposed exact test based on the partial ordering, we randomly selected 1000 permutations of the treatment groups (subject to Monte Carlo sampling) and recomputed the test statistic. As we mentioned in Chapter 4, this set of 1000 values serves as the null reference distribution for this exact permutation test. For simplicity, $(\underline{\pi}_a, \underline{\pi}_c) = (\underline{\pi}_i, \underline{\pi}_j)$ is abbreviated as $i - j$. Power performance comparisons for {1-1, …, 1-8}, {2-2, 2-3, …, 2-9}, …, {6-6, 6-7, …, 6-9} for $\alpha = 0.1$ (similar curves were observed for $\alpha = 0.05$) are shown in Figure 1.

The results show that when the active experiment is slightly better than the control treatment, the proportional odds model gives the highest power among these four tests and the partial ordering test is more powerful than ANCOVA and Fisher's exact test; when the active experiment becomes much better than the control

Table 6: Simulation 1, Balanced Data, N=30



Transition	Power 1 (significance level = 0.1)				Power 2 (significance level = 0.05)			
probabilities	Partial Ordering	Fisher's Exact	ANCOVA	Proportional Odds	Partial Ordering	Fisher's Exact	ANCOVA	Proportional Odds
1-1	0.016	0	0	0	0	0	0	0
1-2	0.02	0	0	0.068	0.004	0	0	0
1-3	0.134	0	0.046	0.214	0.034	0	0	0.036
1-4	0.428	0.02	0.576	0.52	0.15	0	0.064	0.4
1-5	0.826	0.432	0.934	0.898	0.454	0.092	0.686	0.676
1-6	0.852	0.74	0.986	0.948	0.644	0.286	0.884	0.792
1-7	0.994	0.98	1	0.998	0.958	0.942	0.998	0.964
1-8	1	1	1	1	1	1	1	1
2-2	0.002	0	0	0	0	0	0	0
2-3	0.056	0	0	0.022	0.008	0	0	0
2-4	0.164	0.002	0.006	0.23	0.058	0	0	0.09
2-5	0.332	0.082	0.612	0.58	0.17	0.01	0	0.41
2-6	0.512	0.232	0.772	0.716	0.276	0.06	0.228	0.486
2-7	0.98	0.9	1	0.992	0.832	0.57	0.996	0.936
2-8	1	1	1	1	0.998	1	1	0.994
2-9	1	1	1	1	1	1	1	1
3-3	0	0	0	0	0	0	0	0
3-4	0.03	0	0	0.044	0.004	0	0	0.01
3-5	0.152	0.004	0.028	0.198	0.028	0	0	0.096
3-6	0.306	0.054	0.27	0.52	0.094	0.002	0.044	0.294
3-7	0.832	0.6	0.99	0.942	0.52	0.194	0.806	0.684
3-8	0.966	0.968	1	0.996	0.902	0.786	0.996	0.936
3-9	1	1	1	1	0.998	1	1	1
4-4	0.002	0	0	0	0	0	0	0
4-5	0.026	0	0	0.036	0	0	0	0.036
4-6	0.068	0.004	0.04	0.168	0.022	0	0	0.132
4-7	0.496	0.174	0.734	0.702	0.232	0	0.196	0.42
4-8	0.792	0.73	1	0.956	0.68	0.264	0.93	0.83
4-9	0.998	1	1	1	0.982	1	1	1
5-5	0.004	0	0	0	0	0	0	0
5-6	0.026	0	0	0.036	0.006	0	0	0.036
5-7	0.168	0.012	0.152	0.362	0.068	0.006	0.004	0.084
5-8	0.558	0.25	0.896	0.81	0.326	0.014	0.586	0.662
5-9	0.98	0.998	1	0.998	0.924	0.974	0.998	0.994
6-6	0.01	0	0	0	0	0	0	0
6-7	0.13	0	0.02	0.304	0.042	0	0	0.026
6-8	0.436	0.096	0.632	0.682	0.238	0	0.286	0.502
6-9	0.902	0.978	1	1	0.77	0.918	1	0.988
7-7	0.006	0	0	0	0.002	0	0	0
7-8	0.078	0	0	0.114	0.008	0	0	0.002
7-9	0.668	0.598	0.972	0.978	0.436	0.548	0.838	0.922
8-8	0.014	0	0	0	0	0	0	0
8-9	0.344	0.002	0.684	0.758	0.162	0.002	0.378	0.374
9-9	0.038	0	0	0	0.014	0	0	0

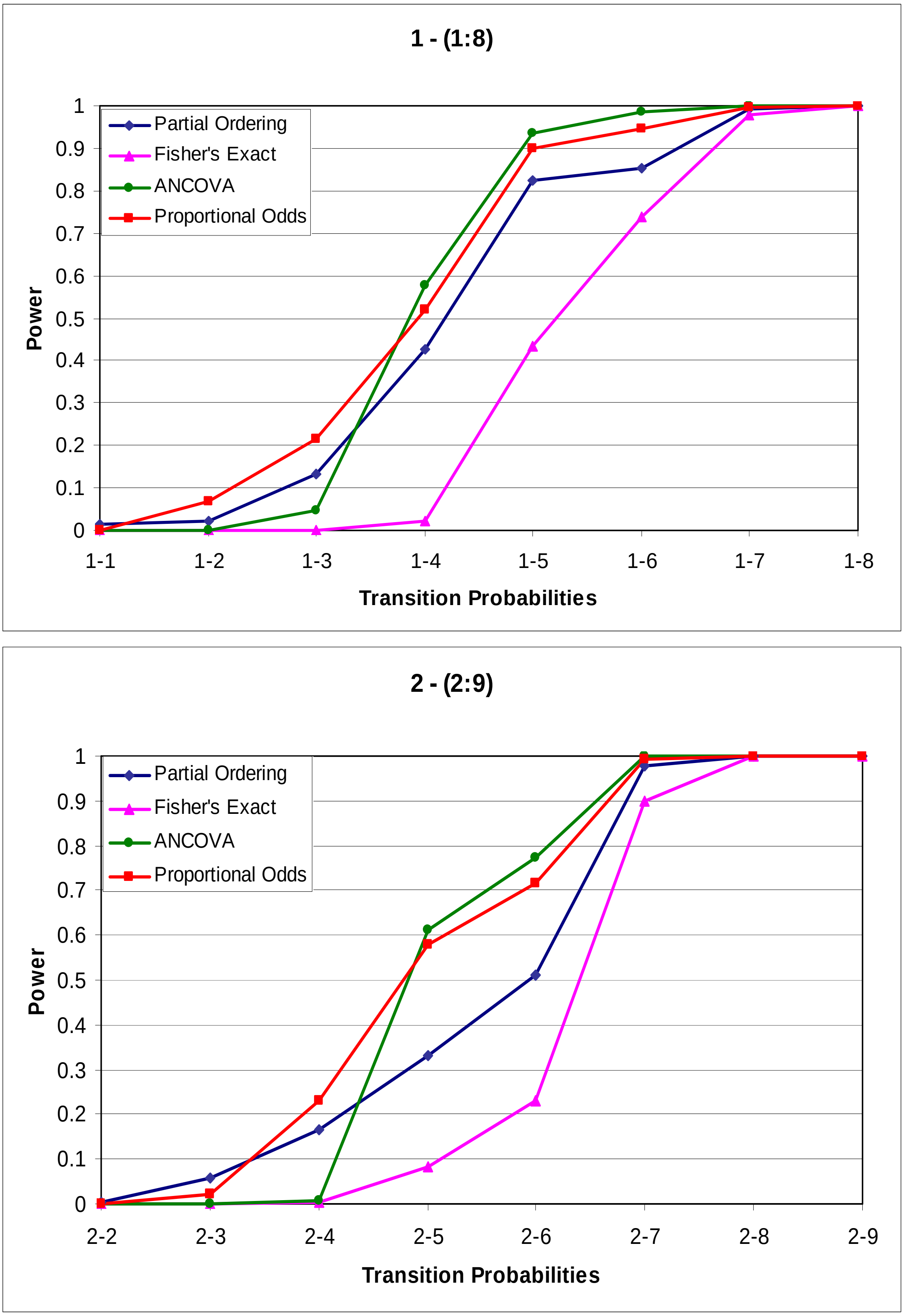


Figure 1(a): Simulation 1, balanced data, N=30, $\alpha = 0.1$

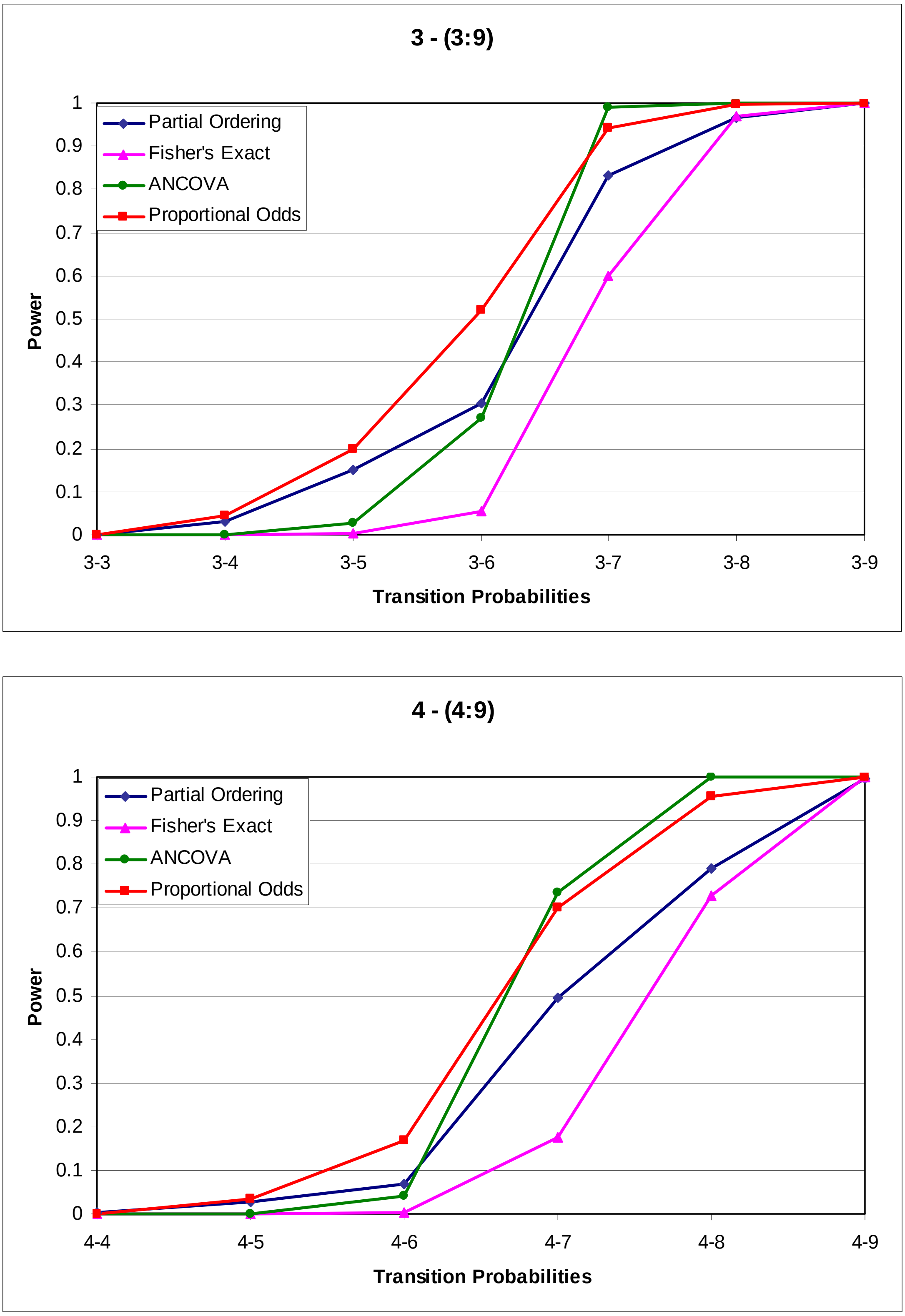
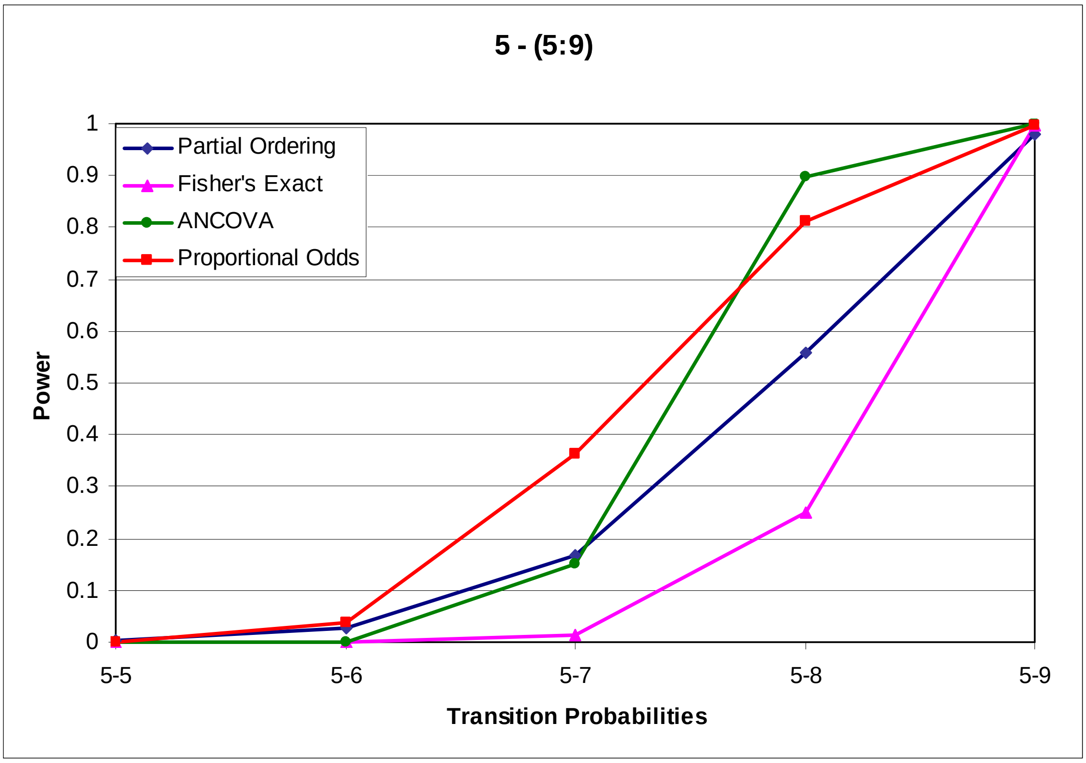


Figure 1(b): Simulation 1, balanced data, N=30, α = 0.1



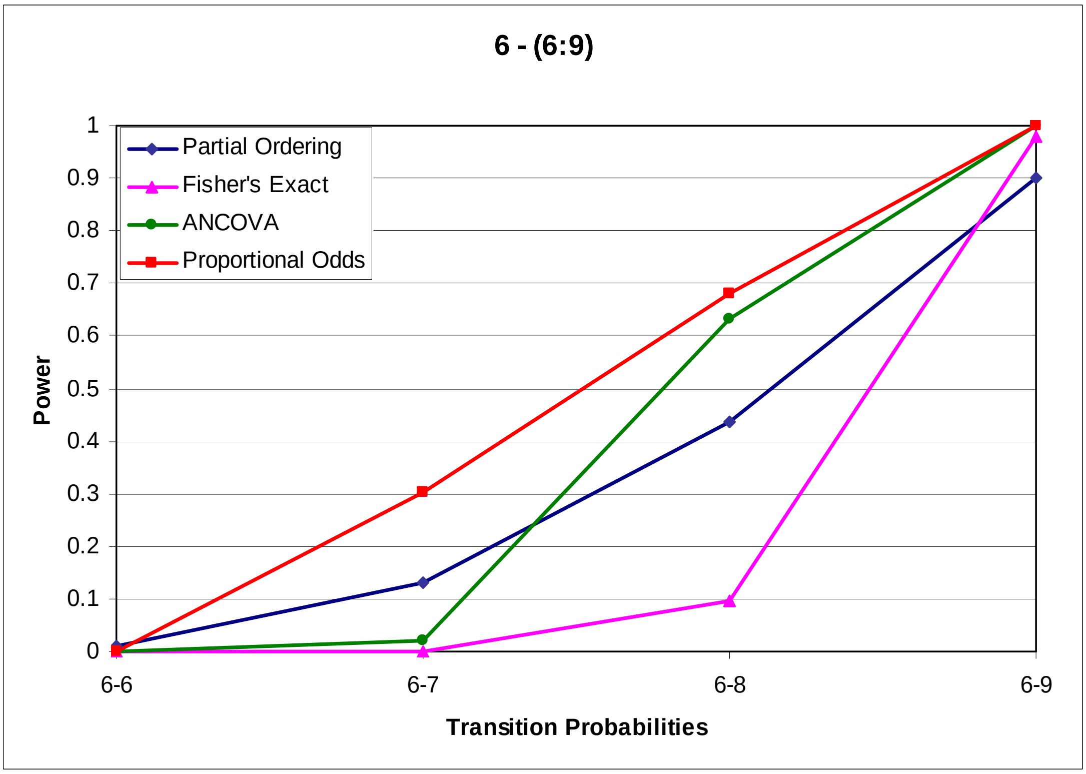


Figure 1(c): Simulation 1, balanced data, N=30, α = 0.1

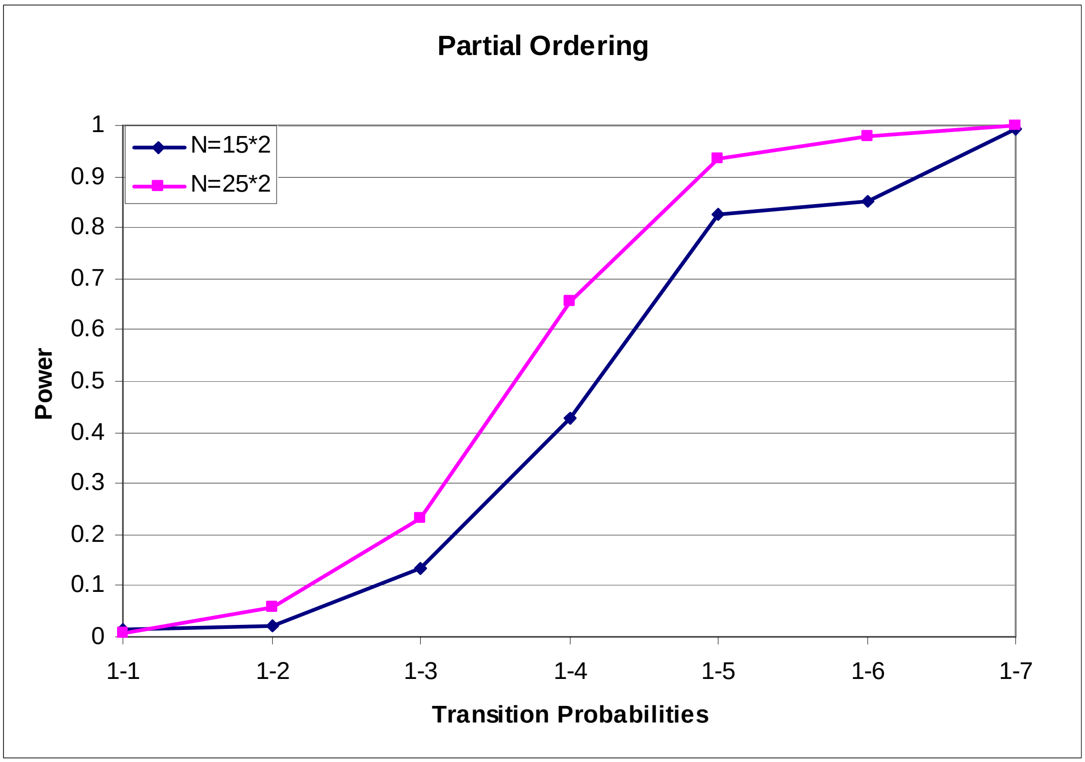
treatment, the power of ANCOVA grows fastest and becomes most powerful while the partial ordering is less powerful than ANCOVA and proportional odds model. As we expected, Fisher’s exact test is the worst one. Additionally, all these tests are conservative and partial ordering test is kind of less conservative than others.

Simulation 2:

Power increase when sample size increases. Next we evaluate the effects of sample size on the tests. One thousand permutations were selected to compute the p-value for our proposed exact test with α = 0.1. Table 7 shows the power of the four tests with transition probabilities {1-1, …, 1-7}, for sample size $n_{1.} + n_{2.} = \{30, 50\}$ in balanced ($n_{1.} = n_{2.}$) samples. The plot for each test is shown in Figure 2.

Table 7: Simulation 2, Balanced Data, N= {30, 50}

	Transition	Power (significance level = 0.1)			
Sample Size	probabilities	Partial Ordering	Fisher's Exact	ANCOVA	Proportional Odds
25*2	1-1	0.008	0	0	0
	1-2	0.058	0	0	0.124
	1-3	0.232	0.018	0.158	0.516
	1-4	0.656	0.46	0.922	0.892
	1-5	0.934	0.92	1	0.998
	1-6	0.978	0.994	1	1
	1-7	1	1	1	1
15*2	1-1	0.016	0	0	0
	1-2	0.02	0	0	0.068
	1-3	0.134	0	0.046	0.214
	1-4	0.428	0.02	0.576	0.52
	1-5	0.826	0.432	0.934	0.898
	1-6	0.852	0.74	0.986	0.948
	1-7	0.994	0.98	1	0.998



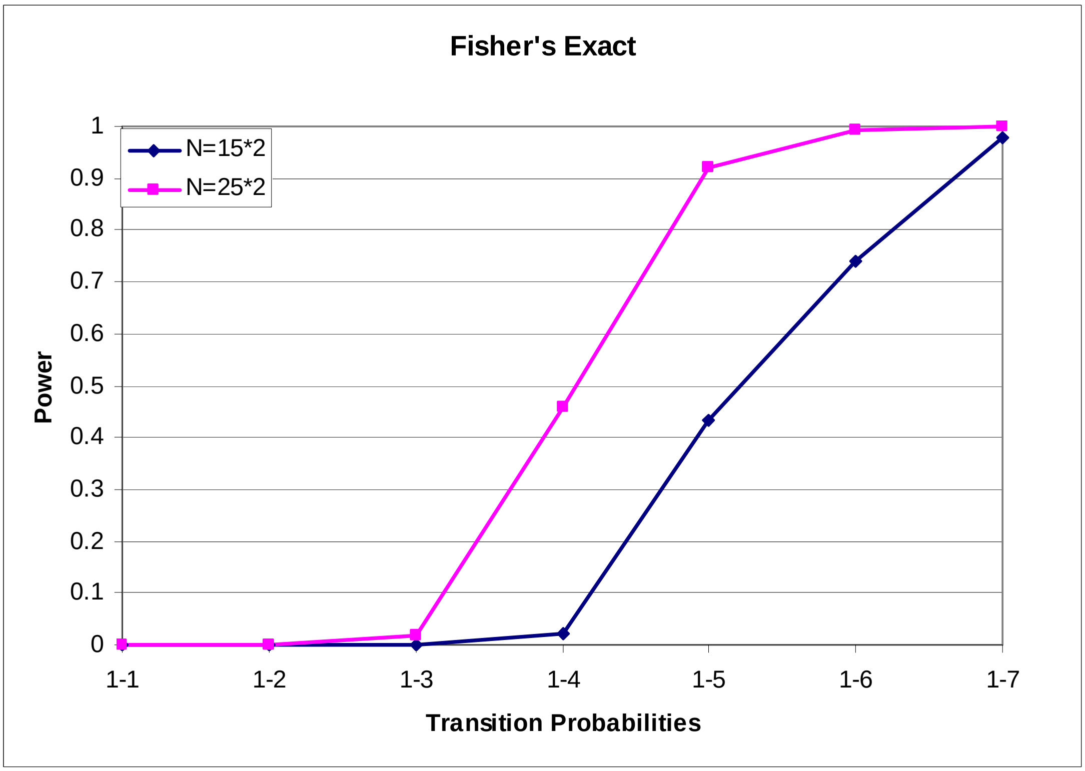
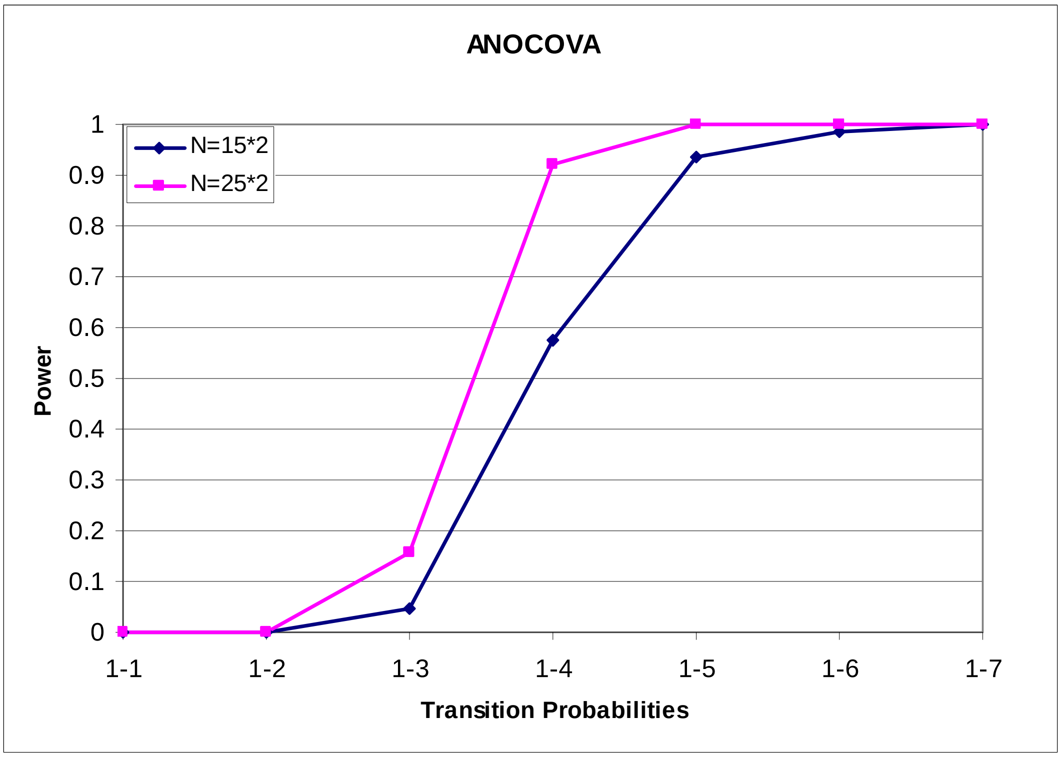


Figure 2(a): Simulation 2, balanced samples, N= {30, 50}, α = 0.1



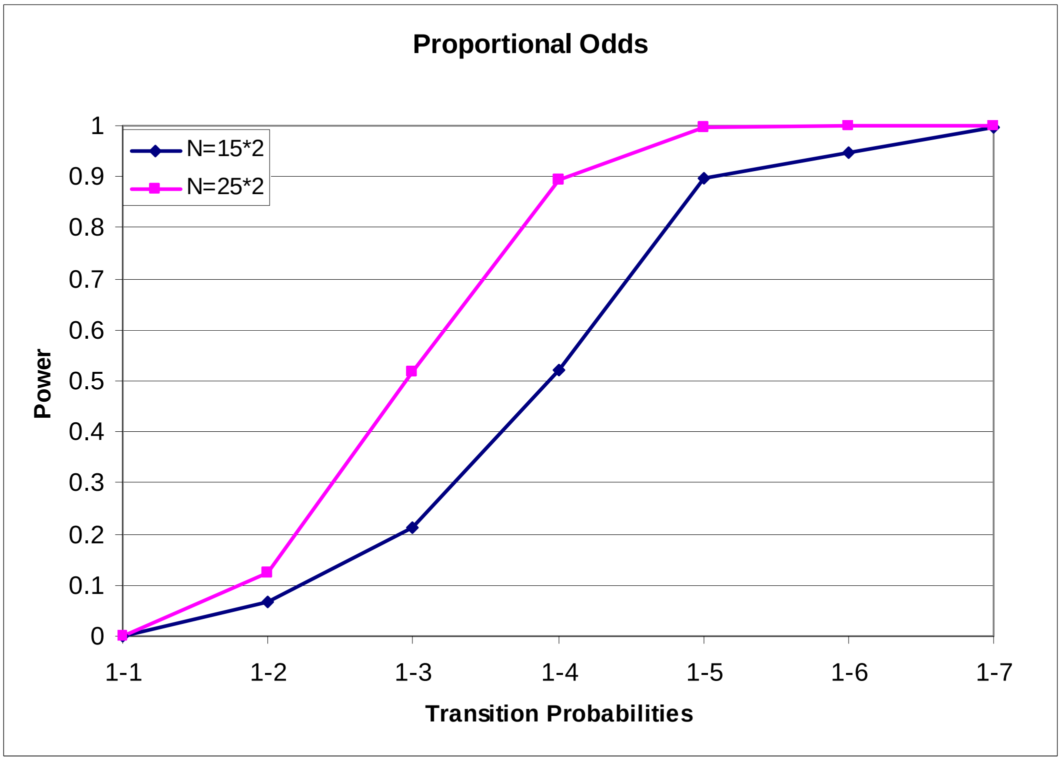


Figure 2(b): Simulation 2, balanced samples, N = {30, 50}, α = 0.1

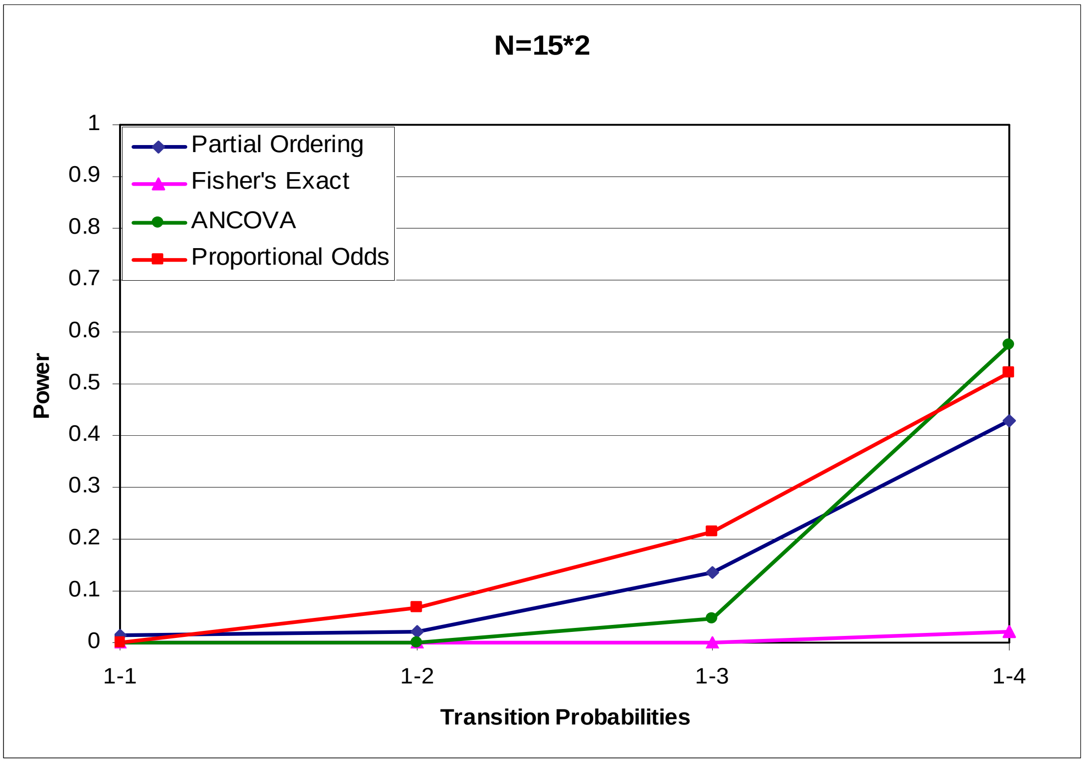
Obviously, the power of the proportional odds test increases more quickly when the difference of two treatments is small and partial ordering is the second position; when the active experiment outperforms control treatment in a large scale, the ANCOVA test rises in power most quickly and becomes the most powerful test. Proportional odds test is still powerful. The power curves of the partial ordering and Fisher's exact test seem to have the similar slopes and their power does not change as dramatically as the other two with larger sample sizes.

Simulation 3:

From Simulation 1 and 2, we can conclude that partial ordering test tends to be less powerful than ANCOVA and proportional odds test when the active experiment is much better than the control treatment. Therefore, in the following simulations we only focus on the case that the difference between the two treatments is small.

Based on the results from Simulation 1, we slightly increase the sample size from 30 to 40 in balanced samples with transition probabilities {1-1, 1-2, 1-3, 1-4}, {2-2, 2-3, …, 2-6}, …, {4-4, 4-5, 4-6, 4-7} and $\alpha = 0.1$. One thousand permutations for partial ordering were selected for each sample. Figure 3, 4, 5, 6 show the differences in power between these tests.

The comparison results show that for the case of small difference between the two treatments, proportional odds model is overall most powerful. The superiority of



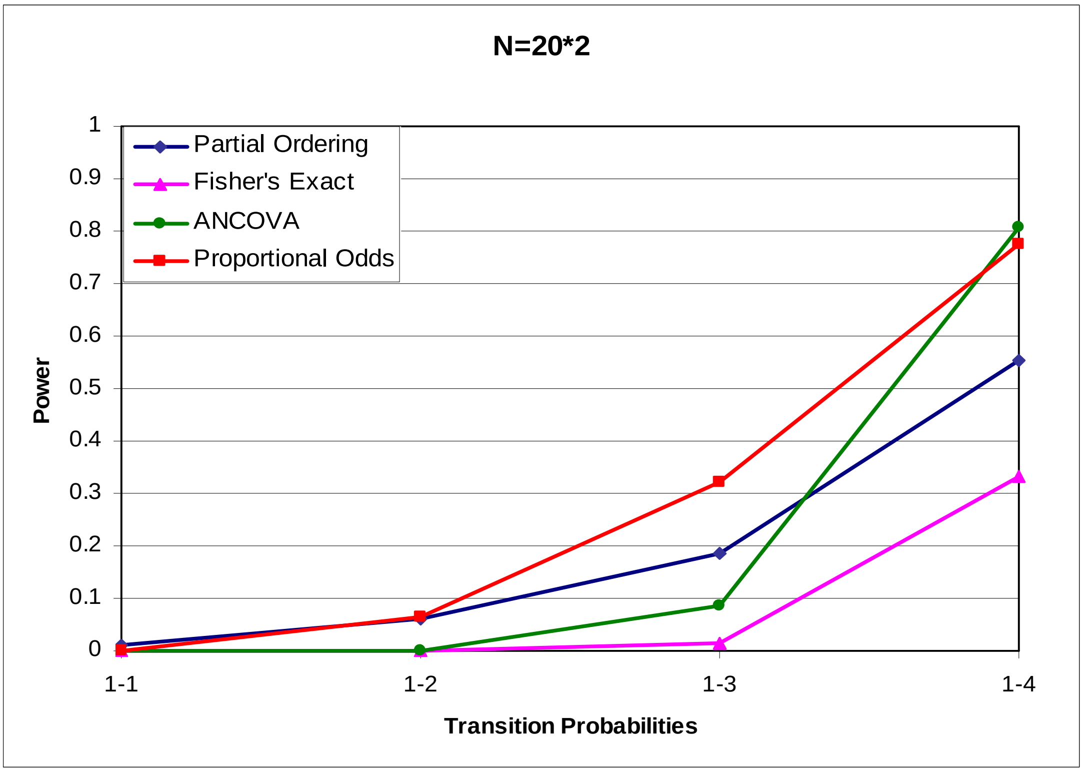
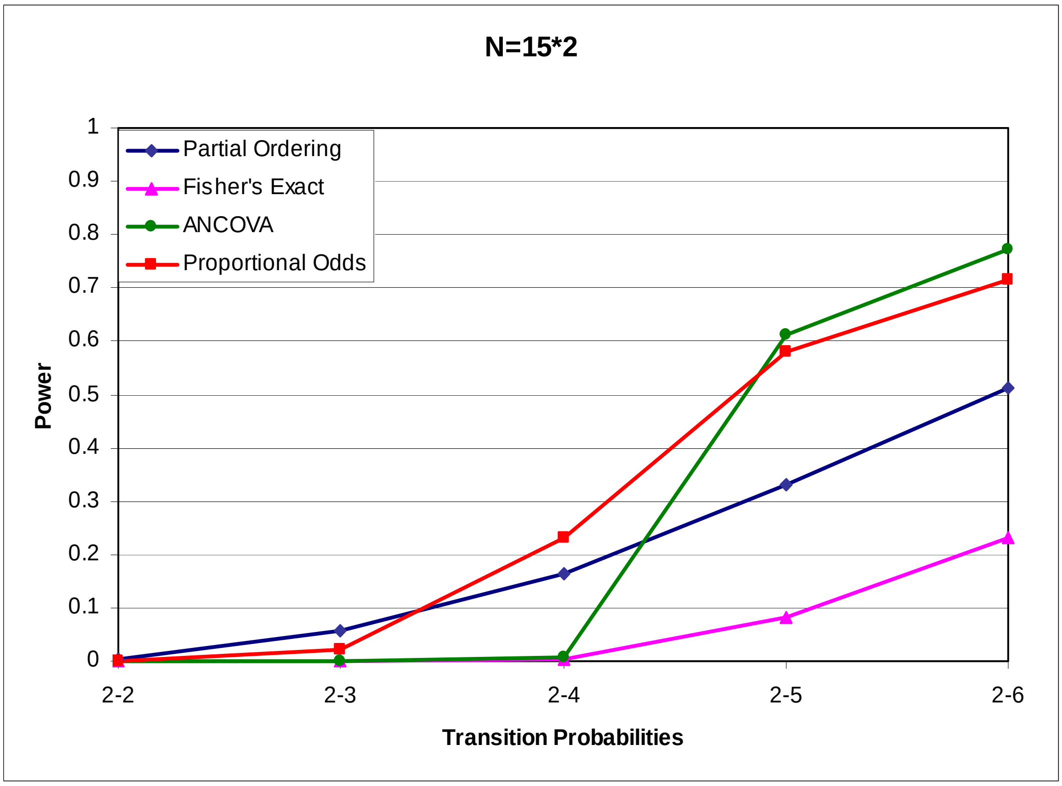


Figure 3: Simulation 3, {1-1,…, 1-4}, balanced samples, N= {30, 40}, α = 0.1



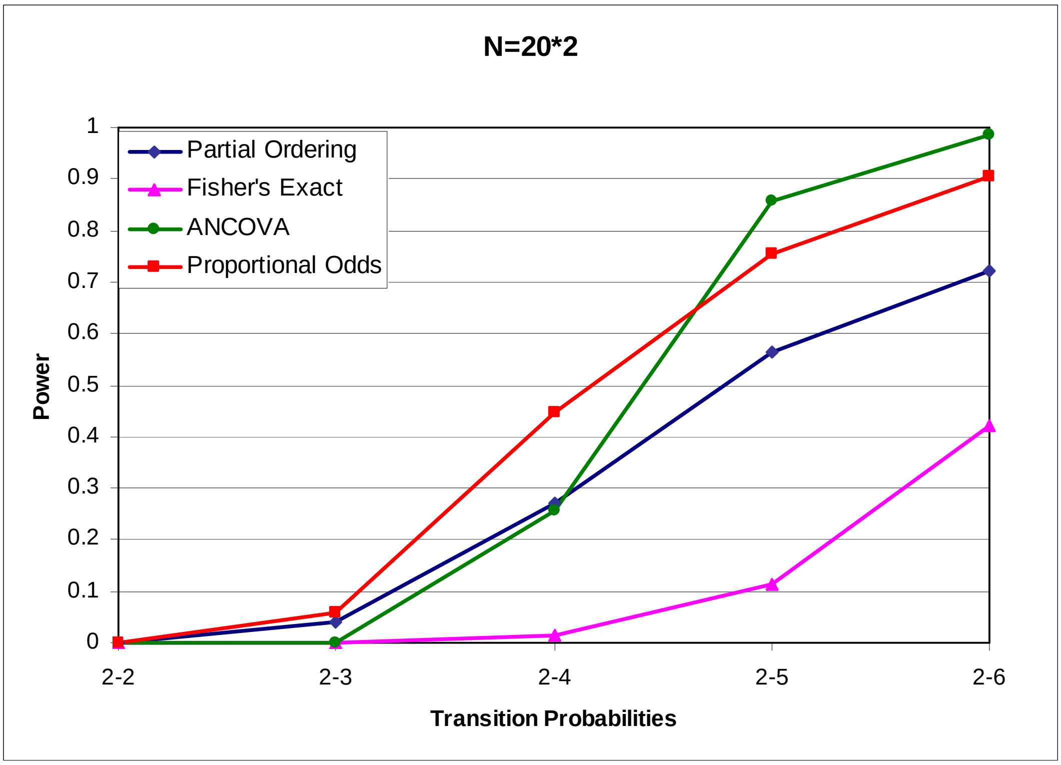
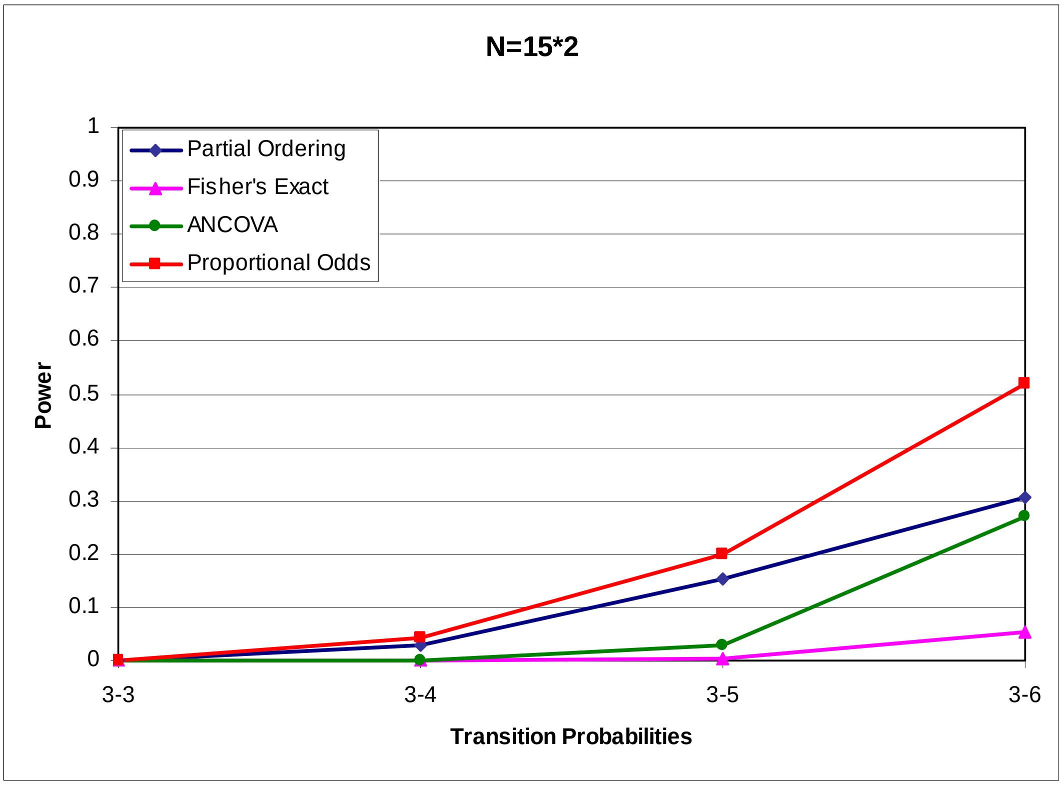


Figure 4: Simulation 3, {2-2,…, 2-6}, balanced samples, N= {30, 40}, α = 0.1



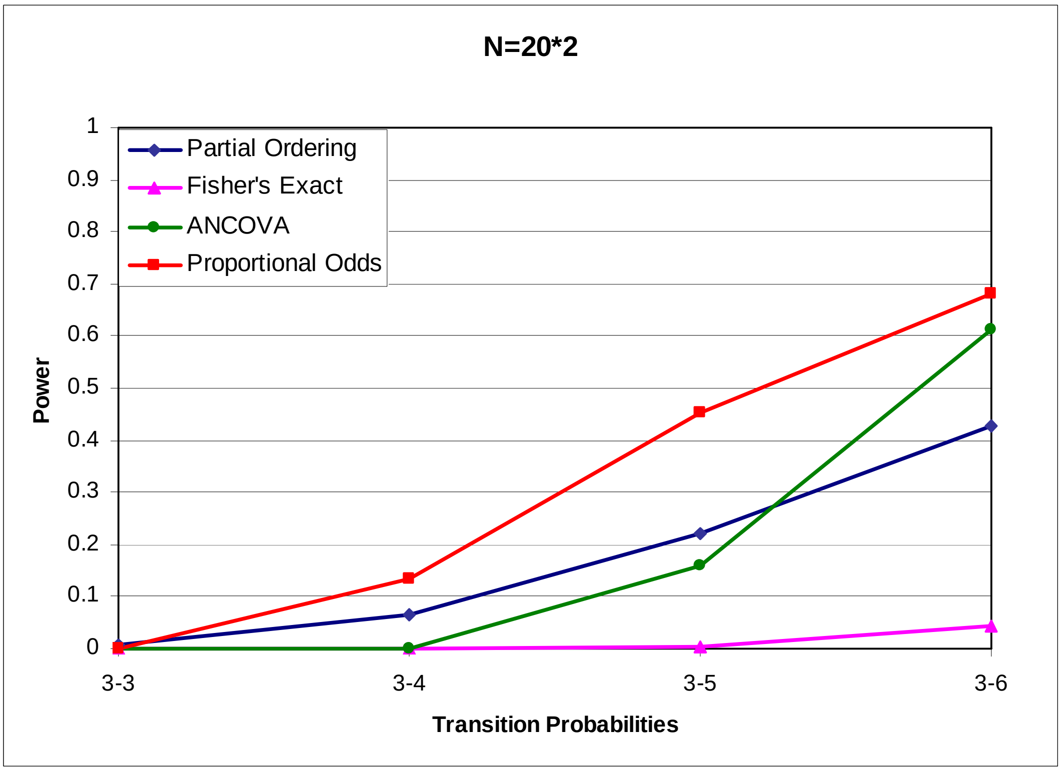
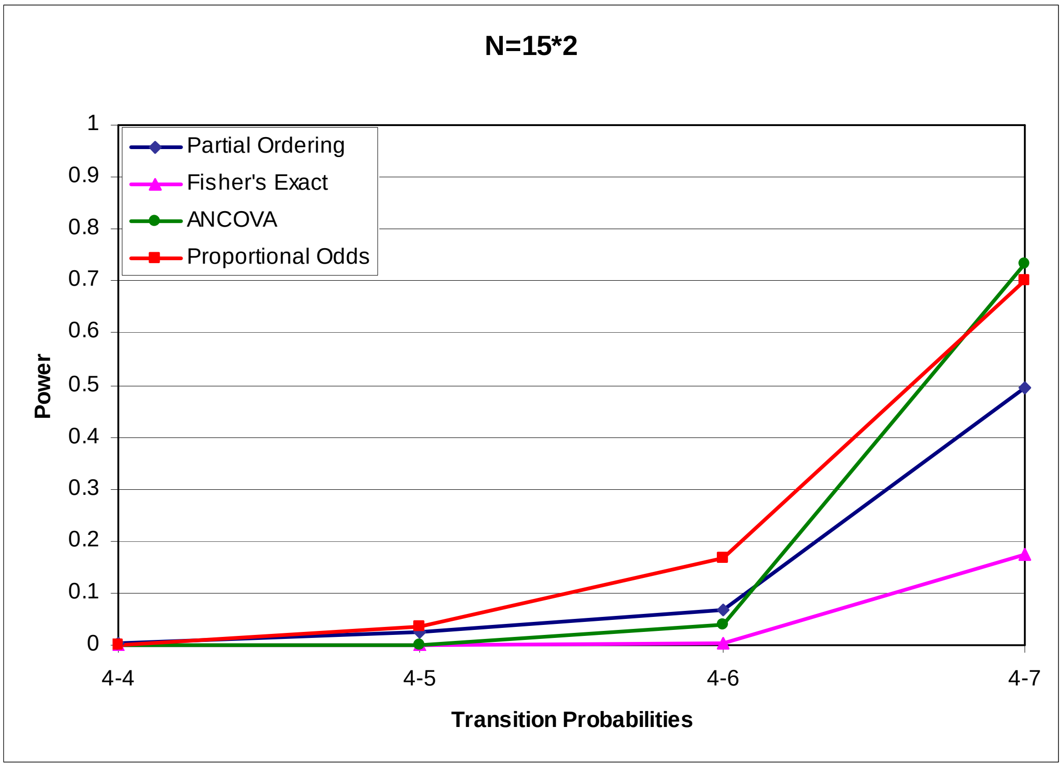


Figure 5: Simulation 3. {3-3,…, 3-6}, balanced samples, N= {30, 40}, α = 0.1



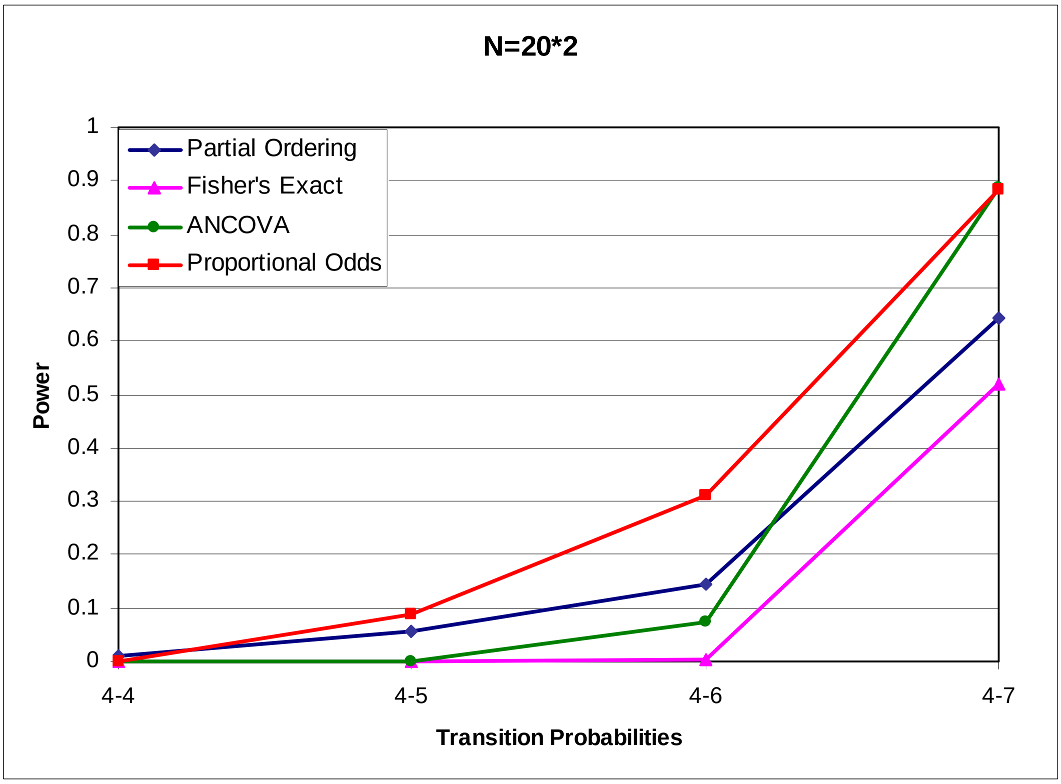


Figure 6: Simulation 3, {4-4,…, 4-7}, balanced samples, N= {30, 40}, α = 0.1

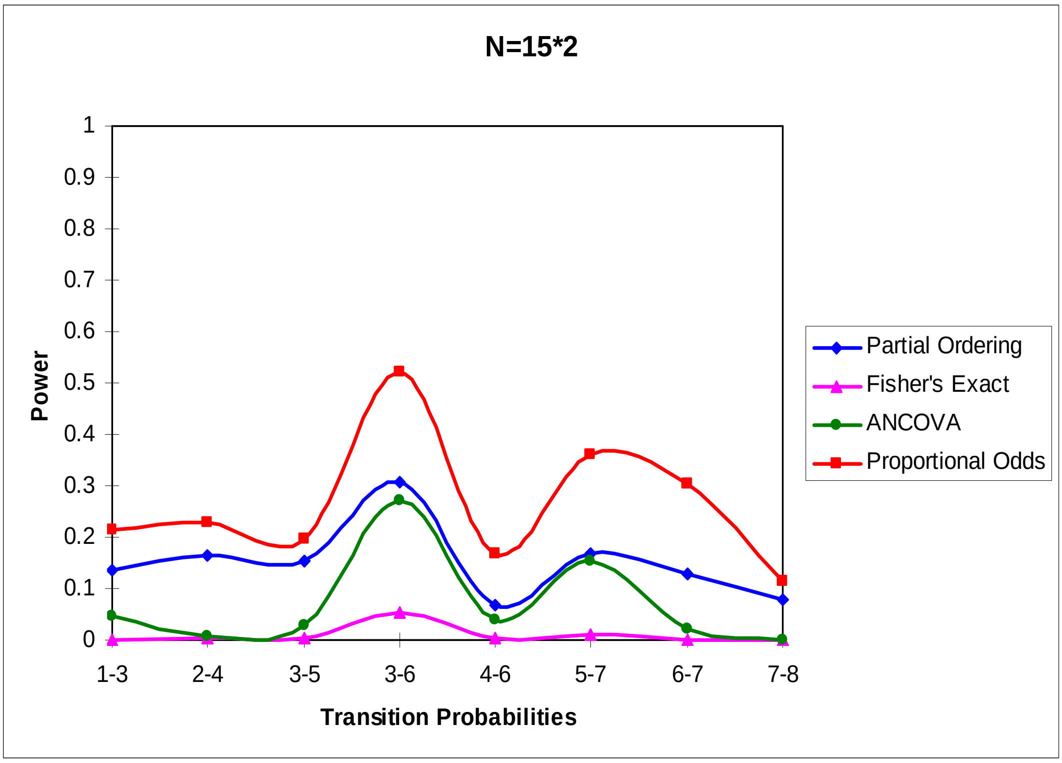
partial ordering over ANCOVA is increased for some cases, and decreased for other cases. The possible reason is that the slight increase of the sample size may be not big enough to improve the power performance significantly for ANCOVA method.

<u>Simulation 4:</u>

In this simulation, we increase the sample size from 30 to 80, in balanced samples, α=0.1. From the results of Simulation 1, we choose some points of $(\underline{\pi}_a, \underline{\pi}_c)$ where the power of partial ordering is higher than ANCOVA for small sample size. The results are shown in Table 8 and Figure 7.

Table 8: Simulation 4, balanced samples, N = {30, 80}

Sample	**Transition**	**Power (significance level = 0.1)**			
Size	**probabilities**	**Partial Ordering**	**Fisher's Exact**	**ANCOVA**	**Proportional Odds**
15*2	1-3	0.134	0	0.046	0.214
	2-4	0.164	0.002	0.006	0.23
	3-5	0.152	0.004	0.028	0.198
	3-6	0.306	0.054	0.27	0.52
	4-6	0.068	0.004	0.04	0.168
	5-7	0.168	0.012	0.152	0.362
	6-7	0.13	0	0.02	0.304
	7-8	0.078	0	0	0.114
40*2	1-3	0.532	0.284	0.768	0.89
	2-4	0.528	0.266	0.842	0.902
	3-5	0.45	0.242	0.698	0.866
	3-6	0.718	0.63	1	0.976
	4-6	0.236	0.048	0.174	0.612
	5-7	0.618	0.55	1	0.958
	6-7	0.37	0.152	0.528	0.782
	7-8	0.286	0.048	0.252	0.63



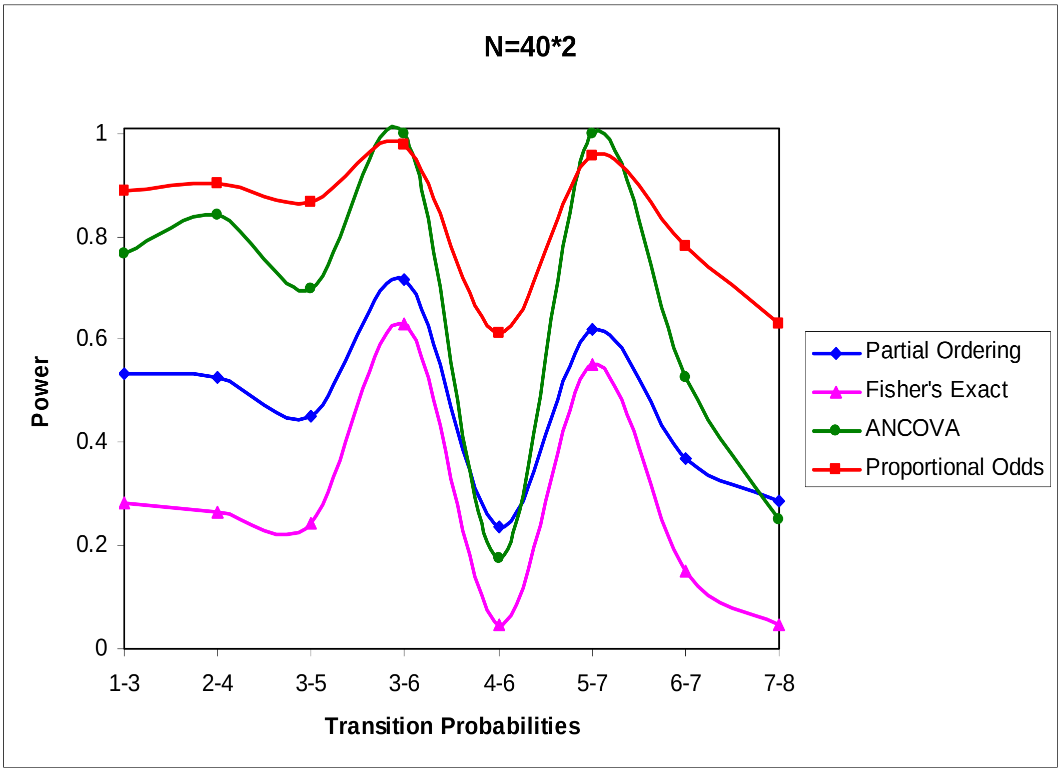


Figure 7: Simulation 4, balanced samples, N= {30, 80}, α = 0.1

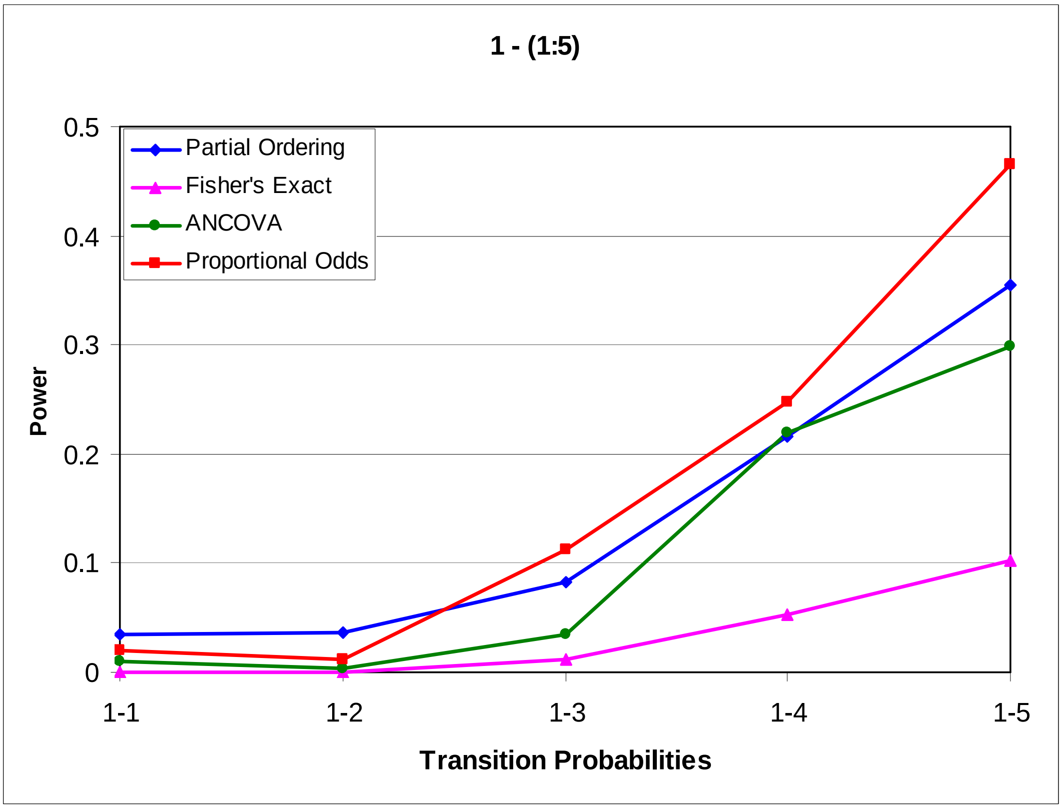
Form the plots, it can be seen that for most of the points, ANCOVA is more powerful than partial ordering in large sample N = 80. This means that for large sample size, even if the difference of two treatments is small, partial ordering is less powerful than ANCOVA and proportional odds test. Fisher's exact test always gives the lowest power.

Simulation 5:

Since our partial ordering does not show the obvious advantage over ANCOVA and proportional odds model, we decrease the sample size to 20 in balanced samples, to evaluate the power performance of our test. Five hundred permutations were selected to compute the p-value of partial ordering for each sample with α = 0.1. Table 9 and Figure 8 display the results.

Table 9: Simulation 5, balanced sample, N=20

Sample Size	Transition probabilities	Power (significance level = 0.1)			
		Partial Ordering	Fisher's Exact	ANCOVA	Proportional Odds
10*2	1-1	0.034	0	0.01	0.02
	1-2	0.036	0	0.004	0.012
	1-3	0.082	0.012	0.034	0.112
	1-4	0.216	0.052	0.22	0.248
	1-5	0.354	0.102	0.298	0.466
	2-2	0	0	0	0
	2-3	0.03	0	0	0
	2-4	0.144	0.004	0.098	0.132
	2-5	0.338	0.06	0.174	0.404
	2-6	0.416	0.078	0.304	0.446
	3-3	0.008	0	0	0
	3-4	0.026	0.002	0.02	0.06
	3-5	0.036	0.008	0.024	0.094
	3-6	0.064	0.008	0.03	0.104
	4-4	0.008	0	0	0
	4-5	0.03	0	0	0
	4-6	0.03	0	0	0.048
	4-7	0.286	0.156	0.522	0.434



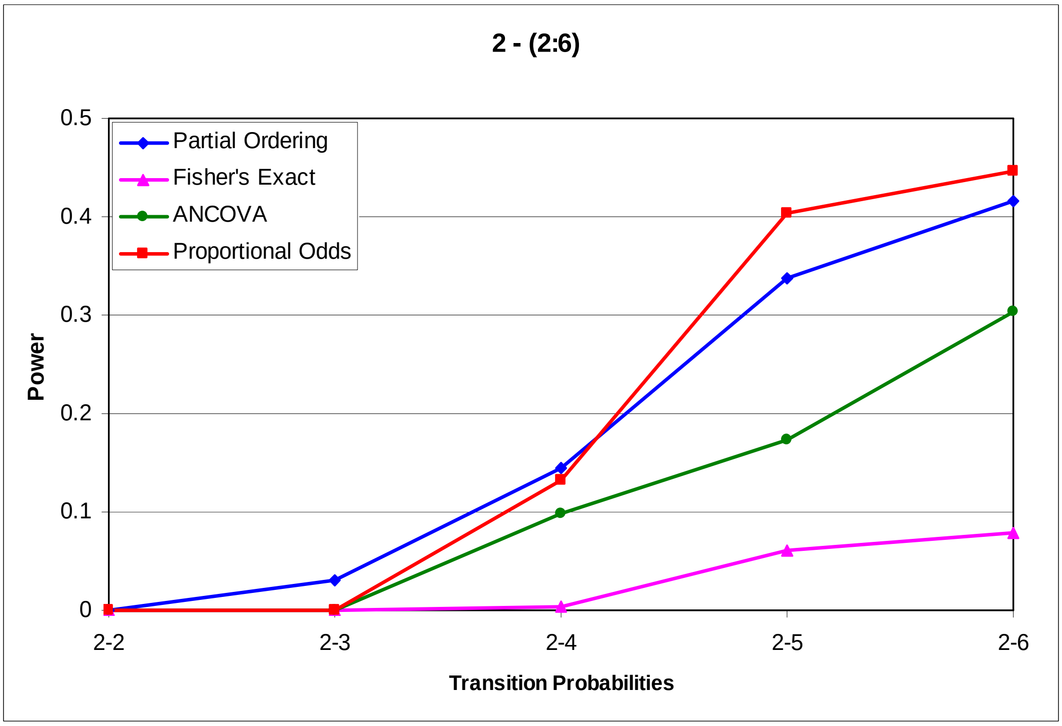
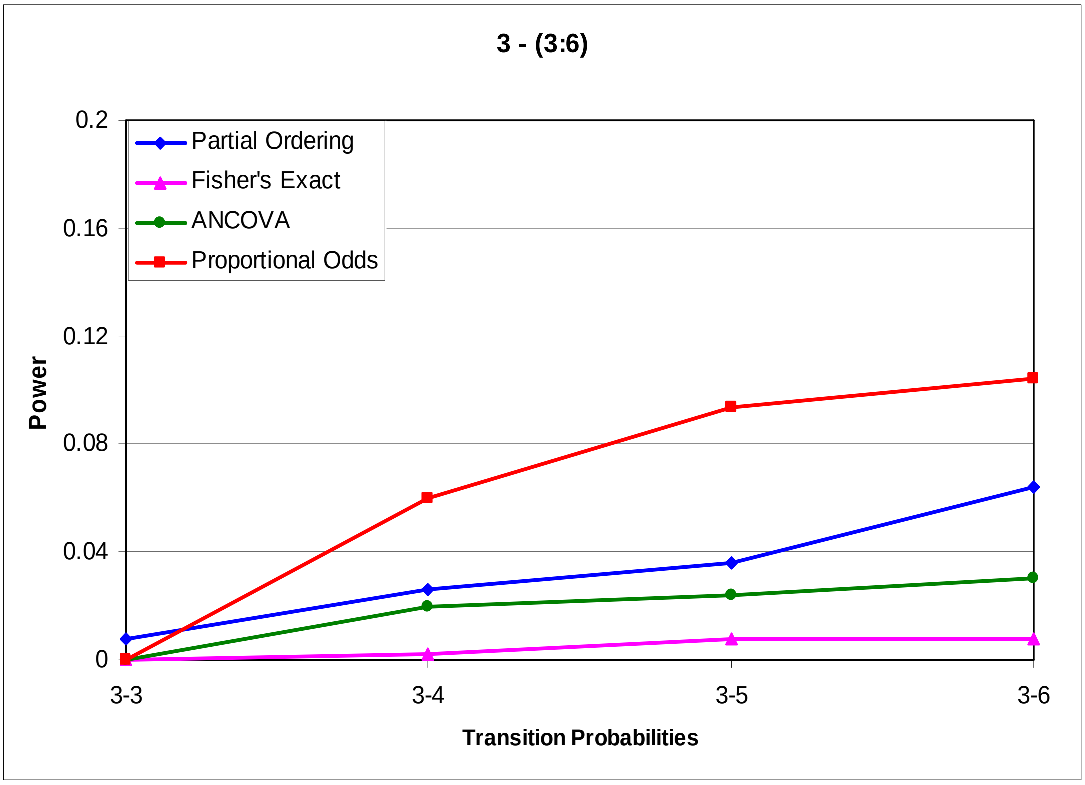


Figure 8(a): Simulation 5, balanced samples, N=20, α = 0.1



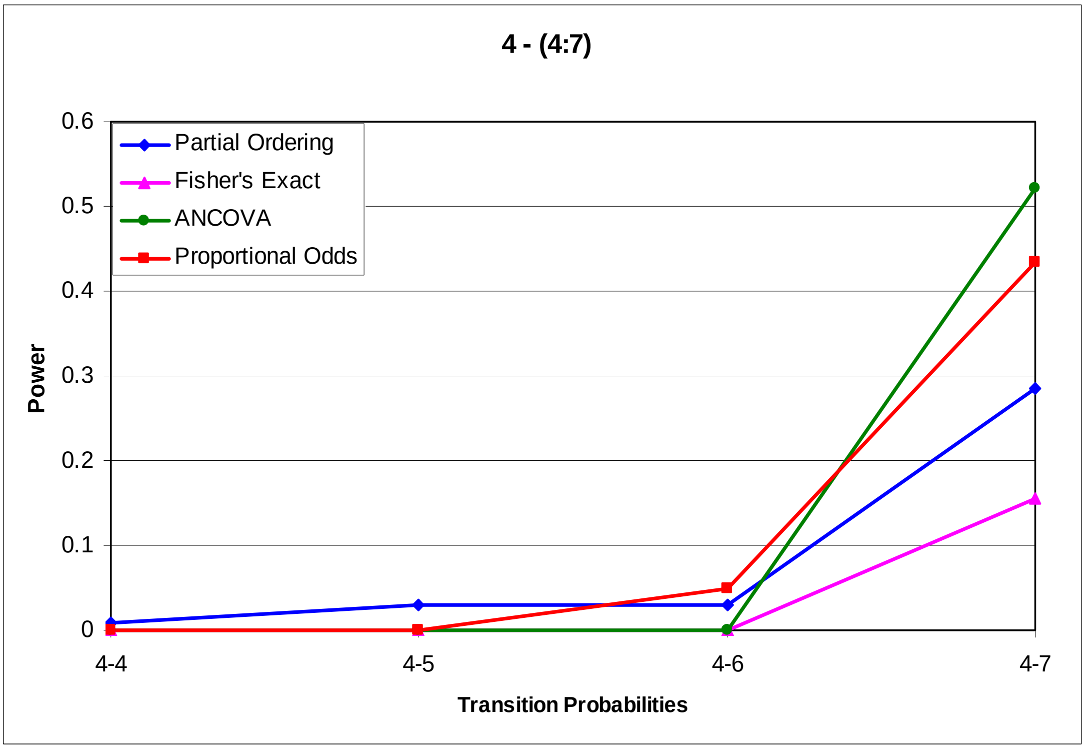


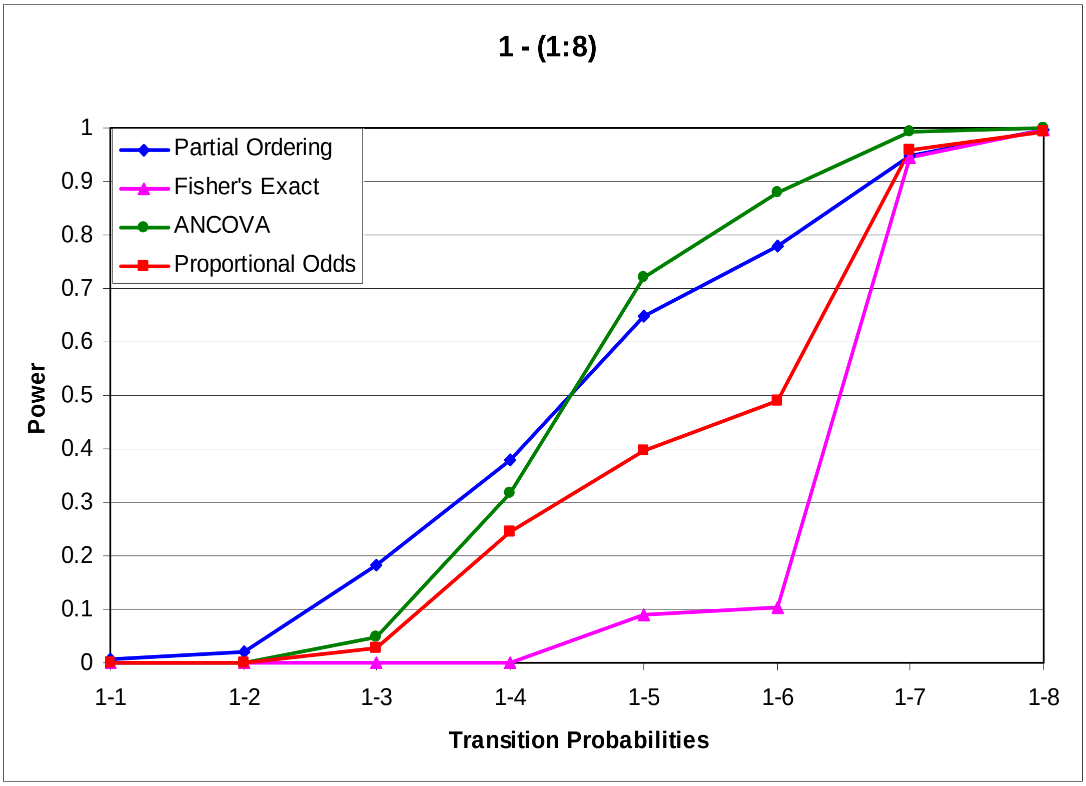
Figure 8(b): Simulation 5, balanced samples, N=20, α = 0.1

From the results, we can conclude that our partial ordering method shows great superiority over other tests in very small sample if the tiny difference between two treatments exists. Partial ordering is overall more powerful than ANCOVA and even better than proportional odds for some cases. It is slightly less conservative than others.

Simulation 6:

Table 10: Simulation 6, unbalanced samples, n1. =15, n2.=10

Sample Size	**Transition probabilities**	**Power (significance level = 0.1)**			
		Partial Ordering	**Fisher's Exact**	**ANCOVA**	**Proportional Odds**
Act=15 Con=10	1-1	0.006	0	0	0
	1-2	0.02	0	0	0
	1-3	0.182	0	0.05	0.026
	1-4	0.38	0	0.318	0.246
	1-5	0.648	0.09	0.72	0.398
	1-6	0.78	0.102	0.88	0.488
	1-7	0.948	0.944	0.994	0.958
	1-8	0.998	0.998	1	0.994
	2-2	0.006	0	0	0
	2-3	0.028	0	0	0
	2-4	0.156	0	0.214	0.056
	2-5	0.318	0.036	0.27	0.12
	2-6	0.396	0.058	0.43	0.208
	2-7	0.892	0.828	0.964	0.858
	2-8	0.99	0.992	0.996	0.952
	2-9	1	1	1	1
	3-3	0.004	0	0	0
	3-4	0.048	0	0.04	0
	3-5	0.122	0.006	0.062	0.036
	3-6	0.116	0	0.164	0.026
	3-7	0.786	0.46	0.93	0.698
	3-8	0.948	0.846	0.998	0.878
	3-9	1	1	1	1



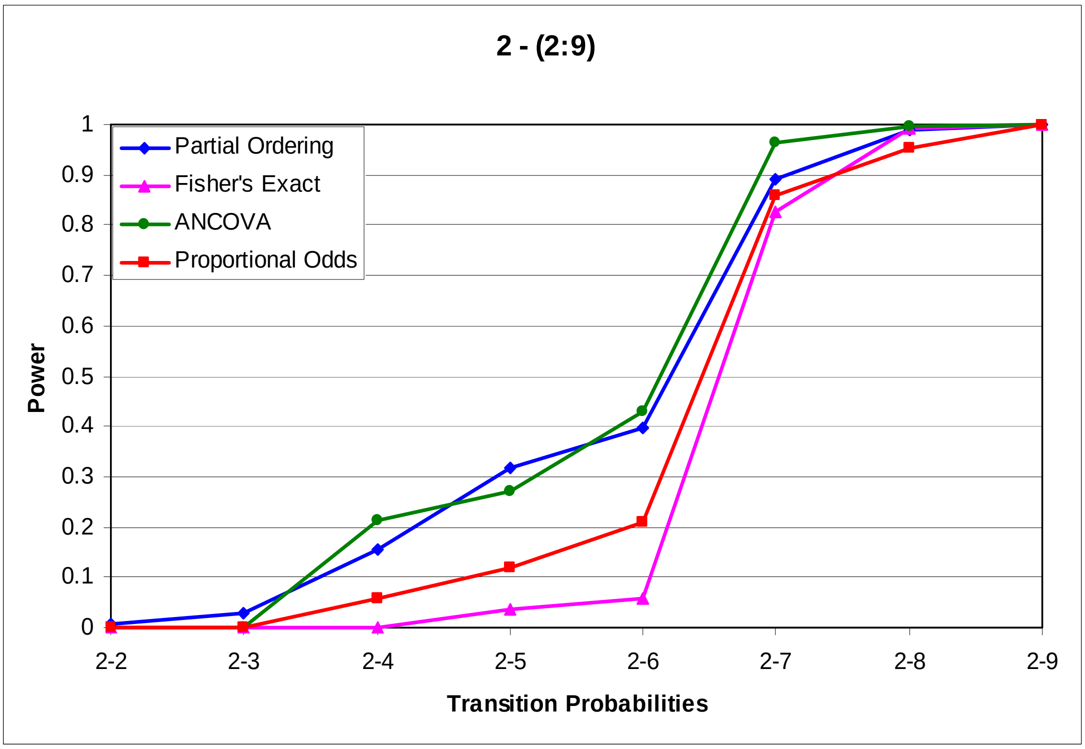


Figure 9(a): Simulation 6, unbalanced samples, $n_{1.}$ =15, $n_{2.}$ = 10, α = 0.1

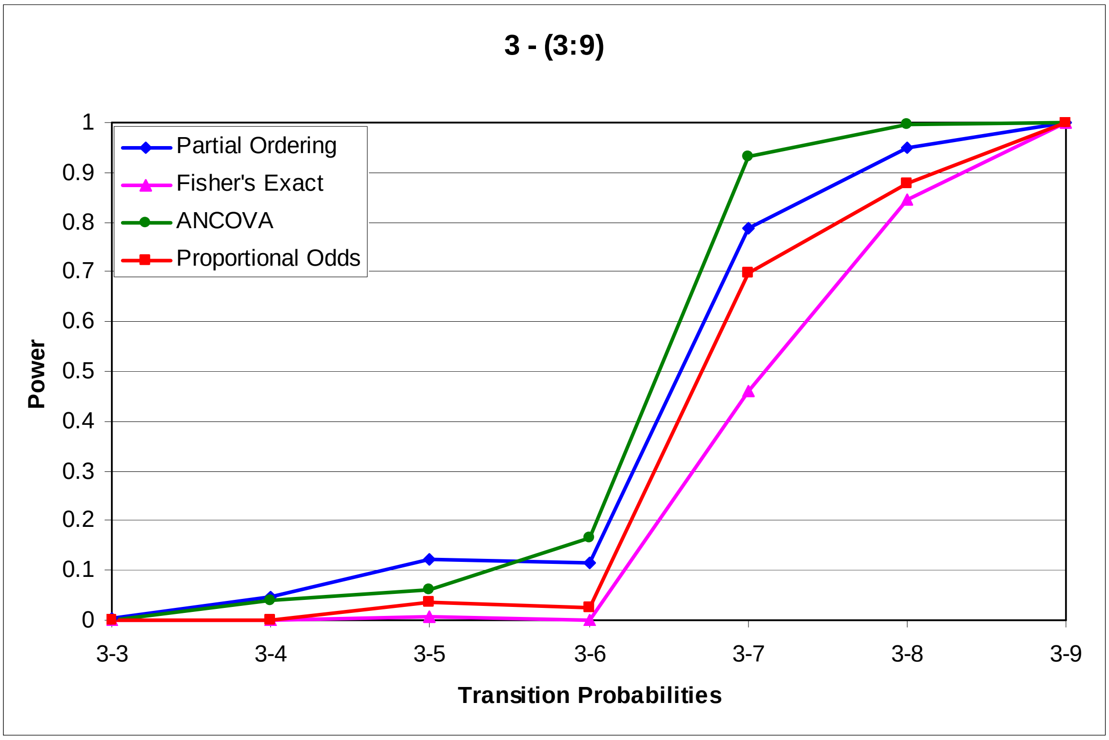


Figure 9(b): Simulation 6, unbalanced samples, $n_{1.}$ =15, $n_{2.}$ = 10, α = 0.1

Our previous simulations are based on balanced samples. A test based on balanced samples will be powerful than the unbalanced ones. Table 10 displays the power comparison for unbalanced small samples with size of active group $n_{1.}$ =15 and size of control group $n_{2.}$ = 10, given α = 0.1. The plots are shown in Figure 9.

By contrast, proportional odds model is always less powerful than partial ordering and ANCOVA. When the difference between two treatments is small, our partial ordering has the best performance in terms of power; otherwise, ANCOVA is the most powerful test among these four tests. Not surprisingly, the power of Fisher's exact test is lowest.

Chapter 6: Conclusion

6.1 Applications

Ordinal categorical data arise frequently in a wide variety of experimental studies. For example, in clinical trials comparing anti-inflammatory drugs in Rheumatoid arthritis, joint pain is often measured on an ordinal scale such as severe, moderate, little or none. A multi clinic study of analgesics relief of headache is recorded in ordinal responses as well. We can keep on going because the number of examples is really unlimited.

Such abundance of real examples makes statistical analysis of ordered categorical data an important field of study. In this thesis we focus on two-armed clinical trials designed to compare a new treatment to a control where for each patient an ordered categorical response is observed on entry to the study and at a single follow-up evaluation. Typically in an RCT where data arise from independent samples from two populations, each with same k distinct ordered categories, the problem is to test whether there is difference in the two multinomial populations. We have developed a new class of ordinal statistics that adjust for baseline differences based on partial ordering the baseline, post-treatment pairs. We have devoted a significant proportion of this thesis to compare the power performance of our test to those commonly used tests, Fisher's exact test, ANCOVA and proportional odds model, for categorical data analysis. Through a variety of simulation studies based on

2 × 6 contingency tables, we conclude that in small balanced or unbalanced samples, out test tends to outperform other three tests when the new experiment is slightly better than the control treatment, which is the normal case in practical studies. Thus, we provide a good option to conduct the statistical analysis if the sample size is limited or unbalanced. Also, clinical trials would have much greater sensitivity if this simple, but comprehensive, nonparametric methods can be used.

6.2 Future Work

One big concern about our simulation is the conservativeness of four tests involved. The mid-P-value, described by Lancaster (1961), modifies the exact test so that it rejects more often. For a test statistic T with observed value t_{obs} and one-sided H_A such that large T contradicts H_0,

$$Mid-P-value = P(T > t_{obs} \mid c) + \frac{1}{2} P(T = t_{obs}) ,$$

with probabilities calculated from the null distribution. Thus, the mid-*P*-value is lee than the ordinary p-value by half the probability of the observed result. Unlike an exact test with ordinary p-value, a test using the mid-*P*-value is less conservative. Although this test was not evaluated here, it may be particularly useful for our proposed test. Also, the conservativeness of other tests needs to be investigated deeply.

Secondly, it has been shown that our proposed test is not the most powerful test for all cases. Also it is really hard to find an optimal test for our problem (Berger and Ivanova, 2002). Our research based on the power performance in some common situations may then motivate future extensive investigation.

Bibliography

Agresti, A. (1990). *Categorical Data Analysis,* Wiley, New York.

Berger, V. W. and Exner, D. V. (1999). Detecting selection bias in randomized clinical trials. *Controlled Clinical Trials*, Vol. 20, 319-327.

Berger, V. W. (2000). Pros and cons of permutation tests in clinical trials. *Statistics in Medicine*, Vol. 19, 1319-1328.

Berger, V. W. and Ivanova, A. (2001). Drawbacks to integer scoring for ordered categorical data. *Biometrics*, Vol. 57, 567-570.

Berger, V. W. (2002). Improving the information content of categorical clinical trials endpoints. *Controlled Clinical Trials,* Vol. 23, 502-514.

Berger, V. W. and Ivanova, A. (2002). The bias of linear rank tests when testing for stochastic order in ordered categorical data. *Journal of Statistical Planning and Inference*, Vol. 107, 237-247.

Berger, V. W., Zhou, Y. Y., Ivanova, A. and Tremmel, L. (2004). Adjusting for ordinal covariates by inducing a partial ordering. *Biometrical Journal,* Vol.46 (1), 48-55.

Diaconis, P. and Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distribution. *The Annals of Statistics*, Vol. 26 (1), 363-397.

Laird, N. M. and Wang, F. (1990). Estimating rates of change in randomized clinical trials. *Controlled Clinial Trials*, Vol. 11, 405-419.

Lancaster, H. O. (1961). Significance tests in discrete distributions. *Journal of the American Statistical Association,* Vol. 56, 223-234.

Liang, K. Y., Zeger, S. L. and Qaqish, B. (1992). Multivariate Regression Analysis for categorical data (with discussion). *Journal of the Royal Statistical Society,* B 54, 3-40.

Lipsitz, S. R., Kim, K. and Zhzo, L. (1994). Analysis of repeated categorical data using generalized estimation equations. *Statistics in Medicine,* Vol. 13, 1149-1163.

Maurer, W. and Commenges, D. (1988). Choice and analysis of judgment criteria, Chapter 3 (pages 29-55). *Methods in Clinical Trials in Neurology,* R. Cupildeo and J. M. Urgoyoze, Editors, Stockton Press.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society*, Series B, Vol. 42, 109-142.

Mehta, C. R. and Patel, N. R. (1983). A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. *Journal of the American Statistical Association*, Vol. 78(382), 427-434

Morgan, W. M. and Blumenstein, A. B. (1991). Exact conditional tests for hierarchical models in multidimensional contingency tables. *Applied Statistics*, Vol. 40(3), 435-442..

Moses, L. E., Emerson, J. D., and Hosseini, H. (1984). Analyzing data from ordered categories. *New England Journal of Medicine*, Vol. 311, 442-448.

Munzel U. and Tamhane A. C. (2002). Nonparametric multiple comparisons in repeated measures design for data with ties. *Biometrical Journal*, Vol. 44(6), 762-779.

Pagano, M. and Halvorsen, K. T. (1981). An algorithm for finding the exact significance levels of $r \times c$ contingency tables. *Journal of the American Statistical Association*, Vol. 76(376), 931-934.

Pagano, M. and Tritchler, D. (1983). On obtaining permutation distributions in polynomial time. *Journal of the American Statistical Association*, Vol. 78(382), 435-440.

Patefield, W. M. (1981). Algorithm AS 159: An efficient method of generating random $R \times C$ tables with given row and column totals. *Applied Statistics*, Vol. 30, 91-97.

Rahlfs, V. W. and Zimmermann, H. (1993). Scores: ordinal data with few categories – How should they be analyzed? *Drug Information Journal*, Vol**.** 27, 1227-1240.

Rencher, A. C. (2000). *Linear Models in Statistics*. Wiley, New York.

Senn, S. J. (1989).The use of baselines in clinical trials. *Statistics in Medicine,* Vol.13, 1715-1726.

Senn, S. J. (1994). Testing for baseline balance in clinical trials. *Statistics in Medicine,* Vol. 13, 1715-1726.

Serfling R. J. (1988). U-statistics. *The Encyclopedia of Statistical Science*, Vol. 9, S. Kotz and N. L. Johnson, Editors, Wiley, New York.

Stevens, S. S. (1951). Mathematics, measurement and psychophysics. In S.S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1-49). New York: Wiley.