学位論文

Various Methods for Analysis in Clinical Trials with Categorical Endpoints

(カテゴリカル変数を評価項目とした臨床試験における種々の解析法)

2023年3月

石原拓磨

Contents

Chapter 1

Introduction

This thesis presents testing procedures focusing on multiple binary endpoints in clinical trials and a measure in the analysis of ordinal categorical variables using contingency tables. This chapter discusses the background and previous research. Additionally, it describes the purpose of this thesis and outlines the main results of this study.

1.1 Background

Discussion of the choice of the primary endpoints is essential in the preliminary steps of a confirmatory clinical trial. In pilot studies and exploratory phases of clinical trials, which of several endpoints will be used as the primary endpoints for the confirmatory phase of the trial will be discussed. While the CONSORT statement provides guidance on how to report the results of randomized controlled trials, there is no guidance to inform the optimal choice of endpoints or methodologies available to quantify endpoints as best. There is a increasing recognition that trial endpoints do not always capture outcomes that are clinically meaningful to patients and clinicians, and therefore, patient preference information must also be considered (McLeod, 2019). Therefore, the choice of endpoints could be better enhanced by understanding their characteristics and properties, and the importance related to this discussion is growing. If the type of primary endpoints used in the trials is a categorical variable, using contingency tables to understand characteristics among multiple endpoints may provide good information to stimulate discussion.

Contingency tables are a basic tool used to examine the relationship between row

and column of ordinal categorical variables. For example, the Pearson χ^2 statistic is commonly used to test the null hypothesis of statistical independence (Agresti, 2013). When statistical independence is rejected, we are interested in describing the association between the row and column categories. Summary measures of association have been proposed, such as the Cramér V, gamma, and uncertainty coefficient. For details, see for instance Agresti (2013, Sec. 2.4) and Bishop et al. (1975, Sec. 11.3). Additionally, the recent development of association measures is described, for example, in Beh et al. (2007), Lombardo (2011), Wei and Kim (2017, 2021), Zhang et al. (2021), and Wei et al. (2022).

Contingency tables with the same row and column classifications are called square contingency tables. These tables are used for unaided distance vision data, social mobility data, and longitudinal data in biomedical research. The analysis of square contingency tables considers the issue of symmetry rather than independence because it is not sensible to treat these data as independent.

Bowker (1948) introduced the simple symmetry model and proposed a test for the hypothesis of symmetry. When the symmetry model fits the given data poorly, we are interested in measuring the degree of departure from symmetry. Tomizawa (1994) proposed a measure that represents the degree of departure from symmetry expressed using the Shannon entropy or Kullback–Leibler information. In the real world, the Shannon entropy is widely applied as a measure of complexity, for example in Fernandes and Araújo (2020). The measure lies between 0 and 1, and its value equals 0 if and only if the symmetry model holds. Additionally, the degree of departure from symmetry increases as the value of the measure increases.

In clinical trials, it is important to statistically evaluate the similarity of the measures and to plan in advance whether multiple endpoints should be used. The identification of one statistical property, symmetry, will contribute to the characterization of key variables.

Consider a confirmatory clinical trial in which the result of variable choices was to use several correlated binary response variables. For example, in recent clinical trials for patients with psoriasis, the percentage of patients with at least 75 percent improvement in the psoriasis area-and-severity index (PASI) score and the percentage with a physician's global assessment score of 0 or 1 at weeks 24 and 52 were used as primary endpoints to demonstrate the superiority of a test treatment (Reich et al., 2011). In clinical trials of rheumatoid arthritis, the percentage of patients achieving a

20 percent short-term improvement in the American College of Rheumatology criteria (ACR20) and the percentage achieving long-term low disease activity (Disease Activity Score: DAS28-ESR ≤ 3.2) are often used as primary endpoints (Smolen et al., 2016). In IgA nephropathy affecting multiple areas, full clinical remission and a decrease in estimated Glomerular Filtration Rate (eGFR) of at least 15 mL per minute per 1.73 m^2 from baseline are sometimes used in primary assessment (Rauen et al., 2015). In the above trials, all of the primary endpoints were binary and often have a continuous latent distribution.

In clinical trials for a variety of diseases, it is often useful to evaluate efficacy using multiple primary endpoints. Most trials use multiple endpoints only to evaluate non-inferiority or superiority, but some trials have been conducted to confirm the non-inferiority and superiority of all endpoints. For example, a clinical trial to confirm the efficacy of four-factor prothrombin complex concentrate (4F-PCC) included two primary endpoints (Goldstein et al., 2015), namely the percentage of patients with a hemostatic effect and the percentage with a decrease in the international normalized ratio (INR). In the above trial, superiority was only evaluated if there was non-inferiority for both endpoints. When we confirm not only non-inferiority but also superiority, the use of the closed testing procedure (Marcus et al., 1976) for the primary analysis is reasonable, and in this case, no adjustment is needed to control the type I error rate. In general, however, it is difficult to demonstrate the superiority of two or more endpoints because the power decreases as the number of endpoints increases.

For multiple continuous endpoints, Perlman and Wu (2004) proposed a testing procedure that is applicable to the framework mentioned above. Moreover, Tamhane and Logan (2004) suggested the Union-intersection(UI)-Intersection-union(IU) test and showed that their method could be controlled Type I error in the same way as the Perlman and Wu's procedure. Nakazuru et al. (2014) proposed a more powerful procedure by modifying Perlman and Wu's procedure using the approximate likelihood ratio test (ALRT) defined by Glimm et al. (2002). However, methods using multiple binary endpoints in the above frameworks are not well enough developed.

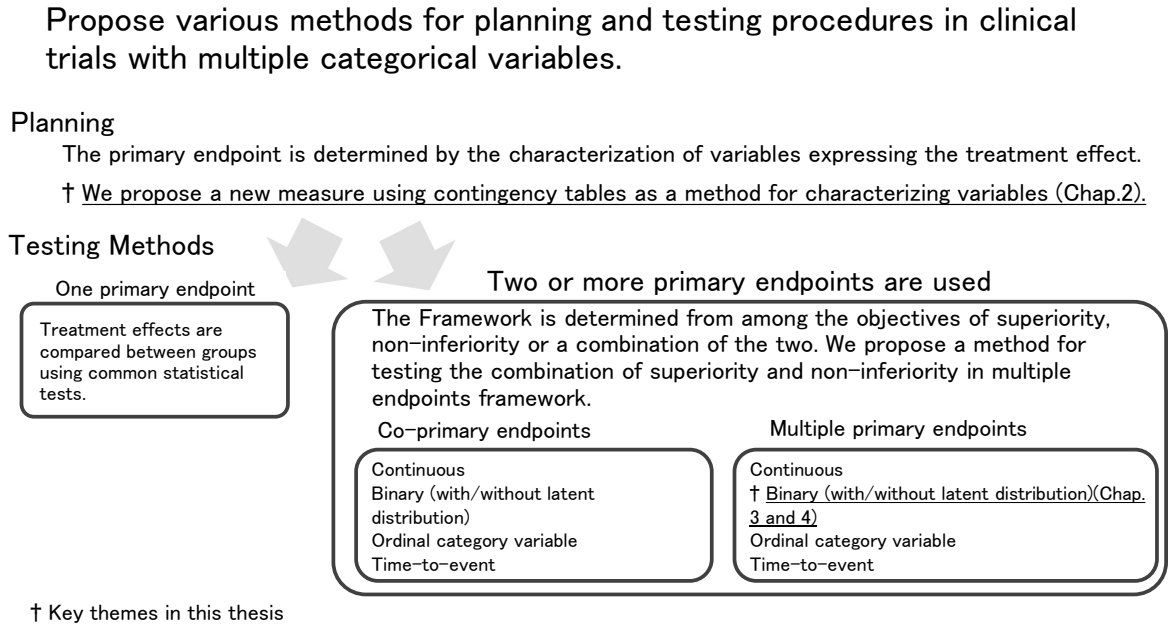
The a priori evaluation of the characteristics of multiple endpoints and the use of statistical methods appropriate to the purpose of the study will contribute to the conduct of high-quality clinical trials.

1.2 Research objectives

The objective of this thesis is to develop a methodology for the design and analysis of clinical trials with multiple categorical variables as primary endpoints. Figure 1 shows the schema of the study. When planning a clinical trial, it is necessary to characterize the categorical variables used in the trial. There are few reports of procedures that evaluate the characteristics among categorical variables using partial measures to express the degree of departure from partial symmetry in contingency tables. Furthermore, the accuracy of estimation is essential for methods applied in clinical trials. We should also consider proposing a measure with small variance for application to the design of clinical trials. The first objective is to propose an idea for determining which category of the two categorical variables is asymmetric, using confidence intervals for the partial measures. This objective also includes proposing a new symmetry measure with minimum variance that combines those partial measures in the class of weighted averages.

In addition, new testing procedures need to be developed that can be applied in clinical trials with multiple binary endpoints as primary endpoints. In the framework that recognizes a treatment effect only if there is superiority on at least one endpoint and non-inferiority on the remaining endpoints, no testing procedure has been developed that uses multiple binary variables as the primary endpoints. Therefore, the second objective is to propose a testing procedure that is appropriate when all endpoints are binary with a framework in which the treatment effect is confirmed only when there is superiority of at least one endpoint and non-inferiority of the remaining endpoints. In addition, the objective includes proposing a testing procedure for the case where the binary endpoints have a latent distribution.

Figure 1 Schema of thesis



1.3 Analysis of Contingency Table

1.3.1 Contingency Table

When an endpoint consists of several categories and an individual's data falls into one of the categories, the endpoint can be considered categorical variable. Categorical data consists of the frequency of observations occurring in each category. Let Y_j be a categorical variable and have R levels. When considering two endpoints with R levels, we can represent the data Y_1, Y_2 in a rectangular table with $R \times R$ cells. A table in this format, where the cells contain the frequency counts of the results, is called a contingency table. In particular, a contingency table in which the rows and columns consist of the same level is called a square contingency table.

1.3.2 Measures of departure from symmetry in contingency tables

Consider an $R \times R$ square contingency table having the same row and column classifications. Let p_{rc} denote the probability that an observation will fall in the (r, c)th cell of the table $(r = 1, \ldots, R; c = 1, \ldots, R)$. The simple symmetry model introduced by Bowker (1948) is defined by

$$p_{rc} = p_{cr} \quad (r \neq c).$$

When the symmetry model does not hold for a given dataset, we are interested in evaluating the degree of departure from symmetry. Assuming $p_{rc} + p_{cr}$ is not equal to zero for $r < c$, the measure is defined as

$$\Phi_T = \frac{1}{\delta \log 2} \sum \sum_{r \neq c} p_{rc} \log \frac{2p_{rc}}{p_{rc} + p_{cr}},$$

where $\delta = \sum \sum_{r \neq c} p_{rc}$.

Let $\pi_{rc} = p_{rc}/(p_{rc} + p_{cr})$ for $r = 1, \ldots, R; c = 1, \ldots, R; r \neq c$. The conditional probability that an observation falls in cell (r, c) or (c, r) in the table is π_{rc}. The measure Φ_T can be expressed as

$$\Phi_T = \sum \sum_{r<c} \left(\frac{p_{rc} + p_{cr}}{\delta} \right) \phi_{rc},$$

where

$$\phi_{rc} = \frac{1}{\log 2} \left(\pi_{rc} \log \frac{\pi_{rc}}{1/2} + \pi_{cr} \log \frac{\pi_{cr}}{1/2} \right).$$

It should be noted that ϕ_{rc} is the normalized Kullback–Leibler information between (π_{rc}, π_{cr}) and $(1/2, 1/2)$. That is, the measure Φ_T is the weighted average of ϕ_{rc}.

1.4 Multiple testing procedure

This section introduces the problems in statistical inference with multiple endpoints and the test hypotheses. We focus on a randomized clinical trial comparing p (≥ 2) endpoints with two treatment groups. There are n_1 subjects in the test group and n_2

subject in the control group. Let Y_{ijk} $(i = 1, 2; j = 1, \ldots, p; k = 1, \ldots, n_i)$ denote the response variable of the jth primary endpoint of the ith treatment in the kth subject. Set $\boldsymbol{X} = (X_1, \ldots, X_p)^t$ with $X_j = (\overline{Y}_{1j} - \overline{Y}_{2j})$ $(j = 1, \ldots, p)$, where $\overline{Y}_{ij}$ $(i = 1, 2; j = 1, \ldots, p)$ is the sample mean or proportion for the jth endpoint to the ith treatment.

1.4.1 Multiple testing principles

(a) Familywise error rate

The concept of a Type I error rate originated from the problem of testing a single hypothesis. The Type I error rate is defined as the probability of rejecting the null hypothesis when it is true. For example, consider testing each of the p endpoints in the treatment and control groups. Let μ_{1j} be the true mean of the response to the jth endpoint in the treatment group and μ_{2j} be the true mean of the response to the jth endpoint in the control group. The treatment effect of the jth endpoint in each group is assumed to be positive if μ_{ij} is positive. The null hypothesis that the difference in treatment effects of jth endpoint is not greater than 0

$$H_{0(j)} : \mu_{1j} - \mu_{2j} \leq 0$$

is tested versus a one-sided alternative

$$H_{1(j)} : \mu_{1j} - \mu_{2j} > 0.$$

The Type I error rate for $H_{0(j)}$ is the probability of concluding that hypothesis $H_{0(j)}$ is rejected when the difference in treatment effects actually no grater than 0. Each of the p null hypotheses is tested so that the proportion of incorrect rejections does not exceed the significance level α, i.e. $\alpha = 0.05$. There is interest for the family of null hypotheses when concluding treatment effects using multiple endpoints.

Familywise error rate (FWER) is defined as the probability of rejecting at least one true null hypothesis. In particular, the FWER calculated under the assumption that several p null hypotheses are simultaneously true is known as *weak control*. On the other hand, if T is the index set of true null hypotheses, *strong control of the FWER* is

require that

$$\sup \text{FWER} = \max_{T} \sup_{\substack{\{\mu_{1j}(T)\} \\ \{\mu_{2j}(T)\}}} P\left(\text{Reject at least one } H_{0(j)}, j \in T\right) \leq \alpha$$

where the supremum is taken over all $\mu_{1j} - \mu_{2j}$ satisfying less than 0 for $j \in T$, over 0 for $j \notin T$, and the maximum is taken over all index sets T. Strong control of the FWER for primary objectives is mandated by regulators in all confirmatory clinical trials.

(b) Union-Intersection testing

In clinical trials, multiple testing can generally be formulated as a union-intersection problem. In the union-intersection framework, if there is evidence of a treatment effect for at least one individual objective, the global hypothesis of no effect is rejected.

Let $H_{0(h)}$ denote the null hypothesis, and $H_{1(h)}$ denote the alternative hypothesis corresponding to the hth objective, $h = 1, \ldots, M$. H_0 defined as the intersection of the hypotheses, is tested versus the union of the alternative hypotheses H_1:

$$H_0 : \bigcap_{h=1}^{M} H_{0(h)} \text{ versus } H_1 : \bigcup_{h=1}^{M} H_{1(h)}.$$

In union-intersection testing, conducting individual tests at unadjusted α levels leads to an inflated probability of rejecting H_0 and compromises the validity of statistical inference. To address this problem, a multiplicity adjustment method needs to be utilized to control the appropriately defined probability of a Type I error.

(c) Intersection-Union testing

Intersection-union testing is a test where significant results for two or more objectives are required simultaneously in order to recognize a treatment effect. The intersection-union method involves testing the union of the null hypotheses H_0 against the intersection of the alternative hypotheses H_1:

$$H_0 : \bigcup_{h=1}^{M} H_{0(h)} \text{ versus } H_1 : \bigcap_{h=1}^{M} H_{1(h)}.$$

When global hypothesis H_0 is rejected, one concludes that all H_1's are true, i.e., there is evidence of positive effect with respect to all of the M objectives.

An interesting feature of the intersection-union test is that no multiplicity adjustment is needed to control for FWER, i.e. individual tests can be performed at nominal significance levels.

(d) Hypothesis based on a combination of superiority and non-inferiority tests

In clinical trials, we assume that superiority of test treatment is recognized when the responses to the test treatment is greater than that to the control treatment. That is, a maximum value of $\delta_j (j = 1, \ldots, p)$ greater than 0 indicates an improvement of at least one endpoint of the test treatment compared to the control treatment. Furthermore, we consider the combined hypothesis for the superiority of at least one endpoint and the non-inferiority of the remaining endpoints. Therefore, we consider a null hypothesis H_0 and an alternative hypothesis expressed by

$$H_0 : \left\{ \max_{1 \leq j \leq p} \delta_j \leq 0 \right\} \cup \left\{ \min_{1 \leq j \leq p} (\delta_j + \epsilon_j) \leq 0 \right\} \text{ versus } H_1 : \text{not } H_0,$$

where $\epsilon_j > 0$ $(j = 1, \ldots, p)$ is the non-inferiority margin of the jth endpoint that denotes a prespecified positive constant. H_0 is also expressed as

$$H_0 \equiv H_0^{(0)} \cup \left\{ H_0^{(1)} \cup \cdots \cup H_0^{(p)} \right\},$$

which defines the sub hypothesis of superiority "$H_0^{(0)} : \max_{1 \leq j \leq p} \delta_j \leq 0$" and the sub hypothesis of non-inferiority "$H_0^{(j)} : \delta_j \leq -\epsilon_j$", for $j = 1, \ldots, p$. $H_0^{(0)}$ is adaptable to the one-sided ALRT, and the IUT (Berger, 1982) can be applied to test H_0.

1.5 Outline of Chapters

The rest of this thesis is organized as follows. Chapter 2 proposes a measure that represents the degree of departure from symmetry in the class of weighted averages that has the smallest variance. This chapter is based on Ishihara, Yamamoto, Tahata and Tomizawa (2022). Chapter 3 proposes a testing procedure for multiple primary endpoints that are binary with a framework in which the treatment effect is confirmed only when there is the superiority of at least one endpoint and non-inferiority of the re-

maining endpoints. This chapter is based on Ishihara and Yamamoto (2021a). Chapter 4 proposes a testing procedure for multiple binary endpoints with latent distribution with a framework in which the treatment effect is confirmed only when there is the superiority of at least one endpoint and non-inferiority of the remaining endpoints. This chapter is based on Ishihara and Yamamoto (2021b). Finally, Chapter 5 presents the conclusions of this study and suggests future research.

Chapter 2

Regarding Chapter 2, we propose a symmetry measure in the class of weighted averages.

Let n_{rc} denote the observed frequency in the (r, c)th cell of the contingency table $(r = 1, \ldots, R;\ c = 1, \ldots, R)$. We consider the weighted average of estimator of $\{\phi_{rc}\}$, that is

$$\Phi = \sum_{r<c} \sum w_{rc} \hat{\phi}_{rc},$$

where $\hat{\phi}_{rc}$ is the estimator of ϕ_{rc}, the weights $\{w_{rc}\}$ satisfy all $w_{rc} > 0$, and $\sum_{r<c} \sum w_{rc} = 1$. In addition, $\hat{\phi}_{rc}$ $(r < c)$ is asymptotically distributed normal as $N(\phi_{rc}, \sigma_{rc}^2)$ independently. Thus, the measure Φ has an asymptotically normal distribution with mean

$$\sum_{r<c} \sum w_{rc} \phi_{rc},$$

and variance

$$\sigma^2 = \sum_{r<c} \sum (w_{rc})^2 \sigma_{rc}^2.$$

In an analogous manner to Agresti (1984, p.170), we derive the weights $\{w_{rc}^*\}$ so as to minimize the variance of Φ with the constraint that all $w_{rc} > 0$ and $\sum_{r<c} \sum w_{rc} = 1$. We obtain

$$\boldsymbol{w}_{rc}^* = \frac{1/\sigma_{rc}^2}{\sum_{s<t} \sum 1/\sigma_{st}^2},$$

$$\Phi_S = \sum_{r<c}\sum w^*_{rc}\hat{\phi}_{rc}.$$

This measure has the smallest variance among measures in the class of weighted averages. It should be noted that we should estimate the variances $\{\sigma^2_{rc}\}$ because these are unknown.

We propose the estimated measure as follows:

$$\hat{\Phi}_S = \sum_{r<c}\sum \hat{w}^*_{rc}\hat{\phi}_{rc},$$

where $\hat{w}^*_{rc}$ is given by w^*_{rc} with $\{p_{rc}\}$ replaced by $\{\hat{p}_{rc}\}$. The proposed measure approximates the measure in the class of weighted averages that has the smallest variance.

Chapter 3

Regarding Chapter 3, we propose the intersection-union test (IUT) statistic when multiple endpoints are binary. From the multivariate central limit theorem, $\boldsymbol{X} = (X_1, \ldots, X_p)^t$ is approximately normally distributed with mean $\boldsymbol{\Delta}$ and the covariance matrix $\boldsymbol{\Sigma}$. Let $\boldsymbol{A}$ be the positive definite matrix such that $\boldsymbol{A}^t\boldsymbol{A} = \hat{\boldsymbol{\Sigma}}^{-1}$ where $\boldsymbol{\Sigma}^{-1}$ is the inverse matrix of $\boldsymbol{\Sigma}$, and $\hat{\boldsymbol{\Sigma}}$ is the estimated covariance matrix. Furthermore, according to the procedure of Nakazuru et al. (2014), $\boldsymbol{B}$ is defined as the matrix substituting the off-diagonal elements of $\boldsymbol{A}$ with their absolute values. Consider the two transformations such that

$$\begin{aligned}
\boldsymbol{u}_A \equiv (u_{A1}, \ldots, u_{Ap})^t &= \boldsymbol{AX} \quad \text{and} \\
\boldsymbol{u}_B \equiv (u_{B1}, \ldots, u_{Bp})^t &= \left(\frac{\det\boldsymbol{A}}{\det\boldsymbol{B}}\right)^{2/p}\boldsymbol{BX}.
\end{aligned}$$

The proposed IUT rejects H_0 if and only if

$$\begin{aligned}
T^{(0)} &: \min(\overline{u}^2_A, \overline{u}^2_B) > d \quad \text{and} \\
T^{(j)} &: \frac{X_j + \epsilon_j}{\sqrt{\frac{\acute{\pi}_{1j}(1-\acute{\pi}_{1j})}{n_1} + \frac{\acute{\pi}_{2j}(1-\acute{\pi}_{2j})}{n_2}}} > z_\alpha \quad \text{for} \quad j = 1, \ldots, p,
\end{aligned}$$

where $\overline{u}_A^2$ and $\overline{u}_B^2$ are defined by

$$\begin{aligned}\overline{u}_A^2 &= \sum_{j=1}^{p} \max(u_{Aj}, 0)^2, \\ \overline{u}_B^2 &= \sum_{j=1}^{p} \max(u_{Bj}, 0)^2.\end{aligned}$$

z_α represents the upper 100α th percentile of the standard normal distribution. $\acute{\pi}_{ij}$ is the MLE of π_{ij} derived under the sub null hypothesis of non-inferiority $H_0^{(j)}$, and $T^{(j)}$ is a statistic commonly used in non-inferiority tests of binary endpoints (Farrington and Manning, 1990). Further, d is a constant determined by

$$\sum_{j=0}^{p} \frac{p!}{j!\,(p-j)!} \frac{1}{2^p} Pr(\chi_j^2 > d) = \alpha.$$

Here, χ_j^2 denotes the χ^2 distribution with j degrees of freedom, χ_0^2 is defined as the constant zero, and α is the nominal significant level.

$T^{(0)}$ and $T^{(j)}$ are test statistics corresponding to the null hypotheses for superiority for at least one endpoint $H_0^{(0)}$ and non-inferiority for all endpoints $H_0^{(1)}, \ldots, H_0^{(p)}$, respectively. Two types of estimators for $T^{(0)}$ are proposed by obtaining the estimated covariance matrix $\hat{\boldsymbol{\Sigma}}$, which can be derived under the sub null hypothesis or the sub alternative hypothesis. In particular, $T_0^{(0)}$ is defined as the statistic for testing $H_0^{(0)}$ using an estimator obtained under the sub null hypothesis, and $T_1^{(0)}$ is the statistic for testing $H_0^{(0)}$ using an estimator obtained under the sub alternative hypothesis $H_1^{(0)}$ for testing superiority for at least one endpoint.

Chapter 4

Regarding Chapter 4, we propose the IUT statistic when multiple endpoints are binary and those have latent continuous distribution.

We assume that $\boldsymbol{Y}_{ik}$ are dichotomized random variables of continuous unobservable response $\boldsymbol{Z}_{ik} = (Z_{i1k}, \ldots, Z_{ipk})^t$ $(i = 1, 2; k = 1, \ldots, n_i)$. We also assume that $\boldsymbol{Z}_{ik}$ are independently distributed as a standardized p-variate normal distribution with $\mathrm{Corr}(Z_{ijk}, Z_{ij'k}) = \gamma_{(i)jj'}$ for all $j \neq j'$. For each variable $\boldsymbol{Z}_{ik}$, there is a single threshold $g_{ij} = \Phi^{-1}(1 - \pi_{ij})$ $(i = 1, 2; j = 1, \ldots, p)$ that partitions the latent distribution, where Φ^{-1} is the inverse function of the standard normal cumulative distribution function.

Then, the binary response Y_{ijk} $(i = 1, 2; j = 1, \ldots, p; k = 1, \ldots, n_i)$ can be defined as

$$Y_{ijk} = \begin{cases} 1, & Z_{ijk} \geq g_{ij} \\ 0, & Z_{ijk} < g_{ij}. \end{cases}$$

Let the true proportion vector be the ith treatment $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{ip})^t$ with difference of proportion $\boldsymbol{\Delta} = (\delta_1, \ldots, \delta_p)^t = \boldsymbol{\pi}_1 - \boldsymbol{\pi}_2$ and the covariance matrix $\boldsymbol{\Sigma}$ that is defined as follows:

$$\begin{aligned} \boldsymbol{\Sigma} &= \boldsymbol{\Sigma}^{(1)} + \boldsymbol{\Sigma}^{(2)} \\ &= \frac{1}{n_1} \begin{pmatrix} \pi_{11}(1-\pi_{11}) & \cdots & \rho_{(1)1p}\sqrt{\pi_{11}(1-\pi_{11})}\sqrt{\pi_{1p}(1-\pi_{1p})} \\ \vdots & \ddots & \vdots \\ \rho_{(1)p1}\sqrt{\pi_{11}(1-\pi_{11})}\sqrt{\pi_{1p}(1-\pi_{1p})} & \cdots & \pi_{1p}(1-\pi_{1p}) \end{pmatrix} \\ &+ \frac{1}{n_2} \begin{pmatrix} \pi_{21}(1-\pi_{21}) & \cdots & \rho_{(2)1p}\sqrt{\pi_{21}(1-\pi_{21})}\sqrt{\pi_{2p}(1-\pi_{2p})} \\ \vdots & \ddots & \vdots \\ \rho_{(2)p1}\sqrt{\pi_{21}(1-\pi_{21})}\sqrt{\pi_{2p}(1-\pi_{2p})} & \cdots & \pi_{2p}(1-\pi_{2p}) \end{pmatrix}, \end{aligned}$$

where $\boldsymbol{\Sigma}^{(i)}(i = 1, 2)$ is the covariance matrix of $(\overline{Y}_{i1}, \ldots, \overline{Y}_{ip})^t$. Note that $\boldsymbol{X}$ is approximately normally distributed with mean $\boldsymbol{\Delta}$ and covariance matrix $\boldsymbol{\Sigma}$. The correlation coefficient $\rho_{(i)jj'}$ is expressed as

$$\rho_{(i)jj'} = \frac{\phi_{(i)jj'} - \pi_{ij}\pi_{ij'}}{\sqrt{\pi_{ij}(1-\pi_{ij})}\sqrt{\pi_{ij'}(1-\pi_{ij'})}},$$

where $\phi_{(i)jj'}$ is the joint probability of two response variables $(Y_{ijk}, Y_{ij'k})$.

To determine the statistics of IUT, we need to estimate several parameters. For the sake of simplicity, the process of constructing statistics is divided into the following four steps.

Step1. Estimating the cut-off point g_{ij}

We assume that $\hat{g}_{ij}$ is the estimator of the latent cut-off point g_{ij}, and is estimated as $\hat{g}_{ij} = \Phi^{-1}(1 - \widetilde{\pi}_{ij})$, where $\widetilde{\pi}_{ij}$ is the maximum likelihood estimator (MLE) of the marginal probability derived by the p-variate Bernoulli distribution. $\widetilde{\pi}_{ij}$ (and $\hat{g}_{ij}$) can be given in two ways depending on the estimation under the sub-null hypothesis $H_0^{(0)}$ or the sub alternative hypothesis $H_1^{(0)}$: not $H_0^{(0)}$.

Step2. Estimate the joint and marginal probabilities

The estimator of the joint probability $\phi_{(i)jj'}$ is given in two ways, depending on the estimatiors $\widetilde{\pi}_{ij}$ that determines $\hat{g}_{ij}$. Furthermore, the estimator of the marginal probability π_{ij} constructing Σ and $\rho_{(i)jj'}$ should not be $\widetilde{\pi}_{ij}$, which is obtained from the p-variate Bernoulli distribution, but should instead take into account the latent distribution function. Let $\hat{\pi}_{ij}$ denote the estimator of π_{ij} and be given by $\hat{\pi}_{ij} = \text{Prob}(Z_{ij} \geq \hat{g}_{ij})$.

Step3 and 4. Estimate the test statistics of an ALRT and IUT

In the same way as the assumptions in Chapter 3, we propose the ALRT rejects $H_0^{(0)}$ if and only if

$$T^{(0)} \quad : \quad \min(\overline{u}_A^2, \overline{u}_B^2) > c.$$

Furthermore, we consider the statistics of an IUT to test hypothesis H_0 versus H_1. More precisely, the proposed IUT rejects H_0 if and only if

$$\begin{aligned} T^{(0)} \quad &: \quad \min(\overline{u}_A^2, \overline{u}_B^2) > d \quad \text{and} \\ T^{(j)} \quad &: \quad \frac{X_j + \epsilon_j}{\sqrt{\frac{\acute{\pi}_{1j}(1-\acute{\pi}_{1j})}{n_1} + \frac{\acute{\pi}_{2j}(1-\acute{\pi}_{2j})}{n_2}}} > z_\alpha \quad \text{for} \quad j = 1, \ldots, p. \end{aligned}$$

Chapter 2

Partial Asymmetry Measures for Square Contingency Tables

2.1 Introduction

In categorical data analysis, contingency tables are a basic tool used to examine the relationship between row and column categories. For example, the Pearson χ^2 statistic is commonly used to test the null hypothesis of statistical independence (Agresti, 2013, p.75).

When statistical independence is rejected, we are interested in describing the association between the row and column categories. Summary measures of association have been proposed, such as the Cramér V, gamma, and uncertainty coefficient. For details, see for instance Agresti (2013, Sec. 2.4) and Bishop et al. (1975, Sec. 11.3). Additionally, the recent development of association measures is described, for example, in Beh et al. (2007), Lombardo (2011), Wei and Kim (2017, 2021), Zhang et al. (2021), and Wei et al. (2022).

Contingency tables with the same row and column classifications are called *square* contingency tables. These tables are used for unaided distance vision data, social mobility data, and longitudinal data in biomedical research. The analysis of square contingency tables considers the issue of symmetry rather than independence because it is not sensible to treat these data as independent.

Bowker (1948) introduced the simple symmetry model and proposed a test for the hypothesis of symmetry. When the symmetry model fits the given data poorly, we are interested in measuring the degree of departure from symmetry. Tomizawa (1994) pro-

posed a measure that represents the degree of departure from symmetry expressed using the Shannon entropy or Kullback–Leibler information. In the real world, the Shannon entropy is widely applied as a measure of complexity, for example in Fernandes and Araújo (2020). The measure lies between 0 and 1, and its value equals 0 if and only if the symmetry model holds. Additionally, the degree of departure from symmetry increases as the value of the measure increases.

In the present chapter, we propose a measure that represents the degree of departure from symmetry using a different approach. We also consider a partial measure that represents the degree of departure from symmetry for each of several pairs. If the asymmetry appears to be similar in the various pairs, it may be useful to pool the values of the measure into a single summary measure of partial asymmetry. In an analogous manner to Agresti (1984, p.170), we consider taking a weighted average of the sample values as a summary measure. The properties of the proposed measure are given, and it has a characteristic that is different from that of Tomizawa's measure.

2.2 Review of Previous Research

Consider an $R \times R$ square contingency table having the same row and column classifications. Let p_{rc} denote the probability that an observation will fall in the (r, c)th cell of the table $(r = 1, \ldots, R; c = 1, \ldots, R)$. The simple symmetry model introduced by Bowker (1948) is defined by

$$p_{rc} = p_{cr} \quad (r \neq c).$$

This model indicates the symmetry structure with respect to the cell probabilities. Bowker (1948) proposed a test for the hypothesis of symmetry.

When the symmetry model does not hold for a given dataset, we are interested in evaluating the degree of departure from symmetry. Tomizawa (1994) proposed a measure that represents the degree of this departure expressed using Shannon entropy or Kullback–Leibler information. Assuming $p_{rc} + p_{cr}$ is not equal to zero for $r < c$, the measure is defined as

$$\Phi_T = \frac{1}{\delta \log 2} \sum \sum_{r \neq c} p_{rc} \log \frac{2p_{rc}}{p_{rc} + p_{cr}},$$

where $\delta = \sum \sum_{r \neq c} p_{rc}$. The measure Φ_T has three properties: (i) $0 \leq \Phi_T \leq 1$; (ii) the table has a symmetrical structure if and only if $\Phi_T = 0$; (iii) there is a structure for which either $p_{rc} = 0$ or $p_{cr} = 0$ for $r \neq c$ if and only if $\Phi_T = 1$.

Let $\pi_{rc} = p_{rc}/(p_{rc} + p_{cr})$ for $r = 1, \ldots, R; c = 1, \ldots, R; r \neq c$. The conditional probability that an observation falls in cell (r, c) or (c, r) in the table is π_{rc}. It should be noted that the symmetry model can be expressed as

$$\pi_{rc} = \pi_{cr} \left(= \frac{1}{2}\right) \quad (r < c).$$

The measure Φ_T can be expressed as

$$\Phi_T = \sum_{r<c} \sum \left(\frac{p_{rc} + p_{cr}}{\delta} \right) \phi_{rc},$$

where

$$\phi_{rc} = \frac{1}{\log 2} \left(\pi_{rc} \log \frac{\pi_{rc}}{1/2} + \pi_{cr} \log \frac{\pi_{cr}}{1/2} \right).$$

It should be noted that ϕ_{rc} is the normalized Kullback–Leibler information between (π_{rc}, π_{cr}) and $(1/2, 1/2)$. That is, the measure Φ_T is the weighted average of ϕ_{rc}.

We review ϕ_{rc} in Φ_T. The partial measure ϕ_{rc} represents the degree of departure from symmetry for a pair of symmetric cells because: (i) $0 \leq \phi_{rc} \leq 1$; (ii) there is a symmetrical structure for the pair of (r, c) and (c, r) cells if and only if $\phi_{rc} = 0$; (iii) there is a structure for which either $p_{rc} = 0$ or $p_{cr} = 0$ for the pair of (r, c) and (c, r) cells if and only if $\phi_{rc} = 1$. That is, the measure ϕ_{rc} expresses partial asymmetry.

2.3 The Proposed Measure

Let n_{rc} denote the observed frequency in the (r, c)th cell of the table $(r = 1, \ldots, R; c = 1, \ldots, R)$. We assume that $\{n_{rc}\}$ have a multinomial distribution:

$$\frac{n!}{\prod_{r=1}^{r} \prod_{c=1}^{r} n_{rc}!} \prod_{i=1}^{r} \prod_{j=1}^{r} p_{rc}^{n_{rc}},$$

where $n = \sum_{r=1}^{r} \sum_{c=1}^{r} n_{rc}$. Let $\hat{\boldsymbol{p}}^t$ be the $1 \times R^2$ vector

$$\hat{\boldsymbol{p}}^t = (\hat{\boldsymbol{p}}^t_{(12)}, \hat{\boldsymbol{p}}^t_{(13)}, \ldots, \hat{\boldsymbol{p}}^t_{(R-1,R)}, \hat{p}_{11}, \ldots, \hat{p}_{RR}),$$

where

$$\hat{\boldsymbol{p}}^t_{(rc)} = (\hat{p}_{rc}, \hat{p}_{cr}), \quad \hat{p}_{rc} = \frac{n_{rc}}{n}, \quad \hat{p}_{rr} = \frac{n_{rr}}{n},$$

and the superscript "t" denotes transpose.

Furthermore, let us define the vector $\boldsymbol{p}$ in terms of p_{rc}s in the same way as $\hat{\boldsymbol{p}}$. Let $\hat{\phi}_{rc}$ denote the sample version of ϕ_{rc}. Namely, the estimated ϕ_{rc} is given as

$$\hat{\phi}_{rc} = \frac{1}{\log 2} \left(\hat{\pi}_{rc} \log \frac{\hat{\pi}_{rc}}{1/2} + \hat{\pi}_{cr} \log \frac{\hat{\pi}_{cr}}{1/2} \right),$$

where $\hat{\pi}_{rc} = \hat{p}_{rc}/(\hat{p}_{rc} + \hat{p}_{cr})$ and $\hat{\pi}_{cr} = \hat{p}_{cr}/(\hat{p}_{rc} + \hat{p}_{cr})$. Let $\hat{\boldsymbol{\phi}}$ be the $R(R-1)/2 \times 1$ vector:

$$\hat{\boldsymbol{\phi}} = (\hat{\phi}_{12}, \hat{\phi}_{13}, \ldots, \hat{\phi}_{R-1,R})^t,$$

and we define the vector $\boldsymbol{\phi}$ in terms of ϕ_{rc}s in a similar manner to $\hat{\boldsymbol{\phi}}$. From Appendix (a), $\hat{\boldsymbol{\phi}}$ is asymptotically distributed as normal with mean $\boldsymbol{\phi}$ and covariance matrix

$$\sigma^2[\hat{\boldsymbol{\phi}}] = \begin{pmatrix} \sigma^2_{12} & 0 & \cdots & 0 \\ 0 & \sigma^2_{13} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \sigma^2_{R-1,R} \end{pmatrix},$$

where

$$\sigma^2_{rc} = \frac{\pi_{rc}\pi_{cr}}{n(p_{rc} + p_{cr})} \Big(\frac{1}{\log\ 2} (\log\ \pi_{rc} - \log\ \pi_{cr}) \Big)^2 \quad (r < c).$$

It should be noted that the set $\hat{\phi}_{12}, \ldots, \hat{\phi}_{R-1,R}$ is asymptotically mutually independent for large n.

We consider the weighted average of $\{\hat{\phi}_{rc}\}$, that is

$$\Phi = \sum_{r<c}\sum w_{rc}\hat{\phi}_{rc}, \tag{2.1}$$

where the weights $\{w_{rc}\}$ satisfy all $w_{rc} > 0$ and $\sum_{r<c}\sum w_{rc} = 1$. We see from Appendix (a) that $\hat{\phi}_{rc}$ $(r < c)$ is asymptotically distributed normal as $N(\phi_{rc}, \sigma^2_{rc})$ independently. Thus, the measure Φ has an asymptotically normal distribution with mean

$$\sum_{r<c}\sum w_{rc}\phi_{rc},$$

and variance

$$\sigma^2 = \sum_{r<c}\sum (w_{rc})^2\sigma^2_{rc}.$$

In an analogous manner to Agresti (1984, p.170), we derive the weights $\{w^*_{rc}\}$ so as to minimize the variance of Φ with the constraint that all $w_{rc} > 0$ and $\sum_{r<c}\sum w_{rc} = 1$. From Appendix (b), we obtain

$$\boldsymbol{w}^*_{rc} = \frac{1/\sigma^2_{rc}}{\sum_{s<t}\sum 1/\sigma^2_{st}}.$$

Then, we consider the following measure, which represents the degree of departure from symmetry:

$$\Phi_S = \sum_{r<c}\sum w^*_{rc}\hat{\phi}_{rc}.$$

The measure Φ_S has the smallest variance among measures in the class of weighted averages given in Equation (2.1). It should be noted that we should estimate the variances $\{\sigma^2_{rc}\}$ because these are unknown.

We propose the estimated measure as follows:

$$\hat{\Phi}_S = \sum_{r<c}\sum \hat{w}^*_{rc}\hat{\phi}_{rc},$$

where $\hat{w}^*_{rc}$ is given by w^*_{rc} with $\{p_{rc}\}$ replaced by $\{\hat{p}_{rc}\}$. The proposed measure approximates the measure in the class of weighted averages that has the smallest variance. The estimated measure $\hat{\Phi}_T$ is the weighted average of $\hat{\phi}_{rc}$ using the weights $\{\hat{w}_{rc} = (\hat{p}_{rc} + \hat{p}_{cr})/\hat{\delta}\}$, where $\hat{\delta} = \sum\sum_{r \neq c} \hat{p}_{rc}$. On the other hand, the proposed measure $\hat{\Phi}_S$ is the weighted average of $\hat{\phi}_{rc}$ using the weights $\{\hat{w}^*_{rc}\}$. It should be noted that (i) $\hat{w}_{rc} = (\hat{p}_{rc} + \hat{p}_{cr})/\hat{\delta}$ indicates the estimated conditional probability that the observation falls in (r, c) or (c, r) cells on the condition that the observation falls in off-diagonal cells and (ii) the weight $\hat{w}^*_{rc}$ becomes larger as the variance of partial measure $\hat{\phi}_{rc}$ decreases.

2.4 Numerical Examples

The objective is to confirm the difference in the single summary measure for symmetry by comparing the weights $\{\hat{w}_{rc}\}$ and $\{\hat{w}^*_{rc}\}$. Consider the artificial data in Table 2.1(a)–(d) with $n = 1000$ and Table 2.1(e) with $n = 200$. Table 2.1(a)–(d) are generated from the random numbers of the multinomial distribution based on the cell probability tables (a), (b), (c), and (d) in Table 2.2, respectively. Table 2.1(e) is generated from the random numbers of the multinomial distribution based on the cell probability table in Table 2.2(a). The artificial cell probability tables of Table 2.2 focus in particular on the probabilities of cells $(1, 2)$ and $(2, 1)$, and the four patterns (a), (b), (c), and (d) are set according to the combination of partial symmetry/asymmetry. We shall apply the partial measure ϕ_{rc}. Table 2.3 shows the estimated partial measure $\hat{\phi}_{rc}$, estimated variance $\hat{\sigma}^2_{rc}$, Bonferroni corrected confidence interval for ϕ_{rc}, estimated weights $\{\hat{w}_{rc}\}$ for $\hat{\Phi}_T$, and estimated weights $\{\hat{w}^*_{rc}\}$ for $\hat{\Phi}_S$. Figure 2.1 visualizes the estimated partial measure $\hat{\phi}_{rc}$ and Bonferroni corrected confidence interval for ϕ_{rc}. The Bonferroni corrected confidence interval for ϕ_{12} applied to the data in Table 2.1(a) does not contain zero, indicating that there is a partially asymmetric structure in cells $(1, 2)$ and $(2, 1)$. Furthermore, the Bonferroni corrected confidence interval for ϕ_{12} does not overlap with the Bonferroni corrected confidence intervals for ϕ_{rc} for any other pair of cells, indicating that cells $(1, 2)$ and $(2, 1)$ are partially asymmetric compared to every other pair of cells. The $\hat{w}^*_{12}$ of $\hat{\Phi}_S$ is remarkably smaller than the $\hat{w}_{12}$ of $\hat{\Phi}_T$, and the value of $\hat{\Phi}_S$ is smaller than that of $\hat{\Phi}_T$. The value of $\hat{\phi}_{12}$ applied to the data in Table

2.1(b) is as large as the value of $\hat{\phi}_{12}$ applied to the data in Table 2.1(a). However, it cannot be shown that there is a partially asymmetric structure in cells $(1,2)$ and $(2,1)$ in Table 2.1(b) because the Bonferroni corrected confidence interval for ϕ_{12} is wide and contains zero. Both $\hat{\Phi}_S$ and $\hat{\Phi}_T$ in Table 2.1(b) have small values of $\hat{w}_{12}$ and $\hat{w}^*_{12}$, and in particular, the $\hat{w}^*_{12}$ for $\hat{\Phi}_S$ is remarkably small.

On the other hand, the value of $\hat{\phi}_{12}$ applied to the data in Table 2.1(c) indicates that there is a partially symmetric structure in cells $(1,2)$ and $(2,1)$ because the value of $\hat{\phi}_{12}$ is small and the Bonferroni corrected confidence interval for the ϕ_{12} contains zero. In addition, the $\hat{w}^*_{12}$ in Table 2.1(c) is large, indicating that the weight of $\hat{\Phi}_S$ is larger when the pair of cells is more frequent than others and has a partially symmetric structure. Both $\hat{\Phi}_S$ and $\hat{\Phi}_T$ applied to the data in Table 2.1(c) are close to zero because of the greater weight of the pair of cells $(1,2)$ and $(2,1)$ that show partial symmetry compared to the pairs of cells $(4,5)$ and $(5,4)$ and $(3,4)$ and $(4,3)$ that show partial asymmetry. The values of $\hat{\phi}_{rc}$ applied to the data in Table 2.1(d) indicate that the pair of cells $(1,2)$ and $(2,1)$, the pair of cells $(1,4)$ and $(4,1)$, the pair of cells $(2,3)$ and $(3,2)$ and the pair of cells $(3,5)$ and $(5,3)$ have a partially symmetric structure because the Bonferroni corrected confidence intervals of ϕ_{12}, ϕ_{14}, ϕ_{23} and ϕ_{35} include zero. It can be seen that the values of $\hat{w}_{12}$ and $\hat{w}_{14}$ for $\hat{\Phi}_T$ are similar, while the value of $\hat{w}^*_{14}$ for $\hat{\Phi}_S$ is large compared to $\hat{w}^*_{12}$.

The value of $\hat{\phi}_{12}$ applied to the data in Table 2.1(e) is about the same as the value applied to the data in Table 2.1(a), but the Bonferroni corrected confidence interval is wider due to the smaller sample size, making it relatively difficult to conclude partial asymmetry for the pair of cells $(1,2)$ and $(2,1)$. The magnitude of $\hat{\phi}_{12}$ applied to Table 2.1(e) does not differ much from the results applied to Table 2.1(a). However, the values of $\hat{w}^*_{12}$ and $\hat{\Phi}_S$ are greater when applied to Table 2.1(e) than Table 2.1(a). It should be noted that the weight $\hat{w}^*_{rc}$ becomes larger as the variance of the partial measure $\hat{\phi}_{rc}$ decreases, and the weight $\hat{w}_{rc}$ becomes larger as the proportion $(n_{rc}+n_{cr})/n$ increases.

2.5 Example

Consider the data in Table 2.4, derived from the national survey on educational attitudes of high school students and their mothers in Japan in 2012. To clarify the structure of educational inequalities in contemporary Japanese society and the actual educational awareness of parents and children, a postal survey was conducted among

second-year high school students and their mothers throughout Japan, using the same framework as the national survey on the educational awareness of high school students and their mothers conducted in November 2002. The data describe the cross-classification of mothers' and fathers' birth orders. For example, for the 179 high school students whose mothers' birth order is "First" and whose fathers' birth order is "Second", the mother is the eldest daughter and the father is the second son. The partial symmetry of cells $(1, 2)$ and $(2, 1)$ means that the probability of high school students whose mother is the eldest daughter and whose father is the second son is equal to that of high school students whose mother is the second daughter and whose father is the eldest son.

Let G_S^2 and χ_S^2 denote the likelihood ratio and Pearson's chi-squared statistics for testing the goodness of fit of the symmetry model, i.e.,

$$G_S^2 = 2\sum\sum_{r\neq c} n_{rc}\log(2n_{rc}/(n_{rc}+n_{cr})),$$

and

$$\chi_S^2 = \sum\sum_{r<c}(n_{rc}-n_{cr})^2/(n_{rc}+n_{cr}).$$

For large samples, G_S^2 and χ_S^2 have a chi-squared null distribution with $R(R-1)/2$ degrees of freedom. From $G_S^2 = 14.58$ and $\chi_S^2 = 14.17$ with six degrees of freedom for the data in Table 2.4, these values indicate the lack of a symmetrical structure. Note that the exact test introduced by West (2008) is well known as a test for the contingency table including structural zeros. As the proposed measure does not require the frequency of the diagonal components, West's test was also conducted assuming that the diagonal components are structural zeros. The simulated p-value from West's test is 0.085, which indicates that the rows and columns are independent. The value of $\hat{\Phi}_T$ is 0.0184, and the Bonferroni corrected confidence interval is (0.000003, 0.036719), which does not include zero.

Next, we measured the degree of departure from partial symmetry for each pair of cells. We shall apply the partial measure ϕ_{rc} for the data in Table 2.4. Table 2.5 shows the estimated values for ϕ_{rc} and σ_{rc}^2, Bonferroni corrected confidence intervals for ϕ_{rc}, estimated weights $\{\hat{w}_{rc}\}$ and $\{\hat{w}_{rc}^*\}$, and estimated measures $\hat{\Phi}_S$ and $\hat{\Phi}_T$. According to the magnitudes of the estimates, ϕ_{rc} can explain the partial symmetry for each pair of

cells in Table 2.4. The Bonferroni corrected confidence interval for the ϕ_{rc} for all pairs of cells contains zero, which indicates that there is a partially symmetrical structure in each birth order category in the mother–father pairs.

Furthermore, the estimated departure from symmetry is smaller with $\hat{\Phi}_S$, which uses different weights than $\hat{\Phi}_T$. Figure 2.2 plots estimated weights $\hat{w}_{rc}$ and $\hat{w}^*_{rc}$. Cells $(1,2)$ and $(2,1)$ have similar frequencies and are more frequent than the other cells. Then, weights $\hat{w}_{rc}$ and $\hat{w}^*_{rc}$ are similar and large. On the other hand, the pair of cells $(2,3)$ and $(3,2)$ have similar frequencies, but are less frequent than the pair of cells $(1,2)$ and $(2,1)$. In such cases, $\hat{w}^*_{rc}$ is larger than $\hat{w}_{rc}$. Therefore, $\hat{\Phi}_S$ has a higher weight than $\hat{\Phi}_T$ when the pair of cells has a lower frequency than another pair of cells and when the cells have similar frequencies. Conversely, $\hat{w}^*_{rc}$ is smaller than $\hat{w}_{rc}$ when the frequencies are different, as in the pair of cells $(1,3)$ and $(3,1)$. Since the weights $\{\hat{w}^*_{rc}\}$ and $\{\hat{w}_{rc}\}$ take different values, the single summary measures $\hat{\Phi}_S$ and $\hat{\Phi}_T$ also take different values. As mentioned above, there is a partially symmetrical structure in each birth order category in the mother–father pairs. Then, the proposed measure may be reasonable to express the degree of departure from symmetry.

2.6 Concluding Remarks

We proposed a partial measure to express the degree of departure from partial symmetry. The measure was constructed as the weighted average of partial measures expressed using the Shannon entropy or Kullback–Leibler information. The composition of the proposed measure Φ_S is similar to that of the measure proposed by Tomizawa (1994) in the sense that they are classes of weighted averages. However, they differ in that the weights multiplied by the partial measure are constructed so as to minimize the measure's variance. This measure increase with the degree of departure from symmetry, allowing us to see how far away the probability structure of the contingency table is from complete asymmetry.

The measures $\hat{\Phi}_S$ and $\hat{\Phi}_T$ are invariant under the arbitrary simultaneous permutations of row and column categories, and therefore, it is possible to apply these measures to analyze the data on a nominal scale, as well as on an ordinal scale if one cannot use the information about the order in which the categories are listed.

We compared the weights used to construct the measures $\hat{\Phi}_S$ and $\hat{\Phi}_T$. Those used to construct $\hat{\Phi}_T$ are large when the frequency of the pair of cells is high compared to

others. On the other hand, the weight of $\hat{\Phi}_S$ is higher when the frequency of the pair of cells is higher than others and when the structure is partially symmetric. Conversely, when the frequency of the pair of cells is lower than others and the structure is partially asymmetric, the weights of $\hat{\Phi}_S$ are smaller than those of $\hat{\Phi}_T$.

In the present study, Bonferroni corrected confidence intervals for partial measures were used to interpret the partially asymmetric structure of the data. Alternatively, global tests for the null hypothesis that all ϕ_{rc} are equally zero, and multiplicity correction for paired comparisons also need to be considered and are left as future works.

We should note, however, that $\hat{\Phi}_S$ cannot be calculated if any of the off-diagonal cells are zero. As such, the proposed measure should be used for contingency tables with large sample sizes.

Appendix

(a) : Derivation of σ_{rc}^2

From the central limit theorem, $\hat{\boldsymbol{p}}$ is asymptotically distributed as normal $N(\boldsymbol{p}, \Sigma_1(\boldsymbol{p}))$, where $\Sigma_1(\boldsymbol{p})$ is the $R^2 \times R^2$ matrix

$$\Sigma_1(\boldsymbol{p}) = \frac{1}{n}(D(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^t),$$

where $D(\boldsymbol{p})$ denotes a diagonal matrix with the ith element of $\boldsymbol{p}$ as the rth diagonal element. Then, we also obtain

$$\hat{\boldsymbol{\phi}} = \boldsymbol{\phi} + d_1(\boldsymbol{p})(\hat{\boldsymbol{p}} - \boldsymbol{p}) + o(||\hat{\boldsymbol{p}} - \boldsymbol{p}||),$$

where $d_1(\boldsymbol{p}) = \partial\boldsymbol{\phi}/\partial\boldsymbol{p}^t$ is the $R(R-1)/2 \times R^2$ matrix. Thus, $\hat{\boldsymbol{\phi}}$ is asymptotically distributed as normal $N(\boldsymbol{\phi}, \sigma^2[\hat{\boldsymbol{\phi}}])$, where

$$\sigma^2[\hat{\boldsymbol{\phi}}] = d_1(\boldsymbol{p})\Sigma_1(\boldsymbol{p})d_1(\boldsymbol{p})^t.$$

Let $\hat{\boldsymbol{\pi}}^t$ be the $1 \times R(R-1)$ vector:

$$\hat{\boldsymbol{\pi}}^t = (\hat{\boldsymbol{\pi}}_{(12)}^t, \hat{\boldsymbol{\pi}}_{(13)}^t, \ldots, \hat{\boldsymbol{\pi}}_{(R-1,R)}^t),$$

where $\hat{\boldsymbol{\pi}}_{(rc)}^t = (\hat{\pi}_{rc}, \hat{\pi}_{cr})$. Noting that $\hat{\boldsymbol{\phi}}$ is a function of only $\{\pi_{rc}\}$, we obtain

$$d_1(\boldsymbol{p}) = \frac{\partial\boldsymbol{\phi}}{\partial\boldsymbol{\pi}^t} \cdot \frac{\partial\boldsymbol{\pi}}{\partial\boldsymbol{p}^t}.$$

It should be noted that $\partial\boldsymbol{\phi}/\partial\boldsymbol{\pi}^t$ is the $R(R-1)/2 \times R(R-1)$ matrix and $\partial\boldsymbol{\pi}/\partial\boldsymbol{p}^t$ is the $R(R-1) \times R^2$ matrix. By obtaining $\partial\boldsymbol{\pi}/\partial\boldsymbol{p}^t$, we can see that $\sigma^2[\hat{\boldsymbol{\phi}}]$ is expressed as

$$\sigma^2[\hat{\boldsymbol{\phi}}] = \left(\frac{\partial\boldsymbol{\phi}}{\partial\boldsymbol{\pi}^t}\right) \cdot \Sigma_2(\boldsymbol{p}) \cdot \left(\frac{\partial\boldsymbol{\phi}}{\partial\boldsymbol{\pi}^t}\right)^t,$$

where $\Sigma_2(\boldsymbol{p})$ is the $R(R-1) \times R(R-1)$ matrix:

$$\Sigma_2(\boldsymbol{p}) = \begin{pmatrix} \Sigma_{12}(\boldsymbol{p}) & 0 & \cdots & 0 \\ 0 & \Sigma_{13}(\boldsymbol{p}) & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \Sigma_{R-1,R}(\boldsymbol{p}) \end{pmatrix},$$

where

$$\Sigma_{rc}(\boldsymbol{p}) = \frac{1}{n(p_{rc}+p_{cr})} \begin{pmatrix} \pi_{rc}(1-\pi_{rc}) & -\pi_{rc}\pi_{cr} \\ -\pi_{rc}\pi_{cr} & \pi_{cr}(1-\pi_{cr}) \end{pmatrix} \quad (r < c).$$

Thus, $\sigma^2[\hat{\boldsymbol{\phi}}]$ is also expressed as follows:

$$\sigma^2[\hat{\boldsymbol{\phi}}] = \begin{pmatrix} \sigma^2_{12} & 0 & \cdots & 0 \\ 0 & \sigma^2_{13} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \sigma^2_{R-1,R} \end{pmatrix},$$

where

$$\sigma^2_{rc} = \frac{\pi_{rc}\pi_{cr}}{n(p_{rc}+p_{cr})} \Big(\frac{1}{\log\ 2} (\log\ \pi_{rc} - \log\ \pi_{cr}) \Big)^2 \quad (r < c).$$

(b) : Derivation of w^*_{rc}

Let $\boldsymbol{w}$ be the $R(R-1)/2 \times 1$ vector:

$$\boldsymbol{w} = (w_{12}, w_{13}, \ldots, w_{R-1,R})^t.$$

Then, the measure Φ is expressed as $\Phi = \boldsymbol{w}^t \hat{\boldsymbol{\phi}}$.

From Appendix (a), the mean and variance of Φ are approximately calculated as follows:

$$\begin{aligned} E(\Phi) &= \boldsymbol{w}^t \boldsymbol{\phi}, \\ Var(\Phi) &= \boldsymbol{w}^t \sigma^2[\hat{\boldsymbol{\phi}}] \boldsymbol{w}. \end{aligned}$$

Then, we can obtain the following $\boldsymbol{w}^*$ so as to minimize $Var(\Phi)$ with the constraint that $\boldsymbol{w}^t \mathbf{1}_d$ ($d = R(R-1)/2$) is unity ($\mathbf{1}_d$ is the $d \times 1$ vector of 1 elements):

$$w^*_{rc} = \frac{1/\sigma^2_{rc}}{\sum\sum_{s<t} 1/\sigma^2_{st}} \quad (r < c).$$

Table 2.1 Artificial data.

(a)	(1)	(2)	(3)	(4)	(5)
(1)	37	544	12	7	8
(2)	102	26	15	15	12
(3)	9	8	29	10	11
(4)	9	9	12	40	12
(5)	14	9	10	11	29
(b)	(1)	(2)	(3)	(4)	(5)
(1)	47	11	37	44	48
(2)	3	38	34	37	49
(3)	44	44	52	56	48
(4)	38	25	55	45	47
(5)	35	25	51	43	44
(c)	(1)	(2)	(3)	(4)	(5)
(1)	33	316	13	18	18
(2)	321	37	20	18	20
(3)	7	6	26	16	14
(4)	5	5	1	30	19
(5)	5	10	5	2	35
(d)	(1)	(2)	(3)	(4)	(5)
(1)	39	4	70	42	50
(2)	5	34	45	110	84
(3)	17	12	54	103	63
(4)	31	14	20	39	48
(5)	9	29	26	6	46
(e)	(1)	(2)	(3)	(4)	(5)
(1)	7	103	1	1	4
(2)	19	10	2	2	4
(3)	2	1	6	2	2
(4)	3	4	4	5	2
(5)	1	2	4	1	8

Table 2.2 Artificial cell probability tables.

(a)	(1)	(2)	(3)	(4)	(5)
(1)	0.030	0.570	0.010	0.010	0.010
(2)	0.010	0.030	0.010	0.010	0.010
(3)	0.010	0.010	0.030	0.010	0.010
(4)	0.010	0.010	0.010	0.030	0.010
(5)	0.010	0.010	0.010	0.010	0.030
(b)	(1)	(2)	(3)	(4)	(5)
(1)	0.040	0.008	0.040	0.040	0.040
(2)	0.004	0.040	0.040	0.040	0.040
(3)	0.040	0.040	0.050	0.050	0.050
(4)	0.040	0.040	0.050	0.040	0.049
(5)	0.040	0.040	0.050	0.049	0.040
(c)	(1)	(2)	(3)	(4)	(5)
(1)	0.030	0.320	0.015	0.015	0.015
(2)	0.320	0.030	0.021	0.019	0.024
(3)	0.005	0.007	0.030	0.020	0.016
(4)	0.005	0.006	0.005	0.030	0.018
(5)	0.005	0.007	0.004	0.003	0.030
(d)	(1)	(2)	(3)	(4)	(5)
(1)	0.040	0.003	0.080	0.050	0.050
(2)	0.003	0.040	0.050	0.100	0.080
(3)	0.020	0.010	0.050	0.100	0.064
(4)	0.030	0.020	0.020	0.040	0.050
(5)	0.010	0.020	0.020	0.020	0.040

Table 2.3 Estimate of measure ϕ_{rc}, estimated approximate variance for ϕ_{rc}, Bonferroni corrected confidence interval for ϕ_{rc}, weights for measures Φ_S and Φ_T, and estimates of measures Φ_S and Φ_T, applied to Table 4.1(a)–(d).

Applied Data	Cells	$\hat{\phi}_{rc}$	$\hat{\sigma}^2_{rc}$	Confidence Interval for ϕ_{rc}	$\hat{w}^*_{rc}$	$\hat{w}_{rc}$	$\hat{\Phi}_S$	$\hat{\Phi}_T$
(a)	(1,2), (2,1)	0.37075	0.0012005	(0.273, 0.468)	0.057987	0.769964		
	(1,3), (3,1)	0.01477	0.0020088	(−0.111, 0.141)	0.034653	0.025030		
	(1,4), (4,1)	0.01130	0.0020219	(−0.115, 0.138)	0.034428	0.019070		
	(1,5), (5,1)	0.05434	0.0068561	(−0.178, 0.287)	0.010153	0.026222		
	(2,3), (3,2)	0.06789	0.0081116	(−0.185, 0.321)	0.008582	0.027414	0.02624	0.29124
	(2,4), (4,2)	0.04557	0.0053039	(−0.159, 0.25)	0.013124	0.028605		
	(2,5), (5,2)	0.01477	0.0020088	(−0.111, 0.141)	0.034653	0.025030		
	(3,4), (4,3)	0.00597	0.0007797	(−0.072, 0.084)	0.089277	0.026222		
	(3,5), (5,3)	0.00164	0.0002246	(−0.04, 0.044)	0.309966	0.025030		
	(4,5), (5,4)	0.00136	0.0001710	(−0.035, 0.038)	0.407178	0.027414		
(b)	(1,2), (2,1)	0.25041	0.0422558	(−0.327, 0.827)	0.000032	0.018088		
	(1,3), (3,1)	0.00539	0.0001914	(−0.033, 0.044)	0.006979	0.104651		
	(1,4), (4,1)	0.00387	0.0001357	(−0.029, 0.037)	0.009848	0.105943		
	(1,5), (5,1)	0.01777	0.0006101	(−0.052, 0.087)	0.002190	0.107235		
	(2,3), (3,2)	0.01189	0.0004362	(−0.047, 0.071)	0.003063	0.100775	0.00036	0.01843
	(2,4), (4,2)	0.02719	0.0012416	(−0.072, 0.126)	0.001076	0.080103		
	(2,5), (5,2)	0.07727	0.0028494	(−0.073, 0.227)	0.000469	0.095607		
	(3,4), (4,3)	0.00006	0.0000015	(−0.003, 0.004)	0.877858	0.143411		
	(3,5), (5,3)	0.00066	0.0000193	(−0.012, 0.013)	0.069221	0.127907		
	(4,5), (5,4)	0.00143	0.0000457	(−0.018, 0.02)	0.029264	0.116279		
(c)	(1,2), (2,1)	0.00004	0.0000002	(−0.001, 0.001)	0.999900	0.759237		
	(1,3), (3,1)	0.06593	0.0090727	(−0.201, 0.333)	0.000022	0.023838		
	(1,4), (4,1)	0.24463	0.0252616	(−0.202, 0.691)	0.000008	0.027414		
	(1,5), (5,1)	0.24463	0.0252616	(−0.202, 0.691)	0.000008	0.027414		
	(2,3), (3,2)	0.22065	0.0205989	(−0.182, 0.624)	0.000010	0.030989	0.00006	0.06270
	(2,4), (4,2)	0.24463	0.0252616	(−0.202, 0.691)	0.000008	0.027414		
	(2,5), (5,2)	0.08170	0.0074074	(−0.16, 0.323)	0.000027	0.035757		
	(3,4), (4,3)	0.67724	0.0521067	(0.036, 1.318)	0.000004	0.020262		
	(3,5), (5,3)	0.16853	0.0225185	(−0.253, 0.59)	0.000009	0.022646		
	(4,5), (5,4)	0.54628	0.0432851	(−0.038, 1.13)	0.000005	0.025030		
(d)	(1,2), (2,1)	0.00892	0.0028433	(−0.141, 0.159)	0.113903	0.011421		
	(1,3), (3,1)	0.28736	0.0075340	(0.044, 0.531)	0.042986	0.110406		
	(1,4), (4,1)	0.01644	0.0006424	(−0.055, 0.088)	0.504107	0.092640		
	(1,5), (5,1)	0.38383	0.0134101	(0.059, 0.709)	0.024150	0.074873		
	(2,3), (3,2)	0.25751	0.0106028	(−0.032, 0.547)	0.030545	0.072335	0.11518	0.28830
	(2,4), (4,2)	0.49139	0.0071440	(0.254, 0.729)	0.045333	0.157360		
	(2,5), (5,2)	0.17837	0.0039745	(0.001, 0.355)	0.081484	0.143401		
	(3,4), (4,3)	0.35950	0.0061895	(0.139, 0.58)	0.052323	0.156091		
	(3,5), (5,3)	0.12854	0.0037881	(−0.044, 0.301)	0.085494	0.112944		
	(4,5), (5,4)	0.49674	0.0164609	(0.137, 0.857)	0.019674	0.068528		
(e)	(1,2), (2,1)	0.37599	0.0064089	(0.151, 0.601)	0.486330	0.743902		
	(1,3), (3,1)	0.08170	0.0740741	(−0.682, 0.846)	0.042077	0.018293		
	(1,4), (4,1)	0.18872	0.1177550	(−0.775, 1.152)	0.026469	0.024390		
	(1,5), (5,1)	0.27807	0.1280000	(−0.726, 1.282)	0.024350	0.030488		
	(2,3), (3,2)	0.08170	0.0740741	(−0.682, 0.846)	0.042077	0.018293	0.23244	0.30922
	(2,4), (4,2)	0.08170	0.0370370	(−0.459, 0.622)	0.084155	0.036585		
	(2,5), (5,2)	0.08170	0.0370370	(−0.459, 0.622)	0.084155	0.036585		
	(3,4), (4,3)	0.08170	0.0370370	(−0.459, 0.622)	0.084155	0.036585		
	(3,5), (5,3)	0.08170	0.0370370	(−0.459, 0.622)	0.084155	0.036585		
	(4,5), (5,4)	0.08170	0.0740741	(−0.682, 0.846)	0.042077	0.018293		

Table 2.4 Cross-classification of mothers' and fathers' birth orders.

	Fathers' Birth Order				
Mothers' Birth Order	First	Second	Third	Fourth or More	Total
First	224	179	53	22	478
Second	162	153	35	15	365
Third	37	37	18	11	103
Fourth or more	12	7	3	5	27
Total	435	376	109	53	973

Table 2.5 Estimate of measure ϕ_{rc}, estimated approximate variance for ϕ_{rc}, Bonferroni corrected confidence interval for ϕ_{rc}, estimates of measures of Φ_S and Φ_T, and weights for measures of Φ_S and Φ_T, applied to Table 2.4.

Cells	$\hat{\phi}_{rc}$	$\hat{\sigma}^2_{rc}$	Confidence Interval for ϕ_{rc}	$\hat{w}^*_{rc}$	$\hat{w}_{rc}$	$\hat{\Phi}_S$	$\hat{\Phi}_T$
(1,2), (2,1)	0.0018	0.00002	(−0.008, 0.012)	0.5864	0.5951		
(1,3), (3,1)	0.0229	0.00072	(−0.048, 0.094)	0.0123	0.1571		
(1,4), (4,1)	0.0633	0.00514	(−0.126, 0.252)	0.0017	0.0593		
(2,3), (3,2)	0.0006	0.00002	(−0.012, 0.013)	0.3986	0.1257	0.0018	0.0184
(2,4), (4,2)	0.0976	0.01192	(−0.190, 0.386)	0.0007	0.0384		
(3,4), (4,3)	0.2504	0.04226	(−0.292, 0.793)	0.0002	0.0244		

Figure 2.1 Estimate of measure ϕ_{rc} and Bonferroni corrected confidence interval for ϕ_{rc} applied to Table 2.1.

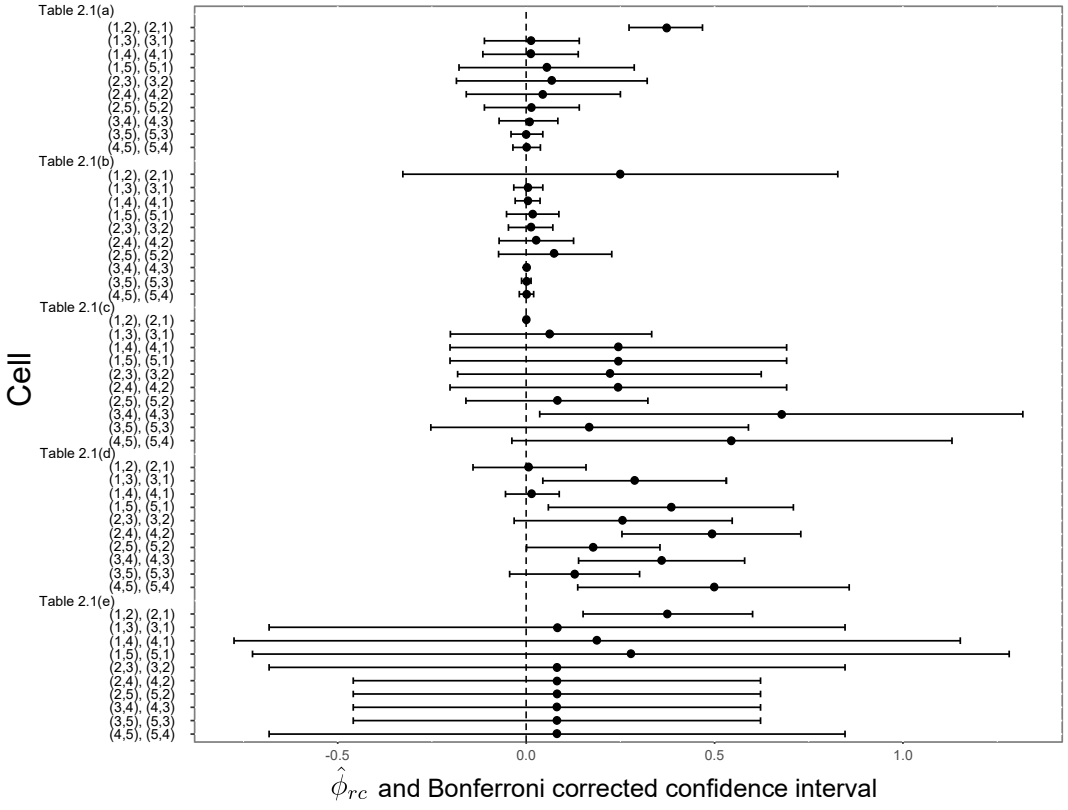
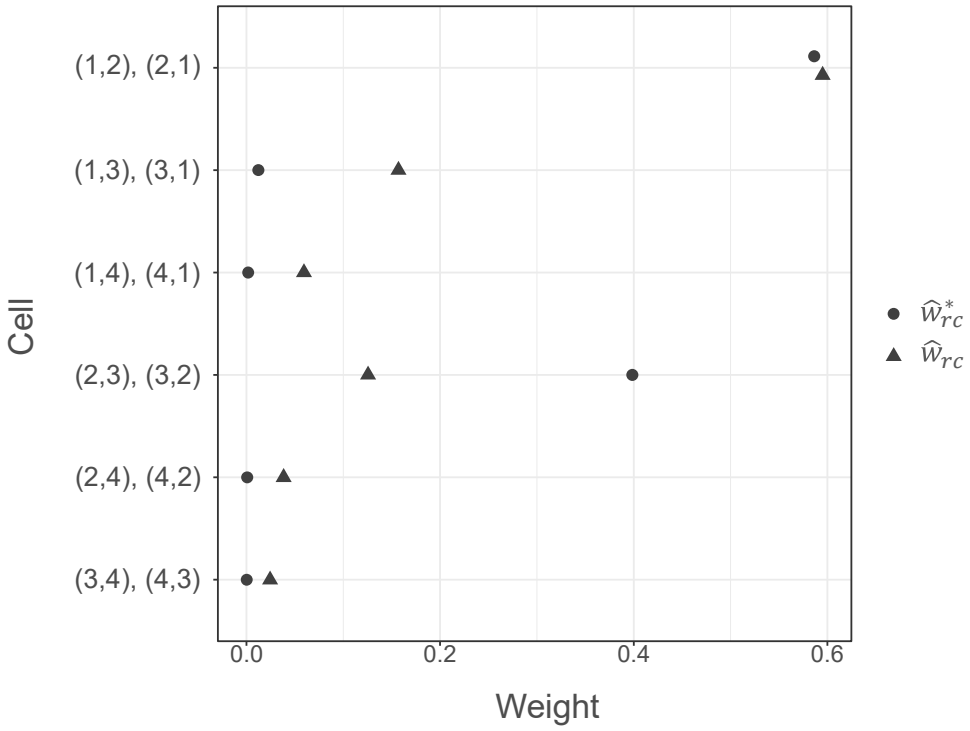


Figure 2.2 The weight for each pair of symmetric cells obtained by applying the proposed method and Tomizawa (1994) to Table 2.4.



Chapter 3

A Testing Procedure in Clinical Trials with Multiple Binary Endpoints

3.1 Introduction

In confirmatory clinical trials, the efficacy of a test treatment is sometimes assessed using multiple primary endpoints. For example, in recent clinical trials for patients with psoriasis, the percentage of patients with at least 75 percent improvement in the psoriasis area-and-severity index (PASI) score and the percentage with a physician's global assessment score of 0 or 1 at weeks 24 and 52 were used as primary endpoints to demonstrate the superiority of a test treatment (Reich et al., 2011). In clinical trials of rheumatoid arthritis, the percentage of patients achieving a 20 percent short-term improvement in the American College of Rheumatology criteria (ACR20) and the percentage achieving long-term low disease activity (Disease Activity Score: DAS28-ESR ≤ 3.2) are often used as primary endpoints (Smolen et al., 2016). In IgA nephropathy affecting multiple areas, full clinical remission and a decrease in estimated Glomerular Filtration Rate (eGFR) of at least 15 mL per minute per 1.73 m^2 from baseline are sometimes used in primary assessment (Rauen et al., 2015). In the above trials, all of the primary endpoints were binary, and most were used only to evaluate superiority, even though there was evidence of non-inferiority for some endpoints whose superiority could not be confirmed. Since it is difficult to demonstrate that each endpoint is statistically significant when the number of endpoints is not small, this chapter deals

with the case where the test treatment is superior for at least one of the endpoints and is not clinically inferior for the remaining endpoints.

For multiple continuous endpoints, Perlman and Wu (2004) proposed a testing procedure that is applicable to the framework mentioned above. Moreover, Tamhane and Logan (2004) suggested the Union-intersection(UI)-Intersection-union(IU) test and showed that their method could be controlled Type I error in the same way as the Perlman and Wu's procedure. Nakazuru et al. (2014) proposed a more powerful procedure by modifying Perlman and Wu's procedure using the approximate likelihood ratio test (ALRT) defined by Glimm et al. (2002). However, no method for multiple binary endpoints has been developed yet. Therefore, we herein propose a testing procedure that is appropriate when all endpoints are binary.

This chapter is structured as follows. In Section 2, we define several notations to formulate the problem. In Section 3, we consider new statistics in order to test a hypothesis that includes superiority of at least one endpoint and non-inferiority of the remaining endpoints when all endpoints are binary. In Section 4, we provide a numerical experiment using Monte Carlo simulation to illustrate the behavior of the power and type I error rate of the proposed test. In addition, we compared the power of the test of only superiority for at least one endpoint with that of the test of superiority for at least one endpoint plus non-inferiority for all endpoints. Finally, in Section 5, we summarize our findings and present concluding remarks.

3.2 Notations

To simplify the case, we consider comparing p endpoints with two treatment groups comprising n_1 and n_2 subjects. Without loss of generality, we assume that efficacy is recognized when the proportion of responses to the test treatment is greater than that to the control treatment. Let Y_{ijk} $(i = 1, 2; j = 1, \ldots, p; k = 1, \ldots, n_i)$ denote the binary response of the kth subject to the ith treatment at the jth endpoint. Suppose that the random vectors of responses $\boldsymbol{Y}_{ik} = (Y_{i1k}, \ldots, Y_{ipk})^t$ $(i = 1, 2; k = 1, \ldots, n_i)$ are independently distributed as a p-variate Bernoulli distribution with $\mathrm{E}(Y_{ijk}) = \pi_{ij}$, $\mathrm{V}(Y_{ijk}) = \pi_{ij}(1 - \pi_{ij})$, and $\mathrm{Corr}(Y_{ilk}, Y_{il'k}) = \rho_{(i)ll'}$ for all $l \neq l'$, where the superscript "t" denotes transpose. Generally, the probability mass function for $\boldsymbol{Y}_{ik}$ $(i = 1, 2; k =$

$1, \ldots, n_i$) can be written as

$$\begin{aligned}\mathrm{P}(Y_{i1k} = y_{i1k}, \ldots, Y_{ipk} = y_{ipk}) = \theta_{(i)0,0,\ldots,0}^{\prod_{j=1}^{p}(1-y_{ijk})} \times \theta_{(i)1,0,\ldots,0}^{y_{i1k}\prod_{j=2}^{p}(1-y_{ijk})} \\ \times \theta_{(i)0,1,\ldots,0}^{(1-y_{i1k})y_{i2k}\prod_{j=3}^{p}(1-y_{ijk})} \times \cdots \times \theta_{(i)1,1,\ldots,1}^{\prod_{j=1}^{p} y_{ijk}},\end{aligned}$$

where $\theta_{(i)0,0,\ldots,0}, \ldots, \theta_{(i)1,1,\ldots,1}$ are joint probabilities when $\boldsymbol{Y}_{ik}$ takes values from $(0, \ldots, 0), \ldots, (1, \ldots, 1)$, respectively, and $\theta_{(i)0,0,\ldots,0} + \cdots + \theta_{(i)1,1,\ldots,1} = 1$. Note that the following equations hold:

$$\pi_{ij} = \sum_{\substack{(s_1,s_2,\ldots,s_p)\in S \\ s_j=1}} \theta_{(i)s_1,s_2,\ldots,s_p},$$

where $S = \{(s_1, s_2, \ldots, s_p) | s_j = 0, 1, j = 1, \ldots, p\}$ is a set with all pair of response values as elements. Set $\boldsymbol{X} = (X_1, \ldots, X_p)^t$ with $X_j = (\overline{Y}_{1j} - \overline{Y}_{2j})$ $(j = 1, \ldots, p)$, where $\overline{Y}_{ij}$ $(i = 1, 2; j = 1, \ldots, p)$ is the sample proportion for the jth endpoint to the ith treatment. Let the true proportion vector be the ith treatment $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{ip})^t$ with difference of proportion $\boldsymbol{\Delta} = \boldsymbol{\pi}_1 - \boldsymbol{\pi}_2$, and the covariance matrix $\boldsymbol{\Sigma}$ is defined as follows:

$$\begin{aligned}\boldsymbol{\Sigma} &= \boldsymbol{\Sigma}^{(1)} + \boldsymbol{\Sigma}^{(2)} \\ &= \frac{1}{n_1}\begin{pmatrix} \pi_{11}(1-\pi_{11}) & \cdots & \rho_{(1)1p}\sqrt{\pi_{11}(1-\pi_{11})}\sqrt{\pi_{1p}(1-\pi_{1p})} \\ \vdots & \ddots & \vdots \\ \rho_{(1)p1}\sqrt{\pi_{11}(1-\pi_{11})}\sqrt{\pi_{1p}(1-\pi_{1p})} & \cdots & \pi_{1p}(1-\pi_{1p}) \end{pmatrix} \\ &+ \frac{1}{n_2}\begin{pmatrix} \pi_{21}(1-\pi_{21}) & \cdots & \rho_{(2)1p}\sqrt{\pi_{21}(1-\pi_{21})}\sqrt{\pi_{2p}(1-\pi_{2p})} \\ \vdots & \ddots & \vdots \\ \rho_{(2)p1}\sqrt{\pi_{21}(1-\pi_{21})}\sqrt{\pi_{2p}(1-\pi_{2p})} & \cdots & \pi_{2p}(1-\pi_{2p}) \end{pmatrix},\end{aligned}$$

where $\boldsymbol{\Sigma}^{(i)} (i = 1, 2)$ is covariance matrix of $(\overline{Y}_{i1}, \ldots, \overline{Y}_{ip})^t$. In addition, the correlation between the jth and j'th endpoints of the ith treatment $(i = 1, 2; j, j' = 1, \ldots, p; j \neq j')$ is written using the joint probability $\phi_{(i)jj'}$ as

$$\rho_{(i)jj'} = \frac{\phi_{(i)jj'} - \pi_{ij}\pi_{ij'}}{\sqrt{\pi_{ij}(1-\pi_{ij})}\sqrt{\pi_{ij'}(1-\pi_{ij'})}}.$$

$\phi_{(i)jj'}$ can be simply written as the summation of joint probability $\theta_{(i)s_1,s_2,\ldots,s_p}$ where $s_j = s_{j'} = 1$ as follows:

$$\phi_{(i)jj'} = \sum_{\substack{(s_1,s_2,\ldots,s_p)\in S \\ s_j = s_{j'} = 1}} \theta_{(i)s_1,s_2,\ldots,s_p}. \tag{3.1}$$

On the other hand, $\phi_{(i)jj'}$ can be also determined using the odds ratio $\psi_{(i)jj'}$ as follows (Dale, 1986; le Cessie and van Houwelingen, 1994):

$$\phi_{(i)jj'} = \begin{cases} \frac{1+(\pi_{ij}+\pi_{ij'})(\psi_{(i)jj'}-1)-S(\pi_{ij},\pi_{ij'},\psi_{(i)jj'})}{2(\psi_{(i)jj'}-1)} & \text{if } \psi_{(i)jj'} \neq 1 \\ \pi_{ij}\pi_{ij'} & \text{if } \psi_{(i)jj'} = 1, \end{cases}$$

where $S(\pi_{ij}, \pi_{ij'}, \psi_{(i)jj'}) = \sqrt{(1+(\pi_{ij}+\pi_{ij'})(\psi_{(i)jj'}-1))^2 + 4\psi_{(i)jj'}(1-\psi_{(i)jj'})\pi_{ij}\pi_{ij'}}$ and $\psi_{(i)jj'}$ is described as

$$\psi_{(i)jj'} = \frac{\phi_{(i)jj'}(1-\pi_{ij}-\pi_{ij'}+\phi_{(i)jj'})}{(\pi_{ij}-\phi_{(i)jj'})(\pi_{ij'}-\phi_{(i)jj'})}.$$

Since $0 < \pi_{ij} < 1$ and $0 < \pi_{ij'} < 1$, $\rho_{(i)jj'}$ is not free to range over $(-1, 1)$ (Bahadur, 1961). That is, $\rho_{(i)jj'}$is bounded below by

$$\max\left(-\sqrt{\frac{\pi_{ij}\pi_{ij'}}{(1-\pi_{ij})(1-\pi_{ij'})}}, -\sqrt{\frac{(1-\pi_{ij})(1-\pi_{ij'})}{\pi_{ij}\pi_{ij'}}}\right), \tag{3.2}$$

and above by

$$\min\left(\sqrt{\frac{\pi_{ij}(1-\pi_{ij'})}{\pi_{ij'}(1-\pi_{ij})}}, \sqrt{\frac{\pi_{ij'}(1-\pi_{ij})}{\pi_{ij}(1-\pi_{ij'})}}\right). \tag{3.3}$$

From the multivariate central limit theorem, $\boldsymbol{X} = (X_1, \ldots, X_p)^t$ is approximately normally distributed with mean $\boldsymbol{\Delta}$ and the covariance matrix $\boldsymbol{\Sigma}$.

Consider a hypothesis H_0 and an alternative hypothesis expressed by

$$H_0 : \left\{\max_{1\le j\le p} \delta_j \le 0\right\} \cup \left\{\min_{1\le j\le p} (\delta_j + \epsilon_j) \le 0\right\} \text{ versus } H_1 : \text{not } H_0,$$

where $\delta_j = \pi_{1j} - \pi_{2j}$ $(j = 1, \ldots, p)$ is true difference of proportions at the jth endpoint, and $\epsilon_j > 0$ $(j = 1, \ldots, p)$ is the non-inferiority margin of the jth endpoint that denotes

a prespecified positive constant. H_0 is also expressed as

$$H_0 \equiv H_0^{(0)} \cup \left\{ H_0^{(1)} \cup \cdots \cup H_0^{(p)} \right\},$$

which defines the sub hypothesis of superiority for at least one endpoint "$H_0^{(0)} : \max_{1 \le j \le p} \delta_j \le 0$" and the sub hypothesis of non-inferiority for all endpoints "$H_0^{(j)} : \delta_j \le -\epsilon_j$", for $j = 1, \ldots, p$. The intersection-union test (IUT) can be applied to test H_0 (Berger, 1982).

3.3 Proposed IUT

We propose a new IUT for testing H_0 versus H_1. Let $\boldsymbol{A}$ be the positive definite matrix such that $\boldsymbol{A}^t\boldsymbol{A} = \hat{\boldsymbol{\Sigma}}^{-1}$ where $\boldsymbol{\Sigma}^{-1}$ is the inverse matrix of $\boldsymbol{\Sigma}$, and the estimated covariance matrix $\hat{\boldsymbol{\Sigma}}$ is defined as follows:

$$\begin{aligned}
\hat{\boldsymbol{\Sigma}} &= \hat{\boldsymbol{\Sigma}}^{(1)} + \hat{\boldsymbol{\Sigma}}^{(2)} \\
&= \frac{1}{n_1} \begin{pmatrix} \hat{\pi}_{11}(1-\hat{\pi}_{11}) & \cdots & \sum_{\substack{(s_1,s_2,\ldots,s_p)\in S \\ s_1=s_p=1}} \hat{\theta}_{(1)s_1,s_2,\ldots,s_p} - \hat{\pi}_{11}\hat{\pi}_{1p} \\ \vdots & \ddots & \vdots \\ \sum_{\substack{(s_1,s_2,\ldots,s_p)\in S \\ s_1=s_p=1}} \hat{\theta}_{(1)s_1,s_2,\ldots,s_p} - \hat{\pi}_{11}\hat{\pi}_{1p} & \cdots & \hat{\pi}_{1p}(1-\hat{\pi}_{1p}) \end{pmatrix} \\
&+ \frac{1}{n_2} \begin{pmatrix} \hat{\pi}_{21}(1-\hat{\pi}_{21}) & \cdots & \sum_{\substack{(s_1,s_2,\ldots,s_p)\in S \\ s_1=s_p=1}} \hat{\theta}_{(2)s_1,s_2,\ldots,s_p} - \hat{\pi}_{21}\hat{\pi}_{2p} \\ \vdots & \ddots & \vdots \\ \sum_{\substack{(s_1,s_2,\ldots,s_p)\in S \\ s_1=s_p=1}} \hat{\theta}_{(2)s_1,s_2,\ldots,s_p} - \hat{\pi}_{21}\hat{\pi}_{2p} & \cdots & \hat{\pi}_{2p}(1-\hat{\pi}_{2p}) \end{pmatrix}.
\end{aligned}$$

$\hat{\pi}_{ij}$ is the estimator of π_{ij}, and is obtained by the maximum likelihood estimator (MLE) of $\theta_{(i)0,0,\ldots,0}, \theta_{(i)0,1,\ldots,0}, \ldots, \theta_{(i)1,1,\ldots,1}$. Let $\hat{\theta}_{(i)0,0,\ldots,0}, \hat{\theta}_{(i)0,1,\ldots,0}, \ldots, \hat{\theta}_{(i)1,1,\ldots,1}$ denote the MLE of $\theta_{(i)0,0,\ldots,0}$,
$\theta_{(i)0,1,\ldots,0}, \ldots, \theta_{(i)1,1,\ldots,1}$, which can be obtained under the sub null hypothesis $H_0^{(0)}$ or the sub alternative hypothesis $H_1^{(0)}$: not $H_0^{(0)}$. In particular, under the sub null hypothesis $H_0^{(0)}$, the Lagrange multiplier method is useful to obtain the MLE. On the other hand, under the sub alternative hypothesis $H_1^{(0)}$, $\hat{\theta}_{(i)0,0,\ldots,0}, \hat{\theta}_{(i)0,1,\ldots,0}, \ldots, \hat{\theta}_{(i)1,1,\ldots,1}$ are obtained

in a closed form as sample proportions.

The statistic $\boldsymbol{u}_A = (u_{A1}, \ldots, u_{Ap})^t = \boldsymbol{AX}$ is approximately distributed as a p-variate normal distribution with mean $\boldsymbol{A\Delta}$ and covariance matrix $\boldsymbol{I}$ (the identity matrix). For simplicity, we use the set of eigen vectors multiplied by the square root of the corresponding eigenvalue to represent $\boldsymbol{A}$, because $\boldsymbol{A}$ is not uniquely determined. Furthermore, according to the procedure of Nakazuru et al. (2014), $\boldsymbol{B}$ is defined as the matrix substituting the off-diagonal elements of $\boldsymbol{A}$ with their absolute values. Consider the two transformations such that

$$\begin{aligned} \boldsymbol{u}_A \equiv (u_{A1}, \ldots, u_{Ap})^t &= \boldsymbol{AX} \quad \text{and} \\ \boldsymbol{u}_B \equiv (u_{B1}, \ldots, u_{Bp})^t &= \left(\frac{\det \boldsymbol{A}}{\det \boldsymbol{B}}\right)^{2/p} \boldsymbol{BX}. \end{aligned}$$

In these assumption, the proposed IUT rejects H_0 if and only if

$$\begin{aligned} T^{(0)} &: \min(\overline{u}_A^2, \overline{u}_B^2) > c \quad \text{and} \\ T^{(j)} &: \frac{X_j + \epsilon_j}{\sqrt{\frac{\acute{\pi}_{1j}(1-\acute{\pi}_{1j})}{n_1} + \frac{\acute{\pi}_{2j}(1-\acute{\pi}_{2j})}{n_2}}} > z_\alpha \quad \text{for} \quad j = 1, \ldots, p, \end{aligned}$$

where $\overline{u}_A^2$ and $\overline{u}_B^2$ are defined by

$$\begin{aligned} \overline{u}_A^2 &= \sum_{j=1}^{p} \max(u_{Aj}, 0)^2, \\ \overline{u}_B^2 &= \sum_{j=1}^{p} \max(u_{Bj}, 0)^2. \end{aligned}$$

z_α represents the upper 100α th percentile of the standard normal distribution. $\acute{\pi}_{ij}$ is the MLE of π_{ij} derived under the sub null hypothesis of non-inferiority $H_0^{(j)}$, and $T^{(j)}$ is a statistic commonly used in non-inferiority tests of binary endpoints (Farrington and Manning, 1990). Further, c is a constant determined by

$$\sum_{j=0}^{p} \frac{p!}{j!\,(p-j)!} \frac{1}{2^p} Pr(\chi_j^2 > c) = \alpha.$$

Here, χ_j^2 denotes the χ^2 distribution with j degrees of freedom, χ_0^2 is defined as the constant zero, and α is the nominal significant level.

$T^{(0)}$ and $T^{(j)}$ are test statistics corresponding to the null hypotheses for superiority for at least one endpoint $H_0^{(0)}$ and non-inferiority for all endpoints $H_0^{(1)}, \ldots, H_0^{(p)}$, respectively. Two types of estimators for $T^{(0)}$ are proposed by obtaining the estimated covariance matrix $\hat{\boldsymbol{\Sigma}}$, which can be derived under the sub null hypothesis or the sub alternative hypothesis. In particular, $T_0^{(0)}$ is defined as the statistic for testing $H_0^{(0)}$ using an estimator obtained under the sub null hypothesis, and $T_1^{(0)}$ is the statistic for testing $H_0^{(0)}$ using an estimator obtained under the sub alternative hypothesis $H_1^{(0)}$ for testing superiority for at least one endpoint.

3.4 Simulation study

3.4.1 Type I error rate and power

We use a Monte Carlo study to compare the type I error rate and power of test types using $T_0^{(0)}$ vs. $T_1^{(0)}$ in the case $p = 2$. We consider $n_1 = n_2 = 50, 100, 200$, $\epsilon_1 = \epsilon_2 = 0.2$, and $\alpha = 0.05$. The random numbers are generated from a two-variate Bernoulli distribution with various mean vectors $\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2})^t$. The correlation between the endpoints assumes $\rho = 0, 0.4, 0.6$ with reference to Offen et al. (2007) and Sankoh et al. (1999), which indicated the correlation coefficients between multiple endpoints in clinical trials are approximately equal to 0.4. The generation of simulated data is repeated 1,000,000 times to calculate the type I error rates and 100,000 times to calculate the powers.

Table 3.1 shows the type I error rates for the two test types. These are cases in which the type I error rate is greater than the nominal significance level for the $T_1^{(0)}$ test type. The $T_0^{(0)}$ test type is more conservative than the $T_1^{(0)}$ test type. Table 3.2 shows the empirical powers of the two tests. The power of the $T_1^{(0)}$ test type is better when the proportional difference of at least one endpoint is large, but because of α-violation, only the $T_0^{(0)}$ test type can be used in practice.

3.4.2 Example-based simulation

We performed the other simulation based on a trial confirming the efficacy of intensive supportive care plus immunosuppression in IgA nephropathy (Rauen et al., 2015). We

are interested in the degree to which power is reduced by simultaneously demonstrating superiority of at least one endpoint and non-inferiority of the remaining endpoints. To explore this issue, we compare the power calculated by above example confirming superiority for at least one endpoint using two primary endpoints with the power adding the non-inferiority test for all endpoints to the superiority test for at least one endpoint. We considered $n_1 = n_2 = 50, 100, 200$, $\epsilon_1 = \epsilon_2 = 0.2$, $\alpha = 0.05$ and using $T_0^{(0)}$ test type. Random numbers were generated from a two variate Bernoulli distribution with mean vectors $\boldsymbol{\pi}_1 = (0.18, 0.26)^t$ and $\boldsymbol{\pi}_2 = (0.05, 0.28)^t$ according to actual proportions obtained in the trial. The correlation between the endpoints assumed $\rho = -0.1, 0, 0.3$ according to restrictions in Section 3.2. The generation of simulated data was repeated 100,000 times. Table 3.3 shows the power comparison between the proposed testing procedure and test of only superiority for at least one endpoint. As sample size increased, the difference in power between test of only superiority for at least one endpoint and that of superiority for at least one endpoint plus non-inferiority for all endpoints disappeared. On the other hand, the power is remarkably decrease when sample size is small and the correlation between two endpoints is increased.

3.4.3 Power reduction by adding non-inferiority test to superiority test

We also confirm the performance in the case of $p = 3$. We considered $n_1 = n_2 = 50, 100, 200$, $\epsilon_1 = \epsilon_2 = 0.2$, and $\alpha = 0.05$. Random numbers were generated from a three-variate Bernoulli distribution with various mean vectors $\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2}, \pi_{i3})^t$. The correlation common to any two of the three endpoints assumed $\rho = 0, 0.4, 0.6$. The generation of simulated data was repeated 100,000 times.

Table 3.4 shows the power comparison between the proposed testing procedure and test of only superiority for at least one endpoint for the case of $p = 3$. Similarly as in the case $p = 2$, adding a non-inferiority test did not reduce the power so much when the sample size is large. Conversely, the power was greatly reduced when the treatment effects were partially within the non-inferiority margin for a small sample size.

3.5 Concluding remarks

In this chapter, we developed a testing procedure that demonstrated efficacy when superiority of at least one binary endpoint and non-inferiority of the remaining binary endpoints were confirmed. We derived two types of test statistics using estimators obtained under the sub null hypothesis $H_0^{(0)}$ and the sub alternative hypothesis $H_1^{(0)}$, and these procedures were compared in a numerical experiment using Monte Carlo simulation.

The numerical experiment clearly demonstrated that the $T_0^{(0)}$ test type was always more conservative than the $T_1^{(0)}$ test type. Not only that, non-negligible α-violation occurred in the $T_1^{(0)}$ test type in studies with large sample sizes. Because α-violation is a serious issue in confirmatory clinical trials, we should consider the study size and employ test statistics using $T_0^{(0)}$. We would also like to note that section 5.6 of the ICH E9 guideline contains information related to the need to adjust the Type I error rate. Furthermore, it has also been found that the $T_1^{(0)}$ test type has significantly lower power than the $T_0^{(0)}$ test type when the proportional difference of at least one endpoint is small. Based on the performance of the power and controlling the type I error rate, we recommend the use of the $T_0^{(0)}$ test type in the clinical trial. In addition, this study showed that if at least one treatment effect was within the non-inferiority margin, remarkably conservative results were obtained. When the correlation coefficient increases, type I error rates tend to increase, as Nakazuru et al. (2014) pointed out in their paper. In particular, this problem is caused by the fact that the area where the sample space exceeds the rejection region is considered to be large when the mean difference for one endpoint is zero and that for the other endpoint is less than zero. To avoid this problem, correlations between outcomes should be investigated before the planning stages of trials. Therefore, we believe that the proposed IUT can be used in practice with certain correlation coefficients when the study has a large sample size.

Simulations also show that there is only a minimal decrease in power if the proportion of responses for the test treatment were completely or partly better than the control treatment. By contrast, the power was greatly reduced when the treatment effects were partially within the non-inferiority margin for a small sample size. In a clinical trial, the testing procedure should be chosen based on the scientific objective of what evidence should be presented for the effect of the intervention. If the objective of the trial is to show not only the superiority of at least one endpoint but also the non-inferiority of remaining endpoints, and if all endpoints are binary variables, the

proposed method would be practical.

Table 3.1 Type I error rates.

			$(\pi_{11}, \pi_{12}), (\pi_{21}, \pi_{22})$	
n	ρ	type	$(0.5, 0.5), (0.5, 0.5)$	$(0.4, 0.5), (0.5, 0.5)$
50	0.6	$T_0^{(0)}$	0.041	0.023
		$T_1^{(0)}$	0.047	0.027
	0.4	$T_0^{(0)}$	0.039	0.015
		$T_1^{(0)}$	0.045	0.017
	0	$T_0^{(0)}$	0.033	0.007
		$T_1^{(0)}$	0.037	0.008
100	0.6	$T_0^{(0)}$	0.045	0.027
		$T_1^{(0)}$	0.047	0.028
	0.4	$T_0^{(0)}$	0.044	0.017
		$T_1^{(0)}$	0.046	0.018
	0	$T_0^{(0)}$	0.045	0.009
		$T_1^{(0)}$	0.046	0.009
200	0.6	$T_0^{(0)}$	0.045	0.023
		$T_1^{(0)}$	0.046	0.024
	0.4	$T_0^{(0)}$	0.044	0.015
		$T_1^{(0)}$	0.045	0.016
	0	$T_0^{(0)}$	0.049	0.012
		$T_1^{(0)}$	0.050	0.012

Table 3.2 Estimated powers.

			$(\pi_{11}, \pi_{12}), (\pi_{21}, \pi_{22})$	
n	ρ	type	$(0.6, 0.5), (0.5, 0.5)$	$(0.7, 0.5), (0.5, 0.5)$
50	0.6	$T_0^{(0)}$	0.165	0.434
		$T_1^{(0)}$	0.184	0.472
	0.4	$T_0^{(0)}$	0.149	0.399
		$T_1^{(0)}$	0.163	0.413
	0	$T_0^{(0)}$	0.138	0.365
		$T_1^{(0)}$	0.149	0.379
100	0.6	$T_0^{(0)}$	0.311	0.782
		$T_1^{(0)}$	0.320	0.791
	0.4	$T_0^{(0)}$	0.278	0.748
		$T_1^{(0)}$	0.285	0.753
	0	$T_0^{(0)}$	0.283	0.737
		$T_1^{(0)}$	0.288	0.741
200	0.6	$T_0^{(0)}$	0.538	0.969
		$T_1^{(0)}$	0.542	0.970
	0.4	$T_0^{(0)}$	0.497	0.969
		$T_1^{(0)}$	0.501	0.970
	0	$T_0^{(0)}$	0.524	0.975
		$T_1^{(0)}$	0.527	0.975

Table 3.3 Evaluation of the power based on a real trial and the power reduction by adding the non-inferiority test.

		$(\pi_{11}, \pi_{12}), (\pi_{21}, \pi_{22}) : (0.18, 0.26), (0.05, 0.28)$	
n	ρ	Only superiority	Superiority + non-inferiority (proposed)
50	0.3	0.567	0.379
	0	0.485	0.351
	-0.1	0.606	0.495
100	0.3	0.814	0.735
	0	0.811	0.724
	-0.1	0.806	0.721
200	0.3	0.996	0.989
	0	0.994	0.986
	-0.1	0.990	0.983

Table 3.4 Evaluation of power reduction by adding the non-inferiority test.

n	ρ	type	$(\pi_{11}, \pi_{12}, \pi_{13}), (\pi_{21}, \pi_{22}, \pi_{23})$		
			(0.45,0.45,0.45), (0.4,0.4,0.4)	(0.45,0.45,0.38), (0.4,0.4,0.4)	(0.45,0.38,0.38), (0.4,0.4,0.4)
50	0.6	Superiority + non-inferiority	0.132	0.110	0.076
		Only superiority	0.140	0.130	0.091
	0.4	Superiority + non-inferiority	0.151	0.108	0.069
		Only superiority	0.165	0.131	0.085
	0	Superiority + non-inferiority	0.163	0.089	0.044
		Only superiority	0.229	0.157	0.093
100	0.6	Superiority + non-inferiority	0.221	0.208	0.135
		Only superiority	0.222	0.214	0.137
	0.4	Superiority + non-inferiority	0.248	0.196	0.115
		Only superiority	0.250	0.209	0.122
	0	Superiority + non-inferiority	0.423	0.271	0.138
		Only superiority	0.447	0.321	0.188
200	0.6	Superiority + non-inferiority	0.335	0.369	0.220
		Only superiority	0.335	0.370	0.220
	0.4	Superiority + non-inferiority	0.772	0.687	0.475
		Only superiority	0.772	0.692	0.478
	0	Superiority + non-inferiority	0.944	0.869	0.661
		Only superiority	0.945	0.885	0.681

Chapter 4

A method for testing multiple binary endpoints with continuous latent distribution in clinical trials

4.1 Introduction

In confirmatory clinical trials, several correlated binary response variables are used to assess the efficacy and safety of new treatments. The ICH E9 guideline recommends that the primary endpoint should consist of only one variable that provides strong scientific evidence of treatment efficacy. However, in clinical trials for a variety of diseases, it is often useful to evaluate efficacy using multiple primary endpoints. For example, in clinical trials of patients with rheumatoid arthritis, the percentage of patients achieving short-term improvement of 20 percent in the American College of Rheumatology criteria (ACR20) and the percentage achieving long-term low disease activity (Disease Activity Score (DAS28-ESR) ≤ 3.2) are often used as primary endpoints (e.g., Smolen et al., 2016). In clinical trials of patients with psoriasis, short- and long-term improvements are simultaneously assessed based on the percentage of patients with at least 75 percent improvement in the psoriasis area-and-severity index (DASI) score (e.g., Reich et al., 2011). In particular, binary endpoints are often used when it is more meaningful to diagnose improvement beyond clear standards rather than to assess the disease state using continuous variables. In such trials, we can consider that all primary endpoints are binary and often have a continuous latent distribution.

Most trials use multiple endpoints only to evaluate non-inferiority or superiority,

but some trials have been conducted to confirm the non-inferiority and superiority of all endpoints. For example, a clinical trial to confirm the efficacy of four-factor prothrombin complex concentrate (4F-PCC) included two primary endpoints (Goldstein et al., 2015), namely the percentage of patients with a hemostatic effect and the percentage with a decrease in the international normalized ratio (INR). In the above trial, superiority was only evaluated if there was non-inferiority for both endpoints. When we confirm not only non-inferiority but also superiority, the use of the closed testing procedure (Marcus et al., 1976) for the primary analysis is reasonable, and in this case, no adjustment is needed to control the type I error rate. In general, however, it is difficult to demonstrate the superiority of two or more endpoints because the power decreases as the number of endpoints increases. Therefore, developing a procedure that can confirm the superiority of at least one binary endpoint with latent distribution is a challenge for statisticians in the design and analysis of clinical trials. The aim of this chapter was thus to define a testing procedure within a framework in which the efficacy of a test treatment is confirmed only when the superiority of the treatment relative to control is evidenced for at least one endpoint, and non-inferiority is demonstrated for the remaining endpoints.

For multiple continuous endpoints, Perlman and Wu (2004) proposed a testing procedure that is applicable to the framework mentioned above. Nakazuru et al. (2014) proposed a more powerful testing procedure using the approximate likelihood ratio test (ALRT) defined by Glimm et al. (2002). However, there has been inadequate development of methods for multiple binary endpoints. Therefore, we herein propose a testing procedure that is appropriate when all endpoints are binary and have a latent distribution.

4.2 Assumption and Hypotheses

4.2.1 Statistical setting

We focus on a randomized clinical trial comparing p (≥ 2) endpoints with two treatment groups. There are n_1 subjects in the test group and n_2 subject in the control group. Let Y_{ijk} ($i = 1, 2; j = 1, \ldots, p; k = 1, \ldots, n_i$) denote the binary response variable of the jth primary endpoint of the ith treatment in the kth subject. Suppose that

the vectors of binary response variables $\boldsymbol{Y}_{ik} = (Y_{i1k}, \ldots, Y_{ipk})^t$ $(i = 1, 2; k = 1, \ldots, n_i)$ are independently distributed as a p-variate Bernoulli distribution with $\mathrm{E}(Y_{ijk}) = \pi_{ij}$, $\mathrm{V}(Y_{ijk}) = \pi_{ij}(1 - \pi_{ij})$, and $\mathrm{Corr}(Y_{ijk}, Y_{ij'k}) = \rho_{(i)jj'}$ for all $j \neq j'$, where the superscript "t" denotes transpose. In this setting, the correlation coefficient $\rho_{(i)jj'}$ of the multivariate Bernoulli distribution is expressed as

$$\rho_{(i)jj'} = \frac{\phi_{(i)jj'} - \pi_{ij}\pi_{ij'}}{\sqrt{\pi_{ij}(1 - \pi_{ij})}\sqrt{\pi_{ij'}(1 - \pi_{ij'})}}, \tag{4.1}$$

where $\phi_{(i)jj'}$ is the joint probability of two response variables $(Y_{ijk}, Y_{ij'k})$. Note that the range of $\rho_{(i)jj'}$ is equal to or less than $(-1, 1)$ depending on the value of π_{ij} and $\pi_{ij'}$ (Bahadur, 1961). That is, $\rho_{(i)jj'}$is bounded below by

$$\max\left(-\sqrt{\frac{\pi_{ij}\pi_{ij'}}{(1 - \pi_{ij})(1 - \pi_{ij'})}}, -\sqrt{\frac{(1 - \pi_{ij})(1 - \pi_{ij'})}{\pi_{ij}\pi_{ij'}}}\right), \tag{4.2}$$

and above by

$$\min\left(\sqrt{\frac{\pi_{ij}(1 - \pi_{ij'})}{\pi_{ij'}(1 - \pi_{ij})}}, \sqrt{\frac{\pi_{ij'}(1 - \pi_{ij})}{\pi_{ij}(1 - \pi_{ij'})}}\right). \tag{4.3}$$

Furthermore, we assume that $\boldsymbol{Y}_{ik}$ are dichotomized random variables of continuous unobservable response $\boldsymbol{Z}_{ik} = (Z_{i1k}, \ldots, Z_{ipk})^t$ $(i = 1, 2; k = 1, \ldots, n_i)$. We also assume that $\boldsymbol{Z}_{ik}$ are independently distributed as a standardized p-variate normal distribution with $\mathrm{Corr}(Z_{ijk}, Z_{ij'k}) = \gamma_{(i)jj'}$ for all $j \neq j'$. For each variable $\boldsymbol{Z}_{ik}$, there is a single threshold $g_{ij} = \Phi^{-1}(1 - \pi_{ij})$ $(i = 1, 2; j = 1, \ldots, p)$ that partitions the latent distribution, where Φ^{-1} is the inverse function of the standard normal cumulative distribution function. Then, the binary response Y_{ijk} $(i = 1, 2; j = 1, \ldots, p; k = 1, \ldots, n_i)$ can be defined as

$$Y_{ijk} = \begin{cases} 1, & Z_{ijk} \geq g_{ij} \\ 0, & Z_{ijk} < g_{ij}. \end{cases}$$

Set $\boldsymbol{X} = (X_1, \ldots, X_p)^t$ with $X_j = (\overline{Y}_{1j} - \overline{Y}_{2j})$ $(j = 1, \ldots, p)$, where $\overline{Y}_{ij}$ $(i = 1, 2; j = 1, \ldots, p)$ is the sample proportion for the jth endpoint of the ith treatment. Let the true proportion vector be the ith treatment $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{ip})^t$ with difference of proportion

$\boldsymbol{\Delta} = (\delta_1, \ldots, \delta_p)^t = \boldsymbol{\pi}_1 - \boldsymbol{\pi}_2$ and the covariance matrix $\boldsymbol{\Sigma}$ that is defined as follows:

$$
\begin{aligned}
\boldsymbol{\Sigma} &= \boldsymbol{\Sigma}^{(1)} + \boldsymbol{\Sigma}^{(2)} \\
&= \frac{1}{n_1} \begin{pmatrix} \pi_{11}(1-\pi_{11}) & \cdots & \rho_{(1)1p}\sqrt{\pi_{11}(1-\pi_{11})}\sqrt{\pi_{1p}(1-\pi_{1p})} \\ \vdots & \ddots & \vdots \\ \rho_{(1)p1}\sqrt{\pi_{11}(1-\pi_{11})}\sqrt{\pi_{1p}(1-\pi_{1p})} & \cdots & \pi_{1p}(1-\pi_{1p}) \end{pmatrix} \\
&+ \frac{1}{n_2} \begin{pmatrix} \pi_{21}(1-\pi_{21}) & \cdots & \rho_{(2)1p}\sqrt{\pi_{21}(1-\pi_{21})}\sqrt{\pi_{2p}(1-\pi_{2p})} \\ \vdots & \ddots & \vdots \\ \rho_{(2)p1}\sqrt{\pi_{21}(1-\pi_{21})}\sqrt{\pi_{2p}(1-\pi_{2p})} & \cdots & \pi_{2p}(1-\pi_{2p}) \end{pmatrix},
\end{aligned}
$$

where $\boldsymbol{\Sigma}^{(i)} (i = 1, 2)$ is the covariance matrix of $(\overline{Y}_{i1}, \ldots, \overline{Y}_{ip})^t$. Note that $\boldsymbol{X}$ is approximately normally distributed with mean $\boldsymbol{\Delta}$ and covariance matrix $\boldsymbol{\Sigma}$.

4.2.2 Hypotheses

Without loss of generality, we assume that test treatment superiority is recognized when the proportion of responses to the test treatment is greater than that to the control treatment. That is, a maximum value of $\delta_j (j = 1, \ldots, p)$ greater than 0 indicates an improvement of at least one endpoint of the test treatment compared to the control treatment. Furthermore, we consider the combined hypothesis for the superiority of at least one endpoint and the non-inferiority of the remaining endpoints. Therefore, we consider a hypothesis H_0 and an alternative hypothesis expressed by

$$
H_0 : \left\{ \max_{1 \le j \le p} \delta_j \le 0 \right\} \cup \left\{ \min_{1 \le j \le p} (\delta_j + \epsilon_j) \le 0 \right\} \text{ versus } H_1 : \text{not } H_0,
$$

where $\epsilon_j > 0$ $(j = 1, \ldots, p)$ is the non-inferiority margin of the jth endpoint that denotes a prespecified positive constant. H_0 is also expressed as

$$
H_0 \equiv H_0^{(0)} \cup \left\{ H_0^{(1)} \cup \cdots \cup H_0^{(p)} \right\},
$$

which defines the sub hypothesis of superiority "$H_0^{(0)} : \max_{1 \le j \le p} \delta_j \le 0$" and the sub hypothesis of non-inferiority "$H_0^{(j)} : \delta_j \le -\epsilon_j$", for $j = 1, \ldots, p$. $H_0^{(0)}$ is adaptable to the one-sided ALRT, and the IUT (Berger, 1982) can be applied to test H_0.

4.3 Estimation procedure of test statistics

To determine the statistics of ALRT and IUT, we need to estimate several parameters. For the sake of simplicity, the process of constructing statistics is divided into the following four steps.

Step1. Estimating the cut-off point g_{ij}

We assume that $\hat{g}_{ij}$ is the estimator of the latent cut-off point g_{ij}, and is estimated as $\hat{g}_{ij} = \Phi^{-1}(1-\widetilde{\pi}_{ij})$, where $\widetilde{\pi}_{ij}$ is the maximum likelihood estimator (MLE) of the marginal probability derived by the p-variate Bernoulli distribution. Let the probability mass function of the p-variate Bernoulli distribution be

$$\begin{aligned}\mathrm{P}(Y_{i1k} = y_{i1k}, \ldots, Y_{ipk} = y_{ipk}) = \theta_{(i)0,0,\ldots,0}^{\prod_{j=1}^{p}(1-y_{ijk})} \times \theta_{(i)1,0,\ldots,0}^{y_{i1k}\prod_{j=2}^{p}(1-y_{ijk})} \\ \times \theta_{(i)0,1,\ldots,0}^{(1-y_{i1k})y_{i2k}\prod_{j=3}^{p}(1-y_{ijk})} \times \cdots \times \theta_{(i)1,1,\ldots,1}^{\prod_{j=1}^{p} y_{ijk}},\end{aligned}$$

where $\theta_{(i)0,0,\ldots,0}, \ldots, \theta_{(i)1,1,\ldots,1}$ are joint probabilities when $\boldsymbol{Y}_{ik}$ takes values from $(0,\ldots,0), \ldots, (1,\ldots,1)$, respectively, and $\theta_{(i)0,0,\ldots,0} + \cdots + \theta_{(i)1,1,\ldots,1} = 1$. $\widetilde{\pi}_{ij}$ can be expressed as $\widetilde{\pi}_{ij} = \sum_{(s_1,s_2,\ldots,s_p)\in S, s_j=1} \hat{\theta}_{(i)s_1,s_2,\ldots,s_p}$ using the estimator of $\theta_{(i)s_1,s_2,\ldots,s_p}$, where $S = \{(s_1, s_2, \cdots s_p) | s_j = 0, 1, j = 1, \ldots, p\}$ is a set whose elements consist of all pairs of response values.

In addition, $\widetilde{\pi}_{ij}$ (and $\hat{g}_{ij}$) can be given in two ways depending on the estimation of $\theta_{(i)s_1,s_2,\ldots,s_p}$ under the sub-null hypothesis $H_0^{(0)}$ or the sub alternative hypothesis $H_1^{(0)}$: not $H_0^{(0)}$. In particular, under the sub null hypothesis $H_0^{(0)}$, the Lagrange multiplier method is useful to obtain the MLE. On the other hand, under the sub-alternative hypothesis $H_1^{(0)}$, the estimator $\hat{\theta}_{(i)s_1,s_2,\ldots,s_p}$ is obtained in a closed form as a sample proportion.

Step2. Estimate the joint and marginal probabilities

The estimator of the joint probability $\phi_{(i)jj'}$ in (4.1) is also given in two ways depending on $\hat{g}_{ij}$, which is obtained by the estimation of $\theta_{(i)s_1,s_2,\ldots,s_p}$ constructing $\widetilde{\pi}_{ij}$. $\hat{\phi}_{(i)jj'}$ can be given by

$$\begin{aligned}\hat{\phi}_{(i)jj'} &= \mathrm{Prob}(Z_{ij} \geq \hat{g}_{ij}, Z_{ij'} \geq \hat{g}_{ij'}) \\ &= \int_{-\infty}^{\infty} \cdots \int_{\hat{g}_{ij}}^{\infty} \cdots \int_{\hat{g}_{ij'}}^{\infty} \cdots \int_{-\infty}^{\infty} f(z_1, \ldots, z_p; \hat{\gamma}_{(i)jj'}) \mathrm{d}z_1 \cdots \mathrm{d}z_p \text{ for all } j \neq j',\end{aligned}$$

where $f(z_1, \ldots, z_p; \hat{\gamma}_{(i)jj'})$ is the joint density function of $\boldsymbol{Z}_{ik}$ and $z_1, \ldots, z_p$ are random variables following the standard p-variate normal distribution wherein $\hat{\gamma}_{(i)jj'}$ is the Person's tetrachoric correlation (Pearson, 1900) calculated from $(Y_{ij1}, \ldots, Y_{ijn_i})$ and $(Y_{ij'1}, \ldots, Y_{ij'n_i})$. Therefore, if the latency of the binary response is assumed to have a standardized multivariate normal distribution, $\hat{\phi}_{(i)jj'}$ is determined by $\hat{\gamma}_{(i)jj'}$ and the cut-off point given in Step 1. Furthermore, the estimator of the marginal probability π_{ij} constructing Σ and $\rho_{(i)jj'}$ in (4.1) should not be $\widetilde{\pi}_{ij}$, which is obtained from the p-variate Bernoulli distribution, but should instead take into account the latent distribution function. Let $\hat{\pi}_{ij}$ denote the estimator of π_{ij} and be given by

$$\begin{aligned}\hat{\pi}_{ij} &= \text{Prob}(Z_{ij} \geq \hat{g}_{ij}) \\ &= \int_{-\infty}^{\infty} \cdots \int_{\hat{g}_{ij}}^{\infty} \cdots \int_{-\infty}^{\infty} f(z_1, \ldots, z_p; \hat{\gamma}_{(i)jj'}) \mathrm{d}z_1 \cdots \mathrm{d}z_p \text{ for all } j.\end{aligned}$$

For example, with $p = 2$ endpoints, the estimator of $\phi_{(i)12}$ is written as

$$\hat{\phi}_{(i)12} = \int_{g_{i1}}^{\infty} \int_{g_{i2}}^{\infty} f(z_1, z_2; \hat{\gamma}_{(i)12}) \mathrm{d}z_1 \mathrm{d}z_2.$$

Furthermore, the marginal probabilities π_{i1} and π_{i2} are described as follows:

$$\begin{aligned}\hat{\pi}_{i1} &= \text{Prob}(Z_{i1} \geq \hat{g}_{i1}) = \int_{-\infty}^{\infty} \int_{\hat{g}_{i1}}^{\infty} f(z_1, z_2; \hat{\gamma}_{(i)12}) \mathrm{d}z_1 \mathrm{d}z_2, \\ \hat{\pi}_{i2} &= \text{Prob}(Z_{i2} \geq \hat{g}_{i2}) = \int_{-\infty}^{\infty} \int_{\hat{g}_{i2}}^{\infty} f(z_1, z_2; \hat{\gamma}_{(i)12}) \mathrm{d}z_1 \mathrm{d}z_2.\end{aligned}$$

Along with the estimation of the joint and marginal probabilities, the estimated covariance matrix $\hat{\boldsymbol{\Sigma}}$ is defined as follows:

$$\begin{aligned}\hat{\boldsymbol{\Sigma}} &= \hat{\boldsymbol{\Sigma}}^{(1)} + \hat{\boldsymbol{\Sigma}}^{(2)} \\ &= \frac{1}{n_1} \begin{pmatrix} \hat{\pi}_{11}(1-\hat{\pi}_{11}) & \cdots & \hat{\phi}_{(1)1p} - \hat{\pi}_{11}\hat{\pi}_{1p} \\ \vdots & \ddots & \vdots \\ \hat{\phi}_{(1)p1} - \hat{\pi}_{11}\hat{\pi}_{1p} & \cdots & \hat{\pi}_{1p}(1-\hat{\pi}_{1p}) \end{pmatrix} \\ &+ \frac{1}{n_2} \begin{pmatrix} \hat{\pi}_{21}(1-\hat{\pi}_{21}) & \cdots & \hat{\phi}_{(2)1p} - \hat{\pi}_{21}\hat{\pi}_{2p} \\ \vdots & \ddots & \vdots \\ \hat{\phi}_{(2)p1} - \hat{\pi}_{21}\hat{\pi}_{2p} & \cdots & \hat{\pi}_{2p}(1-\hat{\pi}_{2p}) \end{pmatrix}.\end{aligned}$$

Step3. Estimate the test statistics of an ALRT

We propose an ALRT for $H_0^{(0)}$. Let $\boldsymbol{A}$ be the positive definite matrix such that $\boldsymbol{A}^t\boldsymbol{A} = \hat{\boldsymbol{\Sigma}}^{-1}$, where $\boldsymbol{\Sigma}^{-1}$ is the inverse matrix of $\boldsymbol{\Sigma}$. The statistic $\boldsymbol{u}_A = (u_{A1}, \ldots, u_{Ap})^t = \boldsymbol{AX}$ is approximately distributed as a p-variate normal distribution with mean $\boldsymbol{A\Delta}$ and covariance matrix $\boldsymbol{I}$ (the identity matrix). For simplicity, to represent $\boldsymbol{A}$ we use the set of eigenvectors multiplied by the square root of the corresponding eigenvalue, because $\boldsymbol{A}$ is not uniquely determined. Furthermore, according to the procedure of Nakazuru et al. (2014), $\boldsymbol{B}$ is defined as the matrix substituting the off-diagonal elements of $\boldsymbol{A}$ with their absolute values. Consider the two transformations such that

$$\begin{aligned} \boldsymbol{u}_A \equiv (u_{A1}, \ldots, u_{Ap})^t &= \boldsymbol{AX} \quad \text{and} \\ \boldsymbol{u}_B \equiv (u_{B1}, \ldots, u_{Bp})^t &= \left(\frac{\det\boldsymbol{A}}{\det\boldsymbol{B}}\right)^{2/p} \boldsymbol{BX}. \end{aligned}$$

In these assumptions, the proposed ALRT rejects $H_0^{(0)}$ if and only if

$$T^{(0)} \ : \ \min(\overline{u}_A^2, \overline{u}_B^2) > c,$$

where $\overline{u}_A^2$ and $\overline{u}_B^2$ are defined by

$$\begin{aligned} \overline{u}_A^2 &= \sum_{j=1}^{p} \max(u_{Aj}, 0)^2, \\ \overline{u}_B^2 &= \sum_{j=1}^{p} \max(u_{Bj}, 0)^2. \end{aligned}$$

c is a constant determined by

$$\sum_{j=0}^{p} \frac{p!}{j!\,(p-j)!} \frac{1}{2^p} Pr(\chi_j^2 > c) = \alpha.$$

Here, χ_j^2 denotes the χ^2 distribution with j degrees of freedom, χ_0^2 is defined as the constant zero, and α is the nominal significance level. Note that test statistics $T^{(0)}$ can be estimated in two ways. One is provided by $\hat{\boldsymbol{\Sigma}}$ derived under the sub-null hypothesis $H_0^{(0)}$, and the other is provided by $\hat{\boldsymbol{\Sigma}}$ derived under the sub-alternative hypothesis $H_1^{(0)}$.

Step4. Estimate the test statistics of an IUT

Based on the Steps 1 to 3 shown above, we consider the statistics of an IUT to test hypothesis H_0 versus H_1. More precisely, the proposed IUT rejects H_0 if and only if

$$\begin{aligned} T^{(0)} &: \min(\overline{u}_A^2, \overline{u}_B^2) > c \quad \text{and} \\ T^{(j)} &: \frac{X_j + \epsilon_j}{\sqrt{\frac{\acute{\pi}_{1j}(1-\acute{\pi}_{1j})}{n_1} + \frac{\acute{\pi}_{2j}(1-\acute{\pi}_{2j})}{n_2}}} > z_\alpha \quad \text{for} \quad j = 1, \ldots, p, \end{aligned}$$

where $\acute{\pi}_{ij}$ is the MLE of π_{ij} derived under the sub null hypothesis of non-inferiority $H_0^{(j)}$, and $T^{(j)}$ is a statistic commonly used in non-inferiority tests of binary endpoints (Farrington and Manning, 1990). $T^{(0)}$ and $T^{(j)}$ are test statistics corresponding to the null hypotheses for superiority $H_0^{(0)}$ and non-inferiority $H_0^{(1)}, \ldots, H_0^{(p)}$, respectively.

In Steps 1 to 3, we stated that there are two types of $T^{(0)}$, and thus we consider that there are also two types of IUT statistics. Let the IUT statistics using $T^{(0)}$ estimated under the sub-null hypothesis $H_0^{(0)}$ be the $T_0^{(0)}$ test type, and the those using $T^{(0)}$ estimated under the sub-null hypothesis $H_1^{(0)}$ be the $T_0^{(1)}$ test type.

4.4 Simulation study

4.4.1 Type I error rate

We use a Monte Carlo simulation to compare the type I error rate of the $T_0^{(0)}$ test type and the $T_1^{(0)}$ test type in the case $p = 2$. We consider $n_1 = n_2 = 50, 100, 200$, $\epsilon_1 = \epsilon_2 = 0.2$, and $\alpha = 0.05$. The random numbers are generated from a standardized bivariate normal distribution, and response variables are obtained by dichotomizing random numbers using $\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2})^t$. The correlation between the latent variables assumes $\rho = 0, 0.4, 0.8$. The generation of simulated data is repeated 1,000,000 times.

Table 4.1 shows the type I error rates for the two test types. The type I error rate is greater than the nominal significance level for the $T_1^{(0)}$ test type when the correlation between the endpoints is zero with a large sample size. The $T_0^{(0)}$ test type is more conservative than the $T_1^{(0)}$ test type.

4.4.2 Power

We use a Monte Carlo simulation to compare the powers of the $T_0^{(0)}$ and the $T_1^{(0)}$ test type for the proposed IUT, in the case $p = 2$. We consider $n_1 = n_2 = 50, 100, 200$, $\epsilon_1 = \epsilon_2 = 0.2$, and $\alpha = 0.05$. The random numbers are generated from a standardized bivariate normal distribution, and response variables are obtained by dichotomizing random numbers using $\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2})^t$. The correlation between the latent variables assumes $\rho = 0, 0.4, 0.6$ according to restrictions in Section 4.2. We also compare the power of the proposed IUT with that of a closed testing procedure that confirms the superiority of at least one of the two endpoints after the non-inferiority of both of the two endpoints is confirmed. The Bonferroni-corrected p-value (Bonferroni, 1936) is used to test for superiority in the closed testing procedure. The generation of simulated data is repeated 100,000 times.

Table 4.2 shows the empirical powers of the proposed IUT and the closed testing procedure. The power of the $T_1^{(0)}$ test type is greater than that of the $T_0^{(0)}$ test type, and it becomes larger as the correlation between the endpoints increases with the small sample size. Even when the difference between the endpoints increases, the relationship between the power of the $T_0^{(0)}$ test type and that of the $T_1^{(0)}$ test type does not change much. On the other hand, as the sample size increases, the power of the $T_0^{(0)}$ test type becomes similar to that of the $T_1^{(0)}$ test type. Furthermore, the power of the proposed IUT is always greater than that of the closed testing procedure. The results for $p = 2$ and $p = 3$ in Section 4.4.3 below show that as the number of endpoints that differ between the two groups increases, the power of the closed testing procedure is noticeably lower than that of the proposed IUT.

4.4.3 Power of non-inferiority test added to superiority test

We also compare the performance of the proposed IUT and a test excluding the non-inferiority in the case of $p = 2$ and $p = 3$. We consider $n_1 = n_2 = 50, 100, 200$, $\epsilon_1 = \epsilon_2 = 0.2$, $\alpha = 0.05$, and test type $T_0^{(0)}$. Random numbers are generated from a bivariate Bernoulli distribution with various mean vectors $\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2})^t$ in the case $p = 2$, and a three-variate Bernoulli distribution with various mean vectors $\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2}, \pi_{i3})^t$ in the case of $p = 3$. The correlation common to any two endpoints is assumed to be $\rho = 0, 0.4, 0.6$. The generation of simulated data is repeated 100,000 times.

Table 4.3 shows a power comparison between the proposed IUT and the superiority test alone for the case of $p = 2$. As the sample size increases, regardless of the value of the correlation coefficient, the powers of the superiority test alone and the IUT remain similar. Even when the differences between the two groups are partially within the non-inferiority margin, the relationship between the power of the IUT and that of the superiority test alone remains comparable.

Table 4.4 shows a power comparison between the proposed IUT and the superiority test alone for the case of $p = 3$. As with the case of $p = 2$, adding a non-inferiority test does not reduce the power much when the sample size is large. Conversely, the power of the IUT is greatly reduced when the differences are partially within the non-inferiority margin for a small sample size.

4.5 Concluding remarks

In this chapter, we developed a testing procedure for studies with multiple binary endpoints and a latent distribution. This was performed within a framework in which the efficacy of a test treatment is recognized when at least one endpoint demonstrates superiority and the remaining endpoints demonstrate non-inferiority. We derived two types of test statistics using cut-off points estimated under the sub null hypothesis $H_0^{(0)}$ and the sub-alternative hypothesis $H_1^{(0)}$, and these procedures were compared in a numerical experiment using a Monte Carlo simulation.

The numerical experiment clearly demonstrated that the $T_0^{(0)}$ test type was always more conservative than the $T_1^{(0)}$ test type. Furthermore, α-violation occurred in the $T_1^{(0)}$ test type when a sample size was large and the correlation coefficient was zero. Because α-violation is a serious issue in confirmatory clinical trials, we should choose a non-inflationary test. Since there was not a large difference in power between the $T_0^{(0)}$ and $T_1^{(0)}$ test types, it may be reasonable to preferentially use the $T_0^{(0)}$ test type, especially if the correlation coefficients between the endpoints have not been investigated. Furthermore, this study showed that an increased number of endpoints within the non-inferiority margin resulted in markedly conservative results, and type I error rates tended to increase when the correlation coefficient decreased. To avoid this problem, correlations between outcomes should be investigated before the planning stages of trials. According to Offen et al. (2007) and Sankoh et al. (1999), the correlation

coefficients between multiple endpoints in clinical trials are approximately equal to 0.4 and range from 0.2 to 0.8. Therefore, we believe that the proposed IUT can be used in practice with certain correlation coefficients when the study has a large sample size.

Incidentally, like the proposed testing procedure, a closed testing procedure can be used for multiple endpoints where the familywise error rate is kept below the nominal significance level. In the framework of this study, the proposed IUT was shown to be more powerful than the closed testing procedure regardless of the correlation coefficient between endpoints, the difference between the endpoints, the number of noticeably different endpoints, and the sample size. Although the closed testing procedure has a significant advantage in that it does not require control of the type I error rate in individual tests when there are inclusion relationships between null hypotheses, it may be more reasonable to use the proposed IUT in the framework of this study, where the superiority of at least one endpoint and the non-inferiorities of the remaining endpoints are confirmed simultaneously.

We also demonstrated a power reduction when the non-inferiority test was added to the superiority test. Our simulations showed that there was only a minimal decrease in power when the proportions of responses to the test treatment were all or somewhat higher than that to the control treatment. By contrast, the power was reduced when the treatment effects were partially within the non-inferiority margin for a small sample size. In particular, the power decreased remarkably with increasing numbers of variables within the non-inferiority margin. Furthermore, the smaller the correlation coefficient, the lower the power of the proposed method in comparison to a procedure that tested only superiority. Therefore, in a primary analysis using the proposed testing procedure for certain sample sizes, assuming differences in proportions and correlations between endpoints, if all endpoints are binary and have a continuous latent distribution then it is ideal in practice to confirm not only the superiority of at least one endpoint, but also the non-inferiority of all remaining endpoints.

Table 4.1 Type I error rates.

			$(\pi_{11}, \pi_{12}), (\pi_{21}, \pi_{22})$	
n	ρ	Type of IUT	$(0.5, 0.5), (0.5, 0.5)$	$(0.4, 0.5), (0.5, 0.5)$
50	0.8	$T_0^{(0)}$	0.042	0.023
		$T_1^{(0)}$	0.047	0.026
	0.4	$T_0^{(0)}$	0.038	0.011
		$T_1^{(0)}$	0.043	0.013
	0	$T_0^{(0)}$	0.034	0.007
		$T_1^{(0)}$	0.038	0.008
100	0.8	$T_0^{(0)}$	0.045	0.026
		$T_1^{(0)}$	0.047	0.027
	0.4	$T_0^{(0)}$	0.044	0.013
		$T_1^{(0)}$	0.045	0.013
	0	$T_0^{(0)}$	0.045	0.009
		$T_1^{(0)}$	0.046	0.009
200	0.8	$T_0^{(0)}$	0.045	0.022
		$T_1^{(0)}$	0.046	0.022
	0.4	$T_0^{(0)}$	0.044	0.013
		$T_1^{(0)}$	0.045	0.013
	0	$T_0^{(0)}$	0.049	0.012
		$T_1^{(0)}$	0.050	0.012

Table 4.2 Estimated powers.

n	ρ	Type of test	$(\pi_{11},\pi_{12}),(\pi_{21},\pi_{22})$		
			(0.6,0.5), (0.5,0.5)	(0.7,0.5), (0.5,0.5)	(0.6,0.6), (0.5,0.5)
50	0.6	$T_0^{(0)}$(proposed)	0.139	0.343	0.270
		$T_1^{(0)}$(proposed)	0.150	0.352	0.288
		Closed testing procedure	0.086	0.280	0.168
	0.4	$T_0^{(0)}$(proposed)	0.132	0.334	0.285
		$T_1^{(0)}$(proposed)	0.143	0.346	0.304
		Closed testing procedure	0.079	0.270	0.171
	0	$T_0^{(0)}$(proposed)	0.126	0.322	0.321
		$T_1^{(0)}$(proposed)	0.134	0.332	0.335
		Closed testing procedure	0.062	0.234	0.162
100	0.6	$T_0^{(0)}$(proposed)	0.275	0.703	0.452
		$T_1^{(0)}$(proposed)	0.280	0.707	0.460
		Closed testing procedure	0.208	0.652	0.342
	0.4	$T_0^{(0)}$(proposed)	0.265	0.686	0.489
		$T_1^{(0)}$(proposed)	0.269	0.691	0.496
		Closed testing procedure	0.203	0.638	0.358
	0	$T_0^{(0)}$(proposed)	0.268	0.684	0.575
		$T_1^{(0)}$(proposed)	0.271	0.689	0.581
		Closed testing procedure	0.186	0.618	0.379
200	0.6	$T_0^{(0)}$(proposed)	0.497	0.961	0.713
		$T_1^{(0)}$(proposed)	0.501	0.962	0.716
		Closed testing procedure	0.403	0.953	0.579
	0.4	$T_0^{(0)}$(proposed)	0.493	0.960	0.757
		$T_1^{(0)}$(proposed)	0.497	0.961	0.760
		Closed testing procedure	0.402	0.951	0.605
	0	$T_0^{(0)}$(proposed)	0.524	0.963	0.841
		$T_1^{(0)}$(proposed)	0.526	0.964	0.842
		Closed testing procedure	0.404	0.950	0.643

Table 4.3 Evaluation of power reduction when adding the non-inferiority test for $p = 2$.

n	ρ	Type of test	$(\pi_{11}, \pi_{12}), (\pi_{21}, \pi_{22})$		
			(0.45,0.45), (0.4,0.4)	(0.45,0.4), (0.4,0.4)	(0.45,0.38), (0.4,0.4)
50	0.8	Superiority + non-inferiority	0.121	0.088	0.082
		Only superiority	0.124	0.090	0.086
	0.4	Superiority + non-inferiority	0.152	0.090	0.074
		Only superiority	0.161	0.100	0.086
	0	Superiority + non-inferiority	0.205	0.108	0.081
		Only superiority	0.232	0.133	0.107
100	0.8	Superiority + non-inferiority	0.716	0.592	0.548
		Only superiority	0.719	0.601	0.566
	0.4	Superiority + non-inferiority	0.771	0.632	0.572
		Only superiority	0.788	0.673	0.637
	0	Superiority + non-inferiority	0.804	0.650	0.579
		Only superiority	0.835	0.719	0.680
200	0.8	Superiority + non-inferiority	0.821	0.696	0.660
		Only superiority	0.821	0.696	0.660
	0.4	Superiority + non-inferiority	0.870	0.745	0.706
		Only superiority	0.871	0.748	0.713
	0	Superiority + non-inferiority	0.905	0.777	0.725
		Only superiority	0.905	0.781	0.738

Table 4.4 Evaluation of power reduction when adding the non-inferiority test for $p = 3$.

n	ρ	Type of test	$(\pi_{11}, \pi_{12}, \pi_{13}), (\pi_{21}, \pi_{22}, \pi_{23})$		
			(0.45,0.45,0.45), (0.4,0.4,0.4)	(0.45,0.45,0.38), (0.4,0.4,0.4)	(0.45,0.38,0.38), (0.4,0.4,0.4)
50	0.8	Superiority + non-inferiority	0.314	0.244	0.132
		Only superiority	0.318	0.294	0.154
	0.4	Superiority + non-inferiority	0.410	0.225	0.100
		Only superiority	0.435	0.322	0.162
	0	Superiority + non-inferiority	0.461	0.192	0.063
		Only superiority	0.523	0.344	0.161
100	0.8	Superiority + non-inferiority	0.548	0.536	0.303
		Only superiority	0.548	0.562	0.310
	0.4	Superiority + non-inferiority	0.673	0.485	0.240
		Only superiority	0.674	0.549	0.281
	0	Superiority + non-inferiority	0.803	0.492	0.207
		Only superiority	0.807	0.598	0.292
200	0.8	Superiority + non-inferiority	0.787	0.809	0.481
		Only superiority	0.787	0.812	0.481
	0.4	Superiority + non-inferiority	0.951	0.876	0.449
		Only superiority	0.951	0.915	0.457
	0	Superiority + non-inferiority	0.968	0.833	0.484
		Only superiority	0.968	0.854	0.507

Chapter 5

Discussion and Conclusions

This thesis provided proposals for a methodology for clinical trials with multiple categorical variables. Especially in the preliminary steps of confirmatory clinical trials, the characterization of categorical variables can provide valuable information to the end-users of the study, such as patients and clinicians.

Chapter 2 proposed a symmetry measure of contingency tables that combines partial measures in the class of weighted averages. The partial measure allowed us to identify the categories that cause the symmetric structure not to be hold.

A marginal homogeneity model (Stuart, 1955) focusing on marginal probabilities is also considered a model of homogeneity between ordinal categorical variables. A measure of homogeneity of marginal probabilities has been proposed by, for example, Tomizawa, Miyamoto, and Ashihara (2003), and this measure can also be used to characterize ordinal categorical variables. The method proposed in this thesis has the advantage of being able to check the structure between cells in detail when symmetry between categories does not hold. The measures pooled as weighted averages were compared and characterized with existing measures of symmetry. The choice of primary endpoints cannot be made only from a statistical perspective and must be discussed from a clinical perspective as well. The use of symmetry measures and that confidence intervals may be useful in discussing whether there is a discrepancy between clinical perceptions and the characteristics of the endpoints.

The methods proposed in Chapters 3 and 4 may be useful when the primary endpoints, selected through various clinical, statistical, and ethical discussions, are multiple and binary. Chapter 3 describes two statistics in order to test a hypothesis that includes superiority of at least one endpoint and non-inferiority of the remaining endpoints when

all endpoints are binary. We showed that using the test statistic estimated under the sub null hypothesis of superiority is a more conservative result in the sense of α error. We also show that there is only a minimal decrease in power if the proportion of responses for the test treatment were completely or partly better than the control treatment. By contrast, the power was greatly reduced when the treatment effects were partially within the non-inferiority margin for a small sample size. Chapter 4 proposes two test statistics for studies with multiple binary endpoints and a latent distribution. The procedure for obtaining estimates of the test statistic was presented, and two types of statistics were proposed depending on how the cut-off points of the latent variables were determined.

In general, the closed testing procedure is commonly used for testing when the null hypothesis of inclusion holds. It is also applicable to the framework addressed in this thesis, which recognizes for a treatment effect only when at least one of the endpoints is superior and remaining endpoints are non-inferior. It would also be important to compare the performance of the proposed method in Chapter 3 with that of the proposed method in Chapter 4 in the case where the binary endpoints have latent distributions. Therefore, we compared the power of the closed testing procedure with the methods treated in Chapters 3 and 4, assuming that when multiple binary endpoints have a latent continuous distribution.

We consider $p = 2$, $n_1 = n_2 = 50, 100, 200$, $\epsilon_1 = \epsilon_2 = 0.2$, $\rho = 0, 0.4, 0.8$, and $\alpha = 0.05$. The random numbers are generated from a standardized bivariate normal distribution, and response variables are obtained by dichotomizing random numbers using $\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2})^t$. The Pearson's χ^2 test and the Bonferroni-corrected p-value (Bonferroni, 1936) are used to test for superiority in the closed testing procedure. The generation of simulated data is repeated 100,000 times. Table 5.1 shows the empirical powers from the closed testing procedure and the power of the IUTs proposed in Chapters 3 and 4. In all scenarios, the proposed method had higher power than the closed testing procedure. We also found that the proposed method in Chapter 4 is slightly higher than the proposed method in Chapter 3 under the assumption of a continuous latent distribution for the binary endpoints. However, the difference in their performance is slight, so in practical terms there is no harm in applying the Chapter 3 method even if the binary variables have a potentially continuous distribution.

We also evaluated the performance of the methods in Chapter 4 when the latent distribution is misidentified. Random numbers are generated from a bivariate lognormal

distribution, and response variables are obtained by dichotomizing random numbers using $\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2})^t$. Other settings were the same as in Table 5.1. The power of each method is shown in Table 5.2. Compared to the closed testing procedure, the methods in Chapters 3 and 4 have higher power, similar to the results in Table 5.1. We also found that even when the distribution was misidentified, the Chapter 4 method was slightly more powerful.

More recently, methods using a mix of multiple binary and continuous variables as the primary endpoint in situations where latent distributions are assumed have also been discussed (McMenamin, 2022). In addition, the proposed testing procedure and the closed testing procedure do not allow the user to specify the order in which the variables are tested. Considering practicality, the development of a theory that can be tested according to the user-specified priorities within the framework addressed in this thesis may be necessary in the future. The types of clinical trials are diversifying, and there are more and more situations that cannot be handled by simply using a single primary endpoint as in the past. The results of this study may contribute to the development categorical analyses in clinical trials because the proposed procedure and measures can provide a new interpretation.

Table 5.1 Comparison of the power of proposed methods in Chapter 3 and Chapter 4 and the closed testing procedure.

			$(\pi_{11}, \pi_{12}), (\pi_{21}, \pi_{22})$		
n	ρ	Type of test	(0.6,0.5), (0.5,0.5)	(0.7,0.5), (0.5,0.5)	(0.4,0.5), (0.5,0.5)
50	0.8	$T_0^{(0)}$(Chapter 3)	0.1392	0.3038	0.0175
		$T_1^{(0)}$(Chapter 3)	0.1498	0.3117	0.0191
		$T_0^{(0)}$(Chapter 4)	0.1408	0.3053	0.0179
		$T_1^{(0)}$(Chapter 4)	0.1498	0.3117	0.0191
		Closed testing procedure	0.0786	0.2386	0.0067
	0.4	$T_0^{(0)}$(Chapter 3)	0.1294	0.3253	0.0085
		$T_1^{(0)}$(Chapter 3)	0.1415	0.3368	0.0098
		$T_0^{(0)}$(Chapter 4)	0.1303	0.3258	0.0087
		$T_1^{(0)}$(Chapter 4)	0.1415	0.3368	0.0098
		Closed testing procedure	0.0785	0.2606	0.0042
	0	$T_0^{(0)}$(Chapter 3)	0.1185	0.3114	0.0052
		$T_1^{(0)}$(Chapter 3)	0.1272	0.3230	0.0059
		$T_0^{(0)}$(Chapter 4)	0.1196	0.3132	0.0052
		$T_1^{(0)}$(Chapter 4)	0.1272	0.3230	0.0059
		Closed testing procedure	0.0604	0.2274	0.0019
100	0.8	$T_0^{(0)}$(Chapter 3)	0.2995	0.6976	0.0249
		$T_1^{(0)}$(Chapter 3)	0.3077	0.7005	0.0259
		$T_0^{(0)}$(Chapter 4)	0.3032	0.6990	0.0257
		$T_1^{(0)}$(Chapter 4)	0.3077	0.7005	0.0259
		Closed testing procedure	0.2093	0.6342	0.0121
	0.4	$T_0^{(0)}$(Chapter 3)	0.2646	0.6810	0.0127
		$T_1^{(0)}$(Chapter 3)	0.2699	0.6866	0.0128
		$T_0^{(0)}$(Chapter 4)	0.2652	0.6813	0.0127
		$T_1^{(0)}$(Chapter 4)	0.2699	0.6866	0.0128
		Closed testing procedure	0.2009	0.6320	0.0043
	0	$T_0^{(0)}$(Chapter 3)	0.2714	0.6851	0.0066
		$T_1^{(0)}$(Chapter 3)	0.2755	0.6892	0.0068
		$T_0^{(0)}$(Chapter 4)	0.2719	0.6852	0.0067
		$T_1^{(0)}$(Chapter 4)	0.2755	0.6892	0.0068
		Closed testing procedure	0.1879	0.6199	0.0043
200	0.8	$T_0^{(0)}$(Chapter 3)	0.5378	0.9531	0.0226
		$T_1^{(0)}$(Chapter 3)	0.5415	0.9537	0.0229
		$T_0^{(0)}$(Chapter 4)	0.5396	0.9533	0.0228
		$T_1^{(0)}$(Chapter 4)	0.5415	0.9537	0.0229
		Closed testing procedure	0.4012	0.9475	0.0132
	0.4	$T_0^{(0)}$(Chapter 3)	0.4883	0.9604	0.0115
		$T_1^{(0)}$(Chapter 3)	0.4938	0.9610	0.0119
		$T_0^{(0)}$(Chapter 4)	0.4887	0.9603	0.0115
		$T_1^{(0)}$(Chapter 4)	0.4938	0.9610	0.0119
		Closed testing procedure	0.3997	0.9520	0.009
	0	$T_0^{(0)}$(Chapter 3)	0.5256	0.9645	0.0100
		$T_1^{(0)}$(Chapter 3)	0.5280	0.9649	0.0101
		$T_0^{(0)}$(Chapter 4)	0.5262	0.9645	0.0099
		$T_1^{(0)}$(Chapter 4)	0.5280	0.9649	0.0101
		Closed testing procedure	0.4058	0.9509	0.0062

Table 5.2 Comparison of the power when the latent distribution of the binary variable is log-normal.

			$(\pi_{11}, \pi_{12}), (\pi_{21}, \pi_{22})$		
n	ρ	Type of test	(0.6,0.5), (0.5,0.5)	(0.7,0.5), (0.5,0.5)	(0.4,0.5), (0.5,0.5)
50	0.8	$T_0^{(0)}$(Chapter 3)	0.1397	0.2982	0.0184
		$T_1^{(0)}$(Chapter 3)	0.1506	0.3052	0.0207
		$T_0^{(0)}$(Chapter 4)	0.1425	0.2999	0.0191
		$T_1^{(0)}$(Chapter 4)	0.1505	0.3052	0.0207
		Closed testing procedure	0.0814	0.2315	0.0077
	0.4	$T_0^{(0)}$(Chapter 3)	0.1219	0.3265	0.0104
		$T_1^{(0)}$(Chapter 3)	0.1338	0.3398	0.0119
		$T_0^{(0)}$(Chapter 4)	0.1226	0.3271	0.0105
		$T_1^{(0)}$(Chapter 4)	0.1338	0.3398	0.0119
		Closed testing procedure	0.0725	0.2602	0.0059
	0	$T_0^{(0)}$(Chapter 3)	0.1164	0.3198	0.0051
		$T_1^{(0)}$(Chapter 3)	0.1273	0.3307	0.0059
		$T_0^{(0)}$(Chapter 4)	0.1179	0.3215	0.0051
		$T_1^{(0)}$(Chapter 4)	0.1273	0.3307	0.0059
		Closed testing procedure	0.0582	0.2305	0.0022
100	0.8	$T_0^{(0)}$(Chapter 3)	0.2995	0.6935	0.0238
		$T_1^{(0)}$(Chapter 3)	0.3067	0.6966	0.0252
		$T_0^{(0)}$(Chapter 4)	0.3027	0.6953	0.0245
		$T_1^{(0)}$(Chapter 4)	0.3067	0.6966	0.0252
		Closed testing procedure	0.2103	0.6288	0.0113
	0.4	$T_0^{(0)}$(Chapter 3)	0.2626	0.6867	0.0125
		$T_1^{(0)}$(Chapter 3)	0.2683	0.6915	0.0126
		$T_0^{(0)}$(Chapter 4)	0.2631	0.6873	0.0125
		$T_1^{(0)}$(Chapter 4)	0.2683	0.6915	0.0126
		Closed testing procedure	0.2009	0.6364	0.0041
	0	$T_0^{(0)}$(Chapter 3)	0.2702	0.6810	0.0061
		$T_1^{(0)}$(Chapter 3)	0.2750	0.6854	0.0062
		$T_0^{(0)}$(Chapter 4)	0.2715	0.6815	0.0061
		$T_1^{(0)}$(Chapter 4)	0.2750	0.6854	0.0062
		Closed testing procedure	0.1882	0.6168	0.0041
200	0.8	$T_0^{(0)}$(Chapter 3)	0.5292	0.9544	0.0206
		$T_1^{(0)}$(Chapter 3)	0.5337	0.9553	0.0211
		$T_0^{(0)}$(Chapter 4)	0.5306	0.9547	0.0209
		$T_1^{(0)}$(Chapter 4)	0.5337	0.9553	0.0211
		Closed testing procedure	0.4015	0.9466	0.0107
	0.4	$T_0^{(0)}$(Chapter 3)	0.4874	0.9604	0.0142
		$T_1^{(0)}$(Chapter 3)	0.4922	0.9608	0.0143
		$T_0^{(0)}$(Chapter 4)	0.4880	0.9604	0.0141
		$T_1^{(0)}$(Chapter 4)	0.4922	0.9608	0.0143
		Closed testing procedure	0.4011	0.9491	0.0088
	0	$T_0^{(0)}$(Chapter 3)	0.5168	0.9646	0.0089
		$T_1^{(0)}$(Chapter 3)	0.5197	0.9651	0.0089
		$T_0^{(0)}$(Chapter 4)	0.5171	0.9646	0.0089
		$T_1^{(0)}$(Chapter 4)	0.5197	0.9651	0.0089
		Closed testing procedure	0.3988	0.9491	0.0052

Acknowledgments

This thesis includes the content of the article "A testing procedure in clinical trials with multiple binary endpoints" written by the author. This is an Accepted Manuscript of an article published by Taylor & Francis in *Communications in Statistics - Theory and Methods* on 30 Mar 2021, available online: https://www.tandfonline.com/doi/abs/10.1080/03610926.2021.1912766.

I wish to acknowledge all those who supported me to complete this doctoral thesis. Firstly and above all, I would like to show my greatest appreciation to Associate Professor Kouji Yamamoto for the continuous support of the thesis, for his motivation, and immense knowledge. He was very kind in carefully reading the thesis and giving many useful suggestions to improve the thesis. His guidance helped me in all the time of research.

I would like to express my appreciation to Professor Kouji Tahata, Professor Shigeo Akashi, Professor Nobuko Miyamoto, Professor Tomomichi Suzuki and Professor Takashi Sozu, who were my doctoral committee members, for their fruitful discussions and helpful comments.

I also owe my deepest gratitude to Professor Kouji Tahata of Tokyo University of Science. He steered me in the right the direction whenever he thought I needed it. Without his guidance and persistent help this thesis would not have been possible. In addition, I would like to acknowledge sincere thanks to Professor Sadao Tomizawa of Meisei University, who gave valuable advice and supports every time. I am also deeply grateful to Gifu University Hospital for providing me with research opportunities, and for their warm encouragements.

Last but not least, I would like to express my deepest gratitude to my wife Yumi and my daughter Akari for their constant support and encouragement through the process of researching the thesis.

Thank you.

Gifu University Hospital March, 2023

Takuma Ishihara

References

Agresti, A. (1984). *Analysis of Ordinal Categorical Data*. Wiley, New York.

Agresti, A. (2013). *Categorical Data Analysis*, 3rd edition. Wiley, Hoboken.

Bahadur, R. R. (1961). A representation of the joint distribution of responses to n dichotomous items. *In Studies in Item Analysis and Prediction, Vol. VI, Stanford Mathematical Studies in the Social Sciences,* Solomon H (ed.). Stanford University Press: Stanford, CA, 158–168.

Beh, E. J., Simonetti, B., and D'Ambra, L. (2007). Partitioning a non-symmetric measure of association for three-way contingency tables. *Journal of Multivariate Analysis*, **98**, 1391–1411.

Berger, R. L. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics,* **24**, 295–300.

Bhatti, M. I. and Kim, J. H. (2021). Towards a New Paradigm for Statistical Evidence in the Use of p-Value. *Econometrics*, **9(1)**, 2.

Bishop, Y. M. M., Fienberg, S. E., and Holl, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*; The MIT Press, Cambridge.

Bonferroni, C. E. (1936). *Teoria statistica delle classi e calcolo delle probabilit?*, Firenze.

Bowker, A. H. (1948). A test for symmetry in contingency tables. *Journal of the American Statistical Association*, **43**, 572–574.

Dale, J. R. (1986). Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics,* **42**, 909–917.

Farrington, C. P. and Manning, G. (1990). Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Statistics in Medicine*, **9**, 1447–1454.

Fernandes, L. H. S. and Araújo, F. H. A. (2020). Taxonomy of commodities assets via the complexity-entropy causality plane. *Chaos Solitons Fractals*, **137**, 109909.

Glimm, E., Srivastava, M., and Lauter, J. (2002). Multivariate tests of normal mean vectors with restricted alternatives. *Communications in Statistics,* **B 31**, 589–604.

Goldstein, J. N., Refaai, M. A., Jr. Milling, T. J., Lewis, B., Goldberg-Alberts, R., Hug, B. A., and Sarode, R. (2015). Four-factor prothrombin complex concentrate versus plasma for rapid vitamin K antagonist reversal in patients needing urgent surgical or invasive interventions: a phase 3b, open-label, non-inferiority, randomised trial. *The Lancet*, **385(9982)**, 2077–2087.

International Conference on Harmonization (ICH) of Technical Requirements for Regulations of Pharmaceuticals for Human use. (1998). *ICH Tripartite Guideline E-9 Documents*.

Ishihara, T. and Yamamoto, K. (2021). A testing procedure in clinical trials with multiple binary endpoints. *Communications in Statistics - Theory and Methods*, Online Publication.

le Cessie, S. and van Houwelingen, J. C. (1994). Logistic regression for correlated binary data. *Applied Statistics,* **43**, 95–108.

Lombardo, R. (2011). Three-way association measure decompositions: The Delta index. *Journal of Statistical Planning and Inference*, **141**, 1789–1799.

Marcus, R., Peritz, E., and Gabriel, K. R. (1976). On Closed Testing Procedures with Special Reference to Ordered Analysis of Variance. *Biometrika*, **63**, 655–660.

McLeod, C., Norman, R., Litton, E., Saville, B. R., Webb, S., and Snelling, T. L. (2019). Choosing primary endpoints for clinical trials of health care interventions. *Contemporary Clinical Trials Communications*, **16**, 100486.

McMenamin, M. E., Barrett, J. K., Berglind, A., and Wason, J. M. S. (2022). Sample size estimation using a latent variable model for mixed outcome co-primary, multiple primary and composite endpoints. *Statistics in Medicine*, **41(13)**, 2303–2316.

Nakazuru, Y., Sozu, T., Hamada, C., and Yoshimura, I. (2014). A new procedure of one-sided test in clinical trials with multiple endpoints. *Japanese Journal of Biometrics,* **35**, 17–35.

Offen, W., Chuang-Stein, C., Dmitrienko, A., Littman, G., Maca, J., Meyerson, L., Muirhead, R., Stryszak, P., Boddy, A., Chen, K., Copley-Merriman, K., Dere, W., Givens, S., Hall, D., Henry, D., Jackson, J. D., Krishen, A., Liu, T., Ryder, S., Sankoh, A. J., Wang, J., and Yeh, C. H. (2007). Multiple co-primary endpoints: Medical and statistical solutions. *Drug Information Journal*, **41**, 31–46.

Pearson, K. (1900). III. Mathematical contributions to the theory of evolution. VIII. On the inheritance of characters not capable of exact quantitative measurement. Part I. Introductory. Part II. On the inheritance of coat-colour in horses. Part III. On the inheritance of eye-colour in man. *Philosophical transactions of the Royal Society of London. A.*, **195**, 1–47.

Perlman, M. D. and Wu, L. (2004). A note on one-sided tests with multiple endpoints. *Biometrics*, **60**, 276–280.

Rauen, T., Eitner, F., Fitzner, C., Sommerer, C., Zeier, M., Otte, B., Panzer, U., Peters, H., Benck, U., Mertens, P. R., Kuhlmenn, U., Witzke, O., Gross, O., Vielhauer, V., Mann, J. F., Hilgers, R. D., Floege, J., and STOP-lgAN Investigators. (2015). Intensive Supportive Care plus Immunosuppression in IgA Nephropathy. *The New England Journal of Medicine*, **373(23)**, 2225–2236.

Reich, K., Langlay, R. G., Papp, K. A., Ortonne, J. P., Unnebrink, K., Kaul, M., and Valdes, J. M. (2011). A 52-week trial comparing briakinumab with methotrexate in patients with psoriasis. *The New England Journal of Medicine,* **365(17)**, 1586–1596.

Sankoh, A. J., Huque, M. F., Russell, H. K., and D'Agostino, R. B. (1999). Global two-group multiple endpoint adjustment methods applied to clinical trials. *Drug Information Journal*, **33**, 119–140.

Smolen, J. S., Burmester, G. R., Combe, B., Curtis, J. R., Hall, S., Haraoui, B., van,Vollenhoven, R., Cioffi, C., Ecoffet, C., Gervitz, L., Lonescu, L., Peterson, L., and Fleischmann, R. (2016). Head-to-head comparison of certolizumab pegol versus adalimumab in rheumatoid arthritis: 2-year efficacy and safety results from the randomised EXXELERATE study. *The Lancet*, **388(10061)**, 2763–2774.

Sozu, T., Sugimoto, T., and Hamasaki, T. (2010). Sample size determination in clinical trials with multiple co-primary binary endpoints. *Statistics in Medicine*, **29(21)**,

2169–2179.

Sozu, T., Sugimoto, T., and Hamasaki, T. (2011). Sample size determination in superiority clinical trials with multiple co-primary correlated endpoints. *Journal of Biopharmaceutical Statistics*, **21(4)**, 650–668.

Sozu, T., Sugimoto, T., and Hamasaki, T. (2012). Sample size determination in clinical trials with multiple co-primary endpoints including mixed continuous and binary variables. *Biometrical Journal*, **54(5)**, 716–729.

Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika*, **42**, 412–416.

Tamhane, A. C. and Logan, B. R. (2004). A superiority-equivalence approach to one-sided tests on multiple endpoints in clinical trials. *Biometrika*, **91**, 715–727.

Tomizawa, S. (1994). Two kinds of measures of departure from symmetry in square contingency tables having nominal categories. *Statistica Sinica*, **4**, 325–334.

Tomizawa, S., Miyamoto, N., and Ashihara, N. (2003). Measure of departure from marginal homogeneity for square contingency tables having ordered categories. *Behaviormetrika*, **30(2)**, 173–193.

Wei, Z. and Kim, D. (2017). Subcopula-based measure of asymmetric association for contingency tables. *Statistics in Medicine*, **36**, 3875–3894.

Wei, Z. and Kim, D. (2021). Measure of asymmetric association for ordinal contingency tables via the bilinear extension copula. *Statistics and Probability Letters*, **178**, 109183.

Wei, Z., Kim, D., and Conlon, E. M. (2022). A Bayesian approach to the analysis of asymmetric association for two-way contingency tables. *Computational Statistics*, **37**, 1311–1338.

West, L. J. and Hankin, R. K. S. (2008). Exact tests for two-way contingency tables with structural zeros. *Journal of Statistical Software*, **28**, 1–19.

Zhang, L., Lu, D., and Wang, X. (2021). The essential dependence for a group of random vectors. *Communications in Statistics - Theory and Methods*, **50**, 5836–5872.