8-2020

Marginal methods and software for clustered data with cluster- and group-size informativeness.

Mary Elizabeth Gregg
University of Louisville

MARGINAL METHODS AND SOFTWARE FOR CLUSTERED DATA WITH CLUSTER- AND GROUP-SIZE INFORMATIVENESS

By

Mary Elizabeth Gregg
B.A., Bennington College, 2009
M.S., University of Louisville, 2016

A Dissertation
Submitted to the Faculty of the
School of Public Health and Information Science of the
University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy
in Biostatistics

Department of Bioinformatics and Biostatistics
University of Louisville
Louisville, Kentucky

August 2020

MARGINAL METHODS AND SOFTWARE FOR CLUSTERED DATA WITH CLUSTER- AND GROUP-SIZE INFORMATIVENESS

By

Mary Elizabeth Gregg
B.A., Bennington College, 2009
M.S., University of Louisville, 2016

Dissertation approved on

July 14, 2020

by the following dissertation Committee:

Dissertation Chair
Doug Lorenz

Somnath Datta

Ryan Gill

Maiying Kong

KB Kulasekera

ACKNOWLEDGMENTS

I would like to express my gratitude and appreciation to my advisor, Dr. Doug Lorenz, for his support and guidance throughout this dissertation. This work would not have been possible without his mentorship. I also thank all the members of my committee for the time they've devoted to reviewing this work, and for their helpful suggestions which have improved its quality. I am additionally grateful to Dr. KB Kulasekera for his role in my recipience of a University Fellowship, which has allowed me to devote the last two years to this research. I would also like to recognize the University of Louisville Libraries and the interlibrary loan system for providing access to research materials throughout this process. This work was conducted in part using the resources of the University of Louisville's Research Computing group and the Cardinal Research Cluster.

ABSTRACT

MARGINAL METHODS AND SOFTWARE FOR CLUSTERED DATA WITH CLUSTER- AND GROUP-SIZE INFORMATIVENESS

Mary Elizabeth Gregg

July 14, 2020

Clustered data result when observations have some natural organizational association. In such data, cluster size is defined as the number of observations belonging to a cluster. A phenomenon termed informative cluster size (ICS) occurs when observation outcomes vary in a systematic way related to the cluster size. An additional form of informativeness, termed informative within-cluster group size (IWCGS), arises when the distribution of group-defining categorical covariates within clusters similarly carries information related to outcomes. Standard methods for the marginal analysis of clustered data can produce biased estimates and inference when data have informativeness. A reweighting methodology has been developed that is resistant to ICS and IWCGS bias, and this method has been used to establish clustered data analogs of classical hypothesis tests related to ranks and correlation. In this work, we extend the reweighting methodology to develop a versatile collection of marginal hypothesis tests related to proportions, means, and variances in clustered data that are analogous to classical forms. We evaluate the performance of these tests compared to other cluster-appropriate methods through simulation and show that only reweighted tests maintain appropriate size when data have informativeness. We construct reweighted tests of clustered categorical data using several variance estimators, and demonstrate that the method of variance estimation can have substantial effect on these tests. Additionally, we show that when testing simple hypotheses in data lacking informa-

tiveness, reweighted tests can outperform other standard cluster-appropriate methods both in terms of size and power. Combining our novel tests with the existing tests of ranks and correlations, we compile a comprehensive R software package that executes this collection of ICS/IWCGS-appropriate methods through a thoughtful and user-friendly design.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Clustered data are prevalent throughout biomedical research. Some common examples include dental studies and repeated longitudinal measurements on individuals. In such cases, individuals form the clusters and the repeated measurements or distinct teeth represent the observations. Clusters are not only formed through shared or repeated observations within individuals, but also arise through hierarchical structures such as members within family units or patients in hospitals. Here, family units and hospitals represent the clusters and individuals are the observations. Regardless of the clustering structure, intra-cluster observations are more likely to have homogeneous features compared to inter-cluster observations, due to shared genetic or environmental components. This potential correlation between cluster members clearly invalidates standard statistical methods for independent observations.

There are a variety of statistical methods available for the analysis of clustered data, depending on the structure of the data and the research question of interest. Models for clustered data can be grouped broadly into three categories: marginal, cluster-specific, and conditional. The interpretation of these models are generally distinct, though some overlap is possible, and Aerts et al. [1] provide a concise reference to these model families and relevant members. In brief, marginal model parameters are interpreted as a population average, while cluster-specific models include fixed or random effects parameters particular to each cluster and thus have a cluster-specific

interpretation. In conditional models, observations are modeled dependent on other inter- and intra-cluster observations. Marginal analysis of clustered data is perhaps the most prevalent and generalizable, and we restrict our attention to such methods in this document.

Generalized estimating equations (GEE) are among the most widely used methods for the marginal analysis of clustered data. GEE models account for the dependence of observations within clusters through specification of a working correlation matrix and use of a sandwich variance estimate. This method is robust, providing consistent estimates of marginal parameters even under misspecification of the correlation structure. GEEs uses a quasi-likelihood approach and avoid specification of the joint distribution between observations, instead only requiring the univariate distribution for each response. Implementation of GEE models for the marginal analysis of clustered data has become pervasive due to their ease of use and ability to model both discrete and continuous outcomes.

1.2 Informative cluster size

Marginal models such as GEEs provide inference on population-averaged effects, but in clustered data the population can be defined as either the observations within clusters or the clusters themselves. As discerned by Williamson et al. [53], observation-based inference considers associations for the typical member from the population of observations, whereas cluster-based inference considers the associations for a typical observation from a typical cluster. Pavlou [47] provides an excellent illustration of this distinction with an example contrasting two analyses of clinic consultations. In both cases, patients are the clusters and visits are the observations. If the interest lies in estimating the resource use of the clinic, an observation-based marginal analysis (such as estimating costs associated with a typical visit across all clinic visits) might be preferred. If the interest instead lies in the marginal association of patient

health, a cluster-based analysis (typical outcome for a typical patient) would be more pertinent. In the case that the number of observations in a cluster, defined as the cluster size, is fixed or unrelated to the outcome being measured, interpretations for these two models are generally in correspondence. However, when this assumption does not hold, the equivalency between the observation- and cluster-based marginal models is not retained.

Informative cluster size (ICS), also referred to as nonignorable cluster size, is a phenomenon that occurs when the cluster size is a random variable that varies in a systematic way carrying information relating to the response measurement. It can be formally defined to occur when the distribution of the response variable conditioned on the cluster size differs from that of the unconditional distribution. The potential for ICS has been acknowledged in a variety of biostatistical settings, with examples relating to dental diseases, reproductive toxicology, pregnancy studies, and longitudinal treatments all being commonly referenced in the literature. Nevalainen et al. [43] differentiate three methods through which ICS can occur:

1. *Cluster size influences the outcome response.*

   Nevalinen et al. [43] illustrate this scenario with an example from Dunson et al. [13], which measures birth weight in mice pups. The reduced resources and decreased space in larger litters could result in lower birth weights, causing a negative association between outcome and cluster size.

2. *The outcome influences the cluster size.*

   This method of informativeness can be seen in the longitudinal rehabilitation data analyzed by Lorenz et al. [39]. In this study, functional ability was measured in patients with spinal cord injuries enrolled in a rehabilitation program over a series of sessions. Individuals with lower functional ability tend to require a larger number of rehabilitation sessions before disenrollment.

3. *A latent variable influences both cluster size and outcome.*

   The dental studies from Hoffman et al. [30] and Williamson et al. [53], among others, illustrate this third type of informativeness. Here, factors such as oral hygiene affect both the number of teeth an individual possesses and the disease status of the teeth.

1.3 Marginal analysis of clustered data with ICS

Regardless of the underlying mechanism of informativeness, GEE models can be biased in the presence of ICS. A GEE model using an independence working correlation provides observation-based inference and gives equal weight to each observation. If clusters are the unit of interest, marginal parameters from this GEE model may be biased in favor of larger clusters. For illustration, consider the dental study presented in Williamson et al. [53], in which periodontal disease status is measured in each tooth from a sample of individuals. The interest is in estimating the relationship between explanatory variables and disease status of a tooth, but factors related to disease status also affect the number of teeth present in an individual. Individuals with poor oral hygiene are likely to have fewer teeth and worse periodontal health compared to individuals with a higher standard of oral care, thus cluster size is informative. A standard GEE model will accurately estimate the associations between the variables and periodontal disease for the average tooth from the population of all teeth, but will underestimate those associations for the typical tooth for the average person, as individuals with healthier teeth tend to contribute more observations.

1.3.1 Within-cluster resampling

Hoffman et al. [30] addressed the issue of ICS by introducing within-cluster resampling (WCR), a Monte Carlo method that yields unbiased cluster-based marginal estimators under ICS. The WCR process involves forming a pseudo data set by sam-

pling a single observation at random from each cluster. Regular statistical methods can be applied to this data set, as it is a collection of independent observations. An estimator of a marginal parameter calculated from this pseudo data set is consistent for the true marginal parameter, but is unduly random and only uses a fraction of the available data. Therefore, the process is repeated many times and the WCR estimator is defined as the average of the resampled estimators. The WCR method accounts for any informativeness of cluster size by giving equal weight to each cluster through selection of a single observation, preventing over-representation of larger clusters. Hoffman et al. show that WCR estimators are asymptotically normal under mild conditions and give an expression for a consistent variance estimator, allowing inference in the usual manner. While WCR was introduced in the context of generalized linear models, the process can similarly be applied to other methods of parameter estimation.

1.3.2 Reweighted estimating equations

Within-cluster resampling provides an intuitive method for the estimation of marginal parameters, lending natural connotation to the cluster-based interpretation of a "typical observation from a typical cluster". However, WCR is computationally intensive and the estimates it produces are dependent on the resampling realizations. As an alternative to WCR, Williamson et al. [53] introduced cluster-weighted generalized estimating equations (CWGEE), in which standard estimating equations are reweighted by the inverse of the cluster size. Williamson et al. noted that the WCR estimate is an average of a large number of resampled estimates, and will converge to its expected value with respect to the sampling distribution, conditioned on the entire observed data. Rather than estimating this quantity by averaging a large number of replicates, the analytic average can be directly calculated by applying an expectation calculation to a single resampled estimator conditioned on the original data. The

uniform resampling process of WCR leads to an inverse cluster size weight being applied to the estimating equations. Williamson et al. [53] showed the asymptotic equivalence of CWGEE and WCR estimators, and suggested the use of a sandwich estimator for the variance-covariance matrix.

The reweighting approach provides a closed-form estimator of model parameters, removing the inherent randomness of WCR as well as the associated computational expense. Additionally, it was shown through simulation studies that CWGEE methods have less bias than WCR for small samples. This cluster-weighting method has subsequently been applied in the marginal analysis of correlated failure times [8], clustered longitudinal data [52], survival data [54], and ordinal longitudinal data [41].

1.4 Informative within-cluster group size

An additional type of informativeness can occur when the distribution of covariates in a cluster is related to the outcome of interest. In the case of categorical covariates that define groups of observations distinct from the clusters, this is termed informative within-cluster group size (IWCGS). For illustration, consider again a dental study in which the interest is in comparing periodontal disease status of molars and non-molars. IWCGS could occur if factors associated with disease status disproportionately affect the two groups of teeth. In this hypothetical example, poor oral hygiene could affect periodontal disease status in addition to causing attachment loss at a higher proportion in molars compared to non-molars. This would result in individuals with poor oral care tending towards a higher severity of periodontal diseases and having fewer molars compared to individuals with a higher standard of oral care. This secondary type of informativeness can occur independently or alongside ICS, and methods that correct for ICS are susceptible to bias from IWCGS [16, 31, 24]. In the resampling scheme that forms the foundation of the reweighting principal, one observation is selected at random from each cluster. In the hypothetical scenario

above, clusters with fewer molars tend to have worse outcomes than clusters with a higher proportion of molars. The WCR process would select one observation from each cluster, not accounting for the discrepancy in selection probability between the two groups that is also associated with the outcome. This results in molars being disproportionately selected from healthier clusters, leading to a potentially biased estimate of the marginal effect of periodontal disease between molars and non-molars. This bias would likewise be reflected in the estimates obtained from a CWGEE model.

Addressing this issue, Huang and Leroux [31] extended the concept of CWGEE to include cluster-level groups and/or covariates, developing what they term doubly-weighted GEE (DWGEE) that produce estimators invariant to IWCGS. Like cluster reweighting, this method is grounded in a WCR process. The resampling that leads to this secondary reweighting is a two-step process – for each cluster, a group is first selected with equal probability, then an observation from the cluster belonging to the selected group is randomly chosen. As before, regular statistical methods can be applied to this data set. Rather than repeatedly resample and average the estimators, an analytic average can be calculated from a single resampling. The modification in the resampling process leads to estimating equations weighted by the inverse of the intra-cluster group size rather than cluster size.

1.5 Reweighted analogs of classical tests

WCR and the subsequent CWGEE and DWGEE methods initially addressed the issue of informativeness through a model-based approach, and many authors have continued in this vein [8, 33, 41, 52, 54]. However, the reweighting methodology has also been used to derive clustered data analogs of well-known classical statistical tests. Datta and Satten proposed signed-rank [10] and rank-sum [11] tests for clustered data under ICS, and Dutta and Datta extended the rank-sum approach to account for IWCGS [16]. Parametric and non-parametric correlation estimators

for both paired and unpaired data have been proposed, providing clustered versions of the Pearson, Spearman, Kendall, and Phi coefficients [39, 40]. More recently, a clustered log rank test adjusting for informative cluster and group size has also been introduced [24]. Nevalainen et al. [43] have formalized the construction of such test statistics through consideration of statistical functionals and conditional expectation calculation of resampled statistics.

While the estimators and tests above could conceivably be obtained from model-based approaches such as CWGEE, the simplicity offered by non-model-based inferential methods can be advantageous. CWGEE and DWGEE avoid the Monte Carlo resampling of WCR but still require multi-stage computation updating of model parameters until convergence, which can be problematic in certain circumstances [41]. Parametric modeling such as that discussed by Nevalainen [43] and implemented by Neuhaus and McCulloch [42] and Zhang et al. [57] depend on specification of the cluster size distribution and the method of informativeness, and can be computationally burdensome. They are subject to bias from model misspecification [7], and moreover do not necessarily retain a marginal interpretation. In many situations these matters might be of minor concern or necessary for the desired analysis. However, if the research question is simple in nature, as in many preliminary or exploratory analysis, these modeling methods are disproportionately complex and a more straightforward method might be desired. As extensive analyses often evolve from simple hypotheses, the addition of these fundamental marginal tests to the cluster-weighted repertoire is advantageous. These cluster-weighted analogs of classical tests make ideal companions to intricate models, avoiding restrictive assumptions and producing interpretable results that can steer the direction for more extensive methods.

1.6 Objective and structure of the dissertation

The objective of this dissertation is the development of a comprehensive collection of reweighted hypothesis tests for clustered data with potential ICS or IWCGS. This collection integrates and expands upon the limited existing reweighted tests published by other authors. Complimentary to the existing reweighted tests, the novel tests proposed in this work parallel frequently-implemented standard statistical tests. We develop an R software package that executes this collection of tests, modeling the look and feel of the incoporated functions after those functions native to R that perform the analogous classical tests. This single platform of standardized functions administers access to these tests through a user-friendly environment. These methods and software provide researchers the means to perform practical hypothesis tests on clustered data while accounting for informativeness.

This work is organized as follows. Chapter 2 reviews the reweighting methodology and its origin in resampling. In this chapter, we introduce notation that will be used and expanded upon through this document, and discuss how reweighting is related to the structure of the observed data. In Chapter 3, we develop estimators and tests for common categorical data scenarios reweighted to correct for ICS. We focus much of our attention on the performance of these tests under various variance estimation methods. The work in this chapter has been published in manuscript form and is included here with minor edits made for the continuity of this document. Chapter 4 contains reweighted tests for quantitative data. We develop novel tests for hypotheses related to group means and variances, reweighted to correct for IWCGS. Additionally, this chapter summarizes some previously published reweighted tests by other authors, which are included in the comprehensive R package. We discuss the R package in Chapter 5, detailing the intentional resemblance between the functions implimenting the reweighted tests and the endemic R functions that perform their classical analogs. We illustrate the application of each function through examples using a simulated

data set. In Chapter 6, we summarize the work of this document, examine explicit and general limitations of the reweighting methodology, and discuss areas related to informativeness open to future research.

CHAPTER 2

BACKGROUND

2.1 Introduction

The reweighting methods discussed in Chapter 1 were developed to estimate marginal parameters for data with cluster- or group-size informativeness, and this reweighting forms the foundation for the collection of marginal tests in this document. In this chapter, we provide details on the reweighting methodology and its origins in resampling. We begin by establishing some general notation that will be expanded upon throughout subsequent chapters. For simplicity, this notation is presented in the context of quantitative data, though the methods remain generally unchanged for categorical data. We establish reweighting in the context of a marginal parameter correcting for ICS, then detail how the method is adapted to correct for IWCGS. The link between resampling and reweighting is illustrated through the derivation of reweighted marginal means in both the cluster and group informativeness scenarios.

2.2 Notation

Let X_{ij} denote observation j from cluster i. Cluster i contains n_i observations, defined as the cluster size, where $n_i > 0$. The data from cluster i is the set $\boldsymbol{V}_i = \{n_i, X_{ij}\}, i = 1, \ldots, M; j = 1, \ldots, n_i$, and the collection of all observed data is $\boldsymbol{V} = \{\boldsymbol{V}_1, \ldots, \boldsymbol{V}_M\}$. Clusters are assumed to be independent, while observations within cluster are potentially dependent. Cluster size is defined as non-informative

if $P\left(X_{ij} \leq x | n_i = n\right) = P\left(X_{ij} \leq x\right)$, $n = 1, 2, \ldots; j = 1, \ldots, n$; otherwise, it is informative [43].

When observations within clusters belong to one of K distinct groups, let $G_{ij} = k$ represent that observation j from cluster i belongs to the kth group, $k = 1, \ldots, K$. Cluster i has $n_i^{(k)}$ observations in group k, and $n_i = \sum_{k=1}^{K} n_i^{(k)}$. The data from cluster i is now the set $\boldsymbol{V}_i = \{n_i, (X_{ij}, G_{ij})\}$, with observations belonging to group k denoted as the set $\{X_{i1}^{(k)}, \ldots, X_{in_i^{(k)}}^{(k)}\}$. For simplicity, at times we use j to index observations belonging to group k within cluster i, i.e. $j = 1, \ldots, n_i^{(k)}$, in addition to the previous indexing of all observations within a cluster ($j = 1, \ldots, n_i$). In most cases, the indexing of j should be circumstantially evident, such as through the upper bound of a summation. In the event distinction is necessary, we defer to $j^{'}$ to index observations in groups.

When the distribution of X is associated with the probability of group membership, we refer to such data as having informative within-cluster group size. Other authors have referred to this as informative covariate structure [47], sub-cluster covariate informativeness [40], and informative intra-cluster group size [16]. Formally, group size is non-informative when $P\left(X_{ij} \leq x | G_{ij} = k\right) = P\left(X_{ij} \leq x\right)$, and otherwise informative.

When $n_i^{(k)} > 0$ for all i, k, we refer to such data as having complete group structure; i.e., all values of G are observed in all clusters. In practice, data may be collected where not all K groups are observed across all the M clusters. That is, $n_i^{(k)} = 0$ for at least one i, k. We term clusters where $n_i^{(k)} = 0$ for at least one k to be "incomplete clusters", and refer to data containing incomplete clusters as having incomplete group structure. Let $K_i^c = \sum_{k=1}^{K} I[n_i^{(k)} > 0]$ denote the number of distinct groups that contain observations in cluster i.

Let θ represent a marginal parameter for the population of clusters. We are interested in estimating θ and testing hypotheses of the form $H_0 : \theta = \theta_0$, or alternatively

$H_0 : h(\theta) = h(\theta_0)$.

2.3 Within-cluster resampling

The within-cluster resampling algorithm of Hoffman et al. [30] accounts for potential informativeness of cluster size by forming pseudo data sets through resampling of the clusters. This process is in accordance with the marginal analysis of interest, that of a "typical observation from a typical cluster", and is performed as follows. Let X_i^* denote an observation selected at random from cluster i. Resampling across all clusters produces the data set of independent observations $\boldsymbol{X}^* = (X_1^*, \ldots, X_M^*)$. The parameter of interest is then estimated from this resampled data set in the usual manner, $\hat{\theta}^* = g(\boldsymbol{X}^*)$. The WCR process is repeated Q times, where Q is a large number, and the overall WCR estimate is defined as the average of the resampled estimates,

$$\hat{\theta}_{WCR} = \frac{1}{Q} \sum_{q=1}^{Q} \hat{\theta}_q^*,$$

with variance estimated by

$$\hat{\text{var}} \left(\hat{\theta}_{WCR} \right) = \frac{1}{Q} \sum_{q=1}^{Q} \hat{\text{var}}(\hat{\theta}_q^*) - \frac{1}{Q} \sum_{q=1}^{Q} \left(\hat{\theta}_q^* - \hat{\theta}_{WCR} \right)^2$$

Hoffman et al. established the asymptotic normality and consistency of the WCR estimate, and Wald-type tests of H_0 can be constructed with $\hat{\theta}_{WCR}$ and $\hat{\text{var}} \left(\hat{\theta}_{WCR} \right)$ in the usual manner.

2.4 Cluster-weighting

As $M, Q \to \infty$, Williamson et al. [53] note that $\hat{\theta}_{WCR}$ converges to $\hat{\theta} = E \left[\hat{\theta}_q^* | \boldsymbol{V} \right]$ with respect to the sampling distribution. This marginalization is equivalent to averaging the resampled estimator across all realizations of the resampled data. As sampling is uniform across clusters, this expectation can easily be calculated and results in

weighting observations by the inverse of the cluster size. The asymptotic normality of such reweighted estimators has been established under mild regularity conditions by various authors [10, 11, 54].

The connection between resampling and reweighting can be illustrated in the context of a marginal mean. The estimator of interest calculated from a single resampled data set is $\hat{\theta}_q^* = \frac{1}{M}\sum_{i=1}^{M} X_i^*$. Applying the marginalization calculation results in

$$\begin{aligned} \hat{\theta} &= E\left[\hat{\theta}_q^* | \boldsymbol{V}\right] \\ &= \frac{1}{M}\sum_{i=1}^{M} E\left[X_i^* | \boldsymbol{V}\right] \\ &= \frac{1}{M}\sum_{i=1}^{M}\frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij} = \frac{1}{M}\sum_{i=1}^{M}\bar{X}_i \end{aligned} \tag{2.1}$$

The independence of clusters allows the expectation of the resampled estimate to be expressed as the average of the expectations. Conditioned on the observed data, the expectation of a resampled observation from a particular cluster is the cluster average, as the WCR process resamples observations from that cluster with equal probability. This expectation calculation is easily verified empirically and has previously been demonstrated by Lorenz et al. [39] and Nevalainen et al. [43].

We note that an estimate derived in the manner of (2.1) corresponds to an estimate from the marginal distribution

$$F(x) = E_{\boldsymbol{V}}\left\{\sum_{i=1}^{M}\frac{1}{n_i}\sum_{j=1}^{n_i} I\left(X_{ij} \leq x\right)\right\}$$

where $E_{\boldsymbol{V}}$ represents the expectation taken with respect to the distribution of $\boldsymbol{V}$.

2.5 Group-weighting

When observations within clusters belong to distinct groups, alternative weighting to correct for group informativeness may be desired. The link between the reweighting methodology and resampling results in group weights being contingent on the group

structure of the observed data. We illustrate this dependence by first discussing reweighting when data have complete group structure (members belonging to all groups are observed in all clusters), and then detail the subsequent weight adjustments that result when clusters have incomplete group structure.

2.5.1 Weighting under complete group structure

Huang and Leroux [31] first extended WCR methods to correct for IWCGS. They proposed modifying the resampling process at the foundation of reweighting into a two-step procedure that marginalizes the group distributions. In two-step resampling, G_i^* is first selected with uniform probability from the levels of G. Then, conditioned on $G_i^* = k$, X_i^* is sampled from the set $\{X_{i1}^{(k)}, \ldots, X_{in_i^{(k)}}^{(k)}\}$. As in the original WCR scheme, this process is repeated for all clusters, resulting in the resampled data $(\boldsymbol{X}^*, \boldsymbol{G}^*) = \{(X_1^*, G_1^*), \ldots, (X_M^*, G_M^*)\}$, and the parameter of interest calculated from this resampled data. Applying the marginalization principal to data resampled in this manner results in observations weighted by the inverse of the within-cluster group size.

For illustration, consider estimating the mean of observations belonging to one of two distinct groups from data with complete group structure. Let $\theta^{(1)}$ represent the marginal mean from the population of observations that belong to group 1. The estimate of $\theta^{(1)}$ calculated on the resampled data is $\hat{\theta}^{(1)*} = \frac{1}{n^{(1)*}} \sum_{i=1}^{M} X_i^* I[G_i^* = 1]$, where $n^{(1)*} = \sum_{i=1}^{M} I[G_i^* = 1]$ represents the number times group 1 is randomly selected in the realized resampling. Marginalizing $\hat{\theta}^{(1)*}$ across all possible resamplings of $(\boldsymbol{X}^*, \boldsymbol{G}^*)$ is the calculation

$$\hat{\theta}^{(1)} = E\left[\frac{1}{n^{(1)*}} \sum_{i=1}^{M} X_i^* I[G_i^* = 1] \middle| \boldsymbol{V}\right] = \sum_{i=1}^{M} E\left[\frac{1}{n^{(1)*}} X_i^* I[G_i^* = 1] \middle| \boldsymbol{V}\right] \quad (2.2)$$

The randomly-selected group, G_i^*, is an artificial random variable determined by the resampling process. It is thus independent of the observed data values, allowing the

interior expectation to be calculated

$$
\begin{aligned}
E\left[\frac{1}{n^{(1)*}} X_i^* I[G_i^* = 1] \mid \boldsymbol{V}\right] &= E\left[\frac{1}{n^{(1)*}}\right] E\left[X_i^* I[G_i^* = 1] | \boldsymbol{V}\right] \\
&= E\left[\frac{1}{n^{(1)*}}\right] E[X_i^* | \boldsymbol{V}] P[G_i^* = 1] \\
&= E\left[\frac{1}{n^{(1)*}}\right] \frac{1}{2n_i^{(1)}} \sum_{j=1}^{n_i^{(1)}} X_i^{(1)} = E\left[\frac{1}{n^{(1)*}}\right] \frac{1}{2} \bar{X}_i^{(1)}
\end{aligned}
$$

The expectation of $\frac{1}{n^{(1)*}}$ is a non-trivial calculation. However, we have seen through simulation it is well-approximated by $\frac{1}{E\left[\sum_{i=1}^{M} I[G_i^*=1]\right]} = \left(\frac{M}{2}\right)^{-1}$ as M increases. This results in

$$
\hat{\theta}^{(1)} = \frac{1}{M} \sum_{i=1}^{M} \bar{X}_i^{(1)}
$$

Heuristically, it is easy to see that the estimate of the group 1 mean marginalized across all possible resamplings is simply the within-cluster group 1 averages averaged across all clusters. As before, independence of clusters justifies the average of the expected contribution from each cluster. In the resampling process, group 1 is selected from a cluster with probability $\frac{1}{2}$ and X_i^* is selected uniformly from the set $\{X_{i1}^{(1)}, \ldots, X_{in_i^{(1)}}^{(1)}\}$. In the marginalization process, observations initially receive a $\frac{1}{2}$ weight corresponding to the group selection probability; however, this additional weight cancels out since all clusters contribute equally and the expected number of resampled group 1 observations is $\frac{M}{2}$.

Formally, marginal estimates calculated from data with complete group structure can be defined as functionals of the distribution

$$
F(x|k) = E_{\boldsymbol{V}} \left\{ \sum_{i=1}^{M} \frac{1}{n_i^{(k)}} \sum_{j=1}^{n_i} I\left(X_{ij} \leq x, G_{ij} = k\right) \right\} \tag{2.3}
$$

2.5.2 Weighting under incomplete group structure

When data have incomplete group structure, the weighting of (2.3) needs to be modified to account for incomplete clusters. In the two-step resampling process, a group

is selected uniformly from the number of groups available in the given cluster. If a particular group is not observed within a cluster, that group has a selection probability of 0. Similarly, if a cluster only contains observations belonging to a single group, that group is selected with probability 1. This results in observations from clusters being weighted not just by their respective group size, but additionally weighted by the inverse of the number of available groups within the cluster. In contrast to data with complete group structure, these "group selection" probability weights are not equal across clusters and no longer cancel out in the marginalization process.

For illustration, consider the derivation of $\hat{\theta}^{(1)}$ in the scenario of incomplete clusters.

$$\begin{aligned}
\hat{\theta}^{(1)} &= E\left[\hat{\theta}^{(1)*}|\boldsymbol{V}\right] = E\left[\frac{1}{n^{(1)*}}\sum_{i=1}^{M} X_i^* I[G_i^*=1] \mid \boldsymbol{V}\right] \\
&= E\left[\frac{1}{n^{(1)*}}\right]\sum_{i=1}^{M} E[X_i^*|\boldsymbol{V}]P[G_i^*=1] \\
&= \frac{1}{\frac{1}{2}M^{(c)}+M^{(1)}}\sum_{i=1}^{M}\left\{\frac{1}{2}\bar{X}_i^{(1)} I[n_i^{(1)}>0, n_i^{(2)}>0] + \bar{X}_i^{(1)} I[n_i^{(1)}>0, n_i^{(2)}=0]\right\}
\end{aligned} \tag{2.4}$$

where $M^{(c)} = \sum_{i=1}^{M} I[n_i^{(1)}>0, n_i^{(2)}>0]$ and $M^{(1)} = \sum_{i=1}^{M} I[n_i^{(1)}>0, n_i^{(2)}=0]$. Here, $M^{(c)}$ represents the number of complete clusters, and $M^{(1)}$ denotes the number of incomplete clusters that only contain observations belonging to group 1. Note that a correponding quantity $M^{(2)}$ exists, and $M = M^{(c)} + M^{(1)} + M^{(2)}$.

In equation (2.4), the form of $\hat{\theta}^{(1)*}$ remains the same as in (2.2), but the result of the conditional expectation of this quantity is modified based on the adjusted group selection probabilities. That is, $P[G_i^*=1] = \frac{1}{2}$ for complete clusters, but this probability is 1 for incomplete clusters containing only members that belong to group 1, and similarly is 0 for incomplete clusters comprised of only group 2 observations. In the expectation calculation that corresponds to averaging over all possible resamplings, the group 1 averages within clusters are no longer averaged

equally across the M clusters, because the clusters do not contribute equally to the estimate in the resampling process. Instead, the expected group 1 values are averaged over the expected number of clusters that would contribute a group 1 observation. We note that for data with complete group structure, (i.e., $M^{(1)} = M^{(2)} = 0$), (2.4) simplifies to (2.2).

Recall that K_i^c denotes the number of observed groups within cluster i. For data with incomplete group structure, the marginal parameter θ can then be expressed as a functional of the distribution

$$F\left(x|k\right) = E_{\boldsymbol{V}}\left\{\sum_{i=1}^{M} w_{ij} \sum_{j=1}^{n_i} I\left(X_{ij} \leq x, G_{ij} = k\right)\right\} \tag{2.5}$$

where the weight w_{ij} is defined

$$w_{ij} = \begin{cases} \left(K_i^c n_i^{(k)}\right)^{-1}, & \text{if } n_i^{(k)} > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{2.6}$$

While the assignment of a weight corresponding to the group selection probability to observations might be intuitive, we stress that the overall marginalization of the distribution must also consider this weight. In the calculation of $\hat{\theta}^{(1)}$ above, this was evident in the distinction of $M^{(c)}$ and $M^{(1)}$. Even when all clusters are able to contribute to a group estimate, the overall divisor of the estimate is not M, but instead a function of the number of contributing clusters and their contribution probabilities.

2.5.3 A note on incomplete clusters

When observed data have incomplete group structure, care must be taken to accurately define and estimate marginal parameters. Incomplete clusters can arise when observations exist across all groups in the population of all clusters, but some groups were merely not observed for some of the collected clusters. Alternatively, incomplete clusters can also belong to a population in which some of the K groups do not exist. As noted by Seaman et al. [49], marginal parameters for these two population

may not coincide. This paper offers an excellent discussion on the nuances among these two populations and provides a discussion on appropriate reweighting for differing marginal inferences in the context of model-based estimators. For example, Seaman et al. point out the DWGEE2 model of Huang and Leroux [31] applies an expected weight to incomplete clusters. This becomes philosophically and mathematically problematic, as it results in the modeling of values which do not exist (e.g., cognitive function in dead people [49]). We note that parameters estimated by (2.5) are based on observed and not expected weights, and thus are appropriate for observations from either population. However, the importance of thoroughly considering the marginal inference of interest and suitability of methods to achieve that inference cannot be overstated.

CHAPTER 3

ESTIMATION AND TESTING FOR CATEGORICAL DATA[1]

3.1 Introduction

The reweighting methodology detailed in the previous chapter has been used to develop clustered data analogues of the well-known rank sum [10], signed rank [11], correlation [39, 40], and log rank tests [24]. This collection of tests notably excludes tests of categorical responses, analogous to well-known chi-square tests of proportions. Further, while the aforementioned tests for clustered data under ICS were developed using a common motivating principal, there are substantial differences in the variance estimation techniques implemented in these tests. These variance estimation techniques include those based on Hajek projections and empirical, sandwich, and jackknife forms. The performance of reweighted tests under different variance estimation methods has not been explored previously. This is of particular interest in tests of categorical data, as it is well-known that the performance of tests and confidence intervals can depend greatly on the method of construction [6, 22, 45]. Motivated by the absence of tests of clustered categorical data with ICS and evaluations of test performance under different variance estimation methods, we address both topics in this chapter.

[1]Reproduced in part with permission from Gregg, M., Datta, S. and Lorenz, D. (2020) "Variance estimation in tests of clustered categorical data with informative cluster size", Statistical Methods in Medical Research. doi: 10.1177/0962280220928572.

Using the marginalization principle described in Chapter 2, we develop test statistics for common categorical data scenarios appropriate for clustered data under potential ICS. We construct tests for marginal proportion, categorical proportions, independence of bivariate categorical variables, and marginal homogeneity. These tests mimic the classical one-sample proportion, chi square goodness of fit, chi square independence, and McNemar tests, and their construction is detailed in Section 3.2. Each of these tests are composed using a number of different variance estimation methods, and the performance of each are compared through a simulation study in Section 3.3. In Section 3.4, we apply the proposed methods to a data set of functional rehabilitation measurements from patients with spinal cord injuries, and Section 3.5 includes our concluding remarks.

3.2 Reweighted tests for categorical data

3.2.1 Binary univariate data – one-sample proportion tests

Retaining the notation established in Chapter 2, we observe n_i binary outcomes $X_{ij} = \{0, 1\}$ in cluster i, where X_{ij} takes value 1 if the observation is classified as a "success". We are interested in estimating the marginal probability of success, $p = P(X = 1)$, and in testing the null hypothesis $H_0 : p = p_0$ for some null proportion p_0. As discussed previously, this marginal parameter p can be defined as the probability of success for a typical observation from all observations, or a typical observation from a typical cluster. When cluster size is non-informative these two marginal probabilities are equivalent, but this equality does not hold under ICS [49, 53]. When clusters are the primary unit of interest and cluster size is informative, the latter is a more appropriate marginal analysis, and we develop estimators and tests accordingly.

The usual proportion estimate applied to a resampled data set is $\hat{p}^* = \frac{1}{M}\sum_{i=1}^{M} X_i^*$. Applying the marginalization principle of Section 2.5 to this estimate results in the

statistic

$$\begin{aligned}\hat{p} &= E\left[\hat{p}^* \middle| \{\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n\}\right] \\ &= E\left[\frac{1}{M}\sum_{i=1}^{M} I[X_i^* = 1] \middle| \{\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n\}\right] \\ &= \frac{1}{M}\sum_{i=1}^{M}\frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij} = \frac{1}{M}\sum_{i=1}^{M}\hat{p}_i\end{aligned}$$

where $\hat{p}_i$ represents the proportion of successes in cluster i. As this reweighted estimator is asymptotically normally distributed, we can test H_0 using a Wald-type test by comparing the standardized form $z = (\hat{p} - p_0)/\sqrt{(\hat{v}/M)}$ to appropriate percentiles of the standard normal distribution, where $\hat{v}$ is some estimate of the variance of $\hat{p}$. While a number of tests of marginal parameters for clustered data with ICS have been established using the asymptotic normality of cluster-weighted estimators [10, 11, 24, 39, 40], none have been evaluated under competing variance estimation techniques. To this end, we propose four methods of estimation for $\hat{v}$ including two novel methods that have not previously been considered in the cluster-weighted context.

Variance Estimation

One choice for $\hat{v}$ is the empirical variance of the within-cluster proportions, $\hat{v}_{emp} = \frac{1}{M-1}\sum_{i=1}^{M}(\hat{p}_i - p_0)^2$. Previous authors have estimated the variances of clustered-weighted estimators using the empirical variances of certain within-cluster averages in conjunction with appropriate delta method calculations [39].

Williamson et al. [53] suggested a sandwich variance estimator for their clustered data estimator, of the form

$$\widehat{v} = \widehat{l}(p)^{-1}\widehat{V}(p)\widehat{l}(p)^{-1}, \tag{3.1}$$

where

$$\widehat{l}\,(p) = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{n_i}\sum_{j=1}^{n_i}\frac{\delta U_{ij}(p)}{\delta p}, \tag{3.2}$$

$$\widehat{V}(p) = \frac{1}{M}\sum_{i=1}^{M}\left\{\frac{1}{n_i}\sum_{j=1}^{n_i}U_{ij}(p)\right\}\left\{\frac{1}{n_i}\sum_{j=1}^{n_i}U_{ij}(p)\right\}^T, \tag{3.3}$$

and U_{ij} is the estimating function for the parameter(s) of interest. This method has additionally been used for other cluster-weighted estimators [39]. A sandwich estimator can be implemented for the test of a marginal proportion using the scores $U_{ij} = \frac{X_{ij}}{p} - \frac{1-X_{ij}}{1-p}$. Let $\hat{v}_{\hat{p}}$ denote the sandwich variance estimator when the quantities $\widehat{l}(p)$ and $\widehat{V}(p)$ are evaluated at the estimated value of p.

Wald intervals and tests are known to perform poorly in estimating and testing proportions, providing poor interval coverage and failing to maintain test size near the boundaries of the parameter space, even for reasonably large sample sizes [6, 22, 36]. Agresti [2] has suggested that methods that construct variance estimates under a null hypothesis are closely related to score tests, and score tests are known to be resistant to some of the issues suffered by Wald tests of nominal data. In light of this relationship and in the context of testing H_0, an alternative approach can presume the null hypothesis by evaluating $\widehat{l}(p)$ and $\widehat{V}(p)$ at p_0. Let $\hat{v}_{p_0}$ represent the sandwich variance estimator evaluated at p_0. To our knowledge, this technique has never been considered in the cluster-weighted setting.

An additional variance estimator incorporates a method of moments calculation conducted presuming the null hypothesis. This form has previously been employed in tests of paired clustered binary data [14], but has not been used in the development of tests applying the reweighting methodology. In the present context, testing H_0 is equivalent to testing $p - p_0 = 0$, suggesting the variance estimate $var(\hat{p} - p_0) = \frac{1}{M}\sum_{i=1}^{M}\left[(\hat{p}_i - p_0) - (\hat{p} - p_0)\right]^2$. Under the null hypothesis, this simplifies to $var(\hat{p} - p_0) = \frac{1}{M}\sum_{i=1}^{M}\left[\hat{p}_i - p_0\right]^2$. Let $\hat{v}_{MM}$ represent this method of moments estimator evaluated under the null hypothesis.

In the next section, we present the results of simulation studies evaluating tests of $H_0 : p = p_0$ by constructing the test statistic z with the aforementioned four

variance estimators. We compare the performance of these tests to a modification of the Agresti-Coull interval for clustered survey data [12], a GEE intercept-only model with exchangeable correlation structure, and the naive proportion test assuming independent observations.

3.2.2 Categorical univariate data – goodness of fit

We now progress to nominal random variables with more than two categories. We modify our notation so that each observation is a K-dimensional vector, $\boldsymbol{x}_{ij} = \left(x_{ij}^{(1)}, \ldots, x_{ij}^{(K)}\right)^T$, where $x_{ij}^{(k)}$ is an indicator variable that observation j from cluster i belongs to category k, for $k = 1, \ldots, K$. The data from cluster i are $\boldsymbol{V}_i = \{n_i, x_{i1}, \ldots, x_{in_i}\}$, the observed within-cluster proportions are $\hat{\boldsymbol{p}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} \boldsymbol{x}_{ij} = \left(\hat{p}_i^{(1)}, \ldots, \hat{p}_i^{(K)}\right)^T$, and $\hat{p}_i^{(k)} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}^{(k)}$. A hypothesis of interest may be that the marginal proportions, after accounting for individual cluster differences and ICS, are equal to some pre-specified values, corresponding to the chi square goodness of fit test. Specifically, we will test $H_0 : \boldsymbol{p} = \boldsymbol{p}_0$, where $\boldsymbol{p} = \left(p^{(1)}, \ldots, p^{(K)}\right)^T$ are the marginal group proportions and $\boldsymbol{p}_0 = \left(p_0^{(1)}, \ldots, p_0^{(K)}\right)^T$ are the hypothesized values. Following the conditional expectation calculations detailed in Section 2.4, it is not difficult to see that the cluster-weighted estimate of the vector of marginal group proportions is $\hat{\boldsymbol{p}} = \left(\hat{p}^{(1)}, \ldots, \hat{p}^{(K)}\right)^T = \frac{1}{M}\sum_{i=1}^{M}\left(\hat{p}_i^{(1)}, \ldots, \hat{p}_i^{(K)}\right)^T = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{n_{ij}}\sum_{j=1}^{n_i}\left(x_{ij}^{(1)}, \ldots, x_{ij}^{(K)}\right)^T$. That is, the marginal estimator is simply the vector of within-cluster group proportions averaged across all clusters. This estimator is asymptotically normal, and we can calculate the Wald-type quadratic form:

$$X^2 = M\left(\hat{\boldsymbol{p}} - \boldsymbol{p}_0\right)^T \hat{\boldsymbol{\Sigma}}^{-1}\left(\hat{\boldsymbol{p}} - \boldsymbol{p}_0\right) \tag{3.4}$$

where $\hat{\boldsymbol{\Sigma}}$ is an estimate of the variance matrix for $\hat{\boldsymbol{p}}$. Under mild conditions and H_0, this statistic asymptotically follows a chi square distribution with $K - 1$ degrees of freedom.

The empirical, sandwich, and method of moments variance estimates derived above can all be employed to estimate $\hat{\boldsymbol{\Sigma}}$. Let $\hat{\boldsymbol{\Sigma}}_{emp}$ be the empirical variance covariance matrix of $\hat{\boldsymbol{p}}$, $\hat{\boldsymbol{\Sigma}}_{emp} = \frac{1}{M-1}\sum_{i=1}^{M}\left(\hat{\boldsymbol{p}}_i - \hat{\boldsymbol{p}}\right)\left(\hat{\boldsymbol{p}}_i - \hat{\boldsymbol{p}}\right)^T$. The sandwich form can be obtained from Equations (3.1)-(3.3), using multinomial scores $U_{ij}\left(\boldsymbol{p}\right) = \left(\frac{x_{ij}^{(1)}}{p^{(1)}}, \ldots, \frac{x_{ij}^{(K)}}{p^{(K)}}\right)^T$, and replacing p with the vector $\boldsymbol{p}$. We evaluate this sandwich form at both the cluster-weighted estimates $\hat{\boldsymbol{p}}$ and the hypothesized category proportions $\boldsymbol{p}_0$. Denote these estimates as $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{p}}}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{p}_0}$, respectively, for which we omit the details as they are straightforward. The method of moments estimator of $\hat{\boldsymbol{\Sigma}}$ is $var(\hat{\boldsymbol{p}} - \boldsymbol{p}_0) = \frac{1}{M}\sum_{i=1}^{M}\left[\left(\hat{\boldsymbol{p}}_i - \boldsymbol{p}_0\right) - \left(\hat{\boldsymbol{p}} - \boldsymbol{p}_0\right)\right]\left[\left(\hat{\boldsymbol{p}}_i - \boldsymbol{p}_0\right) - \left(\hat{\boldsymbol{p}} - \boldsymbol{p}_0\right)\right]^T$. Under the null hypothesis, $\hat{\boldsymbol{p}} - \boldsymbol{p}_0 = 0$ and the method of moments variance estimate simplifies to $\hat{\boldsymbol{\Sigma}}_{MM} = \frac{1}{M}\sum_{i=1}^{M}\left[\hat{\mathbf{p}}_i - \boldsymbol{p}_0\right]\left[\hat{\mathbf{p}}_i - \boldsymbol{p}_0\right]^T$. We compare the performance of statistic (3.4) under the variance estimates $\hat{\boldsymbol{\Sigma}}_{emp}$, $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{p}}}$, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{p}_0}$, and $\hat{\boldsymbol{\Sigma}}_{MM}$ in the simulation study in Section 3.3.

3.2.3 Bivariate categorical data – test of independence

We now extend to the case of two categorical variables, where the hypothesis of interest is their independence. Denote the two random variables X and Y, where X takes values from 1 to K and Y takes values from 1 to G. Observation j from cluster i is the bivariate pair (X_{ij}, Y_{ij}), where $j = 1, \ldots, n_i$ and $i = 1, \ldots, M$. Let $p^{(k,g)} = P\left(X = k, Y = g\right)$, corresponding to the cell probability for row k and column g of a two-way contingency table. Let $p^{(k,+)} = P\left(X = k\right)$ and $p^{(+,g)} = P(Y = g)$, defining the marginal row and column probabilities. In cluster i, the observed cell frequency in row k and column g is $n_i^{(k,g)} = \sum_{j=1}^{n_i} I\left[X_{ij} = k, Y_{ij} = g\right]$, and the associated marginal row and column frequencies are $n_i^{(k,+)} = \sum_{g=1}^{G} n_i^{(k,g)}$ and $n_i^{(+,g)} = \sum_{k=1}^{K} n_i^{(k,g)}$. The corresponding observed proportion for the cell in row k and column g in cluster i is $\hat{p}_i^{(k,g)} = \frac{n_i^{(k,g)}}{n_i}$, and we collect the observed cell proportions for cluster i in the vector $\hat{\boldsymbol{p}}_i = \left(\hat{p}_i^{(1,1)}, \ldots, \hat{p}_i^{(1,G)}, \hat{p}_i^{(2,1)}, \ldots, \hat{p}_i^{(K,G)}\right)^T$. The observed marginal proportions in cluster i are $\hat{p}_i^{(k,+)} = \frac{n_i^{(k,+)}}{n_i}$ and $\hat{p}_i^{(+,g)} = \frac{n_i^{(+,g)}}{n_i}$ for row k and column g, respectively.

The null hypothesis asserts that X and Y are independent, explicitly stated as $H_0 : p^{(k,g)} = p^{(k,+)}p^{(+,g)}$ for all k and g. Let $\boldsymbol{p}$ denote the vector of probabilities $p^{(k,g)}$ for all KG cells and let $\boldsymbol{e}$ denote the vector of all possible marginal proportion products, $\boldsymbol{e} = \left(p^{(k,+)}p^{(+,g)}\right)^T$ for $k = 1, \ldots, K, g = 1, \ldots, G$. Note that this vector defines the cell proportions of the two-way table under the null hypothesis. By the same conditional expectation calculations as above, the clustered-weighted estimates of the cell and marginal proportions are $\hat{p}^{(k,g)} = \frac{1}{M}\sum_{i=1}^{M} \hat{p}_i^{(k,g)}$, $\hat{p}^{(k,+)} = \frac{1}{M}\sum_{i=1}^{M} \hat{p}_i^{(k,+)}$, and $\hat{p}^{(+,g)} = \frac{1}{M}\sum_{i=1}^{M} \hat{p}_i^{(+,g)}$. We can then define the vectors $\hat{\boldsymbol{p}}$ and $\hat{\boldsymbol{e}}$ as estimates of the vectors $\boldsymbol{p}$ and $\boldsymbol{e}$. A reasonable statistic for measuring departures from the null hypothesis is then $\hat{\boldsymbol{d}} = \hat{\boldsymbol{p}} - \hat{\boldsymbol{e}}$, The vector $\hat{\boldsymbol{d}}$ can be shown to be asymptotically normal and under the null hypothesis, $\boldsymbol{d} = \boldsymbol{0}$. Thus, we can test the null hypothesis using the statistic $X^d = M\hat{\boldsymbol{d}}^T(\hat{\boldsymbol{\Sigma}}^d)^{-1}\hat{\boldsymbol{d}}$, where $\hat{\boldsymbol{\Sigma}}^d$ is some estimate of the variance of $\hat{\boldsymbol{d}}$. Under the null, this statistic asymptotically follows a chi square distribution with $(K-1)(G-1)$ degrees of freedom.

As in the univariate case, $\hat{\boldsymbol{\Sigma}}^d$ can be estimated using an empirical, sandwich, or method of moments estimator. Let $\hat{\boldsymbol{\Sigma}}^d_{emp}$ be the empirical variance-covariance matrix of $\hat{\boldsymbol{d}}$, $\hat{\boldsymbol{\Sigma}}^d_{emp} = \frac{1}{M-1}\sum_{i=1}^{M}(\hat{\boldsymbol{d}}_i - \hat{\boldsymbol{d}})(\hat{\boldsymbol{d}}_i - \hat{\boldsymbol{d}})^T$. The sandwich form can be derived in the same manner as the goodness of fit sandwich estimator, and likewise can be evaluated at the cluster-weighted estimate $\hat{\boldsymbol{p}}$ or be evaluated at the vector of null hypothesis cell proportions $\hat{\boldsymbol{e}}$. Let $\hat{\boldsymbol{\Sigma}}^d_{\hat{p}}$ and $\hat{\boldsymbol{\Sigma}}^d_{\hat{e}}$ denote these variance estimates, respectively. The method of moments estimator is

$$\hat{\boldsymbol{\Sigma}}^d_{MM} = \frac{1}{M}\sum_{i=1}^{M}\left[(\hat{\boldsymbol{p}}_i - \hat{\boldsymbol{e}}_i) - (\hat{\boldsymbol{p}} - \hat{\boldsymbol{e}})\right]\left[(\hat{\boldsymbol{p}}_i - \hat{\boldsymbol{e}}_i) - (\hat{\boldsymbol{p}} - \hat{\boldsymbol{e}})\right]^T$$

which reduces to $\hat{\boldsymbol{\Sigma}}^d_{MM} = \frac{1}{M}\sum_{i=1}^{M}\left[\hat{\boldsymbol{p}}_i - \hat{\boldsymbol{e}}_i\right]\left[\hat{\boldsymbol{p}}_i - \hat{\boldsymbol{e}}_i\right]^T$ under the null. The performances of the chi square statistic derived with each of these variance estimators are compared to a Cochran-Mantel-Haenszel test stratified by cluster in the Section 3.3.

3.2.4 Paired binary data – test of marginal homogeneity

For paired bivariate data, observation j from cluster i is (X_{ij}, Y_{ij}), where X and Y are binary variables with a value of 1 indicating success for the first and second measurement of the observation, respectively. As before, there are n_i observations from cluster i and M total clusters. Observations from cluster i can be summarized in a 2×2 table where the diagonal elements $n_i^{(1,1)} = \sum_{j=1}^{n_i} I\left[X_{ij} = 1, Y_{ij} = 1\right]$ and $n_i^{(0,0)} = \sum_{j=1}^{n_i} I\left[X_{ij} = 0, Y_{ij} = 0\right]$ are the frequencies of concordant successes and failures, respectively. The off-diagonal elements are $n_i^{(1,0)} = \sum_{j=1}^{n_i} I\left[X_{ij} = 1, Y_{ij} = 0\right]$ and $n_i^{(0,1)} = \sum_{j=1}^{n_i} I\left[X_{ij} = 0, Y_{ij} = 1\right]$. Define $p^{(1,1)} = P\left[X = 1, Y = 1\right]$ as the joint probability of success for X and Y across the population of clusters, and define $p^{(1,0)}$, $p^{(1,0)}$, and $p^{(0,0)}$ in a similar fashion.

We will test the hypothesis of marginal homogeneity, $H_0 : p^{(1,+)} = p^{(+,1)}$, where $p^{(1,+)}$ and $p^{(+,1)}$ are the marginal probability of success for the first and second measurements of the random variable. Since $p^{(1,+)}$=$p^{(1,1)} + p^{(1,0)}$ and $p^{(+,1)} = p^{(1,1)} + p^{(0,1)}$, the null hypothesis can be equivalently stated as $H_0 : p^{(1,0)} = p^{(0,1)}$. Again, the resampling and subsequent conditional expectation calculation can be applied to obtain the cluster-weighted estimates of $p^{(1,0)}$ and $p^{(0,1)}$:

$$\hat{p}^{(1,0)} = \frac{1}{M} \sum_{i=1}^{M} \hat{p}_i^{(1,0)}, \tag{3.5}$$

$$\hat{p}^{(0,1)} = \frac{1}{M} \sum_{i=1}^{M} \hat{p}_i^{(0,1)}.$$

We can test H_0 using the statistic

$$X^2 = \frac{\left(\hat{p}^{(1,0)} - \hat{p}^{(0,1)}\right)^2}{\text{var}\left(\hat{p}^{(1,0)} - \hat{p}^{(0,1)}\right)} \tag{3.6}$$

Durkalski et al. [14], by adopting sampling techniques proposed by Obuchowski [46], arrived at the same estimates of $p^{(1,0)}$ and $p^{(0,1)}$ presented in (3.5). Using the form of statistic (3.6), Durkalski et al. proposed testing the hypothesis of marginal

homogeneity by estimating the variance using the method of moments estimator evaluated under the null hypothesis: $\hat{\text{var}}\left(\hat{p}^{(1,0)} - \hat{p}^{(0,1)}\right) = \frac{1}{M^2}\sum_{i=1}^{M}\left[\hat{p}_i^{(1,0)} - \hat{p}_i^{(0,1)}\right]^2$. Alternatively, we can also employ an empirical variance estimate in the construction of test statistic (3.6) by

$$\hat{\text{var}}\left(\hat{p}^{(1,0)} - \hat{p}^{(0,1)}\right) = \frac{1}{M}\left[\hat{\text{var}}\left(\hat{p}^{(1,0)}\right) + \hat{\text{var}}\left(\hat{p}^{(0,1)}\right) - 2\hat{\text{cov}}\left(\hat{p}^{(1,0)}, \hat{p}^{(0,1)}\right)\right]$$

where

$$\hat{\text{var}}\left(\hat{p}^{(k,g)}\right) = \frac{1}{M-1}\sum_{i=1}^{M}\left(\hat{p}_i^{(k,g)} - \hat{p}^{(k,g)}\right)^2$$

and

$$\hat{\text{cov}}\left(\hat{p}^{(k,g)}, \hat{p}^{(k',g')}\right) = \frac{1}{M-1}\sum_{i=1}^{M}\left(\hat{p}_i^{(k,g)} - \hat{p}^{(k,g)}\right)\left(\hat{p}_i^{(k',g')} - \hat{p}^{(k',g')}\right).$$

Additional tests of marginal homogeneity for clustered data have been proposed by Eliasziw and Donner [18], Obuchowski [46], and Yang et al. [56]. We compare the performance of the cluster-weighted test constructed with both the method of moments and empirical variance estimator to these additional methods through simulations in the following section.

3.3 Simulation Study

We evaluated the performance of our cluster-weighted tests under the several proposed variance estimators via simulation studies. To summarize, the reweighted proportion, goodness of fit, and test of independence analogs were constructed using the following variance estimation methods: 1. Empirical (CW-Emp.), 2. Sandwich form evaluated at the estimate(s) (CW-$\text{SW}_{\hat{p}}$). 3. Sandwich form evaluated under the null hypothesis (CW-SW_0). 4. Method of moments evaluated under the null hypothesis (CW-MM). The reweighted test of marginal homogeneity was constructed using variance estimation methods 1 and 4. For each of the testing scenarios, we compared the empirical size of our tests to their analogues for independent data, in addition to any well-known alternatives for clustered data appropriate to the respective test. Specifically,

we compared our cluster-weighted proportion test to a binomial GEE model and the Dean-Pagano [12] modification of the Agresti-Coull [3] method, our cluster-weighted test of independence to the Cochran-Maentel-Haenzel test stratified by cluster, and our cluster-weighted test of marginal homogeneity to the methods of Eliasziw and Donner [18], Obuchowski [46], and Yang et al. [56]. We calculate the empirical size of each test as the proportion of rejections of the null hypothesis over 10 000 Monte Carlo iterations at a nominal level of .05. To evaluate the effect of sample size, all simulations were run for 50, 100, and 200 clusters.

We simulated clustered binary data featuring ICS by first generating a random effect for each cluster, u_i, from the standard normal distribution. We then simulated cluster sizes n_i from Poisson$(10 + 10 * I[u_i > 0]) + 1$ distribution, so that cluster sizes and random effects were positively associated. We then simulated n_i standard normal random variables e_{ij} within each cluster independent of the per-cluster random effects. Under these conditions, the random variable $u_i + e_{ij} + \delta$ follows the $N(\delta, 2)$ distribution. We dichotomized this random variable as $X_{ij} = I[u_i + e_{ij} > c]$, and selected c to satisfy $P(X = 1) = p_0$ for $p_0 = 0.1$, 0.25, and 0.5. The null model corresponds to $\delta = 0$, while power estimates were produced by setting $\delta > 0$. Under this design, larger clusters were more likely to exhibit observations with $X = 1$, as the random effects tended to be larger.

We simulated clustered categorical data with more than two levels in an identical way. We categorized the $N(\delta, 2)$ random variable $u_i + e_{ij} + \delta$ into 3- and 5-level categorical variables under both balanced and unbalanced scenarios. Specifically, under the null model ($\delta = 0$), (1) the levels were marginally uniformly distributed for the balanced scenario and (2) the levels were marginally distributed as (0.25, 0.25, 0.50) in the 3-level cases and (0.10, 0.15, 0.20, 0.25, 0.30) in the 5-level case for the unbalanced scenario. Under this design, clusters of larger size had larger latent random effects, and have observations more likely to land in the last of the categories.

Table 1. Univariate proportion tests; empirical size and power.

		Estimate				Size						
	M	CW	GEE	DP/AC	UW	CW-Emp.	CW-SW$_{\hat{p}}$	CW-SW$_0$	CW-MM	GEE	DP/AC	UW
	50	0.100	0.105	0.108	0.128	0.078	0.079	0.054	0.071	0.065	0.061	0.594
$p = .1$	100	0.100	0.105	0.104	0.128	0.064	0.065	0.050	0.061	0.059	0.056	0.744
	200	0.100	0.105	0.102	0.128	0.052	0.053	0.050	0.051	0.067	0.050	0.905
	50	0.250	0.259	0.257	0.308	0.060	0.063	0.048	0.054	0.060	0.052	0.761
$p = .25$	100	0.250	0.259	0.253	0.308	0.056	0.057	0.049	0.053	0.069	0.051	0.909
	200	0.251	0.259	0.252	0.309	0.054	0.055	0.051	0.053	0.082	0.053	0.986
	50	0.501	0.512	0.501	0.578	0.056	0.057	0.049	0.049	0.071	0.052	0.844
$p = .5$	100	0.500	0.511	0.500	0.578	0.050	0.051	0.047	0.047	0.073	0.049	0.959
	200	0.500	0.511	0.500	0.578	0.056	0.056	0.054	0.054	0.091	0.054	0.997

CW, cluster weighted estimate; CW-Emp., CW-SW$_{\hat{p}}$, CW-SW$_0$, CW-MM are cluster-weighted tests evaluated with empirical, sandwich at $\hat{p}$, sandwich at p_0, and method of moments variance estimates; GEE, GEE model; DP/AC, the Dean and Pagano adaptation of Agresti-Coull method [12]; UW, unweighted.

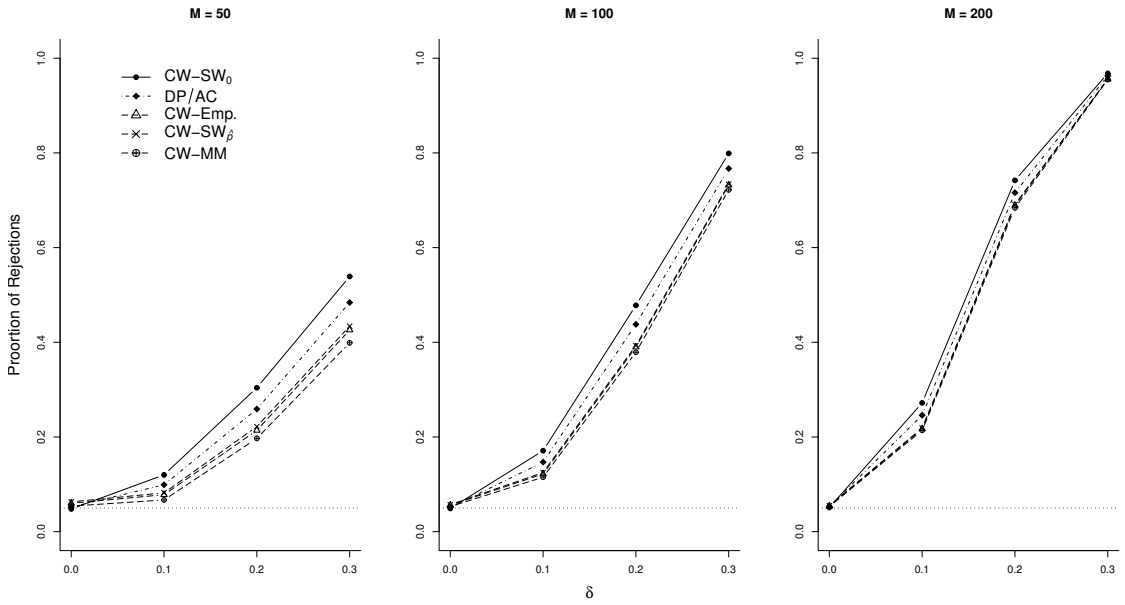

Figure 1. Power curves for reweighted proportion tests; $p = 0.25$.

Table 1 contains the results for the test of a marginal proportion. We begin by noting that our marginal estimator was approximately unbiased under all scenarios, while other methods exhibited varying degrees of bias. Cluster-weighted tests based on $\hat{v}_{emp}$, $\hat{v}_{\hat{p}}$, and $\hat{v}_{MM}$ performed reasonably well at higher sample sizes and null proportions away from 0 and 1, but were slightly biased otherwise. The cluster-weighted test based on $\hat{v}_{p_0}$ maintained appropriate size under all scenarios, even with small samples at p away from .5. We note that when $p = 0.5$, it can be shown that $\hat{v}_{MM}$ and $\hat{v}_{p_0}$ are equivalent. The Dean-Pagano [12] test was slightly biased at $p = 0.1$ and $M = 50$ clusters, but performed well for larger M for all p. The GEE model was biased for all scenarios, and as expected, the naive proportion test exhibited substantially inflated size. Figure 1 displays power curves for five of the tests at null $p = 0.25$. The cluster-weighted test using $\hat{v}_{p_0}$ exhibited consistently higher power across all values of M, while the other cluster-weighted tests had comparable power. Similar behavior occurred for tests at null $p = 0.1$, while for null $p = 0.5$ there was negligible difference in power between the tests (results provided in supplemental

Table 2. Goodness of fit and independence tests; empirical size and power.

Test	Scenario	M	CW-Emp.	CW-SW$_{\hat{p}}$	CW-SW$_0$	CW-MM	UW	MH
GOF	$K = 3$ Balanced	50	0.068	0.072	0.061	0.053	0.860	-
		100	0.055	0.056	0.052	0.048	0.961	-
		200	0.057	0.058	0.055	0.054	0.998	-
	$K = 3$ Unbalanced	50	0.071	0.074	0.062	0.055	0.850	-
		100	0.054	0.055	0.049	0.046	0.961	-
		200	0.056	0.057	0.055	0.052	0.998	-
	$K = 5$ Balanced	50	0.096	0.100	0.082	0.055	0.855	-
		100	0.076	0.078	0.067	0.057	0.960	-
		200	0.065	0.065	0.063	0.055	0.998	-
	$K = 5$ Unbalanced	50	0.104	0.110	0.082	0.064	0.851	-
		100	0.080	0.083	0.065	0.061	0.958	-
		200	0.067	0.068	0.061	0.058	0.998	-
Indep.	2x2	50	0.057	0.072	0.072	0.051	0.300	0.316
		100	0.053	0.060	0.060	0.051	0.310	0.327
		200	0.050	0.054	0.054	0.049	0.311	0.329
	2x3	50	0.059	0.060	0.059	0.044	0.380	0.387
		100	0.058	0.051	0.050	0.050	0.392	0.404
		200	0.053	0.041	0.040	0.049	0.392	0.399
	3x4	50	0.111	0.213	0.200	0.040	0.582	0.592
		100	0.075	0.125	0.123	0.044	0.572	0.578
		200	0.064	0.091	0.089	0.050	0.577	0.589

GOF, goodness of fit test for univariate data with K categories; Balanced (Unbalanced), equal (unequal) marginal category probabilities; Indep., test of independence for bivariate data; MH, Cochran-Maentel-Haenzel test; all other acronyms are as defined in Table 1.

tables in Section 3.6).

Additional simulations exploring the impact of absolute cluster size and the degree of informativeness are presented in Section 3.6. For a large number of clusters ($M = 100$), our test and the Dean-Pagano test performed consistently well across a range of absolute cluster sizes and degrees of informativeness, including when cluster size was not informative. The bias of the GEE approach tended to increase with the degree of informativeness, although this effect was mitigated when absolute cluster size increased.

The top portion of Table 2 provides the results from the marginal goodness of fit tests. The standard goodness of fit test was heavily biased for all scenarios. For the three-group simulation, the cluster-weighted test using $\hat{\mathbf{\Sigma}}_{MM}$ maintained appropriate size under both balanced and unbalanced designs for all values of M. The sandwich forms and empirical variance estimator exhibited slightly inflated size when $M = 50$,

but performed reasonably well for balanced and unbalanced proportions at 100 and 200 clusters. For five groups with balanced proportions, $\hat{\mathbf{\Sigma}}_{MM}$ remained approximated unbiased, while the empirical and sandwich estimators were moderately biased. All forms of the cluster-weighted test exhibited some inflation of size under a five-group unbalanced simulation, with the test based on $\hat{\mathbf{\Sigma}}_{MM}$ exhibiting the least bias. A comparison of power for the cluster-weighted tests (Section 3.6) shows that the cost of maintaining size for the method of moments variance is a slight reduction of power at low sample size.

To simulate bivariate categorical data, we generated multivariate random effects $(u_{i1}, \ldots, u_{iG})$ from a multivariate normal distribution $N_G(\mathbf{0}, I_G)$, where I_G is the GxG identity matrix. Within-cluster group sizes were generated as $n_{ig} \sim POI(10 + 5 * I[u_{ig} > 0]) + 1$, and $n_i = \sum_{g=1}^{G} n_{ig}$. We defined the random variable Y_{ij} to be a categorical, taking values $\{1, 2, \ldots, G\}$, so that n_{ig} observations were in group g. We created the second categorical variable X by discretizing $u_{iY_{ij}} + e_{ij} + \delta * I[Y_{ij} = 1]$ into K categories, where e_{ij} were generated i.i.d. from a standard normal distribution and $\delta = 0$ corresponds to the null hypothesis. The vector of cut points defining X were quantile values from a $N(0, 2)$ distribution, which we selected to produce desired category probabilities. Data from each cluster can be organized into a KxG contingency table, with values of X and Y as row and column variables, respectively. We ran simulations for 2×2, 3×2, and 4×3 tables. For $K = 2$, we selected a cut point for X such that the marginal proportion of observations in the first category was 0.6, and for $K > 2$ we selected cut points for equal probabilities among the categories of X.

The bottom portion of Table 2 contains results for the marginal tests of independence. The Cochran-Maentel-Haenzel test stratified by cluster and the naive chi-square test of independence were substantially biased. The cluster-weighted test using $\hat{\mathbf{\Sigma}}^{d}_{MM}$ maintained size for the 2×2 and 3×2 tables, even under small sample

sizes, and performed well for 4×3 tables under larger samples. The cluster-weighted tests using $\hat{\boldsymbol{\Sigma}}^d_{emp}$, $\hat{\boldsymbol{\Sigma}}^d_{\hat{p}}$, and $\hat{\boldsymbol{\Sigma}}^d_{\hat{e}}$ required a large sample size to exhibit appropriate size for a 2×2 table, and were biased for tables of larger size. Power comparisons for the cluster-weighted tests are included in the supplementary tables in Section 3.6. Similar to the results from the goodness of fit scenario, the method of moments-based test maintains appropriate size at the cost of a minor loss in power.

To test marginal homogeneity of matched pairs, we simulated data from the following multivariate random effects model

$$\begin{pmatrix} X_{ij} \\ Y_{ij} \end{pmatrix} = \begin{pmatrix} u_i \\ v_i \end{pmatrix} + \begin{pmatrix} e_{ij} \\ f_{ij} \end{pmatrix}$$

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix} \right), \begin{pmatrix} e_{ij} \\ f_{ij} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right).$$

Here, (u_i, v_i) represents the random effects for the paired observations in cluster i, and (e_{ij}, f_{ij}) are the random errors for paired observation j in cluster i. We set $\gamma = \rho = 0.8$, simulating positive correlation between both random effects and model errors. X and Y were dichotomized according to cut points corresponding to $P(X = 1) = P(Y = 1) = 0.1$ and 0.5. We simulated cluster size as the following function $n_i \sim POI(10 + 10 * I[u_i * v_i > 0])$. Under this simulation, clusters with concordant random effects tended to be larger than clusters with discordant random effects.

Table 3 shows results for the tests of marginal homogeneity. Both forms of the cluster-weighted test remained unbiased across all scenarios. The clustered tests of Eliasziw and Donner [18], Obuchowski [46], and Yang et al. [56] exhibited inflated size, illustrating the biasing effects of ICS. Unsurprisingly, the naive McNemar test performed poorly. Under the simulated scenario, cluster size was favorably associated with concordant random effects, resulting in larger clusters tending to have fewer discordant observations. Should cluster size be positively associated with discordant

Table 3. Empirical size for tests of marginal homogeneity

	M	CW-Emp.	CW-MM	χ^2_{ED}	χ^2_O	χ^2_Y	UW
	50	0.056	0.048	0.054	0.048	0.058	0.293
$p = .1$	100	0.052	0.048	0.058	0.055	0.059	0.314
	200	0.048	0.046	0.066	0.067	0.069	0.336
	50	0.058	0.050	0.090	0.070	0.081	0.375
$p = .5$	100	0.049	0.046	0.134	0.112	0.120	0.459
	200	0.050	0.049	0.214	0.191	0.198	0.578

CW-Emp., cluster weighted test with empirical variance; CW-MM, cluster weighted test with method of moments variance estimator evaluated under null hypothesis [14]; χ^2_{ED}, test by Eliasziw and Donner [18]; χ^2_O, test by Obuchowski [46]; χ^2_Y, test by Yang et al. [56]; UW, unweighted McNemar.

random effects, we would expect the bias shown by the McNemar test to further increase.

3.4 Application

We applied our cluster-weighted methods to a data set of repeated functional evaluations on 175 individuals with spinal cord injuries (SCI). The data were from patients enrolled in the Christopher and Dana Reeve Foundation's NeuroRecovery Network (NRN), an organization of treatment centers across the USA that provide a standardized, activity-based rehabilitation program to individuals with SCI [26]. While enrolled in the program, patients are periodically evaluated on their functional capability approximately every 20 rehabilitation sessions. One of these assessment instruments is the Neuromuscular Recovery Scale (NRS), which consists of 13 functional tasks developed by NRN researchers designed to measure functional ability of SCI patients in relation to pre-injury capability [4, 27]. For each functional task, patients are given an ordinal rating called the patient's phase, which ranges from 1 to 4 with the highest rating of phase 4 representing return to pre-injury ability. We considered responses on 9 of the NRS items; three items were added to the NRS during data collection and one item that does not incorporate a phase 1 rating were excluded.

There were very few observations of phase 4 across individuals and tasks, so for the purpose of this analysis phases 3 and 4 were combined.

The data set contained 892 evaluations from 175 patients. Cluster size, i.e. the number of evaluations on a patient, ranged from 2 to 24 with a median of 4. From an empirical standpoint, there are two suspected mechanisms of informativeness in a marginal analysis of these data: (1) lower functioning patients tend to occupy lower phases of recovery and also tend to remain enrolled in the NRN longer, (2) enrolled patients are actively receiving rehabilitation, presumably improving functional capacity and increasing their phases of recovery. The former mechanism corresponds to "negative" informativeness, wherein patients with lower phases tend to contribute more observations. The latter indicates "positive" informativeness, wherein patients enrolled longer have more room time over which to improve their phase, thus contributing observations with higher phase with greater frequency. Figure 2 provides a bar chart of the proportion of individuals in each phase category (1, 2, 3/4) for each NRS item by the quartiles of the number of evaluations contributed by each patient – 2, 3, 4-5, and 6+ evaluations. The latter, positive mechanism of informativeness appears to have been more prominent for the treadmill-based items of the NRS (the top row of plots in Figure 2), as the proportion of phase 1 observations decreased over the quartiles for the number of observations. This was also somewhat apparent in other NRS items, but to an extent less clear and direct than for the three treadmill items. To formally evaluate informativeness, we implemented a balanced bootstrap test of ICS proposed by Nevalainen et al. [44] for all 9 NRS items, which suggested that cluster size was informative for these data (Table 4).

A phase 1 rating represents the greatest impairment relative to normal movement patterns, indicating that substantial rehabilitation is required for the given functional task. Given the resource intensiveness of many SCI rehabilitation programs, it can be of value to estimate the load of phase 1 patients enrolled in the rehabilitation

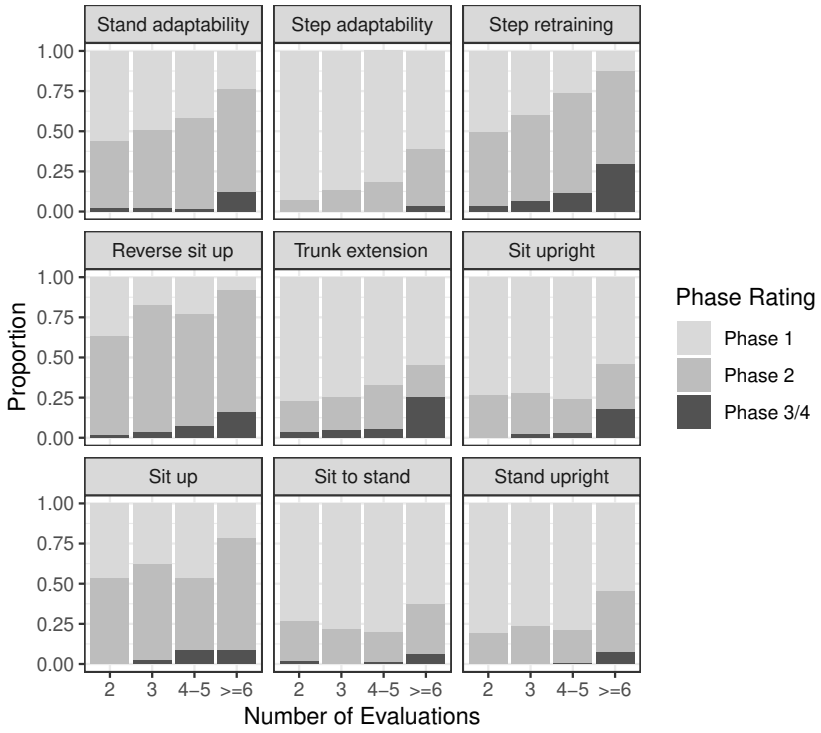

Figure 2. Proportion of phase category by quartile of the number of evaluations contributed by patients for each individual NRS item.

program, i.e. the proportion of patients in phase 1 for NRS tasks at a typical point in time for the program. While longitudinal analyses of patient progress are often of interest for data such as these, the question posed in the previous sentence is marginal in scope for which our proposed method is ideally suited. To this end, we estimated the cluster-weighted proportion of phase 1 patients for each NRS item. Table 4 includes estimates and 95% confidence intervals for this proportion calculated from the cluster-weighted test using a null sandwich variance with $p_0 = 0.5$. We also compared estimates and intervals from the Dean-Pagano adaptation of the Agresti-Coull method, a GEE intercept model, and the unweighted estimate. All tests that account for clustering provided similar estimates and intervals, while the unweighted estimate was consistently lower than the cluster-weighted estimate, as would be expected when cluster size is positively associated with functional capabil-

ity. Moreover, the unweighted interval failed to include the cluster-weighted estimate in every instance.

We can extend the estimation of the proportion of patients in phase 1 for each functional task to estimating proportions in the three phases, as well as test against a null distribution for the three proportions. The estimated cluster-weighted category proportions are presented in Table 5, along with test statistics against the null values $p_0 = (0.5, 0.4, 0.1)$ for the cluster-weighted chi-square test using a method of moments variance and the unweighted chi-square test. Compared against the critical value $\chi^2_2(.95) = 5.99$, both weighted and unweighted tests rejected the hypotheses that the true phase 1, phase 2, and phase 3/4 proportions are 0.5, 0.4, and 0.1, respectively. The unweighted test statistics were noticeably larger, due to their inappropriate handling of clustering, and were inconsistently rank-ordered relative to the cluster-weighted test.

Trunk extension in sitting is an NRS item that measures a patient's ability to return to a seated position from a forward extension. It is generally recognized that abdominal control is necessary for normal function of the upper and lower limbs, and trunk control has been shown to be related to functional recovery of limb movement in stroke patients [20, 21, 27]. However, for SCI patients the importance of seated balance in relation to other functional capabilities has not been established [20, 39]. We used the chi square test of independence to test the association between patients' trunk control, as measured by the trunk extension task, and functionality in other NRS tasks. Table 5 contains the cluster-weighted with method of moments variance, Mantel-Haenszel, and unweighted test statistics. All tests rejected the hypothesis of independence (critical value = $\chi^2_4(.95) = 9.49$), in part due to association that is induced by the escalating difficulty of NRS tasks, e.g. patients rated as phase 1 for Sit upright will tend to also be phase 1 for Sit to stand. The Mantel-Haenszel and unweighted test statistics were considerably larger than the cluster-weighted statistics

Table 4. Application of proportion tests to SCI data

	Proportion in Phase 1				
Task	CW	DP-AC	GEE	UW	T_F
Stand adaptability	0.41 (0.35, 0.46)	0.41 (0.35, 0.46)	0.39 (0.34, 0.45)	0.34 (0.31, 0.37)	0.001
Step adaptability	0.79 (0.75, 0.83)	0.79 (0.74, 0.83)	0.78 (0.73, 0.82)	0.72 (0.69, 0.75)	0.002
Step retraining	0.30 (0.24, 0.35)	0.30 (0.25, 0.35)	0.27 (0.22, 0.32)	0.22 (0.19, 0.25)	0
Reverse sit up	0.18 (0.14, 0.23)	0.19 (0.14, 0.24)	0.17 (0.13, 0.22)	0.15 (0.12, 0.17)	0.007
Trunk extension	0.67 (0.61, 0.73)	0.67 (0.61, 0.73)	0.67 (0.60, 0.72)	0.62 (0.59, 0.65)	0.012
Sit upright	0.68 (0.62, 0.74)	0.68 (0.62, 0.73)	0.68 (0.61, 0.73)	0.63 (0.60, 0.66)	0.030
Sit up	0.36 (0.31, 0.42)	0.37 (0.31, 0.42)	0.36 (0.30, 0.42)	0.31 (0.28, 0.35)	0.005
Sit to stand	0.74 (0.68, 0.80)	0.74 (0.68, 0.79)	0.74 (0.68, 0.79)	0.70 (0.67, 0.73)	0.045
Stand upright	0.72 (0.66, 0.77)	0.71 (0.66, 0.77)	0.71 (0.66, 0.76)	0.65 (0.62, 0.69)	0.002

Estimated proportion and 95% confidence interval of individuals with SCI in Phase 1 using the cluster-weighted method with $\hat{v}_{p_0}$ (CW), the Dean/Pagano adaptation of Agresti/Coull method (DP-AC), GEE model (GEE), and unweighted (UW). T_F, p-value from test statistic using the balanced bootstrap scheme of Nevalainen et al. [44] to test for cluster size informativeness in the NRS task.

Table 5. Application of goodness of fit and independence tests to SCI data

	Estimated Phase Proportions			Goodness of Fit		Independence		
Task	Phase 1	Phase 2	Phase 3/4	X^2_{CW}	X^2_{UW}	X^2_{CW}	X^2_{MH}	X^2_{UW}
Stand adaptability	0.41	0.54	0.05	31.1	129.4	15.6	58.6	268.2
Step adaptability	0.79	0.20	0.01	121.0	181.9	22.3	48.9	218.0
Step retraining	0.30	0.56	0.14	43.2	301.6	36.7	95.3	203.0
Reverse sit up	0.18	0.74	0.08	90.7	486.3	16.7	58.6	256.1
Trunk extension	0.67	0.22	0.11	44.5	133.5	-	-	-
Sit upright	0.68	0.25	0.07	30.7	78.1	28.1	97.4	441.1
Sit up	0.36	0.58	0.06	33.9	168.1	10.9	27.0	144.9
Sit to stand	0.74	0.23	0.03	59.0	147.9	18.0	61.2	449.4
Stand upright	0.72	0.25	0.03	57.9	91.9	18.5	65.2	360.3

Goodness of Fit: Estimated cluster-weighted proportions of individuals with SCI in each phase of recovery, and cluster-weighted using variance $\hat{\mathbf{\Sigma}}_{MM}$ (X^2_{CW}) and unweighted (X^2_{UW}) chi-square goodness of fit statistics testing null phase proportions of 0.5, 0.4, and 0.1. Independence: cluster-weighted using variance $\hat{\mathbf{\Sigma}}^d_{MM}$ (X^2_{CW}), Mantel-Haenazel (X^2_{MH}), and unweighted (X^2_{UW}) chi-square statistics from tests of independence with the Trunk Extension task.

for each task, indirectly indicating that cluster size may have been informative. The rank-ordering of statistics for the cluster-weighted and Mantel-Haenszel tests were in general correspondence, with Step adaptability being a notable exception. Using the magnitude of the chi square statistic as a measure of association, performance on the Trunk extension task appeared to be most strongly related to Step retraining, a task completed on the treadmill as part of a patient's rehabilitation. Among non-treadmill items, Trunk extension was most strongly related to Sit upright, a physiologically sensible finding as sitting upright requires good postural and trunk control.

3.5 Discussion

In the analysis of clustered data, methods that account for potential dependence among observations should be implemented. There are many available methods that properly account for dependence often found in clustered data. However, these methods may be biased if there is a relationship between the outcome and the size of the cluster. In this chapter, we proposed hypotheses tests for marginal parameters of clustered categorical data that adjust for potential informative cluster size. Further, we constructed these tests using competing variance estimation techniques and evaluated their performance through simulations. We then applied these cluster-weighted methods to estimate marginal functional capability proportions and to test marginal association between functional tasks in a longitudinal data set in which SCI patients with higher functional ability contributed more observations due to longer program enrollment. Another potential application of these methods is the analysis tooth-level dental data, for which much of the development of cluster-weighted methodologies has been applied [16, 40, 53].

In our establishment of clustered data analogues of proportion, chi square goodness of fit, and independence tests, we provided an evaluation of different variance estimators for Wald-type procedures of cluster-weighted test statistics. It is well

documented that Wald-based methods can perform poorly in estimating and testing categorical variables when sample sizes are small and when parameters approach their boundaries, i.e. (0, 1) for proportions. Conventional clustered data methods as well as ICS-adjusting methods can utilize sandwich variance estimates, which have been shown to be resistant to model misspecifications, but potentially biased when the number of clusters is small [35, 38]. These issues were apparent in our simulation study, as tests constructed from the sandwich variance estimator evaluated at the estimated parameter showed inflated size. Score tests often have superior performance to Wald tests in categorical data analyses. While a true score test in the context of this paper would require significant assumptions regarding the complex relationship between the variables and cluster size, it has been suggested that tests with variance estimates constructed under a null hypothesis are related to score tests. As such, for each of the three data scenarios we evaluated two tests with variance estimates evaluated under the null hypothesis. Our simulation results showed these score-related tests exhibited sizes closer to nominal compared to tests constructed with empirical variances or sandwich variances evaluated at the parameter estimate. When testing a single marginal proportion, the cluster-weighted test using the null sandwich variance estimator not only maintained appropriate size under all simulation parameters, but also exhibited the highest power of all comparison tests for p away from 0.5. Tests using a method of moments variance estimate constructed under the null hypothesis outperformed tests based on the null sandwich estimator in the goodness of fit and independence scenarios, perhaps due to the increased number of parameters being estimated. While these method of moments tests were not the most efficient, they exhibited superior size to the additional cluster-weighted forms and the loss in power was minimal.

Regardless of variance estimation method, the tests presented in this chapter perform consistently across a range of absolute cluster sizes and varying degree of

informativeness. Additionally, they maintain appropriate size when cluster size is non-informative (additional simulations provided in online supplementary material). However, the methods proposed here are asymptotic in nature, and we recommend their use only when the number of clusters is sufficiently large, approximately 30 or more.

In working with clustered data, careful attention must be given to defining the marginal analysis of interest. As previously mentioned, the application of our methods corresponds to an analysis of a typical member from a typical cluster, in which the cluster is the primary unit of interest. There are several possible marginal analyses of clustered data which do not always correspond, particularly in the presence of ICS. Seaman et al. [49] detail the distinction of interpretations between certain marginal models, and Lorenz et al. [40] provide a comprehensive review of reweighting methods corresponding to different marginal analyses. Careful attention to these distinctions should be considered. In particular, we note for bivariate categorical data, group membership as well as cluster size can be informative. Under such within-cluster group size informativeness, the tests proposed in this chapter could be inappropriate and a weighting method incorporating group membership in the manner presented in Section 2.5 might be considered.

The cluster reweighting methodology provides a closed-form, computationally unburdened method of marginal parameter estimation in clustered data that mitigates bias from informative cluster size. Its computational ease and applicability have resulted in developments of rank-based tests, correlation estimation, and survival methods for clustered data. To our knowledge, this is the first extension of the marginalization principle to hypotheses tests for categorical data analogous to well-known methods for independent data. Additionally, this is the first evaluation of these cluster weighted tests under several methods of variance estimation. Our simulations show that in the context of clustered categorical data, the choice of variance

estimation technique can have profound impact on the performance of these tests.

3.6 Supplemental results

Tables 6-9 contain additional simulation results from the binary univariate (proportion), multi-category univariate (goodness of fit), and bivariate (chi square independence) categorical data scenarios. Tables 6-8 contain size and power for the reweighted proportion, goodness of fit, and chi square independence analogs. For the goodness of fit scenario, size and power are shown for $K = 3$ for both balanced and unbalanced category proportions. Results from the chi square independence simulations are for a 2x2 table. Table 9 contains size estimates for tests of marginal proportion from data simulated with varying degrees of informativeness and average cluster size. For these simulations, cluster size n_i are simulated from Poisson$(b + c * I[u_i > 0]) + 1$. The value of c indicates the degree of informativeness; when $c = 0$, cluster size is non-informative.

For each data scenario, CW-Emp. is the cluster-weighted test constructed with empirical variance, CW-SW$_{\hat{p}}$ is the cluster-weighted test constructed with the sandwich variance evaluated at the estimate(s), CW-SW$_0$ is the cluster-weighted test constructed with sandwich variance evaluated under the null hypothesis, and CW-MM is the cluster-weighted test constructed with method of moments variance evaluated under the null hypothesis. Size and power estimates are calculated as the proportion of rejections over 10 000 Monte Carlo iterations at nominal size of 0.05.

Table 6. Empirical size and power for reweighted proportion tests; $p = .1, .5$

M	Test	$p = .1$				$p = .5$			
		Size	Power			Size	Power		
		$\delta = 0$	$\delta = .1$	$\delta = .2$	$\delta = .3$	$\delta = 0$	$\delta = .1$	$\delta = .2$	$\delta = .3$
50	CW-Emp.	0.078	0.061	0.131	0.298	0.056	0.110	0.259	0.502
	CW-SW$_{\hat{p}}$	0.079	0.063	0.135	0.308	0.057	0.115	0.265	0.509
	CW-SW$_0$	0.054	0.133	0.297	0.525	0.049	0.099	0.244	0.483
	CW-MM	0.071	0.053	0.116	0.274	0.049	0.099	0.244	0.483
	DP/AC	0.061	0.091	0.216	0.421	0.052	0.106	0.252	0.493
100	CW-Emp.	0.064	0.089	0.275	0.590	0.050	0.157	0.452	0.774
	CW-SW$_{\hat{p}}$	0.065	0.090	0.279	0.594	0.051	0.160	0.456	0.778
	CW-SW$_0$	0.050	0.173	0.447	0.753	0.047	0.151	0.441	0.766
	CW-MM	0.061	0.084	0.264	0.576	0.047	0.151	0.441	0.766
	DP/AC	0.056	0.130	0.366	0.686	0.049	0.153	0.446	0.770
200	CW-Emp.	0.052	0.157	0.550	0.896	0.056	0.254	0.727	0.965
	CW-SW$_{\hat{p}}$	0.053	0.158	0.552	0.897	0.056	0.256	0.728	0.965
	CW-SW$_0$	0.050	0.254	0.679	0.947	0.054	0.250	0.723	0.964
	CW-MM	0.051	0.153	0.544	0.892	0.054	0.250	0.723	0.964
	DP/AC	0.050	0.207	0.621	0.928	0.054	0.252	0.726	0.965

DP/AC, Dean and Pagano's adaptation of Agresti-Coull method.

Table 7. Empirical size and power for reweighted goodness of fit tests; $K = 3$

M	Proportions	Test	Size	Power		
			$\delta = 0$	$\delta = .1$	$\delta = .2$	$\delta = .3$
50	Balanced	CW-Emp.	0.068	0.111	0.240	0.445
		CW-SW$_{\hat{p}}$	0.072	0.115	0.248	0.455
		CW-SW$_0$	0.061	0.102	0.228	0.430
		CW-MM	0.053	0.086	0.206	0.394
	Unbalanced	CW-Emp.	0.071	0.128	0.270	0.468
		CW-SW$_{\hat{p}}$	0.074	0.133	0.277	0.476
		CW-SW$_0$	0.062	0.082	0.182	0.358
		CW-MM	0.055	0.103	0.229	0.423
100	Balanced	CW-Emp.	0.055	0.140	0.376	0.715
		CW-SW$_{\hat{p}}$	0.056	0.144	0.382	0.719
		CW-SW$_0$	0.052	0.136	0.373	0.715
		CW-MM	0.048	0.124	0.356	0.696
	Unbalanced	CW-Emp.	0.054	0.153	0.395	0.722
		CW-SW$_{\hat{p}}$	0.055	0.155	0.400	0.726
		CW-SW$_0$	0.049	0.111	0.314	0.648
		CW-MM	0.046	0.139	0.368	0.701
200	Balanced	CW-Emp.	0.057	0.208	0.652	0.955
		CW-SW$_{\hat{p}}$	0.058	0.210	0.655	0.955
		CW-SW$_0$	0.055	0.206	0.656	0.956
		CW-MM	0.054	0.198	0.642	0.952
	Unbalanced	CW-Emp.	0.056	0.216	0.654	0.950
		CW-SW$_{\hat{p}}$	0.057	0.218	0.656	0.951
		CW-SW$_0$	0.055	0.178	0.602	0.935
		CW-MM	0.052	0.207	0.642	0.947

Table 8. Empirical size and power for reweighted independence tests; 2x2 table

M	Test	Size	Power		
		$\delta = 0$	$\delta = .1$	$\delta = .2$	$\delta = .3$
50	CW-Emp.	0.057	0.084	0.176	0.316
	CW-SW$_{\hat{p}}$	0.072	0.104	0.212	0.374
	CW-SW$_0$	0.072	0.104	0.208	0.364
	CW-MM	0.051	0.076	0.162	0.299
100	CW-Emp.	0.053	0.109	0.293	0.544
	CW-SW$_{\hat{p}}$	0.060	0.118	0.316	0.580
	CW-SW$_0$	0.060	0.117	0.312	0.572
	CW-MM	0.051	0.104	0.284	0.533
200	CW-Emp.	0.050	0.169	0.517	0.844
	CW-SW$_{\hat{p}}$	0.054	0.179	0.537	0.859
	CW-SW$_0$	0.054	0.177	0.534	0.856
	CW-MM	0.049	0.166	0.511	0.842

Table 9. Effect of absolute cluster size and degree of informativeness on tests of marginal proportion

	c	CW-Emp.	CW-SW$_{\hat{p}}$	CW-SW$_0$	CW-MM	GEE	DP/AC	UW
$b = 5$	0	0.055	0.056	0.051	0.051	0.051	0.051	0.225
	5	0.059	0.060	0.051	0.055	0.089	0.054	0.789
	10	0.055	0.056	0.048	0.052	0.134	0.049	0.975
	20	0.053	0.055	0.047	0.050	0.204	0.050	0.999
$b = 10$	0	0.056	0.057	0.052	0.052	0.054	0.053	0.328
	5	0.058	0.059	0.051	0.055	0.059	0.053	0.672
	10	0.053	0.055	0.046	0.050	0.066	0.049	0.906
	20	0.059	0.060	0.054	0.056	0.085	0.056	0.993
$b = 20$	0	0.057	0.057	0.049	0.053	0.053	0.052	0.468
	5	0.052	0.054	0.048	0.048	0.050	0.048	0.607
	10	0.056	0.058	0.048	0.053	0.052	0.053	0.778
	20	0.058	0.059	0.050	0.054	0.056	0.054	0.950

Estimated size for tests of marginal proportion across varying average cluster size and degree of informativeness. Data simulated from $M = 100$, $p = .25$, n_i from Poisson($b + c * I[u_i > 0]$) + 1.
GEE, GEE model; DP/AC, Dean and Pagano's adaptation of Agresti-Coull method; UW, unweighted.

CHAPTER 4

ESTIMATION AND TESTING FOR QUANTITATIVE DATA

4.1 Introduction

In this chapter, we present a collection of reweighted tests for clustered quantitative data. We begin this chapter by discussing how incomplete group structure in clustered quantitative values restricts variance estimation methods for reweighted tests, and describe an alternative technique that accounts for this issue. We then apply the reweighting principal to derive novel tests of marginal means and variances that parallel classical forms, with particular attention on the different approaches to assessing the equality of variance across intra-cluster groups. The performance of these novel tests are explored through a simulation study. This chapter additionally includes summaries of the reweighted rank-based tests and tests of correlation that have been developed by other authors. These existing tests are included here for cohesion of the comprehensive R package discussed in the following chapter, and are integrated with the newly developed tests in a logical progression of methods.

4.2 Variance estimation in tests for quantitative data

In the tests of clustered categorical data in the previous chapter, we compared the performance of our tests under a number of variance estimation methods. While all of these methods remain valid for estimators from quantitative data, we encounter

complications related to group structure that were previously precluded by the categorical nature of the data. The variance forms discussed in Section 3.2.1 are functions of group estimates within clusters. If a group estimate does not exist, these functions can not be calculated. Thus, the variance forms are not defined when data have incomplete group structure. However, the reweighted tests of categorical data in Chapter 3 were based on proportions. Group proportions within clusters are defined $\frac{n_i^{(k)}}{n_i}$, so even when $n_i^{(k)} = 0$ these proportions have a defined values. As such, incomplete clusters thus retain a form of "completeness" that allow us the range in variance estimation techniques without assuming complete group structure.

This quality of "completeness" is not retained for statistics of reweighted quantitative data. In order to use the variance estimation methods of the previous chapter in the quantitative tests derived here, we would have to assume complete group structure. This is an unrealistic assumption that would lessen the utility of the proposed tests. Additionally, it might encourage analysts to throw out incomplete clusters in order to implement the tests. This would not only waste valuable data but also raise questions on the data missingness structure. Therefore, we avoid these issues by estimating variance in the novel tests using an alternative variance estimation technique which we describe below.

The delete-one jackknife is a nonparametric method that estimates the variance of a statistic by repeatedly calculating the statistic after systematically removing one observation. It is known to be consistent in large samples for a wide class of estimators, and has been extensively studied [17, 29, 50]. As clusters are the unit of interest for the marginal analysis in question, this process corresponds to systematically removing each cluster. Such a "delete-one-cluster" jackknife method has previously been used to estimate the variance of a statistic for reweighted quantities [16]. As this method is applicable regardless of the observed group structure, we implement this technique in estimating the variance of the reweighted statistics in this chapter.

We briefly summarize the process here.

Let T be the statistic of interest calculated from the full data, and $T_{(i)}$ be the statistic computed when the i^{th} cluster is removed. Define pseudovalues as $p_i = MT - (M-1)T_{(i)}$. The jackknifed estimate of the variance of T is $\hat{V}_{JK} = \frac{1}{M(M-1)} \sum_{i=1}^{M} (p_i - \bar{p})^2$, where $\bar{p} = \frac{1}{M} \sum_{i=1}^{M} p_i$. Defining $\bar{T} = \frac{1}{M} \sum_{i=1}^{M} T_i$, this can be simplified to

$$\hat{V}_{JK} = \frac{M-1}{M} \sum_{i=1}^{M} \left(T_{(i)} - \bar{T} \right)^2 . \tag{4.1}$$

The above form denotes the jackknife variance form when T is a scalar. This is easily extended to a covariance matrix when $\boldsymbol{T}$ is a vector, expressed as $\hat{\Sigma}_{JK} = \frac{M-1}{M} \sum_{i=1}^{M} \left(\boldsymbol{T}_{(i)} - \bar{\boldsymbol{T}} \right) \left(\boldsymbol{T}_{(i)} - \bar{\boldsymbol{T}} \right)^T$. Hinkley [28] showed that the jackknife variance does not account for the unbalanced nature of multiparameter data, and in the context of linear modeling proposed a modification using reweighted pseudovalues. This reweighting is a function of the projection matrix, making it untenable for our purposes. However, it has been demonstrated that this reweighting is closely approximated by the correction factor $\frac{N}{N-P}$ in the bootstrap variance, where N is the sample size and P is the number of model parameters [55]. Therefore, we implement this correction and use the variance estimate $\frac{M}{M-K} \hat{\Sigma}_{JK}$ for tests in which T is a vector of length K.

4.3 Tests of means

The objective of this dissertation is the development of a collection of hypothesis tests analogous to frequently implemented standard statistical methods. Tests of means are perhaps the most obvious candidates for cluster-weighted adaptation due to their prominence in classical statistics. Therefore, it is natural that we begin this chapter on tests of quantitative data with analogs of the classical t-test and ANOVA tests.

4.3.1 One sample - t-test analog

Previous authors have begun the development of a cluster-weighted t-test analog. Datta et al. [9] use a modified cluster-weighted t-test as a benchmark for their signed-rank tests in simulation studies, while Nevalainen et al. [43] provide a formal derivation of this statistic. The construction of this test is straightforward. For $\theta = E[X_{ij}]$, the hypothesis of interest is $H_0 : \theta = \theta_0$. We have previously derived the statistic $\hat{\theta} = \frac{1}{M}\sum_{i=1}^{M} \bar{X}_i$ for this hypothesis in equation (xx) using the marginalization principle, and note its equivalence to the statistic derived by Nevalainen et al. using functionals. H_0 can then be tested by comparing

$$Z = \frac{\hat{\theta} - \theta_0}{\sqrt{\frac{\hat{\sigma}^2}{M}}}$$

to the quantiles of the standard normal distribution. Group notation is not required for this test, allowing $\hat{\sigma}^2$ to be estimated using any of the methods discussed in the previous chapter. We note that a method of moments variance constructed under H_0 is consistent with the second moment variance estimate of Nevalainen et al., and suggest $\hat{\sigma}^2 = \frac{1}{M}\sum_{i=1}^{M} \frac{1}{n_i^2}\left(\sum_{j=1}^{n_i} X_{ij} - \theta_0\right)^2$. The extension of this test to paired data follows in the usual manner.

4.3.2 Two sample - t-test analog

While a reweighted one-sample t-test analog is straightforward and has been previously discussed, an extension to the two-sample test requires additional considerations and, to our knowledge, has not previously been explored.

In an analogue of the two-sample t-test, we wish to test $H_0 : \theta^{(1)} = \theta^{(2)}$, where $\theta^{(k)} = E[X_{ij}^{(k)}]$. In the construction of such a test, two potential issues arise that were irrelevant in the one-sample analog. First, group size as well as cluster size could be informative. If there exists an association between number of observations within a group and the measured values in that cluster (e.g., clusters with more

group 1 observations tend to have larger values), a test that corrects for ICS could still be biased. Second, it's possible that data will be collected in which not all clusters contain observations belonging to both groups. The effect of incomplete group structure on the parameter estimates and variance estimation methods should be considered. We can address both of these issues by applying the reweighting method from Section 2.5.2, which corrects for IWCGS and allows for incomplete clusters. Recall that $M^{(c)}$ is the number of complete clusters, $M^{(k)}$ is the number of incomplete clusters that contains observations belonging only to group k, and $\bar{X}_i^{(k)}$ is the group k average from cluster i. We estimate the marginal group 1 mean as in formula (2.4)

$$\hat{\theta}^{(1)} = \frac{1}{\frac{1}{2}M^{(c)} + M^{(1)}} \sum_{i=1}^{M} \left\{ \frac{1}{2}\bar{X}_i^{(1)} I[n_i^{(1)} > 0, n_i^{(2)} > 0] + \bar{X}_i^{(1)} I[n_i^{(1)} > 0, n_i^{(2)} = 0] \right\}$$

and similarly estimate the marginal group 2 mean as

$$\hat{\theta}^{(2)} = \frac{1}{\frac{1}{2}M^{(c)} + M^{(2)}} \sum_{i=1}^{M} \left\{ \frac{1}{2}\bar{X}_i^{(1)} I[n_i^{(1)} > 0, n_i^{(2)} > 0] + \bar{X}_i^{(2)} I[n_i^{(1)} = 0, n_i^{(2)} > 0] \right\}.$$

Define $T = \hat{\theta}^{(1)} - \hat{\theta}^{(2)}$. With an estimate of the variance of T, $\hat{V}(T)$, we can test H_0 using the standardized statistic $\frac{T - E[T]}{\sqrt{\hat{V}(T)}}$, which asymptotically follows the standard normal distribution. Under H_0, $E[T] = 0$. We can obtain an estimate of $\hat{V}(T)$ using the jackknife approach.

We note that this test is easily extended to testing hypotheses of the form H_0 : $\theta^{(1)} = \theta^{(2)} + c$ by replacing $E[T]$ in the standardized statistic with c.

4.3.3 K-group - ANOVA analog

A natural extension of 4.3.2 is the testing of equality of K group means. Therefore, we propose an omnibus reweighted test analogous to the standard one-way analysis of variance (ANOVA) test for independent observations. While the classical ANOVA method performs a test of equality of group means by comparing the ratio of intra-group to inter-group variability using an F distribution, the analog test for clustered

data is most easily approached using the asymptotic normality of the cluster-weighted group means.

Recall that K_i^c denotes the number of groups observed in cluster i, $K_i^c = 1, \ldots, K$, and $\bar{X}_i^{(k)} = \frac{1}{n_i^{(k)}} \sum_{j'=1}^{n_i^{(k)}} X_{ij'}^{(k)}$ denotes the k^{th} group average in cluster i. Define

$$K_i^{(k)} = \begin{cases} K_i^c, & \text{if } n_i^{(k)} > 0 \\ 0, & \text{otherwise.} \end{cases}$$

By extending (2.4) to the K-group case, the k^{th} group mean is defined

$$\hat{\theta}^{(k)} = \frac{1}{\tilde{M}^{(k)}} \sum_{K_i^c=1}^{K} \sum_{i=1}^{M} \left[\frac{1}{K_i^c} \bar{X}_i^{(k)} I[K_i^{(k)} = K_i^c] \right], \tag{4.2}$$

where $\tilde{M}^{(k)} = \sum_{K_i^c=1}^{K} \frac{1}{K_i^c} \sum_{i=1}^{M} I[K_i^{(k)} = K_i^c]$. The notation of estimate (4.2) belies its true simplicity. For data with complete group structure, $\hat{\theta}^{(k)}$ is simply the overall average of the intra-cluster group k averages. The introduction of K_i^c and $K_i^{(k)}$ in this notation is to account for weighting and marginalization resulting from varying selection probabilities from incomplete clusters.

The vector of group-weighted means for all K groups is $\hat{\boldsymbol{\theta}} = \left(\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(K)} \right)^T$. The hypothesis of interest is $H_0 : \theta^{(1)} = \theta^{(2)} = \ldots = \theta^{(K)}$, equivalently stated with the $K-1$ composite hypotheses $H_0 : \theta^{(1)} - \theta^{(2)} = 0, \theta^{(2)} - \theta^{(3)} = 0, \ldots, \theta^{(K-1)} - \theta^{(K)} = 0$. By applying the $(K-1) \times K$ contrast matrix

$$C = \begin{bmatrix} 1 & -1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & -1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 & -1 \end{bmatrix}$$

the hypothesis of interest can be expressed $H_0 : C\boldsymbol{\theta} = \mathbf{0}$. This can be tested using the statistic

$$X^2 = M \left(C\hat{\boldsymbol{\theta}} \right)^T \left(C\hat{\Sigma}C \right)^{-1} \left(C\hat{\boldsymbol{\theta}} \right)$$

where $\hat{\Sigma}$ is a variance estimate of $\hat{\boldsymbol{\theta}}$. We estimate $\hat{\Sigma}$ using the jackknife technique, applying the correction factor $\frac{M}{M-K}$. X^2 follows a chi square distribution with $K-1$ degrees of freedom.

4.4 Rank-based tests

The reweighted rank-based hypothesis tests that have previously been established include a rank sum test [10], signed rank test [11], and an extension of the rank sum test correcting for IWCGS [16]. These tests use the general form

$$Z = \frac{S - E[S]}{\sqrt{\hat{V}(S)}} \tag{4.3}$$

where S is a statistic, $E[S]$ is the statistic's expected value under the null hypothesis, and $\hat{V}(S)$ is an estimate of the variance of S. The standardized statistic, Z, is asymptotically normal under mild regularity conditions.

While the expression of S varies across these tests, the derivation follows that described in Section 2.4. Let W^* represent the traditional form of the desired statistic applied to a data set formed by one iteration of the WCR process. The cluster-weighted statistic is derived by applying a marginal expectation calculation to W^* with respect to the resampling process, conditioned on the entire collection of original data $\mathbf{V}$. This can be generalized as

$$S = E[W^*|\mathbf{V}]. \tag{4.4}$$

4.4.1 Rank sum test for ICS

For the j^{th} observation from cluster i, we observe $(X_{ij}, G_{ij}), 1 \leq i \leq M, 1 \leq j \leq n_i$, where G_{ij} denotes the group membership (0 or 1) for outcome X_{ij}. Let $n_{i1} = \sum_{j=1}^{n_i} G_{ij}$ denote the number of group 1 observations in cluster i. The null hypothesis is that the two groups follow the same distribution, formalized as $P(X_{ij} \leq x|G_{ij} = 0, n_i, n_{i1}) = P(X_{ij} \leq x|G_{ij} = 1, n_i, n_{i1}) = F(x)$. A single WCR iteration produces the data set

$(X_i^*, G_i^*), i = 1, \ldots, M$. Applied to this data set, the Wilcoxon rank sum statistic is $W^* = \frac{1}{M+1} \sum_{i=}^{M} G_i^* R_i^*$, where $R_i^* = 1 + \frac{1}{2} \left[\sum_{i' \neq i} I \left[X_{i'}^* \leq X_i^* \right] + \sum_{i' \neq i} I \left[X_{i'}^* < X_i^* \right] \right]$. The cluster-weighted rank sum statistic proposed by Datta and Satten [10] is derived by performing the expectation calculation of (4.4) on W^*, and results in the equation

$$S = \frac{1}{M+1} \sum_{i=1}^{M} \sum_{j=1}^{n_i} \frac{G_{ij}}{n_i} \left[1 + \frac{1}{2} \sum_{i' \neq i} \left[F_{i'} \left(X_{ij} \right) + F_{i'} \left(X_{ij} - \right) \right] \right],$$

where $F_i(x) = \frac{1}{n_i} \sum_{j=1}^{n_i} I \left[X_{ij} \leq x \right]$. To test the null hypothesis using the standardized statistic (4.3), Datta and Satten show that $E[S] = \frac{1}{2} \sum_{i=1}^{M} \frac{n_{i1}}{n_i}$, and use a Hájek projection of S to estimate the variance of S. This variance estimate is $\hat{V}(S) = \sum_{i=1}^{M} \left[\hat{W}_i - E[W_i] \right]$, where

$$\hat{W}_i = \frac{1}{2n_i (M+1)} \sum_{j=1}^{n_i} \left[(M-1) G_{ij} - \sum_{i' \neq i} \frac{n_{i'1}}{n_{i'}} \right] \left[\hat{F} \left(X_{ij} \right) + \hat{F} \left(X_{ij} - \right) \right],$$

$$\hat{F} = \frac{\sum_{i=1}^{M} n_i F_i}{n},$$

and

$$E[W_i] = \frac{M}{2(M+1)} \left[\frac{n_{i1}}{n_i} - \frac{1}{M} \sum_{i'=1}^{M} \frac{n_{i'1}}{n_{i'}} \right].$$

4.4.2 Rank sum test for IWCGS

Dutta and Datta [16] demonstrate that the cluster-weighted rank sum test [10] can be biased in the presence of IWCGS, and propose a modified test weighted by the intra-cluster group size. This group-weighted statistic is derived in a similar fashion to that in 4.4.1. The Wilcoxon rank sum statistic W^*, defined as in 4.4.1, is applied to a resampled data set. The only difference is the pseudo data set $(X_i^*, G_i^*), i = 1, \ldots, M$, is based in a two-step resampling method, previously described in Section 2.5. The expectation calculation of (4.4) is then applied to W^*, with respect to the modified resampling. The result is defined below.

Let $\mathbf{X}_i^{(1)} = \left\{X_{i1}^{(1)}, \ldots, X_{in_{i1}}^{(1)}\right\}$ be the set of observations belonging to group 1 in cluster i, $\mathbf{X}_i^{(0)}$ be similarly defined for the set of observations belonging to group 0, and let $n_{i0} = n_i - n_{i1}$ denote the number of group 0 observations in cluster i. In the case of complete clusters, the group-weighted rank sum statistic can be expressed as

$$S = \frac{1}{M+1}\sum_{i=1}^{M}\left(\sum_{j=1}^{n_i}\frac{1}{2n_{i1}}\left[1 + \frac{1}{2}\sum_{i'\neq i}\left[F_{i'}\left(X_{ij}^{(1)}\right) + F_{i'}\left(X_{ij}^{(1)}-\right)\right]\right]\right),$$

where

$$F_i(x) = \frac{1}{2n_{i1}}\sum_{k=1}^{n_{i1}} I\left[X_{ij}^{(1)} \leq x\right] + \frac{1}{2n_{i0}}\sum_{k'=1}^{n_{i0}}\sum_{k'=1}^{n_{i0}} I\left[X_{ik'}^{(0)} \leq x\right].$$

The expected value of S under the null hypothesis is $E[S] = \frac{M}{4}$, and the variance can be estimated using a delete-one-cluster jackknife technique. The standardized form of the statistic can then be calculated using (4.3). In the case of data with incomplete clusters, Dutta and Datta [16] provide a modification of the statistic which is omitted here for brevity.

4.4.3 Signed-rank test

The reweighted signed-rank test [11] is used to test for marginal symmetry of paired clustered observations with ICS. Let X_{ij} be the pair-specific difference in the outcome of interest and $F(x) = E\left[\frac{1}{n_i}\sum_{j=1}^{n_i} I[X_{ij} \leq x]\right]$. The null hypothesis is $H_0 : F$ is symmetric around 0 against the alternative that F is not symmetric. As in 4.4.1 and 4.4.2, the cluster-weighted signed-rank statistic is derived by first considering a resampled data set $X_i^*, i = 1, \ldots, M$, from a single WCR iteration. The traditional signed-rank statistic, $W^* = \sum_{i=1}^{M} R_i^{*+}\text{sign}(X_i^*)$, is then calculated from this resampled data, where $R_i^{*+} = \frac{1}{2}\left[\sum_{i'=1}^{M} I\left[|X_{i'}^*| \leq |X_i^*|\right] + \sum_{i'=1}^{M} I\left[|X_{i'}^* < |X_{i'}^*|\right]\right]$ and $\text{sign}(x) = I[x > 0] - I[x < 0]$. The conditional expectation calculation of (4.4) applied to W^* leads to:

$$S = \sum_{i=1}^{M}\left(\frac{n_i^+ - n_i^-}{n_i}\right) + \sum_{i=1}^{M}\frac{1}{n_i}\sum_{j=1}^{n_i}\text{sign}(X_{ij})\,\hat{D}_i(|X_{ij}|),$$

where

$$\hat{D}_i(x) = \sum_{i' \neq i} \hat{H}_{i'}(x),$$

$$\hat{H}_i(x) = \frac{1}{2n_i}\left[\sum_{j=1}^{n_i} I\left[|X_{ij}| \leq x\right] + \sum_{j=1}^{n_i} I\left[|X_{ij}| < x\right]\right],$$

$$n_i^{*+} = \sum_{j=1}^{n_i} I\left[X_{ij} > 0\right],$$

$$n_i^{*-} = \sum_{j=1}^{n_i} I\left[X_{ij} < 0\right].$$

The variance of S is estimated by the summands of the Hajek projection of S, $\hat{V}(S) = \sum_{i=1}^{M} \hat{\sigma}_i^2$, where

$$\hat{\sigma}_i = \frac{n_i^+ - n_i^-}{n_i} + \left(\frac{M-1}{n_i}\right) \sum_{j=1}^{n_i} \text{sign}\left(X_{ij}\right) \hat{H}\left(|X_{ij}|\right),$$

$$\hat{H}(x) = \sum_{i=1}^{M} \frac{n_i \hat{H}_i(x)}{n}.$$

As $E[S] = 0$ under the null hypothesis, the standardized test statistic of (4.3) simplifies to $Z = \frac{S}{\sqrt{\hat{V}(S)}}$.

4.5 Tests of variance homogeneity

In Sections 4.3 and 4.4, we presented hypothesis tests related to central tendency. Other analyses might be concerned with dispersion of groups defined within clusters. In the i.i.d. setting, tests of central tendency can be contingent on relative variability among groups. In the clustered data setting, assessing equality of variance is an important element in genetic modeling of twin data [32]. In this section, we apply the reweighting methodology to tests of variance of intra-cluster groups. We first focus on tests of equality of variances for two groups, resulting in analogs of the classical F and Levene's tests. We then extend the method to K groups, deriving a reweighted analog of Bartlett's test.

4.5.1 Test for 2 groups using moments - F test analog

In the i.i.d. setting, letting σ_k^2 denote the variance of population k, the classical F test assesses the hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ by leveraging standardization of normally distributed random variables and the relationship between the chi square and F distributions. The distributional foundations of this test make deriving a reweighted analog though conditional expectation calculations difficult. Instead, a reweighted test of variance equality between two groups is more easily approached through the estimation of moments, as in Lorenz et al. [40].

In the clustered setting, the marginal moment for group k is defined $m_a^{(k)} = E\left[(X^{(k)})^a\right]$ and is estimated by

$$\hat{m}_a^{(k)} = \frac{1}{\tilde{M}^{(k)}} \sum_{i=1}^{M} w_{ij} \sum_{j=1}^{n_i^{(k)}} \left(X_{ij}^{(k)}\right)^a$$

where w_{ij} is the weight defined by (2.5) and $\tilde{M}^{(k)}$ is defined as in (4.2). The marginal variance of group k can be expressed $\sigma_k^2 = E\left[(X^{(k)})^2\right] - \left(E\left[X^{(k)}\right]\right)^2$. Let $\hat{\boldsymbol{m}} = \left(\hat{m}_1^{(1)}, \hat{m}_1^{(2)}, \hat{m}_2^{(1)}, \hat{m}_2^{(2)}\right)$. We seek to test $H_0 : \sigma_1^2 = \sigma_2^2$. Letting the vector $\boldsymbol{s} = (s_1, s_2, s_3, s_4)$ represent the four moments, define

$$F_r = g_r(\boldsymbol{s}) = \frac{s_3 - s_1^2}{s_4 - s_2^2} \tag{4.5}$$

We can estimate F_r by applying formula (4.5) to the first and second raw sample moments of groups 1 and 2: $\hat{F}_r = g_r(\hat{\boldsymbol{m}})$. Under the null hypothesis of equality of group variances, $E[F_r] = 1$ and we can test H_0 by comparing $\frac{\hat{F}_r - 1}{\hat{V}\left(\hat{F}_r\right)}$ to the quantiles from the standard normal distribution. $\hat{V}\left(\hat{F}_r\right)$, an estimate of the variance of $\hat{F}_r$, is obtained using the jackknife method.

Note that in contrast to the other statistics presented thus far, $\hat{F}_r$ is not the result of a conditional expectation calculation performed on the traditional F statistic calculated from a single resampling. Instead, the $\hat{m}_a^{(k)}$ estimates are the result of the conditional expectation detailed in Section 2.5.2 applied to the first and second sample

moments calculated from a resampled data set, and $\hat{F}_r$ is a smooth function of these estimators.

$\hat{F}_r$ is the natural functional form for the reweighted F statistic as it mimics the ratio form of the classical test. An alternative statistic can be defined based on differences in variations rather than ratios,

$$F_d = g_d(\boldsymbol{s}) = \left(s_3 - s_1^2\right) - \left(s_4 - s_2^2\right)$$

Under H_0, $E[F_d] = 0$. We can test H_0 by applying this form to the vector of reweighted moments, $\hat{F}_d = g_d(\hat{\boldsymbol{m}})$, once again using the standardized form $\frac{\hat{F}_d}{\sqrt{\hat{V}(F_d)}}$ and calculating $\hat{V}(F_d)$ using the jackknife method.

While testing $H_0 : F_r = 1$ is tantamount to testing $H_0 : F_d = 0$, the convergence in distribution of $\hat{F}_r$ and $\hat{F}_d$ may not occur at the same rate. Therefore, we compare the performance of the respective tests of $\hat{F}_r$ and $\hat{F}_d$ through simulations in Section 4.7.4.

4.5.2 Test for 2 groups using transformations - Levene test analog

The traditional F test is well-known to be heavily reliant on the assumption that observations are normally distributed, and performs poorly when this assumption is violated. A robust alternative is the test by Levene [37], which implements a one-way ANOVA on the centered absolute values of observations. That is, for independent observations $X_i^{(k)}$, Levene's test transforms the data to $Z_i^{(k)} = |X_i^{(k)} - \bar{X}^{(k)}|$ where $\bar{X}^{(k)}$ is a measure of central tendency for group k. An ANOVA-based F test is then performed on these transformed values. Levene originally proposed centering based on group means, but it has been demonstrated that centering around trimmed means or medians can offer improved performance when data are not normally distributed. The ability to assess equality of variances when data is non-normally distributed has made Levene's test a standard addition to statistical practice.

Iachine et al. [32] extended Levene's test to clustered data by noting the correspondence between ANOVA and linear regression. For independent observations, rather than applying ANOVA to the transformed values, a test for variance equality could instead be performed by modeling

$$E[Z^{(k)}] = \beta_0 + \beta_1 I[k = 2], k = 1, 2 \tag{4.6}$$

and testing $H_0 : \beta_1 = 0$ using a Wald test. Iachine et al. leverage the regression form and test variance equality in clustered data by performing the regression (4.6) through a cluster-appropriate modeling method, e.g., GEE. They demonstrate through simulations that, with appropriate sample size, this clustered version of Levene's test closely maintains nominal size for clustered data under both normality and non-normality, with the transformation based on trimmed mean exhibiting the best performance.

As has been previously discussed, GEE models can produce biased estimates when data have cluster- or group-size informativeness. This makes the clustered variant of Levene's test a potentially poor choice for data with plausible ICS/IWCGS. We introduce a Levene's test analog that accounts for informativeness by observing that a Wald test of the group indicator coefficient from regression model (4.6) is equivalent to performing a test of mean equality of the Z-transformed data between the two groups. We previously constructed a reweighted test of mean equality in Section 4.3.2. By similarly transforming our data as is done in the classical and clustered Levene's tests, we can test $H_0 : \sigma^2_{(1)} = \sigma^2_{(2)}$ by applying the reweighted two-sample t-test analog to the transformed data.

This method is conducted as follows. As in the classical and clustered Levene's tests, define a new variable $Z^{(k)}_{ij} = |X^{(k)}_{ij} - \tilde{\theta}^{(k)}|$, where $\tilde{\theta}^{(k)}$ is a measure of central tendency for group k. Let $\mu^{(k)} = E[Z^{(k)}_{ij}]$, and $\hat{\mu}^{(k)}$ represent the reweighted mean of the transformed $Z^{(k)}_{ij}$ values using the reweighting method of Sections 2.5.2. That is, for complete clusters, $\hat{\mu}^{(k)} = \frac{1}{M}\sum_{i=1}^{M} \frac{1}{n_i^{(k)}} \sum_{j'=1}^{n_i^{(k)}} Z^{(k)}_{ij'}$. This form can be similarly

modified under incomplete clusters in the manner previously described. We then test equality of marginal variances by standardizing $T = \hat{\mu}^{(1)} - \hat{\mu}^{(2)}$, $\frac{T}{\hat{V}(T)}$, estimating $\hat{V}(T)$ using the jackknife method. Under the null hypothesis, $\frac{T}{\hat{V}(T)}$ asymptotically follows the standard normal distribution.

In correspondence with the original Levene's test and the methods of Iachine et al. [32], we compare the performance of this method using three forms of $\tilde{\theta}^{(k)}$ corresponding to a mean, trimmed mean, and median. The weighted group means $\tilde{\theta}^{(k)} = \hat{\theta}^{(k)}$ have previously been defined in (4.2). We estimated trimmed mean and median as functionals from the reweighted emperical CDF (2.5) in the manner of Nevalainen et al. [43]. This is conducted by calculating the empirical CDF and identifying the weighted quantiles (median, $\alpha 100\%$ tails), with the α-trimmed mean defined as the weighted mean after removing the $\alpha 100\%$ upper and lower tails.

4.5.3 Extension to K groups

Testing equality of variance among K groups in the classical setting is done through extensions of the F-test (Bartlett's test) or Levene's test. While the clustered Levene test by Iachine et al. [32] was presented in the context of assessing variance equality for 2 groups, it can be used to compare variances across K groups by fitting a GEE model to the group factor and testing significance of that factor. Similarly, we can extend the reweighted method in the previous section to the K group setting. Recognizing the relationship between the reweighted 2-group and K-group tests of means, this is a straightforward process. By the same principal that motivates a 2-sample test of means on the transformed Z values in the two group case, we can test equality of K variances by implementing the test of K mean equality from Section 4.3.3 on the transformed values. Let $\hat{\boldsymbol{\mu}} = (\hat{\mu}^{(1)}, \ldots, \hat{\mu}^{(K)})^T$, where $\hat{\mu}^{(k)}$ represents the weighted group means of the transformed Z-values as defined in the section above. To test

$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_K^2$, we use the statistic

$$X^2 = M\,(C\hat{\boldsymbol{\mu}})^T \left(C\hat{\Sigma}C\right)^{-1} (C\hat{\boldsymbol{\mu}})$$

where C is the contrast matrix defined in Section 4.3.3 and $\hat{\Sigma}$ is the jackknifed variance estimate of $\hat{\boldsymbol{\mu}}$.

4.6 Tests of correlation

A number of marginal correlation estimators for clustered data with potential ICS have been developed. For paired clustered data, Lorenz et al. [39] proposed analogs of the Pearson and Kendall correlation coefficients. An additional paper expanded upon that work [40], generalizing the reweighting of paired correlation estimators and extending the method to develop marginal correlation estimators for data unpaired at the cluster level, including the development of Spearman coefficient analog. We restrict our attention to correlation estimators for paired clustered data as they are the most natural and interpretable. As such, we will adapt the Spearman coefficient proposed in the unpaired case to the paired case. Hypothesis tests based on these estimators and their variance estimates can be applied in the usual manner using the standardized test statistic (4.3).

Let (X_{ij}, Y_{ij}) be the j^{th} bivariate observation from cluster i, and let M and n_i be defined as above. The reweighted Pearson and Kendall coefficients for paired clustered data [39] are based on Equation (4.4) applied to a resampled data set in which one paired observation has been selected at random from each cluster.

The marginal Pearson correlation analog for paired data is expressed using the standard product moment formula

$$\rho = g(\boldsymbol{w}) = \frac{w_3 - w_1 w_2}{\sqrt{(w_4 - w_1^2)(w_5 - w_2^2)}} \tag{4.7}$$

The cluster-weighted estimator of ρ is $\hat{\rho}_p = g(\hat{\boldsymbol{w}})$, where $\hat{\boldsymbol{w}} = (\hat{w}_{10}, \hat{w}_{01}, \hat{w}_{11}, \hat{w}_{20}, \hat{w}_{02})$ and $\hat{w}_{kl} = \frac{1}{M}\sum_{i=1}^{M} \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}^k Y_{ij}^l$. Like $\hat{F}_r$ and $\hat{F}_d$ from Section 4.5.1, $\hat{\rho}_p$ is not

obtained directly from a conditional expectation calculation, but instead is a smooth function of reweighted estimates of the first and second sample moments. Defining $\hat{w}_{kli} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}^k Y_{ij}^l$ and $\hat{\boldsymbol{w}}_i = (\hat{w}_{10i}, \hat{w}_{01i}, \hat{w}_{12i}, \hat{w}_{20i}, \hat{w}_{02i})$, the variance of $\hat{\rho}_p$ can be estimated using the empirical variance-covariance matrix of $\hat{\boldsymbol{w}}_i$ in conjunction with the delta method.

In addition to the cluster-weighted Pearson-type estimator, Lorenz et al. [39] develop a non-parametric marginal correlation estimator for paired clustered data analogous to the standard Kendall correlation estimator. The cluster-weighted Kendall estimator is derived through the U-statistic formulation of the Kendall coefficient, and is expressed as

$$\hat{\tau} = 2\binom{M}{2}^{-1} \sum_{i=1}^{M-1} \sum_{i' \neq i}^{M} \frac{1}{n_i n_{i'}} I\left[\left(Y_{i'n'} - Y_{ij}\right)\left(X_{i'j'} - X_{ij}\right) > 0\right] - 1 \tag{4.8}$$

The asymptotic variance of 4.8 can be estimated using the Hajek projection of $\hat{\tau}$, and takes the form $\hat{\sigma}_\tau^2 = \frac{M}{M-1}\sum_{j=1}^{n_i}\left(\hat{S}_i - \bar{S}\right)^2$, where

$$\hat{S}_i = \frac{4}{M n_i} \sum_{j=1}^{n_i} \left(\hat{F}^l\left(X_{ij}, Y_{ij}\right) + \hat{F}^r\left(X_{ij}, Y_{ij}\right)\right),$$

$$\hat{F}^l(x, y) = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{n_i}\sum_{j=1}^{n_i} I\left[X_{ij} < x, Y_{ij} < y\right],$$

$$\hat{F}^r(x, y) = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{n_i}\sum_{j=1}^{n_i} I\left[X_{ij} > x, Y_{ij} > y\right],$$

and

$$\bar{S} = \frac{1}{M}\sum_{i=1}^{M}\hat{S}_i.$$

The Spearman coefficient can be expressed using the product moment formula (4.7), where the expression is evaluated at the rank moments. While Lorenz et al. [40] only explicitly state a Spearman coefficient analog for the case of unpaired clustered data, their methodology is easily modified to accommodate paired data. The weighted rank functions previously used in the rank sum and signed rank tests from

Sections 4.4.1 and 4.4.3 are used to define the rank of X_{ij} among all X observations as $R_{X_{ij}} = \frac{1}{2}\left[\hat{F}_X\left(X_{ij}\right) + \hat{F}_X\left(X_{ij}-\right)\right]$, where $\hat{F}_X\left(x\right) = \frac{1}{M}\sum_{i=1}^{M} F_{X_i}(x)$ and $F_{X_i}\left(x\right) = \sum_{j=1}^{n_i} I\left[X_{ij} \leq x\right]$. Functions for $R_{Y_{ij}}$, $\hat{F}_Y\left(y\right)$, and $F_{Y_i}\left(y\right)$ are similarly defined. The paired rank moments are expressed as $\hat{r}_{kl} = \sum_{i=1}^{M}\sum_{j=1}^{n_i} \frac{1}{n_i} R_{X_{ij}}^k R_{Y_{ij}}^l$, and the paired marginal Spearman correlation analog is then $\hat{\rho}_s = g\left(\hat{r}_{10}, \hat{r}_{01}, \hat{r}_{11}, \hat{r}_{20}, \hat{r}_{02}\right)$. A variance estimate for $\hat{\rho}_s$ can be obtained using a delta method calculation.

4.7 Simulations

We evaluated the performance of the novel tests of means from Section 4.3 and the tests of variance from Section 4.5 through a simulation study. We compared size and power of these tests to their classical counterparts for independent observations and a cluster-appropriate alternative. Empirical size and power for each method was calculated as the number of rejections of the null hypothesis for each test at a nominal level of .05 over 10,000 Monte Carlo iterations. To evaluate the effect of sample size on the proposed tests, we performed each simulation scenario for $M = 30, 50$, and 100 clusters.

4.7.1 Simulation design for tests of means

To simulate quantitative data with an informativeness structure related to group means, we generated multivariate random effects $(u_i^{(1)}, \ldots, u_i^{(K)})$ from a multivariate normal distribution $N_K(\mathbf{0}, \Sigma_K)$, where Σ_K is a KxK symmetric matrix with 1 on the diagonal and .2 on the off-diagonal. Within-cluster group sizes were generated as $n_i^{(k)} \sim \text{Poisson}(5 + 5 * u_i^{(k)})$, with cluster size defined as $n_i = \sum_{k=1}^{K} n_i^{(k)}$. In the event $n_i = 0$, group sizes were re-simulated from $n_i^{(k)} \sim \text{Poisson}(5+5*u_i^{(k)})+1$. We defined a categorical random variable G_{ij}, taking values $\{1, 2, \ldots, K\}$, so that $n_i^{(k)}$ observations belonged to group k. Let $\boldsymbol{x}_{ij} = (x_{ij}^{(1)}, \ldots, x_{ij}^{(K-1)})^T$, where $x_{ij}^{(k)} = I[G_{ij} = k]$. We defined the quantitative variable $t_{ij} = u_i^{(G_{ij})} + e_{ij} + \sum_{k=1}^{K-1} \delta^{(k)} I[G_{ij} = k]$, where e_{ij}

were simulated from $N(0,1)$ independently of u_i and $n_i^{(k)}$, and $\delta^{(k)}$, $k = 1, \ldots, (K-1)$ are parameters used to induce differences in group means for the purpose of assessing power. Simulations for the 2-sample t-test analog from Section 4.3.2 correspond to $K = 2$, while simulations for the ANOVA analog test of 4.3.3 were run for $K = 3$ and $K = 5$. We performed simulations for size, corresponding to $\delta^{(k)} = 0$ for all k, and three forms of power. To evaluate power when $K = 2$, we set $(\delta_1, \delta_2, \delta_3) = (.25, .5, .75)$. For $K = 3$, $\delta_1 = (\delta_1^{(1)}, \delta_1^{(2)}) = (.14, -.14)$, $\delta_2 = (.28, -.28)$, and $\delta_3 = (.42, -.42)$. For $K = 5$, $\delta_1 = (.10, .15, -.10, -.15)$, $\delta_2 = (.20, .30, -.20, -.30)$, and $\delta_3 = (.30, .45, -.30, -.45)$. Under this design, the data are normally distributed with variance of 2 and equal means when δ is 0, and different means when $\delta \neq 0$. Informativeness is induced through the relationship between the random effects and group sizes. Within each cluster, groups with large positive (negative) random effects tend to have more (fewer) observations, and those observations tend to have larger (smaller) values.

4.7.2 Simulation results for tests of means

Table 10 contains the results for the reweighted 2-group t-test analog from 4.3.2, and Table 11 contains results for the test of K-group equality from Section 4.3.3. In both tables, we compare the reweighted tests (RW) with their classical analog (UW) and a GEE model using exchangeable correlation structure (GEE). The heading δ_0 indicates size, while $\delta_d, d = 1, 2, 3$ denotes the three power simulations. As would be expected, the traditional two-sample t-test and ANOVA tests that ignore cluster membership are heavily biased. The GEE models exhibited mild to moderate bias, with bias increasing with the number of groups and decreasing with sample size. The reweighted tests maintained appropriate size when $K = 2$ and 3 for all M. When $K = 5$, the reweighted ANOVA-analog test was mildly biased under the smallest sample, but exhibited appropriate size as the number of clusters increased.

Table 10. 2-sample test of means; empirical size and power.

M	Test	δ_0	δ_1	δ_2	δ_3
30	RW	0.0535	0.1588	0.4589	0.7897
	GEE	0.0850	0.2008	0.4938	0.7902
	UW	0.3602	0.5001	0.7763	0.9406
50	RW	0.0495	0.2257	0.6656	0.9463
	GEE	0.0730	0.2437	0.6529	0.9287
	UW	0.3539	0.5822	0.8888	0.9865
100	RW	0.0486	0.4005	0.9247	0.9985
	GEE	0.0602	0.3924	0.8985	0.9971
	UW	0.3563	0.7330	0.9801	0.9997

RW, reweighted test; GEE, GEE model; UW, unweighted classical t-test.

Table 11. K-sample test of means; empirical size and power.

M	Test	$K=3$				$K=5$			
		δ_0	δ_1	δ_2	δ_3	δ_0	δ_1	δ_2	δ_3
30	RW	0.0519	0.1370	0.4202	0.7640	0.0605	0.1663	0.5339	0.8713
	GEE	0.1167	0.2160	0.5121	0.8116	0.1903	0.3353	0.6918	0.9296
	UW	0.5189	0.6543	0.8622	0.9729	0.7250	0.8342	0.9663	0.9978
50	RW	0.0515	0.2020	0.6447	0.9430	0.0576	0.2419	0.7639	0.9837
	GEE	0.0871	0.2607	0.6734	0.9432	0.1290	0.3498	0.8134	0.9874
	UW	0.5284	0.7117	0.9424	0.9958	0.7309	0.8825	0.9912	0.9999
100	RW	0.0547	0.3709	0.9236	0.9987	0.0527	0.4649	0.9796	1.0
	GEE	0.0711	0.3889	0.9159	0.9991	0.0848	0.4870	0.9773	1.0
	UW	0.5308	0.8284	0.9944	1.0	0.7241	0.9497	0.9998	1.0

RW, reweighted test; GEE, GEE model; UW, unweighted classical ANOVA test.

4.7.3 Simulation design for tests of variance

We compared performance of the variance tests using a random effects design modified from the previous simulation. Group random effects $(u_i^{(1)}, \ldots, u_i^{(K)})$ were simulated from a multivariate normal distribution $N_K(\mathbf{0}, \Sigma_K)$, where $\Sigma_K = \sigma^2 * I_k$, σ^2 was set to 4, and I_k denotes the $K \times K$ identity matrix. We ran simulations for $K = 2, 3,$ and 5 groups. Group sizes were generated from

$$n_i^{(k)} \sim \text{Poisson}(8 + c_k I[|u_i^{(k)}| > \sigma])$$

where

$$c_k = \begin{cases} 4, \text{if } k \text{ odd} \\ -4, \text{ otherwise.} \end{cases}$$

Cluster size was defined as $n_i = \sum_{k=1}^{K} n_i^{(k)}$. In the event any $n_i = 0$, group sizes for the empty clusters were re-simulated. When $K = 2$, group sizes were re-simulated as $n_i^{(k)} \sim \text{Poisson}(8 + c_k I\left[|u_i^{(k)}| > \sigma\right]) + 1$. For the $K = 3, 5$ scenarios, empty clusters had a random group selected uniformly from $1 : K$, and $n_i^{(k)}$ was simulated from $\text{Poisson}(8) + 1$.

A categorical group variable G_{ij} taking values $1, 2, \ldots, K$ was defined so that in cluster i, $n_i^{(k)}$ observations belonged to group k. We defined $W_{ij} = u_i^{(G_{ij})} + e_{ij}$, where $e_{ij} \sim N(0, 1)$. As $u_i^{(k)}$ and e_{ij} were simulated independently, $W_{ij} \sim N(\sigma^2 + 1)$. Following the simulation design of Iachine et al. [32], we applied a non-linear transformation defined by

$$X_{ij} = a^{(G_{ij})} g(W_{ij}). \tag{4.9}$$

Here, $a^{(k)}$ is a scalar that generates differences in group variances. When the ratio $\left(a^{(k)}\right)^2 : \left(a^{(k')}\right)^2 = 1$ for all k, k', simulations were under the null hypothesis of equal group variances. Power was assessed by varying the $a^{(k)}$ values across the K groups. When $K = 2$, power simulations were evaluated for ratios $1.5, 2.0$, and 2.5. To evaluate power when $K = 3$ and 5, the values $\boldsymbol{a}^2 = \left((a^{(1)})^2, \ldots, (a^{(K)})^2\right)$ were the columns of the matrices

$$\begin{bmatrix} 1.2 & 1.4 & 1.6 \\ 1.1 & 1.2 & 1.3 \\ 1.0 & 1.0 & 1.0 \end{bmatrix},$$

and

$$\begin{bmatrix} 1.3 & 1.6 & 1.9 \\ 1.2 & 1.4 & 1.6 \\ 1.1 & 1.2 & 1.3 \\ 1.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 1.0 \end{bmatrix}.$$

In the classical setting, Levene's test is preferential to the F test due to its robustness against non-normality. To evaluate the performance of the reweighted tests of variance under varied distributions, we ran simulations for three marginal distributions. The data X_{ij}, as defined in (4.9), was generated though a transformation $g(x)$, which induced the distinct distributions. The transformation $g(x)$ was defined

$$g(x) = \begin{cases} x, & \text{if } X_{ij} \sim N(0, \sigma^2 + 1) \\ F_{t_4}^{-1}(\Phi_{\sigma^2+1}(x)), & \text{if } X_{ij} \sim t_4 \\ F_{\chi_4^2}^{-1}(\Phi_{\sigma^2+1}(x)), & \text{if } X_{ij} \sim \chi_4^2 \end{cases}$$

where $\Phi_{\sigma^2+1}(x)$, F_{t_4}, and $F_{\chi_4^2}$ are the cumulative distribution functions for the $N(0, \sigma^2+1)$, Student's t_4, and χ_4^2 distributions, respectively. These distributions allow us to evaluate the relative performance of tests when data are normal, symmetric with fat tails, and skewed, respectively.

Informativeness is induced in this simulation design through the relationship between the random effects and group sizes. Unlike previous designs, the influence of the random effect on group size differs across groups. In the 2-group simulations, clusters with large absolute $u_i^{(1)}$ values, e.g., values of $|u_i^{(1)}| > 2$, have group 1 sizes simulated from Poisson(12) compared to Poisson(8) for clusters with less extreme $u_i^{(1)}$ values. In contrast, large absolute random effects have a diminishing effect on group size for group 2 observations, resulting in group sizes from Poisson(4) compared to Poisson(8). This differing response produces a disproportionate number of group 1 observations from the extremes of the distribution, while the majority of group 2 observations have more moderate values. Similar variation in the random effect-group size relationship was generated in the $K = 3$ and $K = 5$ scenarios. This falsely inflates an appearance of unequal spread across the different groups.

4.7.4 Simulation results for tests of variance

Table 12 contains simulation results for the 2-group tests of variance equality. This table is organized into three sections corresponding to the three marginal distributions. The first column denotes the number of clusters for the simulation, and the second column denotes the size or power setting. The following four columns contain results for the classical F (F) and Levene's tests, where the Levene's test has been constructed under the unweighted group mean (W_0), trimmed mean (W_{10}), and median (W_{50}). The next three columns contain results for the clustered version of Levene's test implemented by GEE [32], which we denote GEE_0, GEE_{10}, and GEE_{50} corresponding to the three unweighted measures of central tendency. The final five columns contain the reweighted tests: both forms of the F test analog, ratio (CF_r) and difference (CF_d); and the three forms of the Levene analog: CW_0, CW_{10}, and CW_{50}, with subscripts denoting the respective reweighted measure of centers. For clarity, we have differentiated these three categories of tests by the headings "classical", for the standard tests; "clustered", denoting the cluster-appropriate variants; and "reweighted" for the reweighted tests that correct for informativeness. All trimmed means were calculated with a 10% trim from each end.

As would be expected, the conventional tests ignoring clustering exhibited extreme bias. Power for these tests have been suppressed for clarity, as they are of little interest. All forms of the clustered Levene's test were consistently biased across distributions. These tests displayed a two to three fold increase in Type I error rate under the smallest sample, and still higher rates of error for larger samples. This contradictory behavior highlights the unsuitability of these methods for data under informativeness.

The reweighted tests performed comparably under the Gaussian distribution, maintaining close to nominal size except for CF_r, which was mildly biased at smaller sample sizes. The ratio form of the reweighted F-test analog retained this bias across

distributions and exhibited poor power compared the alternative reweighted forms. In contrast, the reweighted F-test analog based on a difference, CF_d, not only had the highest power of the reweighted tests under the Gaussian distribution, but was also maintained size under the skewed χ^2_4 distribution. This form was conservative under the t_4 simulations, which is likely related to overestimation of the standard errors due to the heavy-tailed nature of this distribution [48]. The forms of the reweighed Levene's test based on mean and 10% trimmed mean maintained size and performed similarly across power under the t_4 distribution, but both exhibited nominal sizes higher than 5% under the skewed transformation, with the test based on the trimmed mean having the lesser bias. The reweighted Levene test based on the median maintained size with a slight power disadvantage under the symmetric distributions, but offered the best overall performance under the χ^2_4 distribution, closely maintaining size and having higher power than CF_d.

Tables 13 and 14 contain results for the K-group test of variance homogeneity, where $K = 3$ and 5, respectively. These tables are similarly formatted as Table 12, with the column headed by $\boldsymbol{a}$ denoting the size (a_0) and three power scenarios. We compare the classical Bartlett test to the clustered GEE-based Levene tests, and the reweighted tests. As before, the GEE-based tests are denoted by "GEE" and the reweighted tests by "CW". The subscripts indicate the measure of center for the respective data transformation, with 0 being the mean, 10 denoting the mean with 10% trim from each end, and 50 the median. We forgo the classification headings from Table 12 as there are fewer tests to be compared, but continue to suppress empirical power from the conventional Bartlett test.

Results from the tests of variance homogeneity for K groups are similar to those from the 2 groups simulation. The standard Bartlett test is biased for all scenarios, due both to the clustered nature of the data and the informativeness. The cluster-appropriate methods fail to maintain size under informativeness, exhibiting a

minimum four fold increase of Type I error rate, with this bias increasing with sample size and K. When assessing variance equality of three groups under the Gaussian and t_4 distributions, all forms of the reweighted test are appropriate and have comparable size for a large number of clusters. CW_0 and CW_{10} are mildly biased under smaller samples for these distributions, while CW_{50} performs reasonably well even when $M = 30$. For the χ^2_4 distribution, all forms of the reweighted test are biased, with the form using the median as central tendency showing the least bias. Results for $K = 5$ closely mirror those of $K = 3$, with the additional groups exasperating the biasing effect of distribution and small sample size on the reweighted tests.

4.7.5 Supplemental simulations

We ran additional simulations to compare the performance of the reweighted tests to GEE models when data have uninformative cluster or group size. We contrasted the reweighted two-sample test of means and two-sample test of variance to their GEE counterparts. Non-informative designs for both tests were consistent with their respective informative designs previously described, with the exception that group sizes were simulated independently from random effects. For the test of means, $n_i^{(k)}$ was simulated from Poisson(7) for $k = 1, 2$. For the tests of variance, data were simulated from the Gaussian distribution, and group sizes $n_i^{(1)}$, $n_i^{(2)}$ were simulated from Poisson(10) and Poisson(6), respectively.

Table 15 contains the non-informative simulation results. The reweighted and GEE tests of means are presented in the top part of the table, while tests of variance are compared in the lower portion of the table. Values under the heading δ_0 correspond to empirical size, while those under the alternative headings correspond to the three power settings from the respective simulation design. The reweighted tests outperformed GEE-based methods in maintaining nominal size when testing both equality of group means and variances. All reweighted tests were approximately un-

biased for all sample sizes, while GEE-based tests under both simulations were mildly biased for 30 and 50 clusters. Additionally, even when GEE methods controlled the Type I error rate, they offered no advantage in power compared to the corresponding reweighted test(s).

4.8 Discussion

In this chapter, we developed clustered data analogs of well-known and frequently performed tests for independent quantitative observations. These clustered tests mirror the 2-sample t-, one-way ANOVA, F-, and Levene's tests found in the classical statistical literature. The tests developed in this chapter are reweighted to correct for group-size informativeness and avoid assumptions on completeness of group structure by estimating variance through a delete-one-cluster jackknife technique. We demonstrate through simulation that these reweighted tests maintain appropriate size, while other cluster-appropriate methods are biased when data have informativeness.

The multiple forms for 2-group tests of variance in the classical setting result from distributional assumptions. The standard F test requires observations be normally distributed, whereas Levene's test is robust against non-normality but selection of measure of center (i.e., mean, trimmed mean, median) is distributionally dependent. In contrast, the marginalization principle that results in the reweighted analog tests is nonparametric. That is, none of the reweighted tests of variance make any assumptions about the underlying distribution structure of the clustered data. In some cases, this leads to the reweighted tests performing in contrast to what would be expected of their classical forms in an unclustered setting. For example, the robust performance of CF_d to skewed data under the χ^2_4 distribution (Table 12). Despite their nonparametric nature, we see from Table 12 that the reweighted tests are still somewhat distributionally dependent. This is not surprising, as the rate at which reweighted estimators converge to normality would be expected to be related to the

underlying distribution. An additional layer of convergence needs to be considered when tests are executed through functionals of reweighted estimators. This is most evident in the comparison of CF_r and CF_d in Table 12. Both of these tests are reliant on the same vector of reweighted estimators, but test the hypothesis of interest by applying contrasting functions to those estimators. The differing rate of convergence of those functions to their expected values is evident through the consistently superior performance of CF_d to CF_r.

The asymptotic nature of the reweighted tests is highlighted through their improved performance with sample size, with sample size dependance being additionally increased with the number of parameters being tested. This was particularly apparent in the tests of variance. Due to this, these methods can only be recommended when collected data contain at least 30 clusters, but the number of groups being tested should additionally be considered when determining appropriate sample sizes. We note that use of a second order expansion to the jackknife variance [29] offers minor improvement to the bias observed when testing a higher number of parameters (e.g, $K = 5$) under a reduced sample size (e.g., $M = 30$).

Through simulations, we demonstrated that reweighted tests of group means and variance are clearly the optimal method when data have informativeness. Additionally, reweighted tests perform competitively for clustered data when group/cluster size is uninformative, maintaining appropriate Type I error rates and offering comparable or superior power to other cluster-appropriate methods. In contrast, GEE-based tests displayed unacceptability high levels of empirical size in simulations under informativness. In particular, we note that this bias increased with sample size for the tests assessing equality of variance, suggesting the unsuitability of GEE models for data with ICS/IWCGS is not solely related to sample size deficiencies. In summary, when clustered data have variable group or cluster size, reweighting methods that correct for informativeness should be considered.

In Sections 4.4 and 4.6 we summarized the reweighted rank-based tests and tests of correlation developed by other authors. Like the novel tests derived in this dissertation, these previously developed tests are analogous to well-known classical forms for independent observations. Combined with these summarized tests and the categorical tests of Chapter 3, the reweighted tests of group means and variances proposed in this chapter comprise a versatile collection of methods for marginal analysis of clustered data accounting for potential informativeness. As this collection may be of use in applied data analysis, we make these tests available through a comprehensive R software package detailed in the next chapter.

Table 12. 2-group tests of variance homogeneity; empirical size and power.

		Classical				Cluster			Reweighted				
M	$\sigma_1^2 : \sigma_2^2$	F	W_0	W_{10}	W_{50}	GEE_0	GEE_{10}	GEE_{50}	CF_r	CF_d	CW_0	CW_{10}	CW_{50}
Gaussian													
30	1.0 : 1	0.6723	0.7120	0.7102	0.7059	0.1649	0.1615	0.1496	0.0675	0.0496	0.0519	0.0508	0.0471
30	1.5 : 1					0.5813	0.5748	0.5570	0.0936	0.2602	0.2517	0.2500	0.2383
30	2.0 : 1					0.8535	0.8489	0.8387	0.2575	0.5831	0.5639	0.5613	0.5460
30	2.5 : 1					0.9558	0.9538	0.9488	0.4519	0.8153	0.7996	0.7956	0.7868
50	1.0 : 1	0.7971	0.8411	0.8399	0.8386	0.2336	0.2310	0.2208	0.0612	0.0527	0.0545	0.0530	0.0512
50	1.5 : 1					0.7891	0.7848	0.7768	0.1917	0.3944	0.3755	0.3734	0.3665
50	2.0 : 1					0.9701	0.9693	0.9674	0.5634	0.8127	0.7864	0.7853	0.7804
50	2.5 : 1					0.9972	0.9970	0.9962	0.8314	0.9688	0.9566	0.9561	0.9537
100	1.0 : 1	0.9362	0.9594	0.9594	0.9589	0.3929	0.3911	0.3845	0.0544	0.0518	0.0511	0.0509	0.0492
100	1.5 : 1					0.9751	0.9747	0.9741	0.5099	0.6754	0.6465	0.6437	0.6415
100	2.0 : 1					0.9999	0.9999	0.9998	0.9441	0.9826	0.9782	0.9779	0.9776
100	1.5 : 1					1	1	1	0.9976	0.9997	0.9994	0.9994	0.9993
Student's t_4													
30	1.0 : 1	0.7404	0.7022	0.6958	0.6929	0.1589	0.1461	0.1362	0.0810	0.0379	0.0544	0.0493	0.0468
30	1.5 : 1					0.4808	0.4606	0.4467	0.0505	0.1401	0.1861	0.1793	0.1721
30	2.0 : 1					0.7208	0.7049	0.6913	0.0847	0.2973	0.4019	0.3926	0.3806
30	2.5 : 1					0.8625	0.8515	0.8410	0.1498	0.4573	0.5966	0.5865	0.5766
50	1.0 : 1	0.8165	0.8123	0.8094	0.8081	0.2144	0.2045	0.1964	0.0740	0.0402	0.0541	0.0503	0.0479
50	1.5 : 1					0.6557	0.6426	0.6325	0.0642	0.1876	0.2535	0.2480	0.2418
50	2.0 : 1					0.8829	0.8778	0.8744	0.1689	0.4264	0.5825	0.5752	0.5717
50	2.5 : 1					0.9682	0.9659	0.9654	0.2936	0.6245	0.8070	0.8016	0.7984
100	1.0 : 1	0.9035	0.9376	0.9385	0.9383	0.3423	0.3349	0.3301	0.0626	0.0421	0.0508	0.0489	0.0486
100	1.5 : 1					0.9011	0.8984	0.8953	0.1524	0.3110	0.4429	0.4410	0.4381
100	2.0 : 1					0.9910	0.9906	0.9905	0.3905	0.6467	0.8461	0.8455	0.8436
100	2.5 : 1					0.9989	0.9990	0.9989	0.5796	0.8298	0.9725	0.9721	0.9721
$\boldsymbol{\chi_4^2}$													
30	1.0 : 1	0.7035	0.6986	0.6707	0.6453	0.2295	0.1684	0.1204	0.0863	0.0565	0.1101	0.0784	0.0543
30	1.5 : 1					0.5557	0.4770	0.3956	0.0654	0.1709	0.2737	0.2335	0.1768
30	2.0 : 1					0.7684	0.7017	0.6364	0.1193	0.3503	0.5036	0.4562	0.3764
30	2.5 : 1					0.8945	0.8539	0.8068	0.1910	0.5195	0.6933	0.6502	0.5784
50	1.0 : 1	0.7777	0.7936	0.7713	0.7503	0.2893	0.2200	0.1613	0.0777	0.0565	0.1158	0.0811	0.0526
50	1.5 : 1					0.7217	0.6506	0.5789	0.0859	0.2310	0.3571	0.3182	0.2524
50	2.0 : 1					0.9106	0.8797	0.8458	0.2212	0.5025	0.6803	0.6470	0.5755
50	2.5 : 1					0.9792	0.9672	0.9534	0.3569	0.7215	0.8708	0.8518	0.8049
100	1.0 : 1	0.8807	0.9163	0.9073	0.8961	0.4171	0.3323	0.2671	0.0661	0.0516	0.1083	0.0743	0.0512
100	1.5 : 1					0.9276	0.8964	0.8613	0.1929	0.3772	0.5550	0.5213	0.4453
100	2.0 : 1					0.9952	0.9926	0.9877	0.5039	0.7667	0.9048	0.8937	0.8541
100	2.5 : 1					0.9998	0.9994	0.9991	0.7266	0.9357	0.9868	0.9845	0.9772

F, F test; W_m, Levene test; GEE_m, GEE Levene test analog; CF_r, reweighted F test analog, ratio form; CF_d, reweighted F test analog, difference form; CW_m, reweighted Levene test analog. Subscript m denotes measure of center: 0, mean; 10, 10% trimmed mean; 50, median.

Table 13. K-group tests of variance homogeneity; empirical size and power, $K = 3$.

M	$\boldsymbol{a}$	Bartlett	GEE_0	GEE_{10}	GEE_{50}	CW_0	CW_{10}	CW_{50}
				Gaussian				
30	a_0	0.8320	0.2462	0.2398	0.2234	0.0635	0.0614	0.0562
30	a_1		0.2740	0.2687	0.2491	0.0909	0.0882	0.0800
30	a_2		0.3265	0.3177	0.2977	0.1522	0.1512	0.1369
30	a_3		0.3953	0.3872	0.3681	0.2427	0.2412	0.2252
50	a_0	0.9049	0.3285	0.3232	0.3103	0.0578	0.0568	0.0529
50	a_1		0.3713	0.3648	0.3532	0.0990	0.0969	0.0928
50	a_2		0.4556	0.4497	0.4372	0.2114	0.2090	0.2013
50	a_3		0.5675	0.5596	0.5464	0.3803	0.3768	0.3657
100	a_0	0.9780	0.5685	0.5652	0.5578	0.0582	0.0572	0.0561
100	a_1		0.6311	0.6284	0.6234	0.1458	0.1453	0.1432
100	a_2		0.7336	0.7304	0.7247	0.3966	0.3954	0.3892
100	a_3		0.8529	0.8510	0.8474	0.6691	0.6671	0.6620
				Student's t_4				
30	a_0	0.9090	0.2407	0.2245	0.2104	0.0686	0.0639	0.0591
30	a_1		0.2583	0.2422	0.2280	0.0877	0.0810	0.0765
30	a_2		0.2908	0.2694	0.2540	0.1273	0.1211	0.1135
30	a_3		0.3334	0.3135	0.2984	0.1817	0.1710	0.1659
50	a_0	0.9368	0.3110	0.2942	0.2842	0.0642	0.0603	0.0568
50	a_1		0.3333	0.3175	0.3058	0.0908	0.0850	0.0827
50	a_2		0.3771	0.3641	0.3543	0.1541	0.1478	0.1440
50	a_3		0.4512	0.4368	0.4264	0.2523	0.2424	0.2360
100	a_0	0.9721	0.4987	0.4874	0.4824	0.0578	0.0552	0.0547
100	a_1		0.5490	0.5411	0.5353	0.1096	0.1057	0.1041
100	a_2		0.6126	0.6035	0.5986	0.2530	0.2488	0.2458
100	a_3		0.7161	0.7093	0.7042	0.4370	0.4336	0.4304
				χ^2_4				
30	a_0	0.8794	0.3647	0.2774	0.1961	0.1566	0.1145	0.0713
30	a_1		0.3746	0.2825	0.2051	0.1806	0.1345	0.0855
30	a_2		0.4195	0.3218	0.2369	0.2320	0.1793	0.1220
30	a_3		0.4609	0.3658	0.2747	0.3067	0.2493	0.1800
50	a_0	0.9171	0.4208	0.3172	0.2345	0.1439	0.1013	0.0613
50	a_1		0.4523	0.3473	0.2645	0.1835	0.1348	0.0870
50	a_2		0.4974	0.3957	0.3079	0.2723	0.2144	0.1559
50	a_3		0.5727	0.4720	0.3763	0.3975	0.3251	0.2461
100	a_0	0.9661	0.5972	0.4893	0.3921	0.1425	0.0996	0.0617
100	a_1		0.6383	0.5309	0.4371	0.2168	0.1639	0.1098
100	a_2		0.7098	0.6082	0.5187	0.4078	0.3359	0.2601
100	a_3		0.7956	0.7122	0.6250	0.5965	0.5294	0.4420

Bartlett, classical Bartlett test; GEE_m, Levene test analog based on GEE; CW_m, reweighted Levene test analog. Subscript m denotes measure of center: 0, mean; 10, 10% trimmed mean; 50, median.

Table 14. K-group tests of variance homogeneity; empirical size and power, $K = 5$.

M	$\boldsymbol{a}$	Bartlett	GEE_0	GEE_{10}	GEE_{50}	CW_0	CW_{10}	CW_{50}
					Gaussian			
30	a_0	0.9541	0.4258	0.4171	0.3884	0.0760	0.0748	0.0673
30	a_1		0.5351	0.5278	0.4977	0.1383	0.1342	0.1226
30	a_2		0.6845	0.6719	0.6459	0.2762	0.2712	0.2554
30	a_3		0.8118	0.8029	0.7832	0.4646	0.4568	0.4348
50	a_0	0.9856	0.5589	0.5509	0.5360	0.0693	0.0673	0.0645
50	a_1		0.6934	0.6863	0.6717	0.1623	0.1591	0.1531
50	a_2		0.8533	0.8495	0.8388	0.4300	0.4249	0.4118
50	a_3		0.9525	0.9507	0.9456	0.7241	0.7202	0.7068
100	a_0	0.9990	0.8361	0.8344	0.8283	0.0581	0.0566	0.0554
100	a_1		0.9407	0.9388	0.9357	0.2853	0.2849	0.2792
100	a_2		0.9923	0.9921	0.9916	0.7604	0.7589	0.7529
100	a_3		0.9994	0.9994	0.9994	0.9669	0.9668	0.9653
					Student's t_4			
30	a_0	0.9852	0.4259	0.4009	0.3773	0.0912	0.0818	0.0744
30	a_1		0.5028	0.4778	0.4555	0.1330	0.1227	0.1153
30	a_2		0.6011	0.5786	0.5569	0.2224	0.2082	0.1986
30	a_3		0.7073	0.6832	0.6640	0.3441	0.3288	0.3150
50	a_0	0.9945	0.5324	0.5152	0.5029	0.0819	0.0765	0.0734
50	a_1		0.6222	0.6038	0.5918	0.1354	0.1281	0.1235
50	a_2		0.7539	0.7396	0.7296	0.3022	0.2925	0.2841
50	a_3		0.8641	0.8559	0.8466	0.5092	0.4979	0.4906
100	a_0	0.9979	0.7770	0.7707	0.7662	0.0678	0.0648	0.0615
100	a_1		0.8712	0.8643	0.8591	0.1952	0.1896	0.1870
100	a_2		0.9544	0.9533	0.9508	0.5130	0.5069	0.5032
100	a_3		0.9859	0.9847	0.9844	0.8033	0.8001	0.7965
					$\boldsymbol{\chi^2_4}$			
30	a_0	0.9801	0.5659	0.4457	0.3353	0.2218	0.1557	0.1011
30	a_1		0.6338	0.5226	0.4063	0.2808	0.2026	0.1349
30	a_2		0.7214	0.6216	0.5119	0.3914	0.3070	0.2229
30	a_3		0.8089	0.7199	0.6193	0.5276	0.4386	0.3365
50	a_0	0.9885	0.6491	0.5243	0.4102	0.2090	0.1396	0.0837
50	a_1		0.7330	0.6207	0.5038	0.2968	0.2149	0.1410
50	a_2		0.8431	0.7585	0.6697	0.4851	0.3929	0.2927
50	a_3		0.9167	0.8665	0.7971	0.6921	0.6083	0.4947
100	a_0	0.9986	0.8420	0.7454	0.6442	0.1908	0.1178	0.0695
100	a_1		0.9121	0.8481	0.7787	0.3830	0.2832	0.1918
100	a_2		0.9757	0.9482	0.9133	0.7071	0.6181	0.5124
100	a_3		0.9930	0.9847	0.9741	0.9081	0.8672	0.7999

Bartlett, classical Bartlett test; GEE_m, Levene test analog based on GEE; CW_m, reweighted Levene test analog. Subscript m denotes measure of center: 0, mean; 10, 10% trimmed mean; 50, median.

Table 15. Empirical size and power of reweighted tests under no informativeness.

M	Test	δ_0	δ_1	δ_2	δ_3
			Test of means		
30	RW	0.0492	0.1676	0.4829	0.8146
	GEE	0.0625	0.1924	0.5135	0.8257
50	RW	0.0490	0.2462	0.7045	0.9631
	GEE	0.0590	0.2590	0.7073	0.9604
100	RW	0.0519	0.4254	0.9421	0.9999
	GEE	0.0544	0.4173	0.9348	0.9999
			Test of variance		
30	CF_d	0.0521	0.2548	0.5682	0.8112
	CW_0	0.0545	0.2463	0.5590	0.7965
	CW_{10}	0.0538	0.2426	0.5502	0.7920
	CW_{50}	0.0501	0.2291	0.5338	0.7806
	GEE_0	0.0716	0.2745	0.5776	0.8089
	GEE_{10}	0.0700	0.2685	0.5728	0.8044
	GEE_{50}	0.0650	0.2570	0.5602	0.7965
50	CF_d	0.0511	0.3922	0.8091	0.9649
	CW_0	0.0545	0.3710	0.7813	0.9572
	CW_{10}	0.0529	0.3694	0.7779	0.9556
	CW_{50}	0.0504	0.3588	0.7711	0.9535
	GEE_0	0.0657	0.3777	0.7798	0.9508
	GEE_{10}	0.0634	0.3760	0.7769	0.9503
	GEE_{50}	0.0598	0.3695	0.7698	0.9493
100	CF_d	0.0501	0.6673	0.9829	1
	CW_0	0.0538	0.6344	0.9755	0.9996
	CW_{10}	0.0538	0.6324	0.9745	0.9995
	CW_{50}	0.0521	0.6290	0.9745	0.9996
	GEE_0	0.0592	0.6251	0.9699	0.9990
	GEE_{10}	0.0586	0.6222	0.9695	0.9990
	GEE_{50}	0.0572	0.6187	0.9690	0.9990

Top: RW, reweighted test of means; GEE, GEE model. Bottom: CF_d, reweighted F test analog, difference form; CW_m, reweighted Levene test analog; GEE_m, Levene test analog based on GEE. Subscript m denotes measure of center: 0, mean; 10, 10% trimmed mean; 50, median.

CHAPTER 5

htestClust: AN R PACKAGE

5.1 Introduction

R is an open-source programing language and software environment that is a commonly used tool for statistical analysis. The native R environment includes core packages that implement general computing, graphing, and statistical methods. One such core package is **stats**, which provides functions for many classical hypothesis tests, including most of the classical analogs to the reweighted tests developed and discussed in this work. The facilities of R can be extended to more complex or specialized methods through secondary packages developed by users and accessible from repositories such as Comprehensive R Archive Network (CRAN). Developing an R package to accompany new methodological work is beneficial as it provides convenient means for analysts to implement the methods on real data.

A number of R packages related to clustered data have been developed. The popular package **geepack** [25] provides a flexible approach to modeling clustered data using generalized estimating equations, with an interface designed to resemble that for generalized linear modeling in the core R environment. Other clustered data packages are designed for more particular analyses. The package **clusrank** [34] contains a collection of Wilcoxon rank-sum and sign-rank tests for clustered data, including the tests of Section 4.4 that correct for potential informativeness (also found in the package **ClusterRankTest** [15]). Various methods for analyzing marginal homogeneity of binary matched pairs in clustered data are available through the package

clust.bin.pair [23], including the test [14] discussed in Chapter 3 that coincides with a reweighted test correcting for informativeness. As individual packages have mostly autonomous authors, there can be significant variation in function usage, syntax, and accepted data structure across packages.

In this chapter, we introduce an R package **htestClust** which contains functions that execute the marginal hypothesis tests from Chapters 3 and 4. As noted above, some of the previously-developed reweighted tests discussed in preceding chapters are available in R through various packages. However, there does not exist a comprehensive package containing ICS/IWCGS-appropriate methods. By including functions for the previously-developed tests, **htestClust** unifies the collection of reweighted tests to a single package, allowing consistency across arguments and data structure. As these reweighted tests mimic classical forms, we have intentionally modeled the syntax and output of **htestClust** functions after their classical analog functions. This makes the usage of **htestClust** functions intuitive to users familiar with the R environment.

In addition to the reweighted tests of Chapters 3 and 4, **htestClust** also includes a function that performs a recently-developed test of informative cluster size [44]. Marginal tests based on the reweighting methodology remain valid when data lack informative cluster size. However, standard methods that account for clustering may offer a power advantage for some analyses when informativeness is not a concern. The test for ICS provides researchers a method of discriminating appropriate analysis methods when data have variable cluster size. Combined, **htestClust** is a suite of functions providing analysts the means to perform various marginal analyses of clustered data with potential informativeness through the convenience and consistency of a single package.

The rest of this chapter is organized as follows. In Section 5.2 we summarize the balanced bootstrap method of Nevalainen et al. [44] that tests for informative

cluster size. In Section 5, we introduce a simulated data set that contains variables appropriate for illustrating the performance of the functions. Section 5.4 contains examples of the application of the functions contained in **htestClust**, and the chapter concludes with a brief discussion.

5.2 Test of informativeness

Nevalainen et al. [44] recently proposed a test for ICS using a novel balanced bootstrap scheme. As it might be desirable to implement this test prior to the application of the marginal methods mentioned thus far, we include this test for ICS in the **htestClust** package and briefly summarize it below.

Let $\boldsymbol{V} = (V_1, \ldots, V_M)$ be a collection of independent clustered observations, where $V_i = (n_i; Y_{i1}, \ldots, Y_{in_i})$ is the data from cluster i. Assuming exchangeability of observations within clusters, the hypothesis of interest is $H_0 : P\left(Y_{ij} \leq y | n_i = k\right) = F(y) = , k = 1, 2, \ldots; j = 1, \ldots, k$, for some unknown distribution F. Nevalainen et al. [44] propose two test statistics for testing H_0; a Kolmogorov-Smirnov type statistic takes the form

$$T_F = sup_y |\hat{F}(y) - \tilde{F}(y)|$$

where $\hat{F}(y) = \frac{1}{n}\sum_{i=1}^{M}\sum_{j=1}^{n_i} I\left[Y_{ij} \leq y\right]$ and $\tilde{F}(y) = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{n_i}\sum_{j=1}^{n_i} I\left[Y_{ij} \leq y\right]$. A potentially more powerful alternative to T_F is a Cramer-von Mises type statstic:

$$T_{CM} = \sum_{k \epsilon \psi}\left[kM_k \int \left(\hat{F}_k\left(y\right) - \hat{F}\left(y\right)\right)^2 dy\right],$$

where ψ represents the set of unique cluster sizes, M_k represents the number of clusters of size k, and $\hat{F}_k(y) = \frac{1}{kM_k}\sum_{i=1}^{M}\sum_{j=1}^{n_i} I\left[n_i = k, Y_{ij} \leq y\right]$. For data with a small number of distinct cluster sizes $T_{CM}$is the suggested statistic as it offers a power advantage over T_F. However, $T_{CM}$is liberal when there are a large number of distinct cluster sizes and the number of observed clusters of each size is small, in which case T_F is the preferred statistic.

Once either T_F or T_{CM} has been selected as the desired test statistic, the following bootstrap scheme is implemented: For bootstrap iteration $b, b = 1, \ldots, B$,

1. Permute observations within each cluster.

2. Resample clusters from the permuted data by performing the following for $i = 1, \ldots, M$:

   a) Randomly select a cluster $i^*, i^* = 1, \ldots, M$.

   b) If $n_{i^*} \geq n_i$, form the i^{th} bootstrapped cluster from the first n_i observation from cluster i^*; e.g., $V_{bi}^* = (n_i; Y_{i^*1}, \ldots, Y_{i^*n_i})$.

   c) If $n_i^* < n_i$, form the i^{th} bootstrapped cluster by merging observations from the resampled cluster i^* and observations from the closest 'matching' cluster to cluster i^*; e.g., $V_{bi}^* = \left(n_i; Y_{i^*1}, \ldots, Y_{i^*n_i^*}, Y_{k(n_i^*+1)}, \ldots, Y_{kn_i}\right)$, where $k = \arg\min_k \{D(V_{i^*}, V_k) : n_k \geq n_i\}$.

3. Calculate the test statistic from the collection of bootstrapped clusters, $T_b^* = T\left(\boldsymbol{V}_b^*\right)$, $\boldsymbol{V}_b^* = (V_{b1}^*, \ldots, V_{bM}^*)$.

The approximate p-value is then obtained from the sample of bootstrapped test statistics by $\frac{1}{B}\sum_{b=1}^{B} I\left[T_b^* \geq T\right]$, where T is the desired test statistic calculated from the original data. In part c of step 2, the closest matching cluster is determined by minimum distance calculated by $D\left(V_i, V_j\right) = \left(\min\{n_i, n_j\}\right)^{-1} \sum_{k=1}^{\min\{n_i, n_j\}} \left(Y_{ik} - Y_{jk}\right)^2$.

5.3 An example data set

To illustrate the reweighted tests in **htestClust**, we simulated a hypothetical data set of clustered observations with informativeness, with variables suited to a number of marginal analyses. This data set is provided in **htestClust** and allows us to illustrate the usage of the various functions in a consistent manner. Details on the simulation of these data are provided in Appendix B.

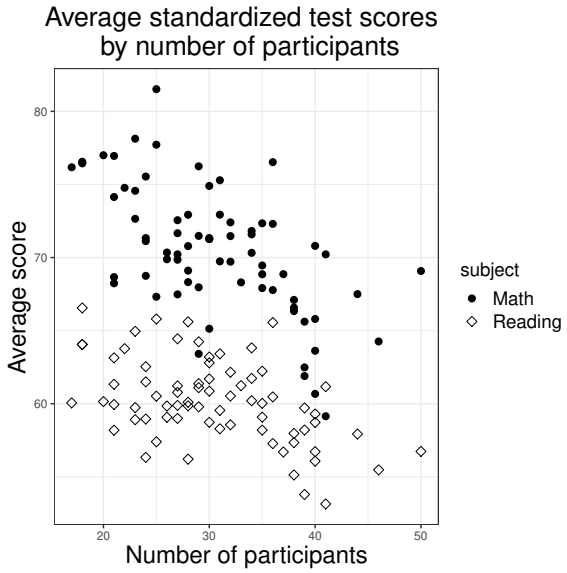

Figure 3. Average scores by cluster size in `screen8` data.

Consider the following hypothetical scenario. Through a voluntary comprehensive exit survey, an urban school district has collected demographic, biometric, and academic performance data from graduating 8th grade students. These data are clustered, with schools forming the clusters and students comprising the observations within clusters. The school district has implemented an incentive program in which schools with higher participation rates are prioritized for classroom and technology upgrades. Cluster size could be informative in these data, as resource-poor schools might have higher participation rates (larger cluster size), but also tend to have worse health metrics and lower standardized test scores.

```
> head(screen8)
  sch.id stud.id age gender height weight math read phq2 qfit qfit.s activity
1      1       1  15      M     65    136   69   75    3   Q2     Q2    other
2      1       2  14      M     66    135   80   57    2   Q4     Q3    other
3      1       3  15      M     65    146   60   85    0   Q2     Q3   sports
4      1       4  15      M     68    156   70   83    1   Q3     Q2    other
5      1       5  15      M     68    170   66   60    1   Q2     Q2   sports
6      1       6  14      M     63    109   84   62    0   Q1     Q1 academic
```

The (hypothetical) data set `screen8` contains data from 2224 students from 73 schools in this district. This data set contains an identification variable for the school,

`sch.id`, and an identification variable for the individual students from each school, `stud.id`. These correspond to the clusters and observations within clusters, respectively. Demographic variables from each student include age in years (`age`), height in inches (`height`), weight in lbs (`weight`), and binary gender (`gender`). The data include each student's standardized test scores in math (`math`) and reading (`read`), which are numeric values ranging from 0 to 100. The variable `phq2` is an ordinal (0-6) score from a mental health screening, in which higher scores correspond to higher levels of depression. Each student has two records from a physical health assessment: `qfit` is the student's (age-adjusted) fitness quartile from the assessment at the time of the exit survey, and `qfit.s` is the student's fitness quartile from the assessment taken at the beginning of the school year. Students may elect to participate in a variety of extracurricular activities, including academic clubs and sports teams. These activities have been broadly categorized into `academic`, `sports`, and `other` (including students with no extracurricular participation), and are recorded in the variable `activity`.

In this data set, cluster size is the number of participants from each school. The number of students participating from each school ranged from 17 to 50, with a median of 30. To examine if cluster size might be informative for academic performance, in Figure 3 we plot the average math and reading scores from each school by their number of participants. From these plots, we see a negative association between cluster size and test scores in both math and reading. Schools that collected data from more students tend to have lower standardized test scores compared to schools that evaluated a smaller number of students.

Cluster size may additionally be related to categorical variables or groups defined within clusters in these data. Figure 4 examines the relationship between gender and student extracurricular activity selections with screening participation rate. The right panel of Figure 4 shows a bar chart of the proportion of students engaged in the

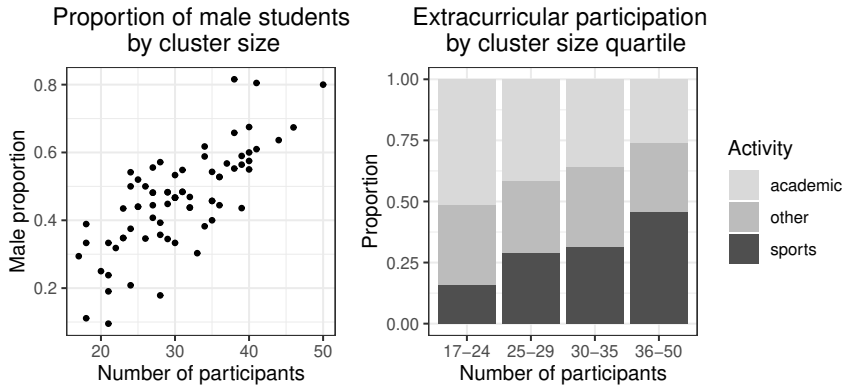

Figure 4. Plots of categorical variables by cluster size in `screen8` data.

three types of extracurricular activities by quartile of cluster size. The proportion of students engaging in sports-related extracurricular activities increased with the number of participants. That is, schools that contributed a larger number of students tended to have more students involved in extracurricular sports activities, whereas schools that contributed fewer observations had higher participation rates in academic activities. The left panel of this figure plots the proportion of male students in each school's sample by the number of participants from that school. Here, we observe a positive association between cluster size and proportion of male students in the sample. If this is feature of the data and not representative of each school's gender ratio, IWCGS weighting should be considered in analyses comparing outcomes between the genders. Cluster-weighted analyses will correct for the relationship between larger cluster sizes and outcomes (such as lower test scores), but will fail to account for the overrepresentation of male students in the larger samples.

5.4 R implementation

The syntax of the functions in **htestClust** largely follows that of common, recognizable hypothesis testing functions in the **stats** package, which is supplied through the fundamental R environment. As the reweighted tests are analogs of classical forms

for non-clustered data, their execution has been designed to mimic that of their conventional counterparts. A summary of the functions that comprise **htestClust** is given in Table 16, along with the reweighted test(s) each function performs and the established R function that executes the analogous test for independent observations.

In the sections below, we outline the usage of the functions available **htestClust** package. Echoing the interface of conventional R functions, most **htestClust** functions accept vector input that designate the response and clustering variables for individual observation. For convenience, many functions are designed with a secondary interface accepting tables or formulas. Minimum function output include the test statistic, p-value, number of clusters, data name, and name of the test, and most functions return additional values such as estimates and confidence intervals. Functions that return confidence intervals produce them in the usual way based on test asymptotics. In the interest of brevity and clarity, rather than detailing every input argument and output value, we instead provide an overview of each function's usage and illustrate its application through examples related to the `screen8` data. Complete information on function arguments and values are provided in their documentation in the R environment.

5.4.1 Test for informative cluster size

The test of ICS from Section 5.2 is implemented in **htestClust** through the function `icstestClust()`, which has the following usage:

```
icstestClust(x, id, test.method = c("TF", "TCM"), B = 1000, print.it = TRUE)
```

The main arguments of this function are `x`, a vector of numeric outcomes potentially related to sample size, and `id`, a vector or factor object which identifies the clusters. The argument `test.method` allows the user to select the desired test statistic, and `B` defines the number of bootstrap iterations to be performed. This test is computationally intensive and can take significant time to perform. By default, the progression

Table 16. List of functions available in the **htestClust** package

htestClust function	Reweighted test(s)	Classical analog function
`chisqtestClust()`	Chi squared goodness of fit, independence	`chisq.test()`
`cortestClust()`	Correlation	`cor.test()`
`icstestClust()`	Test of ICS	NA
`levenetestClust()`	K-group test of variance	`leveneTest()`
`mcnemartestClust()`	Homogeneity	`mcnemar.test()`
`onewaytestClust()`	K-group mean equality	`oneway.test()`
`proptestClust()`	Proportion	`prop.test()`
`ttestClust()`	Test of means (one/two group, paired)	`t.test()`
`vartestClust()`	2-group test of variance	`var.test()`
`wilcoxtestClust()`	Rank sum, signed rank	`wilcox.test()`

Each row gives the name of a **htestClust** function, the reweighted test the function performs, and the R function that executes the corresponding classical analog test. All classical analog functions are available in R through the **stats** package, except for `leveneTest()`, which is included in the **car** package.

of bootstrap iterations is printed to assist the user in estimating the execution time, though this can be suppressed by setting `print.it = FALSE`.

In the `screen8` data, there appeared to be a negative association between average test scores and the number of screening participants from each school, as illustrated in Figure 3. We can test whether cluster size is informative for math scores by performing the test of ICS with the `icstestClust()` function.

```
> set.seed(100)
> test.ics <- icstestClust(screen8$math, screen8$sch.id, B = 1000, print.it=FALSE)
> test.ics

Test of informative cluster size (TF)

data:  screen8$math
TF = 0.029686, p-value < 2.2e-16
```

This function returns the data name, the value of the test statistic, and the approximate p-value. Based on 1000 iterations, there is evidence to suggest that there is a significant association between the number of participants from each school and math scores in the `screen8` data.

5.4.2 Categorical tests

Proportion test

The function `proptestClust()` performs the reweighted test of marginal proportion from Section 3.2.1. This function has the usage

```
proptestClust(x, id, p = NULL, alternative = c("two.sided", "less", "greater"),
          variance = c("sand.null", "sand.est", "emp", "MoM"), conf.level = 0.95)
```

The argument `x` can be a binary vector of indicators denoting the success or failure of each observation, or a two-dimensional table with two columns giving the aggregate counts of failures and successes (respectively) across clusters. If `x` is a vector, a vector of numeric or factor objects denoting the respective cluster membership of the observations must be provided for the argument `id`. If `x` is a table, the rows of the table define the clusters and the `id` argument is ignored. The argument `variance` allows the user to specify the method of variance estimation for the statistic, selecting from the sandwich estimate evaluated at the null hypothesized value (`sand.null`), the sandwich estimate evaluated at the cluster-weighed proportion (`sand.est`), the empirical estimate (`emp`), or the method of moments estimate (`MoM`). If not specified, the function defaults to `sand.null`, as a test constructed with this estimator exhibited the most desirable properties in simulation studies. The argument `p` specifies the null marginal proportion to be tested; if not given, the function defaults to testing a null value of 0.5. The arguments `alternative` and `conf.level` allow the user to specify the alternative hypothesis and confidence level of the returned confidence interval.

In the hypothetical school district that collected the `screen8` data, suppose math proficiency is defined by standardized score of 65 or higher. The district wishes to test whether the marginal proportion of proficient students is higher than 75%. To apply the `proptestClust()` function, we must first define a binary variable denoting whether students have a standardized math score of at least 65. We then submit to the function this binary vector that contains the success/failure status of each student,

along with the `sch.id` vector that defines cluster membership. The appropriate null proportion and alternative hypothesis must also be specified. Based on this analysis, there was insufficient evidence to conclude that the marginal math proficiency in the district was higher than 75%.

```
> screen8$math.p <- 1*(screen8$math>=65)
> proptestClust(x = screen8$math.p, id = screen8$sch.id, p = .75,
+               alternative = "greater")

Cluster-weighted proportion test with variance est: sand.null

data:  screen8$math.p, M = 73
z = 0.70159, p-value = 0.2415
alternative hypothesis: true p is greater than 0.75
95 percent confidence interval:
 0.7311459 1.0000000
sample estimates:
Cluster-weighted proportion
                  0.7640235
```

Alternatively, this test can be performed by defining a table of counts of the non-proficient/proficient students from each school, and submitting the table to the `proptestClust()` function in the following manner. Note that the table must be defined so the counts of failures and successes are in the first and second columns, respectively, which occurs naturally when tabulating an appropriately defined binary variable.

```
> mathp.tab <- table(screen8$sch.id, screen8$math.p)
> head(mathp.tab)

    0  1
  1 12 23
  2  3 29
  3  5 21
  4 13 20
  5  3 20
  6  8 17
> test.tab <- proptestClust(mathp.tab, p = .75, alternative = "greater")
> test.tab

Cluster-weighted proportion test with variance est: sand.null

data:  mathp.tab, M = 73
z = 0.70159, p-value = 0.2415
alternative hypothesis: true p is greater than 0.75
```

```
95 percent confidence interval:
 0.7311459 1.0000000
sample estimates:
Cluster-weighted proportion
                  0.7640235
```

Regardless of the input method, the `proptestClust()` function returns a list of class `htest` containing a number of components. This list includes `statistic`, the test statistic appropriately named with its limiting distribution; `p.value`, the p-value of the test; `estimate`, the estimate of the marginal proportion; `null.value`, the null hypothesized marginal proportion; `conf.int`, the asymptotic confidence interval for the true marginal proportion; and `alternative`, a character string specifying the alternative hypothesis. The name of the test and the method for variance estimation is given by `method`. For summary purposes, the name of the data is paired with the number of clusters in the value `data.name`, which is returned automatically in the function output. The number of clusters can independently be returned through the value `M`.

```
> str(test.tab)
List of 9
 $ statistic  : Named num 0.702
  ..- attr(*, "names")= chr "z"
 $ p.value    : num 0.241
 $ estimate   : Named num 0.764
  ..- attr(*, "names")= chr "Cluster-weighted proportion"
 $ null.value : Named num 0.75
  ..- attr(*, "names")= chr "p"
 $ conf.int   : num [1:2] 0.731 1
  ..- attr(*, "conf.level")= num 0.95
 $ alternative: chr "greater"
 $ method     : chr "Cluster-weighted proportion test with variance est: sand.null"
 $ data.name  : chr "mathp.tab, M = 73"
 $ M          : Named int 73
  ..- attr(*, "names")= chr "M"
 - attr(*, "class")= chr "htest"
```

Output from all of the `htestClust` functions have similar values to those listed here, appropriately modified to the respective hypothesis of interest.

For illustrative purposes, compare the output of the `proptestClust()` function to that of the native R function performing the classical one-sample proportion test,

`prop.test()`. Among other input methods, this function can accept a table of counts of the total number of successes and failures in a sample, respectively. Note that this order is opposite of the natural R tabulation of a standard binary variable.

```
> mathp.tab2 <- rev(table(screen8$math.p))
> mathp.tab2

   1    0
1648  576
> prop.test(mathp.tab2, p = .75, alternative = "greater")

1-sample proportions test with continuity correction

data:  mathp.tab2, null probability 0.75
X-squared = 0.91187, df = 1, p-value = 0.8302
alternative hypothesis: true p is greater than 0.75
95 percent confidence interval:
 0.7252123 1.0000000
sample estimates:
        p
0.7410072
```

The classical proportion test performed by `prop.test()` is an inappropriate analysis of the `screen8` as it ignores clustering. However, the application here highlights the intentional symmetry designed in `proptestClust()` to the `prop.test()` function. This affinity in application and output to that of their classical analog function is similarly reflected in all the **htestClust** functions.

Goodness of Fit test

The reweighted goodness of fit test from Section 3.2.2 is executed through the function `chisqtestClust()`, which has the usage

```
chisqtestClust(x, y = NULL, id, p = NULL,
          variance = c("MoM", "sand.null", "sand.est", "emp"))
```

Similar to `proptestClust()`, this function allows for both vector and table input. If `x` is a vector denoting group membership for individual observations, a corresponding vector for `id` needs to be provided that gives the cluster identification for each observation. Alternatively, `x` can be a table where clusters are defined by rows and columns

contain counts across the categorical outcome levels. The optional argument `p` allows the user to specify the null category proportions to be tested. If `p` is not given, a test of equality of marginal proportions is performed. The argument `variance` allows the user to specify the variance estimation method, with the argument options defined as in `proptestClust()`.

From the `screen8` data, district administrators wish to test if marginal participation levels are equal across the categories of academic, sports, and other extracurricular activities. This can be performed by specifying the activity selection and school for each student through vector input, or by tabulating the counts of students engaged in the three activity types across schools.

```
> chisqtestClust(x=screen8$activity, id=screen8$sch.id)

Cluster-weighted chi-squared test for given probabilities with variance est: MoM

data:  screen8$activity, M = 73
X-squared = 13.101, df = 2, p-value = 0.001429

> head(act.table)

    academic other sports
  1       10    13     12
  2        8    16      8
  3       10     7      9
  4       10     7     16
  5       10     8      5
  6       14     6      5
> chisqtestClust(act.table)

Cluster-weighted chi-squared test for given probabilities with variance est: MoM

data:  act.table, M = 73
X-squared = 13.101, df = 2, p-value = 0.001429
```

The weighted marginal category proportions can be obtained by calling `observed` from the function output. From this analysis, we conclude that the marginal proportion of extracurricular activity participation is not equal across the three selections. Marginal across schools, students in this district participate in academically-oriented extracurricular activities at a higher rate than other forms of activities. This func-

tion has an additional value `expected`, which for the reweighted goodness of fit test returns the category proportions under the null hypothesis.

```
> act.test <- chisqtestClust(act.table)
> act.test$observed
 academic     other    sports
0.3867422 0.3070048 0.3062530

> act.test$expected
 academic     other    sports
0.3333333 0.3333333 0.3333333
```

Test of independence

The function `chisqtestClust()` also performs the reweighted chi squared test of independence (Section 3.2.3). This test is executed with vectors, where the arguments `x` and `y` contain the categorical membership of observations for the variables whose independence is to be assessed. Cluster membership is supplied through the `id` argument. The method of variance estimation can be selected with the `variance` argument.

We perform the reweighted test of independence to test if extracurricular activity selection is independent of gender in the `screen8` data. Based on this analysis, we conclude that marginal participation in extracurricular activities is independent of gender.

```
> marg.indep <- chisqtestClust(x=screen8$activity, y=screen8$gender,
+                              id=screen8$sch.id)
> marg.indep

Cluster-weighted Chi-squared test of independence with variance est: MoM

data:  screen8$activity and screen8$gender, M = 73
X-squared = 1.6131, df = 2, p-value = 0.4464
```

The observed joint reweighted proportions can be obtained by calling the `observed` value, while the expected joint proportions under independence are called with `expected`.

```
> marg.indep$observed
                 F         M
academic 0.2225577 0.1641845
other    0.1599899 0.1470149
sports   0.1562627 0.1499902

> marg.indep$expected
                 F         M
academic 0.2193678 0.1673744
other    0.1662734 0.1407314
sports   0.1531691 0.1530838
```

Test of marginal homogeneity

The test of marginal homogeneity of matched pairs in clustered data from Section 3.2.4 is performed through the function `mcnemartestClust()`.

```
mcnemartestClust(x, y, id, variance = c("MoM", "emp"))
```

The arguments `x` and `y` take vectors with two levels, denoting the success/failure of the first and second measurement from observations. Cluster membership is given through the `id` argument, and variance estimation method selected through `variance`.

Using the `screen8` data, we test whether the marginal proportion of students in the lowest fitness quartile at the end of the school year was the same as at the beginning of the school. Based on this analysis, there was no change in marginal proportion of students evaluated at the lowest quartile of fitness between the start and end of the school year.

```
> screen8$low.start <- 1*(screen8$qfit.s=='Q1')
> screen8$low.end <- 1*(screen8$qfit=='Q1')
> mcnemartestClust(screen8$low.start, screen8$low.end, screen8$sch.id)

       Cluster-weighted test of marginal homogeneity with variance est: MoM

data:  screen8$low.start and screen8$low.end, M = 73
chi-square = 0.013417, df = 1, p-value = 0.9078
```

5.4.3 Quantitative tests

Tests of means

The function `ttestClust()` performs the reweighted one-sample, paired, and two-sample test of means from Section 4.3.

```
ttestClust(x, y = NULL, idx, idy = NULL, alternative = c("two.sided", "less",
    "greater"), mu = 0, paired = FALSE, conf.level = 0.95)
```

To execute the reweighted one-sample test, a numeric vector of outcomes must be provided for `x` and a vector of cluster identifiers must be provided for `idx`. The argument `mu` specifies the hypothesized value of the true marginal mean (or difference in marginal means, if performing a paired or two-sample test). A one or two-sided test can be specified through the `alternative` argument, and a confidence interval with level `conf.level` appropriate to the performed test is returned.

Suppose the national average on the standardized math exam taken by students in the `screen8` data set is 65. Administrators wish to test if the marginal average in this district is equal to the national average. We can perform this test by supplying the math exam scores and school (cluster) identifying variable for each student, and specifying the null hypothesized average score. We conclude that students in this school district have a marginal average math score higher than the national average.

```
> ttestClust(x = screen8$math, idx = screen8$sch.id, mu = 65)

    One sample cluster-weighted test of means

data:  screen8$math, M = 73
z = 6.7164, p-value = 1.863e-11
alternative hypothesis: true mean is not equal to 65
95 percent confidence interval:
 68.91966 72.14999
sample estimates:
cluster-weighted mean of x
                  70.53482
```

For comparison, consider the classical t-test applied to these data using the native R function `t.test()`. While the unweighted estimate is not conspicuously different than the weighted estimate, this test is clearly inappropriate for analysis of the `screen8` data. First, the classical t-test treats the observations as independent, failing to

account for the clustered nature of these data. Second, as evidenced by the degrees of freedom displayed in the function output, this analysis treats students as the unit of interest. When the interest is in the outcome from a typical student from a typical school, the correct marginal analysis is indexed by the clusters (e.g., schools) and not the observations within clusters. One final highlighting of the erroneous nature of this analysis can be seen in the failure of the classical confidence interval to include the reweighted estimate.

```
> t.test(x = screen8$math, mu = 65)

One Sample t-test

data:  screen8$math
t = 26, df = 2223, p-value <2e-16
alternative hypothesis: true mean is not equal to 65
95 percent confidence interval:
 69.5 70.3
sample estimates:
mean of x
     69.9
```

One cluster-appropriate method to estimate the marginal math score of students in this district would be to fit a GEE model. This is easily performed using the `geeglm()` function in the package `geepack`.

```
> library(geepack)
> gee.mod <- geeglm(math ~ 1, id = sch.id, data = screen8, corstr = "exchangeable")
> summary(gee.mod)

Call:
geeglm(formula = math ~ 1, data = screen8, id = sch.id, corstr = "exchangeable")

 Coefficients:
            Estimate Std.err  Wald Pr(>|W|)
(Intercept)   70.465   0.509 19143   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation structure = exchangeable
Estimated Scale Parameters:

            Estimate Std.err
(Intercept)     80.2    3.89
  Link = identity
```

```
Estimated Correlation Parameters:
      Estimate Std.err
alpha    0.225  0.0389
Number of clusters:   73  Maximum cluster size: 50
```

The output from this function includes the estimate of the marginal average math score, along with the standard error for this estimate and the p-value from a two-sided hypothesis test against a null value of 0. However, implementing a hypothesis test against a non-zero null value or obtaining a confidence interval requires additional manual construction by the analyst. Additionally, this GEE model does not account for possible informativeness in the data. In contrast, the `ttestClust()` function executes simple hypothesis tests through a format easily modified by the user, while accounting for clustering and potential informativeness.

A paired test of means can be performed using the one-sample execution method in `ttestClust()` by supplying a vector of paired differences for `x`, accompanied with an appropriate cluster identifier for `idx`. Alternatively, the paired numeric values can be individually input for `x` and `y` along with a cluster identifier vector for `idx`, and specifying `paired = TRUE`.

Suppose at the national level, the average student scores 10 points higher on the math exam compared to the reading exam on the standardized tests taken by the students in the `screen8` data set. To test whether students in the district score higher on math than reading in a manner consistent with the national average, we can implement the reweighted paired test. We conclude that the students in this school district have an average difference in math and reading scores equivalent to the national average.

```
> ttestClust(x = screen8$math, y = screen8$read, idx = screen8$sch.id,
+            paired = TRUE, mu = 10)

    Paired cluster-weighted test of means

data:  screen8$math and screen8$read, M = 73
```

```
z = 0.91303, p-value = 0.3612
alternative hypothesis: true difference in means is not equal to 10
95 percent confidence interval:
  9.611553 11.065973
sample estimates:
cluster-weighted mean of the differences
                                10.33876
```

The reweighted two-sample test can be performed by specifying the vectors `x, y, idx, idy`, where `x` and `y` are the numeric outcomes from the two groups and `idx`, `idy` are their corresponding cluster identification vectors. Alternatively, this test can be implemented using a formula dispatch.

```
ttestClust(formula, id, data, subset, na.action, ...)
```

To implement the formula method, the argument `formula` should be of the form `lhs` $\sim$ `rhs`, where `lhs` is a numeric variable giving the data values and `rhs` is a factor with two levels giving the corresponding groups. A cluster-identifying vector `id` must additionally be specified. The argument `data` is an optional matrix or data frame containing the variables in the formula `formula` and `id`.

Administrators from the `screen8` district wish to test if there's a difference in average math scores between males and females. We perform this test using the formula interface, and conclude that the marginal average math score is not significantly different between male and female students in this district.

```
> ttestClust(math ~ gender, id = sch.id, data = screen8)

Two sample group-weighted test of means

data:  math by gender, M = 73
z = 1.3495, p-value = 0.1772
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2234259  1.2111344
sample estimates:
weighted mean in group F weighted mean in group M
                70.75124                 70.25739
```

The function `onewaytestClust()` performs the test of equality for K-group means discussed in Section 4.3.3. This function operates alternatively on a single input of a table containing the intra-cluster group means, or through vectors supplied via the formula interface. The formula method has the following usage, with inputs structure consistent to that of `ttestClust()`.

```
onewaytestClust(formula, id, data, subset, ...)
```

If applying `onewaytestClust()` to a table, a single argument `x` is submitted, where `x` is a two-dimensional matrix or data frame containing the within-cluster group means, where rows are the clusters and columns are the group means. Note that incomplete clusters, i.e., clusters in which not all groups were observed, should have `NA` in the corresponding empty group column(s).

To illustrate this usage, we use the `screen8` data to test whether students engaged in the three categories of extracurricular activities have the same average reading score. To use the table interface, we first tabulate average reading scores by extracurricular activity for each cluster. This is easily performed, correctly accounting for incomplete clusters, through the `tapply` command.

```
> read.tab <- tapply(screen8$read, list(screen8$sch.id, screen8$activity), mean)
> head(read.tab, n = 8)
  academic    other   sports
1 63.70000 62.00000 61.25000
2 62.00000 60.37500 59.37500
3 58.00000 63.28571 57.00000
4 59.00000 61.42857 62.56250
5 60.60000 55.00000 61.80000
6 56.28571 57.33333 60.60000
7 57.81818 62.55556 59.71429
8 58.20000 58.16667       NA
> onewaytestClust(read.tab)

Reweighted one-way analysis of means for clustered data

data:  read.tab, M = 73
X-squared = 1.3191, df = 2, p-value = 0.5171
sample estimates:
academic    other   sports
60.11498 60.40785 59.69659
```

The same test using the formula interface is performed with the following code.

```
> onewaytestClust(read ~ activity, id = sch.id, data=screen8)

Reweighted one-way analysis of means for clustered data

data:  read and activity, M = 73
X-squared = 1.3191, df = 2, p-value = 0.5171
sample estimates:
academic    other   sports
60.11498 60.40785 59.69659
```

Based on this analysis, we conclude that students engaged in the various types of extracurricular activities have equal performance on the standardized reading test.

Tests of variance

The various methods of assessing variance equality of intra-cluster groups discussed in Section 4.5 can be performed through two functions in **htestClust**. The function `vartestClust()` tests equality of variance between two groups using the reweighted F test analog based on differences in group variances ($\hat{F}_d$) from Section 4.5.1, and the function `levenetestClust()` performs the reweighted Levene test analogs in Sections 4.5.2 and 4.5.3.

`vartestClust()` has the usage:

```
vartestClust(x, y, idx, idy, difference = 0,
  alternative = c("two.sided", "less", "greater"),
  conf.level = 0.95, ...)
```

The arguments `x` and `y` take numeric vectors of outcomes from the two intra-cluster groups, and `idx` and `idy` the respective cluster-identifiers. The function tests the null hypothesis that the marginal difference in variance between `x` and `y` is equal to `difference` against the one or two-sided alternative hypothesis specified by argument `alternative`. The function also has a formula interface:

```
vartestClust(formula, id, data, subset, na.action, ...)
```

The usage of the formula method remains consistent with previous functions, with `formula` taking the form `lhs` $\sim$ `rhs`, where `lhs` is the numeric outcome variable and `rhs` is a grouping variable with exactly two levels. Note that the order of the difference in variance for the null hypothesis will be determined by the order of the levels in the grouping variable `rhs`.

We illustrate these methods using the `screen8` data set and assess whether the variation in math scores is equivalent between male and female students. Note that the signs of the estimate and confidence interval have been reversed when using the formula method compared to the vector input, as the levels of the `gender` variable are `"F" "M"`.

```
> boys <- subset(screen8, gender=='M')
> girls <- subset(screen8, gender=='F')
> vartestClust(x = boys$math, y = girls$math, idx = boys$sch.id,
+              idy = girls$sch.id)

Reweighted test to compare two intra-cluster group variances

data:  boys$math and girls$math, M = 73
z = 0.18089, p-value = 0.8565
alternative hypothesis: true difference of variances is not equal to 0
95 percent confidence interval:
 -8.761322 10.542997
sample estimates:
difference of variances
              0.8908372

>
> vartestClust(math ~ gender, id = sch.id, data = screen8)

Reweighted test to compare two intra-cluster group variances

data:  math by gender, M = 73
z = -0.18089, p-value = 0.8565
alternative hypothesis: true difference of variances is not equal to 0
95 percent confidence interval:
 -10.542997   8.761322
sample estimates:
difference of variances
             -0.8908372
```

Based on this analysis, we conclude there is no difference in the marginal variability of math scores between boys and girls in this school district.

Testing variance equality using the reweighted Levene test analog can be performed through the `levenetestClust()` function. This function has the following default and formula methods:

```
levenetestClust(y, group, id, center = c("median", "mean"), trim = NA, ...)

levenetestClust(formula, id, data, subset, na.action, ...)
```

Using the default method, `y` is the vector of numeric responses, `group` is a vector defining groups, and `id` is a cluster identification vector. The method of centering is specified with the argument `center`. The optional numeric argument `trim` takes a value between [0, 0.5] specifying the percentage trimmed mean, and is only applicable when `center = ‘‘mean’’`. The application of the formula method for `levenetestClust()` remains consistent with that of previous formula methods, where `rhs` of `formula` is a grouping variable with at least two levels. Either `vartestClust()` or `levenetestClust()` can be used to assess variance equality between two groups, while only `levenetestClust()` assesses equality of variance for K intra-cluster groups.

To illustrate, we once again test variance equality in math scores between genders from our example data set. The results of the reweighted Levene test analogue are in concordance with those from the reweighted F test: the marginal variation in math scores between girls and boys in this school district is not significantly different.

```
> levenetestClust(y = screen8$math, group = screen8$gender, id = screen8$sch.id)

Reweighted Levene's Test for Homogeneity of Variance in Clustered Data
(center = median)

data:  screen8$math by screen8$gender, M = 73
X-squared = 0.40921, df = 1, p-value = 0.5224
```

Illustrating the formula method, we test variance equality of math scores between students engaged in the three types of extracurricular activities using a 10% trimmed mean.

```
> levenetestClust(math ~ activity, id = sch.id, data = screen8, center = "mean",
+                 trim = .1)

Reweighted Levene's Test for Homogeneity of Variance in Clustered Data
(center = mean: 0.1)

data:  math by activity, M = 73
X-squared = 0.24726, df = 2, p-value = 0.8837
```

Based on this analysis, we conclude there is no significant difference in marginal variation of math scores between students engaged in the three types of extracurricular activities.

Tests of correlation

The reweighted tests of correlation [39] from Section 4.6 are executed through the function `cortestClust()`, which has both vector and formula methods.

```
cortestClust(x, y, id, method = c("pearson", "kendall", "spearman"),
  alternative = c("two.sided", "less", "greater"),
  conf.level = 0.95, ...)

cortestClust(formula, id, data, subset, na.action, ...)
```

In the default method, `x` and `y` are numeric vectors of outcomes and `id` is a vector denoting cluster membership. The argument `method` allows the user to specify the desired reweighted correlation coefficient. In the formula method, `formula` should be of the form $\sim$ `u + v`, where each of `u` and `v` are numeric variables giving the data values, and `id` is the cluster-denoting vector. The function performs the hypothesis test that the marginal correlation coefficient is equal to 0 against the alternative specified by `alternative`, and returns a confidence interval with confidence level `conf.level`.

To illustrate, we estimate the marginal correlation between math and reading scores in the `screen8` data. There is significant positive marginal correlation between math and reading scores for students in this district.

```
> cortestClust(x = screen8$math, y = screen8$read, id = screen8$sch.id)

Cluster-weighted Pearson's product-moment correlation

data:  screen8$math and screen8$read, M = 73
z = 3.7442, p-value = 0.000181
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.04614837 0.14753645
sample estimates:
cluster-weighted cor
          0.09684241

> cortestClust(~ math + read, id = sch.id, data = screen8, method = "spearman")

Cluster-weighted Spearman's rank correlation rho

data:  math and read, M = 73
z = 4.1313, p-value = 3.607e-05
alternative hypothesis: true  is not equal to 0
95 percent confidence interval:
 0.05112969 0.14343427
sample estimates:
cluster-weighted rho
          0.09728198
```

5.4.4 Rank-based tests

The reweighted rank sum [10, 16] and signed rank tests [11] described in Section 4.4 are implemented through the `wilcoxtestClust()` function.

```
wilcoxtestClust(x, y = NULL, idx, idy = NULL,
  alternative = c("two.sided", "less", "greater"),
  mu = 0, paired = FALSE,
  method = c("cluster", "group"), ...)
```

The arguments `x` and `y` are numeric vectors of responses, and `idx` and `idy` are corresponding cluster-identifier vectors. If only `x` and `idx` are given, a cluster-weighted signed rank test of the null that the distribution of `x` is symmetric about `mu` is performed. If `x` and `y` are both given and `paired = TRUE`, only `idx` is necessary to identify clusters (`idy` is ignored). In this case, a cluster-weighted signed-rank test of the null that the distribution of `x - y` is symmetric about `mu` is performed.

The cluster-weighted rank sum test is performed when `method = cluster`, and the group-weighted rank-sum test is performed when `method = group`. When executing the rank sum tests, the null is that the two groups follow the same marginal distribution and the argument `mu` is ignored.

The reweighted rank sum tests can additionally be executed using a formula method, with the following usage.

```
wilcoxtestClust(formula, id, data, subset, na.action, ...)
```

The argument `formula` is of the form `lhs` $\sim$ `rhs`, where `lhs` is a numeric vector of data values, `rhs` is a factor with two levels giving the groups, and `id` is a vector denoting cluster membership.

Suppose in the `screen8` data we wish to test whether marginal reading scores are symmetric around 60. We implement the reweighted signed rank test, and conclude that the center of the distribution of marginal reading scores is not significantly different than 60.

```
> wilcoxtestClust(x = screen8$read, idx = screen8$sch.id, mu = 60)

One sample cluster-weighted signed rank test

data:  screen8$read, M = 73
z = 0.5741, p-value = 0.5659
alternative hypothesis: true location is not equal to 60
```

To illustrate the paired test, we test whether the distribution of the difference in math and reading scores is symmetric around 10. We conclude that the marginal difference between students' math and reading scores has a symmetric distribution centered around 10.

```
> wilcoxtestClust(x = screen8$math, y = screen8$read,
+                 idx = screen8$sch.id, mu = 10, paired = TRUE)

Paired cluster-weighted signed rank test

data:  screen8$math and screen8$read, M = 73
```

```
z = 0.76137, p-value = 0.4464
alternative hypothesis: true location shift is not equal to 10
```

Now suppose we're interested in determining whether males and females have the same distribution for mental health evaluation scores. We use the formula method to execute the reweighted rank-sum test using group weighting. Based on this analysis we conclude there is no significant difference in the distribution of mental health scores between boys and girls.

```
> wilcoxtestClust(phq2 ~ gender, id = sch.id, data = screen8, method = "group")

Group-weighted rank sum test

data:  phq2 by gender, M = 73
z = 0.14143, p-value = 0.8875
alternative hypothesis: true location shift is not equal to 0
```

5.5 Discussion

htestClust is available from CRAN. The package can be installed by running the following command within the R environment:

```
install.packages("htestClust")
```

Once installed, the **htestClust** package can be loaded in new R sessions using the command `library(htestClust)`.

R's flexible and extensible nature comes at the cost of efficiency. The design of the R-language and processing environment places constraints on performance, resulting in slower execution of complex calculations. Some of the reweighted tests performed by functions in **htestClust** require calculation of computationally expensive empirical CDFs and jackknife variance estimates. As a result, some **htestClust** functions can have lengthy computation times when applied to large data sets. A possible performance-boosting revision to the package would be the integration of portions of computationally expensive functions to a more efficient coding language, such as `C++`.

The **htestClust** package was developed as a tool to aid in the analysis of clustered data with potential ICS/IWCGS, and is comprised of functions that implement the broad collection of reweighted hypothesis tests described in previous chapters. While there exist a number of packages designed for the analysis of clustered data, **htestClust** is the first that is designed with the purpose of addressing informativeness in a comprehensive manner. This novel package implements marginal reweighted tests that are clustered data analogs to well-known classical statistical tests, and the interface of the package has been designed to reflect this relationship. Function interface has been purposefully structured to resemble that of functions available in base R that perform the analogous classical tests, making usage intuitive. Its thoughtful design and expansive collection of methods makes this package an effective tool for researchers to analyze clustered data with varying cluster or group sizes.

CHAPTER 6

DISCUSSION

6.1 Introduction

When analyzing clustered data with varying cluster or group sizes, methods that account for potential informativeness should be considered. GEE and other model-based methods are the standard approach for the marginal analysis of clustered data, but such methods implicitly assume variation in cluster size is ignorable. Failing to account for dependency between response measurements and the number of observations within clusters or intra-cluster groups can lead to over-weighting of larger clusters (or groups), and potentially biased inference. Grounded in resampling procedures, the marginalization principle of Williamson et al. [53] avoids biasing effects of informativeness by reweighting observations by their inverse cluster or group size. Estimators derived through this methodology have been shown to be asymptotically normally distributed, allowing inference to be conducted through Wald-type tests.

The work in this dissertation applied the marginalization principle to estimate marginal parameters related to proportions, means, and variances, and developed clustered data analogs of classic hypothesis tests. We demonstrated the need for these methods by comparing the performance of the reweighted tests to that of otherwise cluster-appropriate methods through simulation, and showed that only the reweighted tests consistently maintain appropriate size for data under informativeness. These tests augmented a small selection of similarly-reweighted established methods related to ranks and correlations in clustered data. Combined, these tests

formed a collection addressing a broad spectrum of general hypotheses. We made this entire collection accessible to analysts through a comprehensive and flexible R software package. Together, the methods and software advanced by this dissertation expand the means for analysis of clustered data with potential cluster- or group-size informativeness.

In this chapter, we summarize the previous chapters of this document, and add to their individual discussions with some specific comments. We then provide a general discussion on the reweighting methodology, its limitations, and areas for further research.

6.2 Summary and additional comments related to previous chapters

In Chapter 2, we detailed the resampling origins of the reweighting methodology, and illustrated how the marginalization principle leads to inverse cluster- and group-weighted estimators. While resampling plays no active roll in the resulting tests, it's important to consider how this process would be performed to ensure accurate weighting. This is of primary concern when applying inverse group-weighting, which is grounded in a two-step resampling process. When data have incomplete group structure, the (theoretical) resampling process needs to reflect this condition, with subsequent weights from the marginalization process being modified accordingly. Failing to recognize how selection probabilities vary across incomplete clusters can result in estimators that are philosophically and mathematically problematic [49].

In Chapter 3, we developed tests of clustered categorical data reweighted to correct for ICS. This work not only advanced reweighted tests to a previously overlooked area, but additionally explored the effects of variance estimation on the performance of such tests. Prior tests of reweighted estimators were constructed using diverse variance estimates, but no comparison of methods had previously been performed. In the context of categorical responses, we demonstrated that the method of variance

estimation significantly affects the performance of reweighted tests. In particular, tests with variance estimates constructed under a null hypothesis consistently outperformed tests using alternative variance estimation methods.

As noted in Section 4.2, the breadth of technique for variance estimation in the tests of Chapter 3 is due to the lack of complications related to incomplete clusters. This convenience results both from the nature of categorical data and from the weighting method chosen for these tests. Observations are counts, so any unobserved categories within clusters have a value of 0. The reweighting applied to these tests corrects for ICS, resulting in observations being weighted by the inverse of the cluster size, which is always a value > 0. Should it be desired to derive a reweighted categorical test that corrects for IWCGS, the methods discussed in Section 2.5 could be applied. However, this weighting applies the inverse group size to observations, so certain intra-cluster parameter estimates would no longer be defined for data with incomplete group structure. As the variance estimates discussed in Chapter 3 are functions of intra-cluster parameter estimates, an alternative method for variance estimation, such as the jackknife form from Section 4.2, must be considered.

We note that IWCGS is of minor concern for the hypotheses addressed by the tests of Chapter 3. The tests of proportion, goodness of fit, and homogeneity are designed for univariate categorical data. Unless an additional variable ancillary to the primary outcome was considered, there are no groups whose distribution could be informative. Moreover, adjusting for informativeness of a secondary variable that is not of direct interest to the hypothesis is incongruous with the essence of this work. Contrary to the other three tests, the reweighted test of independence is performed on bivariate data. However, this test assesses the relationship between the two variables in a cumulative manner; it does not directly compare outcomes between groups. Even in circumstances when variables lend themselves to be defined as "groups" and "outcomes", the test is blind to this distinction. Therefore, the inverse cluster weight

applied in this test should simultaneously address informativeness in either (or both) variables.

In Chapter 4, we developed tests of intra-cluster group means and variances. As these tests compare group parameters, we applied the reweighting method correcting for IWCGS. Prior studies have shown that inverse-group weighting additionally corrects for ICS [16, 24], making these tests appropriate for data with informativeness of either (or both) cluster or group size. Moreover, under non-informative simulations, these reweighted tests exhibited superior nominal size control compared to GEE methods, while maintaining comparable or higher power. This chapter additionally included summaries of the reweighted rank-based tests and tests of correlation that naturally compliment the novel tests in this work. For the consistency of this document, we have altered some of the notation of these tests from their original publication form; however, their nature remains unchanged.

Motivated by distributional assumptions, there are several common tests of variance homogeneity in the classical setting. We paralleled these forms in our reweighted tests, constructing a number of methods for assessing variance homogeneity in clustered data. While the reweighted tests in this work are not subject to the same distributional constraints as their independent-observation analogs, we showed through simulation that their performance is not invariant to the distribution of the data. Much like the classical setting, considering the possible distributions of given data can be helpful in selecting the most appropriate method to assess variance equality of intra-cluster groups.

In Chapter 5, we presented the R package **htestClust**, which was developed to implement the collection of reweighted tests from Chapters 3 and 4. The **htestClust** package is intended to facilitate the analysis of clustered data with potential informativeness, and significant effort was devoted to its functionality. Syntax and nomenclature of functions were designed to be intuitive, mimicking that of recogniz-

able hypothesis testing functions in the native R environment. Most functions have a flexible interface, allowing data to be input either through vectors or, alternatively, formulas or tables.

Included in the **htestClust** package is a function that performs a hypothesis test for the presence of ICS. This test provides analysts a tool for assessing whether there is a relationship in their data between fluctuating cluster sizes and outcome measurements. As methods that fail to account for ICS can produce biased estimates, it would be a natural progression to apply this test prior to selecting the method of analysis when analyzing clustered data. To our knowledge, there is no analog test to assess the presence of IWCGS. As informativeness of group distribution is a separate issue that can occur independently of ICS, future research devoted to the development of such a test would be worthwhile.

6.3 General discussion

The effects of informativeness on otherwise cluster-appropriate methods are seen through the bias of GEE-based methods in the simulation results of Chapters 3 and 4. This bias ranged from mild to severe, depending on circumstance, and it is evident that sample size, number of parameters being tested, and degree of informativeness all play a complex role. In our simulations, GEE methods generally demonstrated heavier bias under smaller samples and when the number of parameters increased. In some cases, the bias increased with the number of clusters, suggesting this vulnerability to informativeness is not simply a matter of adequate sample size. As demonstrated in Table 9, the bias caused by informativeness is related to the degree of informativeness in the data. Informativeness was induced in our simulations through random effect parameters, u_i. Cluster and group sizes were generated through indirect functions of the form $b + c * f(u_i)$, where c serves as the "size modification" parameter. Obviously, as the value of c increases, the greater the degree of informativeness in the

data, which is likewise reflected in GEE performance. In some simulations, informativeness was also partially dependent on a "threshold" parameter t as part of $f(u_i)$, e.g., $f(u_i) = I[u_i > t]$. It is clear that the value of t also plays a role in the amount of informativeness in data, as it determines the frequency of the size variation. The combined effect of these two parameters on otherwise cluster-appropriate methods was not studied, and they represent only a few such parameters that could be related to informativeness. Therefore, while GEE methods resulted in empirical size only slightly higher than the nominal level in Tables 1 and 10, we caution readers to not take this as an indication of GEE robustness to varying cluster or group size when testing simple designs with a large number of clusters.

The methodology behind these reweighted tests rests primarily on the asymptotic normality of reweighted parameters. As previously noted, these methods are appropriate when the clusters are the unit of interest; as such, the asymptotic normality is indexed by the number of clusters. This dependence on sample size is evident throughout our simulation results. Tests of a single parameter performed well under small sample sizes, as shown in Tables 1 and 10. However, consistent across categorical and quantitative data, the reweighted tests required larger samples to maintain appropriate size as the dimension of the parameter vector increased. Classical asymptotic theory advises a threshold of 30 observations to establish normality. But, as the number of parameters increases, or as parameter values approach the boundary space, a larger sample size will be required to ensure the accuracy of these methods. In practice, obtaining a sufficient number of clusters to permit the use of these tests might be of issue. While there has been some development of small-sample inferential methods based on resampling and permutation techniques [19], the advancement of "exact" tests for clustered data under potential informativeness remains an area open to exploration.

The issue of incomplete clusters has played a peripheral theme throughout this

work. We have previously discussed how incomplete group structure changes the weighting assignment in the marginalization process, and how it can restrict variance estimation techniques. Other authors [47, 49] provide a thorough examination into the various distinct populations that produce data with incomplete group structure, and how informativeness in general relates to missing data. Less studied, however, is the effect of incomplete group structure on parameter estimation and testing. In the categorical tests of Chapter 3, incomplete clusters result in observed group proportions of 0, or, at times, 1. As the overall reweighted estimators are the average of within-cluster proportions, it is reasonable to question the biasing effect incomplete clusters have on the overall estimators and tests constructed from such estimators. While this issue is most salient to categorical estimates, it is likewise germane to reweighted quantitative values. To accommodate incomplete clusters, the jackknife method is implemented in estimating the variance of the statistic, and this form can be heavily influenced by outlying observations [51]. The inherent nature of informativeness could make clusters with incomplete group structure more likely to be outliers. As incomplete clusters are of practical concern, more research is needed in this area.

An additional understudied area is the effect of the underlying distribution on the performance of these methods. The tests in this collection have the amenity of being nonparametric and free from assumptions related to the clustering or informativeness structure. This avoids issues of misspecification that would be of concern with model-based methods and results in these tests being broadly applicable. Despite this, these tests depend on the asymptotic normality of the reweighted estimators, which is partially provisory on the marginal distribution of the data. This is evident in the simulations results in Tables 12, 13, and 14, in which the performance of the reweighted tests varied based on the distributional transformation applied to the data. Asymptotic normality is a well-studied area in classical statistics, but the extent to

which that knowledge is directly applicable to clustered data under informativeness remains unknown. Clustered data can have complex dependencies, and these relationships can be further complicated through mechanisms of informativeness. The Wald-type tests of reweighted estimators, both in this work and by other authors, have primarily been evaluated through simulations of symmetric/normal distributions where informativeness is induced by straightforward means. While we would expect this collection of tests to be robust in many settings, the relationship between complex data structure and asymptotic convergence of reweighted estimators has yet to be formally evaluated.

The array of tests in this collection constitute methods for cross-sectional analysis. That is, they provide researchers methods for analyzing data that correspond to a "snapshot in time". This conforms with the original reweighting application in CWGEE models and there are many settings of clustered data where such analysis will be of interest. Alternatively, some clustered data have a natural temporal aspect, the effect of which might be of primary interest. The methods in this collection are clearly incompatible for analyses concerned with changes across time. However, a number of modeling methods that extend the reweighting methodology to longitudinal settings have recently been developed [5, 41, 52].

In this work, we have applied reweighting methods that correct for cluster- and group-size informativeness to develop a comprehensive collection of marginal hypothesis tests for clustered data, and made implementation of these tests accessible through the creation of a software package. Not only do the tests in this collection correct for potential informativeness in clustered data, they provide the means for addressing a number of universal hypotheses without the complexity of model-based methods. These tests maintain nominal size when data have informativeness, where otherwise cluster-appropriate methods can be biased, and have comparable or even higher power to competitor tests when fluctuations of cluster or group size is non-informative. Their

broad applicability and convenience makes these tests the method of choice when informativeness is of a concern, and a legitimate alternative to established methods when performing simple hypotheses of marginal parameters in clustered data when variation of cluster or group size can be discounted.

REFERENCES

[1] Marc Aerts, Geert Molenberghs, Louise M Ryan, and Helena Geys. *Topics in modelling of clustered data*. CRC Press, 2002.

[2] Alan Agresti. Score and pseudo-score confidence intervals for categorical data analysis. *Statistics in Biopharmaceutical Research*, 3(2):163–172, 2011.

[3] Alan Agresti and Brent A Coull. Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, 52(2):119–126, 1998.

[4] Andrea L Behrman, Elizabeth Ardolino, Leslie R VanHiel, Marcie Kern, Darryn Atkinson, Douglas J Lorenz, and Susan J Harkema. Assessment of functional improvement without compensation reduces variability of outcome measures after human spinal cord injury. *Archives of physical medicine and rehabilitation*, 93(9):1518–1529, 2012.

[5] Joe Bible, James D Beck, and Somnath Datta. Cluster adjusted regression for displaced subject data (cards): Marginal inference under potentially informative temporal cluster size profiles. *Biometrics*, 72(2):441–451, 2016.

[6] Colin R Blyth and Harold A Still. Binomial confidence intervals. *Journal of the American Statistical Association*, 78(381):108–116, 1983.

[7] Zhen Chen, Bo Zhang, and Paul S Albert. A joint modeling approach to data with informative cluster size: robustness to the cluster size model. *Statistics in medicine*, 30(15):1825–1836, 2011.

[8] Xiuyu J Cong, Guosheng Yin, and Yu Shen. Marginal analysis of correlated failure time data with informative cluster sizes. *Biometrics*, 63(3):663–672, 2007.

[9] Somnath Datta, Jaakko Nevalainen, and Hannu Oja. A general class of signed-rank tests for clustered data when the cluster size is potentially informative. *Journal of nonparametric statistics*, 24(3):797–808, 2012.

[10] Somnath Datta and Glen A Satten. Rank-sum tests for clustered data. *Journal of the American Statistical Association*, 100(471):908–915, 2005.

[11] Somnath Datta and Glen A Satten. A signed-rank test for clustered data. *Biometrics*, 64(2):501–507, 2008.

[12] Natalie Dean and Marcello Pagano. Evaluating confidence interval methods for binomial proportions in clustered surveys. *Journal of Survey Statistics and Methodology*, 3(4):484–503, 2015.

[13] David B Dunson, Zhen Chen, and Jean Harry. A bayesian approach for joint modeling of cluster size and subunit-specific outcomes. *Biometrics*, 59(3):521–530, 2003.

[14] Valerie L Durkalski, Yuko Y Palesch, Stuart R Lipsitz, and Philip F Rust. Analysis of clustered matched-pair data. *Statistics in medicine*, 22(15):2417–2428, 2003.

[15] Sandipan Dutta and Somnath Datta. *ClusterRankTest: Rank Tests for Clustered Data*, 2016. R package version 1.0.

[16] Sandipan Dutta and Somnath Datta. A rank-sum test for clustered data when the number of subjects in a group within a cluster is informative. *Biometrics*, 72(2):432–440, 2016.

[17] Bradley Efron and Charles Stein. The jackknife estimate of variance. *The Annals of Statistics*, pages 586–596, 1981.

[18] Michael Eliasziw and Allan Donner. Application of the mcnemar test to non-independent matched pair data. *Statistics in medicine*, 10(12):1981–1991, 1991.

[19] Dean Follmann and Michael Fay. Exact inference for complex clustered data using within-cluster resampling. *Journal of biopharmaceutical statistics*, 20(4):850–869, 2010.

[20] Gail F Forrest, Douglas J Lorenz, Karen Hutchinson, Leslie R VanHiel, D Michele Basso, Somnath Datta, Sue Ann Sisto, and Susan J Harkema. Ambulation and balance outcomes measure different aspects of recovery in individuals with chronic, incomplete spinal cord injury. *Archives of physical medicine and rehabilitation*, 93(9):1553–1564, 2012.

[21] FP Franchignoni, L Tesio, C Ricupero, and MT Martino. Trunk control test as an early predictor of stroke rehabilitation outcome. *Stroke*, 28(7):1382–1385, 1997.

[22] BK Ghosh. A comparison of some approximate confidence intervals for the binomial parameter. *Journal of the American Statistical Association*, 74(368):894–900, 1979.

[23] Dan Gopstein. *clust.bin.pair: Statistical Methods for Analyzing Clustered Matched Pair Data*, 2018. R package version 0.1.2.

[24] Mary E Gregg, Somnath Datta, and Doug Lorenz. A log rank test for clustered data with informative within-cluster group size. *Statistics in medicine*, 37(27):4071–4082, 2018.

[25] Ulrich Halekoh, Søren Højsgaard, Jun Yan, et al. The r package geepack for generalized estimating equations. *Journal of Statistical Software*, 15(2):1–11, 2006.

[26] Susan J Harkema, Mary Schmidt-Read, Andrea L Behrman, Amy Bratta, Sue Ann Sisto, and V Reggie Edgerton. Establishing the neurorecovery network: multisite rehabilitation centers that provide activity-based therapies and assessments for neurologic disorders. *Archives of physical medicine and rehabilitation*, 93(9):1498–1507, 2012.

[27] Susan J Harkema, Carrie Shogren, Elizabeth Ardolino, and Douglas J Lorenz. Assessment of functional improvement without compensation for human spinal cord injury: extending the neuromuscular recovery scale to the upper extremities. *Journal of neurotrauma*, 33(24):2181–2190, 2016.

[28] David V Hinkley. Jackknifing in unbalanced situations. *Technometrics*, 19(3):285–292, 1977.

[29] David V Hinkley. Improving the jackknife with special reference to correlation estimation. *Biometrika*, 65(1):13–21, 1978.

[30] Elaine B Hoffman, Pranab K Sen, and Clarice R Weinberg. Within-cluster resampling. *Biometrika*, 88(4):1121–1134, 2001.

[31] Ying Huang and Brian Leroux. Informative cluster sizes for subcluster-level covariates and weighted generalized estimating equations. *Biometrics*, 67(3):843–851, 2011.

[32] Ivan Iachine, Hans Chr Petersen, and Kirsten O Kyvik. Robust tests for the equality of variances for clustered data. *Journal of Statistical Computation and Simulation*, 80(4):365–377, 2010.

[33] Ana-Maria Iosif and Allan R Sampson. A model for repeated clustered data with informative cluster sizes. *Statistics in medicine*, 33(5):738–759, 2014.

[34] Yujing Jiang. *clusrank: Wilcoxon Rank Sum Test for Clustered Data*, 2018. R package version 0.6-2.

[35] Göran Kauermann and Raymond J Carroll. A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, 96(456):1387–1396, 2001.

[36] Edward L Korn and Barry I Graubard. Confidence intervals for proportions with small expected number of positive counts estimated from survey data. *Survey Methodology*, 24:193–201, 1998.

[37] Howard Levene. Contributions to probability and statistics. *Essays in honor of Harold Hotelling*, pages 278–292, 1960.

[38] Peng Li and David T Redden. Small sample performance of bias-corrected sandwich estimators for cluster-randomized trials with binary outcomes. *Statistics in medicine*, 34(2):281–296, 2015.

[39] Douglas J Lorenz, Somnath Datta, and Susan J Harkema. Marginal association measures for clustered data. *Statistics in medicine*, 30(27):3181–3191, 2011.

[40] Douglas J Lorenz, Steven Levy, and Somnath Datta. Inferring marginal association with paired and unpaired clustered data. *Statistical methods in medical research*, 27(6):1806–1817, 2018.

[41] Aya A Mitani, Elizabeth K Kaye, and Kerrie P Nelson. Marginal analysis of ordinal clustered longitudinal data with informative cluster size. *Biometrics*, 75(3):938–949, 2019.

[42] John M Neuhaus and Charles E McCulloch. Estimation of covariate effects in generalized linear mixed models with informative cluster sizes. *Biometrika*, 98(1):147–162, 2011.

[43] Jaakko Nevalainen, Somnath Datta, and Hannu Oja. Inference on the marginal distribution of clustered data with informative cluster size. *Statistical Papers*, 55(1):71–92, 2014.

[44] Jaakko Nevalainen, Hannu Oja, and Somnath Datta. Tests for informative cluster size using a novel balanced bootstrap scheme. *Statistics in medicine*, 36(16):2630–2640, 2017.

[45] Robert G Newcombe. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in medicine*, 17(8):857–872, 1998.

[46] Nancy A Obuchowski. On the comparison of correlated proportions for clustered data. *Statistics in medicine*, 17(13):1495–1507, 1998.

[47] Menelaos Pavlou. *Analysis of clustered data when the cluster size is informative*. PhD thesis, UCL (University College London), 2012.

[48] José Pinheiro and Douglas Bates. *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media, 2006.

[49] Shaun R Seaman, Menelaos Pavlou, and Andrew J Copas. Methods for observed-cluster inference when cluster size is informative: a review and clarifications. *Biometrics*, 70(2):449–456, 2014.

[50] Jun Shao. The efficiency and consistency of approximations to the jackknife variance estimators. *Journal of the American Statistical Association*, 84(405):114–119, 1989.

[51] Arnold J Stromberg. Robust covariance estimates based on resampling. *Journal of Statistical Planning and Inference*, 57(2):321–334, 1997.

[52] Ming Wang, Maiying Kong, and Somnath Datta. Inference for marginal linear models for clustered longitudinal data with potentially informative cluster sizes. *Statistical Methods in Medical Research*, 20(4):347–367, 2011.

[53] John M Williamson, Somnath Datta, and Glen A Satten. Marginal analyses of clustered data when cluster size is informative. *Biometrics*, 59(1):36–42, 2003.

[54] John M Williamson, Hae-Young Kim, Amita Manatunga, and David G Addiss. Modeling survival data with informative cluster size. *Statistics in medicine*, 27(4):543–555, 2008.

[55] Chien-Fu Jeff Wu et al. Jackknife, bootstrap and other resampling methods in regression analysis. *the Annals of Statistics*, 14(4):1261–1295, 1986.

[56] Zhao Yang, Xuezheng Sun, and James W Hardin. A note on the tests for clustered matched-pair binary data. *Biometrical journal*, 52(5):638–652, 2010.

[57] Bo Zhang, Wei Liu, Zhiwei Zhang, Yanping Qu, Zhen Chen, and Paul S Albert. Modeling of correlated data with informative cluster sizes: an evaluation of joint modeling and within-cluster resampling approaches. *Statistical methods in medical research*, 26(4):1881–1895, 2017.

APPENDIX A: Commonly Used Acronyms

CWGEE - cluster-weighted generalized estimating equation

DWGEE - doubly-weighted generalized estimating equation

GEE - generalized estimating equation

ICS - informative cluster size

IWCGS - informative within-cluster group size

WCR - within-cluster resampling

APPENDIX B: Simulation Code for Screen8 Data

```
library(MASS)
set.seed(15000)
M <- 73
u <- rnorm(M, m=0, sd=1)
ni <- rpois(M, 30+5*u)

###   IDENTIFIERS
sch.id <- rep(1:M, ni)
stud.id <- as.numeric(unlist(tapply(sch.id, sch.id, function(x) 1:length(x))))

###   STANDARDIZED TEST SCORES - CLUSTER SIZE NEGATIVELY INFORMATIVE
tmp.math <- rnorm(sum(ni), m=70-rep(4*u, ni), sd=8)
math <- round(pmin(tmp.math,100))
tmp.read <- rnorm(sum(ni), m=60-rep(2*u, ni), sd=10)
read <- round(pmin(tmp.read,100))

###   DEMOGRAPHICS AND BIOMETRICS
###   PROPORTION OF MALES AT SCHOOL INCREASES WITH CLUSTER SIZE
p.male <- .25+.5*(ni-min(ni))/(max(ni)-min(ni))
n.male <- rbinom(M, size=ni, prob=p.male)
gender <- factor(unlist(apply(cbind(n.male,ni-n.male), 1,
function(x) rep(c("M","F"), x))))
###   AGE IS UNRELATED TO LATENT FACTOR
age <- sample(13:15, size=sum(ni), replace=T)
###   HEIGHT AND WEIGHT ATTEMPTED TO FOLLOW 14 YO AVERAGES,
###   HAVE SENSIBLE BMI (703*w/h^2)
height <- weight <- rep(NA, sum(ni))
tmp <- mvrnorm(sum(gender=="M"), mu=c(65,140),
Sigma=cbind(c(8,.7*sqrt(8)*16),c(.7*sqrt(8)*16,256)))
height[gender=="M"] <- round(tmp[,1])
weight[gender=="M"] <- round(tmp[,2])
tmp <- mvrnorm(sum(gender=="F"), mu=c(64,122),
Sigma=cbind(c(7,.6*sqrt(7)*15),c(.6*sqrt(7)*15,225)))
height[gender=="F"] <- round(tmp[,1])
weight[gender=="F"] <- round(tmp[,2])

###   STUDENT-LEVEL CATEGORICAL VARIABLE
###   EXTRACURRICULAR ACTIVITY
###   proportion of students who participates in sports increases
###   with cluster size (opposite for academics)
tmp <- cbind(0.1 + .5*(ni-min(ni))/(max(ni)-min(ni)), 0.1 +
.5*(1-(ni-min(ni))/(max(ni)-min(ni))))
tmp <- cbind(tmp, 1-rowSums(tmp))
```

```
tmp.fun <- function(cl) {
  t(rmultinom(1, ni[cl], prob=tmp[cl,]))
}
aa.ni <- matrix(unlist(lapply(1:length(ni), tmp.fun)), ncol=3, byrow=T)
activity <-  factor(unlist(apply(aa.ni, 1,
function(x) sample(rep(c("sports","academic", "other"), x)))))


###   MENTAL HEALTH VARIABLE - INFORMATIVE; LARGER SCHOOLS, HIGHER SCORES
phq2 <- rbinom(sum(ni), size=6, prob=.5*(ni-min(ni))/(max(ni)-min(ni)))

###   FITNESS QUARTILE - NONINFORMATIVE, RELATED TO BMI
bmi <- 703*weight/height^2
zbmi <- (bmi-mean(bmi))/sd(bmi)
tmp <- rbinom(sum(ni), size=3, prob=2*pnorm(-abs(zbmi)))+1
qfit <- factor(tmp, labels=c("Q1","Q2","Q3","Q4"))

###   SECOND FITNESS QUARTILE - FROM BEGINNING OF SCHOOL YEAR ()
bmi2 <- bmi + rnorm(sum(ni), mean=-0.5, sd=1)
zbmi2 <- (bmi2-mean(bmi2))/sd(bmi2)
tmp2 <- rbinom(sum(ni), size=3, prob=2*pnorm(-abs(zbmi2)))+1
qfit.s <- factor(tmp2, labels=c("Q1","Q2","Q3","Q4"))

###   DATA FRAME
screen8 <- data.frame(sch.id, stud.id, age, gender, height,
    weight, math, read, phq2, qfit, qfit.s, activity)
```

CURRICULUM VITA

Mary Gregg

Education

Ph.D. (August 2020), Biostatistics, University of Louisville, Louisville, KY, USA.

M.S. (May 2016), Biostatistics, University of Louisville, Louisville, KY, USA.

B.A. (June 2009), Music, Bennington College, Bennington, VT, USA.

Publications

Peer Reviewed Methodological Manuscripts

1. **Gregg, M.**, Datta, S., Lorenz, D.J. (2020). Variance estimation in tests of clustered categorical data with informative cluster size. *Statistical Methods in Medical Research.* doi:10.1177/0962280220928572.

2. **Gregg, M.**, Datta, S., Lorenz, D.J. (2018). A log rank test for clustered data with informative within-cluster group size. *Statistics in Medicine.* 37(27), 4071-82.

Peer Reviewed Collaborative Manuscripts

1. Baca, J., Foster, C., Simon, N.J., Lorenz, D., **Gregg, M.**, Schinasi, D. (2020). Children's hospital transfers from referring emergency departments: which patients bypassed the pediatric ED? Submitted.

2. Rosado, N., Charleston, E., **Gregg, M.**, Lorenz, D. (2019). Characteristics of accidental versus abusive pediatric burn injuries in an urban burn center over a 14-year period. *Journal of Burn Care & Research.* 40(4), 437-443.

3. Behrman, A.L., Trimble, S.A., Argetsinger, L.C., Roberts, M.T., Mulcahey, M.J., Clayton, L., **Gregg, M.**, Lorenz, D., Ardolino, E.M. (2019). Interrater reliability of the pediatric neuromuscular recovery scale for spinal cord injury. *Topics in Spinal Cord Injury Rehabilitation*, 25(2), 121-131.

4. Sharp, M.K., **Gregg, M.**, Brock, G., Nir, N., Sahetya, S., Austin, E.H., Mascio, C., Slaughter, M.D., Pantalos, G.M. (2017). Comparison of pediatric and adult blood viscoelasticity. *Cardiovascular Engineering and Technology*, 8(2), 182-192.

Conference Posters/Presentations

Peer Reviewed Presentations

1. **Gregg, M.**, Lorenz, D., htest.clust: An R package for marginal inference of clustered data with informative cluster size. Conference on Statistical Practice, 2020, Sacramento, CA.

2. Johnson, N., Raviv, T., Simonton, K., Rosado, N., Charleston, E., **Gregg, M.**, Hilliard, M., Pierce, M. Documentation of sexual abuse history in patients with psychiatric complaints who are evaluated in a pediatric Emergency Department. Pediatric Academic Societies Meeting, 2019, Baltimore, MD.

3. Baca, J., Simon, N.J., Foster, C., Lorenz, D., **Gregg, M.**, Schinasi, D. Community Emergency Department (ED) patients transferred to a children's hospital for admission: who is routed to the pediatric ED prior to hospitalization? Pediatric Academic Societies Meeting, 2019, Baltimore, MD.

4. Sharp, K., **Gregg, M.**, Brock, G., Pantalos, G. Comparison of pediatric and adult blood viscoelasticity. Summer Biomechanics, Bioengineering and Biotransport Conference, 2016, National Harbor, MD.

5. Redman, R.A., Linden, C., Perez, C.A., Dunlap, N., Silverman, C., Tennant, P., Bumpous, J., **Gregg, M.**, Wu, X., and Rai, S. Effect of angiotensin converting enzyme inhibition on toxicity in patients undergoing radiation with or without chemotherapy for head and neck cancer. American Society of Clinical Oncology Annual Meeting, 2015, Chicago, IL.

Teaching Experience

University of Louisville

2015 – 2018	Graduate Student Assistant REACH (Resources for Academic Achievement) Courses: Special Topics in College Mathematics
2014 – 2015	Math Tutor REACH (Resources for Academic Achievement)

Awards and Honors

University of Louisville

2018 – 2020	University Fellowship
2016 – 2018	University Faculty Favorite (3-time)
2015	Math Tutor of the Year REACH (Resources for Academic Achievement)