# htestClust: A Package for Marginal Inference of Clustered Data Under Informative Cluster Size

*by Mary Gregg, Somnath Datta and Douglas Lorenz*

**Abstract** When observations are collected in/organized into observational units, within which observations may be dependent, those observational units are often referred to as "clustered" and the data as "clustered data". Examples of clustered data include repeated measures or hierarchical shared association (e.g., individuals within families). This paper provides an overview of the R package **htestClust**, a tool for the marginal analysis of such clustered data with potentially informative cluster and/or group sizes. Contained in **htestClust** are clustered data analogues to the following classical hypothesis tests: rank-sum, signed rank, *t*-, one-way ANOVA, F, Levene, Pearson/Spearman/Kendall correlation, proportion, goodness-of-fit, independence, and McNemar. Additional functions allow users to visualize and test for informative cluster size. This package has an easy-to-use interface mimicking that of classical hypothesis-testing functions in the R environment. Various features of this package are illustrated through simple examples.

## 1 Introduction

Observations often occur or can be organized into units called clusters, within which those observations may be dependent. For example, individuals may be repeatedly assessed or naturally belong to some hierarchical structure like a family unit. Potential correlation among intra-cluster observations clearly invalidates the use of classical hypothesis tests for the analysis of such data. Instead, inference is generally performed using model-based methods that capture intra-cluster relationships through parametric or semi-parametric assumptions. Generalized estimating equations (GEEs) are one such approach that fit marginal generalized linear models to clustered data while making a working assumption on the correlation structure. GEE models are appealing for their flexible and robust nature, and several packages in the R environment, such as **gee** (Carey et al., 2019) and **geepack** (Halekoh et al., 2006), offer an implementation of this method. However, GEEs and other standard methods for analysis of clustered data operate under an assumption that the number of observations within the clusters (defined as the cluster size) is ignorable. In practice, this assumption may not hold and cluster size may vary systematically in a way that carries information related to the response of interest. When this occurs data are said to have informative cluster size (ICS). Examples of ICS can be found in data related to dental health (Williamson et al., 2003), pregnancy studies (Chaurasia et al., 2018), and longitudinal rehabilitation (Lorenz et al., 2011), among others. For data with ICS, standard model-based methods can produce biased inference as their estimates may be overweighted in favor of larger clusters.

A related but distinct type of informativeness occurs when the distribution of group-defining covariates varies in a way that carries information on the response. Such phenomenon has been called informative within-cluster group size (IWCGS), as well as informative covariate structure (Pavlou, 2012), sub-cluster covariate informativeness (Lorenz et al., 2018), and informative intra-cluster group size (Dutta and Datta, 2016a). This additional informativeness may occur simultaneously with or separately from ICS, and similarly can result in the failure of standard methods to maintain appropriate nominal size (Huang and Leroux, 2011; Dutta and Datta, 2016a).

Williamson et al. (2003) developed a reweighting methodology that corrects for potential bias from cluster- or group-size informativeness. This reweighting originates from a Monte Carlo resampling process, and leads to weighting observations proportional to their inverse cluster or within-cluster group size. Correction for ICS/IWCGS was originally proposed in the context of modeling, and a number of extensions to this application have been established (Bible et al., 2016; Iosif and Sampson, 2014; Mitani et al., 2019, 2020). However, when adjustment for covariates is not of interest, this reweighting can be directly applied in the estimation of marginal parameters. Under mild conditions, such estimates are asymptotically normal, permitting Wald-type intervals and tests. This methodology has been applied to develop rank-based tests (Datta and Satten, 2005, 2008; Dutta and Datta, 2016a), and tests of correlation (Lorenz et al., 2011), proportions (Gregg et al., 2020), means and variances (Gregg, 2020). This collection of reweighted non-model-based hypothesis tests includes clustered data analogues of the following classical tests: rank-sum, signed rank, *t*-, one-way ANOVA, F, Levene, Pearson/Spearman/Kendall correlation, proportion, goodness-of-fit, independence, and McNemar.

These clustered data analogues to standard hypothesis tests provide simple and intuitive means of

performing exploratory and preliminary analysis of clustered data in which the cluster and/or group size varies and is potentially informative. However, many of these tests are recent developments that are not available in a software environment. We address this deficiency through the package **htestClust**, the first R package designed as a comprehensive collection of direct, non-model-based inferential methods for analysis of clustered data with potential ICS and/or IWCGS. Introduced in this paper, **htestClust** implements the collection of methods by Datta and Satten (2005, 2008); Dutta and Datta (2016a); Lorenz et al. (2011); Gregg et al. (2020) and Gregg (2020), as well as a method by Nevalainen et al. (2017) that tests for the presence of informative cluster size. The syntax and output of functions contained in **htestClust** are intentionally modeled after their corresponding analogous classical function, allowing researchers to assess various marginal analyses through intuitive and user-friendly means. The rest of this paper is organized as follows. We will begin by briefly summarizing the reweighting approach developed by Williamson et al. (2003) and describe how its application has been used in the development of hypothesis tests of marginal parameters in clustered data. We will then provide an overview of the **htestClust** package, describe the features and structure of functions, and describe an illustrative simulated data set with informativeness. Finally, we will demonstrate **htestClust** using the example data set and close with a discussion.

## 2 Methods for clustered data under informativeness

In this section we outline the weighting methodology that corrects for bias from ICS and IWCGS, and describe the general form of the tests in **htestClust** that implement this weighting. We then summarize the balanced bootstrap design implemented in the test of ICS by Nevalainen et al. (2017).

### Notation

Consider a sample of $M$ independent clusters, with each cluster containing $n_i$ potentially correlated observations, $i = 1, \ldots, M$. The $j^{th}$ observation from cluster $i$ is $X_{ij}$, with $j = 1, \ldots, n_i$. The collection of data from cluster $i$ is $\boldsymbol{V}_i = \{n_i, X_{i1}, \ldots, X_{in_i}\}$ and the set of all observed data is $\boldsymbol{V} = \{\boldsymbol{V}_1, \ldots, \boldsymbol{V}_M\}$. Informative cluster size is defined as inequality between the marginal distribution of the response $X$ and the distribution of $X$ conditional on cluster size: $P(X_{ij} \leq x \mid n_i = n) \neq P(X_{ij} \leq x), n = 1, 2, \ldots; j = 1, \ldots, n_i$.

When observations within clusters belong to one of $K$ distinct groups, we define the variable $G_{ij} = k$ to represent that observation $j$ from cluster $i$ belongs to group $k$, $k = 1, \ldots, K$. We let $n_i^{(k)}$ denote the number of observations from cluster $i$ in group $k$, and note that $n_i = \sum_{k=1}^{K} n_i^{(k)}$. We define $K_i^c = \sum_{k=1}^{K} I[n_i^{(k)} > 0]$ to be the number of distinct groups observed in cluster $i$. When $K_i^c < K$, not all groups are observed in cluster $i$, a condition referred to as incomplete group structure. The data from cluster $i$ is now the set $\boldsymbol{V}_i = \{n_i^{(k)}, (X_{ij}, G_{ij})\}$, with observations belonging to group $k$ denoted as the set $\{X_{i1}^{(k)}, \ldots, X_{in_i^{(k)}}^{(k)}\}$. Informative within-cluster group size can be defined as $P\left(X_{ij} \leq x \mid n_i^{(k)}\right) \neq P\left(X_{ij} \leq x\right)$, i.e. that the marginal distribution of $X$ differs from the distribution of $X$ conditional on the within-cluster group size.

### Weighting for ICS/IWCGS

Let $\theta$ denote a marginal parameter to be estimated and/or tested. One approach for estimating $\theta$ is within-cluster resampling (WCR), in which one observation is randomly selected from each cluster (Hoffman et al., 2001). The resulting subset of data, $\boldsymbol{X}^* = \{X_1^*, X_2^*, \ldots, X_M^*\}$, consists of independent observations so an estimate of the parameter, $\hat{\theta}$, can be calculated using standard i.i.d. methods. Clearly, this estimate is inefficient, using only a subset of the data, so the resampling process is repeated many times, creating many pseudo data sets and estimates $\hat{\theta}_q^*$. An overall estimate of $\theta$ is obtained over $Q$ resamplings ($Q$ large) by averaging the resampled estimates, $\hat{\theta}^* = \frac{1}{Q}\sum_{q=1}^{Q} \hat{\theta}_q^*$. This estimator was shown to be asymptotically normal and inference can be conducted using Wald-type intervals and tests.

The method of reweighting proposed by Williamson et al. (2003) derives from WCR by noting that as $M, Q \to \infty$, the overall resampled estimator converges to $\hat{\theta} = E\left[\hat{\theta}_q^* \mid \boldsymbol{V}\right]$ with respect to the resampling distribution. This marginalization is equivalent to averaging the resampled estimator across all realizations of the resampled data. As sampling is uniform across clusters and with equal

probability within each cluster, each observation is weighted by the inverse of the associated cluster size.

The link between WCR and reweighting can be illustrated by a simple example - estimating a marginal mean. For a single resampled data set produced by WCR, the estimate of the marginal mean is the simple average, $\hat{\theta}_q^* = \frac{1}{M}\sum_{i=1}^{M} X_i^*$. Application of the marginalization calculation produces

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

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 $\boldsymbol{V}$, the expectation of a resampled observation from a particular cluster is the average of all observations from the cluster, as the WCR process resamples observations from that cluster with equal probability.

The weighting that corrects for ICS can be adapted to correct for IWCGS by modifying the underlying resampling process into a two-step procedure that marginalizes the within-cluster distribution of groups (Dutta and Datta, 2016a; Huang and Leroux, 2011). In this two-step resampling, we first select a group, $G_i^*$, with uniform probability from the levels of $G$ available in cluster $i$. Second, we select $X_i^*$ from the set of observations in group $k$, $\{X_{i1}^{(k)}, \ldots, X_{in_i^{(k)}}^{(k)}\}$, where $k$ is the group selected in the first step of the process. As in the original WCR methodology, this process is repeated for all clusters, resulting in a resampled data $(\boldsymbol{X}^*, \boldsymbol{G}^*) = \left\{(X_1^*, G_1^*), \ldots, (X_M^*, G_M^*)\right\}$. An estimate of the parameter of interest is calculated from this resampled data. When the marginalization calculation is applied to a single WCR estimate produced by this two-step process, observations are weighted by the product of the two selection probabilities - one for the selection of a group and one for the selection of an observation within the group. Since both of these selections are made with equal probability, the weights in a given cluster are defined by the number of groups available in that cluster and the number of observations within the group:

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

### Hypothesis tests of marginal parameters

The asymptotic normality of the estimators described in the previous section has been established under mild regularity conditions (Datta and Satten, 2005, 2008; Williamson et al., 2003). The tests of ranks, correlation, proportions, means and variances contained in **htestClust** all leverage this asymptotic normality through the general univariate and multivariate Wald-type forms

$$Z = \frac{S - E\left[S\right]}{\sqrt{\hat{V}\left(S\right)}} \qquad \boldsymbol{X} = \left(\boldsymbol{S} - E(\boldsymbol{S})\right)^T \left(\hat{\boldsymbol{V}}(S)\right)^{-1} \left(\boldsymbol{S} - E(\boldsymbol{S})\right).$$

The statistic, $S$, differs across the various tests. However, in each of the tests $S$ is either a reweighted estimator derived through the marginalization calculation or a smooth function of such reweighted estimators. $E[S]$ is the statistic's expected value under the null hypothesis and $\hat{V}\left(S\right)$ is an estimate of the variance of $S$. $Z$ asymptotically follows a standard normal distribution, while $\boldsymbol{X}$ asymptotically follows a chi square distribution with $K-1$ degrees of freedom.

Methods of estimating the variance of $S$ also vary across the tests. The rank-sum and signed rank tests weighted for ICS apply Hajek projections (Datta and Satten, 2005, 2008), while the tests of correlation use an approach based on the empirical variances of within-cluster averages (Lorenz et al., 2011). The rank-sum test weighted for IWCGS and the multi-group tests of means and variances use jackknife estimates (Dutta and Datta, 2016a; Gregg, 2020). The tests of proportions were constructed and evaluated under different variance estimation techniques including sandwich forms, method of moments, and empirical estimates. Gregg et al. (2020) provide a detailed examination by simulation of different variance estimation techniques in the context of estimating and testing proportions, and note that no one variance estimation technique is optimal for different types of tests. Further, the size and power of the tests in **htestClust** previously have been evaluated via simulation in the source manuscripts for each test. Predictably, each has been shown to perform well under the informativeness conditions for which they were designed to adjust.

### Testing for informative cluster size

Nevalainen et al. (2017) proposed a test for ICS using a novel balanced bootstrap scheme. As it might be desirable to perform this test prior to the application of the marginal methods mentioned thus far, we have included 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 $V_i = \left(n_i; X_{i1}, \ldots, X_{in_i}\right)$ is the data from cluster $i$. Assuming exchangeability of observations within clusters, the hypothesis of interest is $H_0 : P\left(X_{ij} \leq x | n_i = k\right) = F(x), k = 1, 2, \ldots; j = 1, \ldots, k$, for some unknown distribution $F$. Two test statistics are proposed for testing $H_0$; a Kolmogorov-Smirnov type statistic takes the form

$$T_F = sup_x |\hat{F}(x) - \tilde{F}(x)|$$

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

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

where $\psi$ represents the set of unique cluster sizes, $M_k$ represents the number of clusters of size $k$, and $\hat{F}_k(x) = \frac{1}{kM_k}\sum_{i=1}^{M}\sum_{j=1}^{n_i} I\left[n_i = k, X_{ij} \leq x\right]$. $T_{CM}$ is suggested for use when there is a small number of distinct cluster sizes, as it tends to be more powerful. $T_F$ is preferred when the number of distinct cluster sizes is large and the number of clusters with those sizes is small, as $T_{CM}$ tends to be too liberal.

The bootstrap scheme, which is employed for either statistic, is as follows. For 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., $V_{bi}^* = \left(n_i; X_{i^*1}, \ldots, X_{i^*n_i}\right)$.
   (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., $V_{bi}^* = \left(n_i; X_{i^*1}, \ldots, X_{i^*n_i^*}, X_{k(n_i^*+1)}, \ldots, X_{kn_i}\right)$, where $k = \arg\min_k\{D(V_{i^*}, V_k) : n_k \geq n_i\}$. The closest matching cluster is determined by minimum distance calculated by $D\left(V_i, V_j\right) = \left(\min\{n_i, n_j\}\right)^{-1}\sum_{k=1}^{\min\{n_i, n_j\}}\left(X_{ik} - X_{jk}\right)^2$.
3. Calculate the test statistic from the collection of bootstrapped clusters, $T_b^* = T\left(\boldsymbol{V}_b^*\right)$, $\boldsymbol{V}_b^* = \left(V_{b1}^*, \ldots, V_{bM}^*\right)$.

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.

## 3 Overview of htestClust

**htestClust** includes ten functions for conducting different hypothesis tests under ICS, one function for visualizing informativeness in cluster size, and a simulated hypothetical data set to illustrate the use of the functions. We first note that, at the time of this publication, we are aware of only two other R packages available on CRAN that provide functions for analyzing data under ICS and IWCGS: **clusrank** (Jiang, 2018) and **ClusterRankTest** (Dutta and Datta, 2016b). Each of these packages provides functionality only for rank-based tests for clustered data, i.e. clustered data analogues of the well-known Wilcoxon signed rank and rank sum tests. We know of no other R package that includes the broad range of tests of means, proportions, variances, and correlations in addition to these rank-based tests that is provided by **htestClust**.

### Package functions, syntax, and output

With the exception of the test of informative cluster size, each of the hypothesis testing functions implemented in **htestClust** has a well-known analogue test for i.i.d. data (Table 1). As such, the syntax and output of the functions in **htestClust** are designed to conform with that of the analogous i.i.d. functions from the R **stats** library. A notable but necessary departure from this correspondence is that

| **htestClust** function | Reweighted test(s) | Classical analogue 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()` |

**Table 1:** Hypothesis testing functions available in the **htestClust** package. Each row gives the name of a **htestClust** function, the reweighted test the function performs, and the R function that executes the corresponding classical analogue test. All classical analogue functions are available in R through the **stats** package, except for `leveneTest()`, which is included in the **car** package.

the **htestClust** functions require as input (1) a variable identifying the clusters as an argument in the data set or (2) a cluster-level summary of the data.

As an example, consider the syntax for the **stats** and **htestClust** functions for conducting a test of a single proportion:

```
prop.test(x, n, p = NULL, alternative = c("two.sided", "less", "greater"),
conf.level = 0.95, correct = TRUE)

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

The **stats** library function `prop.test` does not operate on variables in a data frame, but instead takes summary counts as its input. Argument `x` can be a scalar representing the number of binomial successes, whence `n` is required as the number of binomial trials. Alternatively, `x` can be a one-dimensional table or matrix with two entries, whence `n` is omitted. The remaining arguments customize the test in ways familiar to most users.

The function `proptestClust` from **htestClust** operates on binary variables in a data frame or on cluster-level summary counts. In this function, `x` may be a binary variable measured over clusters, wherein `id` is required as a vector of cluster identifiers. Alternatively, `x` may instead be a two-dimensional table of within-cluster counts of failures and successes, wherein `id` is omitted. As previously noted, several options are available for variance estimation; these may be selected by the user through the `variance` argument. Additional customization of the test is as in `prop.test`.

Each of the testing functions in **htestClust** has been constructed in this vein – parallel to the analogous **stats** function with contingencies necessary for clustered data. **htestClust** functions accept vector input that designates the response, grouping (if necessary), and clustering variables. However, for convenience, many functions are designed with a secondary interface accepting tables or formulas. Like their **stats** package analogues, **htestClust** testing functions produce `list` objects of class `htest` for which the `print` method behaves in the usual way.

`icsPlot` provides a simple method for illustrating informative cluster size, providing a visual supplement to the results of the test of ICS, `icstestClust`. Briefly, `icsPlot` plots a within-cluster summary statistic of a variable, such as a mean, against the size of each cluster. For quantitative variables, `icsPlot` produces a scatterplot of a within-cluster measure of location (mean, median) or variation (SD, variance, IQR, range) against cluster size. For a categorical variable, a barplot of within-cluster proportions is produced.

### Simulated example data set

**htestClust** includes a simulated data set named `screen8` of clustered observations with informativeness, created under a hypothetical scenario we briefly describe here. A large school district has conducted a voluntary comprehensive exit survey for students graduating elementary school, collecting demographic, biometric, and academic performance data. The clustering mechanism for these data are the schools, with students comprising the observations within clusters.

The school district has offered an incentive program to boost participation, wherein schools having

| Variable | Description |
|---|---|
| sch.id | School identification variable |
| stud.id | Student identification variable within school |
| age | Student age in years |
| gender | Student gender |
| height | Student height in inches |
| weight | Student weight in lbs |
| math | Student score on standardized math test |
| read | Student score on standardized reading test |
| phq2 | Ordinal (0-6) score from a mental health screening. Higher scores correspond to higher levels of depression |
| qfit | Age-adjusted fitness quartile from physical health assessment taken at end of school year |
| qfit.s | Age-adjusted fitness quartile from physical health assessment taken at beginning of school year |
| activity | Student after-school activity |

**Table 2:** Variables in screen8 data set. Each row gives the name of a variable included in the screen8 data set and its associated description.

higher participation rates are rewarded with priority status for classroom and technology upgrades for the new academic year. This incentive introduces the potential for ICS – resource-poor schools may exhibit greater participation (larger cluster sizes) but also tend to have students with poorer health metrics and standardized test scores.

screen8 contains data from 2224 students from 73 schools in this district. Cluster sizes – the number of students participating in the exit survey at each school – ranged from 17 to 50, with a median of 30. The first few lines of the data are printed below, followed by the tabulated number of participants from each school and a summary of the cluster sizes. Table 2 provides details on the variables in the data set.

```
R> library(htestClust)
R> data(screen8)
R> 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

R> (tab <- table(screen8$sch.id))
1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
35 32 26 33 23 25 27 21 39 28 32 38 35 24 29 27 36 29 38 39 25 30 36 29 46 27
...

R> summary(as.vector(tab))
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
17.00   25.00   30.00   30.47   36.00   50.00
```

## 4 Examples

In this section, we demonstrate usage of the functions in **htestClust** using the `screen8` data set. Our illustration is not comprehensive, but users can learn more about functions not covered here by browsing the associated help files. To motivate the demonstration, we'll investigate the following questions:

1. Is the proportion of students having "proficient" standardized math test scores (65 or greater) more than 0.75?
2. Are participation in extracurricular activity and gender independent?
3. Are mean standardized math test scores different between male and female students?
4. Are mean standardized reading test scores different among groups defined by extracurricular activities?

### Evaluating informative cluster size

Before addressing these questions, we illustrate how to assess the potential informativeness of cluster size in the data set, starting by visualizing ICS through the `icsPlot` function. The arguments to `icsPlot` specify the variable of interest, a cluster-identifying variable, and a summary function to be applied to the variable within each cluster. This summary can be any of '`obs`', '`mean`', '`median`', '`var`', '`IQR`', '`range`', '`prop`', producing plots of the observations themselves, measure of location, or measures of variation against cluster size. Option '`prop`' can only be used when the variable of interest is a factor, so numerically coded categorical variables must be converted to factors. Standard R graphical parameters can also be specified when calling `icsPlot()`.

```
R> ### Figure 1
R> par(mfrow = c(1,2))
R> icsPlot(x = screen8$math, id = screen8$sch.id, FUN = "mean", pch = 20)
R> icsPlot(x = screen8$read, id = screen8$sch.id, FUN = "mean", pch = 20)

R> ### Figure 2
R> layout(mat = matrix(c(1, 2), nrow = 1, ncol = 2),
+        heights = c(1, 2), # Heights of the two rows
+        widths = c(2, 2.5))
R> par(mar = c(5, 4, 1, 0))
R> icsPlot(x = screen8$gender, id = screen8$sch.id, FUN = "prop",
+        ylab = "P(Female)", pch = 20)
R> par(mar = c(5, 4, 1, 5))
R> icsPlot(x = screen8$activity, id = screen8$sch.id, FUN = "prop",
+         legend = TRUE,
+         args.legend = list(x = "topright", bty = "n", inset=c(-0.32, 0)))
```

Figures 1 and 2 show potential informativeness in cluster size for the `screen8` data. Cluster size appears to be negatively associated with average standardized test scores but positively associated with the proportion of male students and the proportion participating in sports-related extracurricular activities. These empirical results can be verified using the test for ICS, implemented through the function `icstestClust`, as illustrated below. The result of this test suggests that cluster size is informative for standardized math test scores. Cluster size is also informative for standardized reading test scores, gender, and sports as an extracurricular activity ($p < .001$, results not shown).

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

R> ics.math
Test of informative cluster size (TF)
data:  screen8$math
TF = 0.029686, p-value < 2.2e-16
```

Within the `icstestClust` function, the type of test statistic, TF or TCM as detailed earlier, is specified using the `test.method` argument, and the number of bootstrap loops by argument `B`. Argument `print.it` is a logical indicating whether to print the progress of the bootstrap procedure. We note that the need for bootstrap resampling in `icstestClust` can make its implementation computationally expensive.

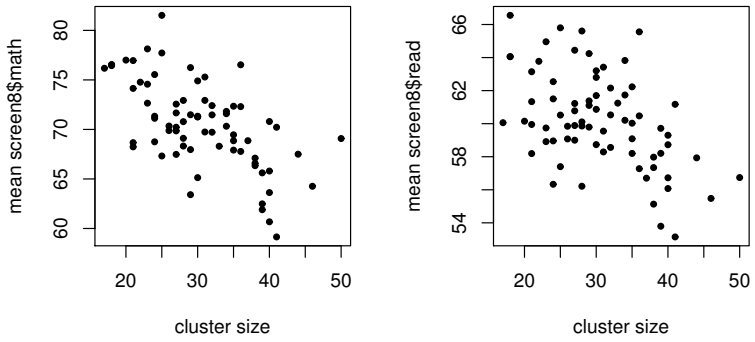


**Figure 1:** Average scores in maths and reading by cluster size in `screen8` data.

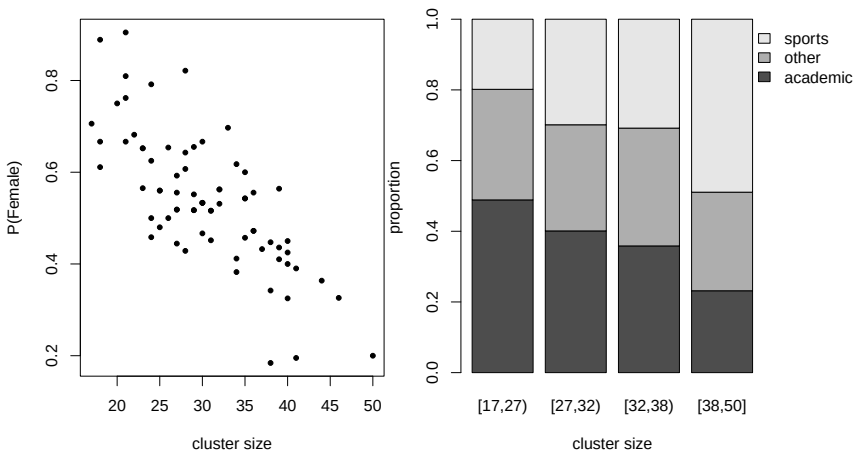


**Figure 2:** Plots of categorical variables by cluster size in `screen8` data. Proportion of female students decreases with cluster size (left), whereas student participation in sports-related extracurricular activities increases with cluster size (right).

### Testing a marginal proportion

The first question of interest suggests a one-sample test of a proportion via `proptestClust`. We specify a one-sided alternative and use the default sandwich variance estimator evaluated at the null value of the proportion (`variance = "sand.null"`), shown to perform best for this test (Gregg et al., 2020).

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

As noted earlier, **htestClust** functions produce objects of class `htest`, producing familiar output through the `print` method for such objects. We conclude that the proportion of students with proficient math test scores is not greater than 0.75.

In the case that all clusters have a size of 1, the results of **htestClust** functions will be in general

correspondence with that of the classical analogue test, though exact results will differ slightly due to the reweighted tests relying on asymptotics. This is demonstrated through the following example.

```
R> set.seed(123)
R> x <- rbinom(100, size = 1, p = 0.7)
R> id <- 1:100
R> proptestClust(x, id)

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

data:  x, M = 100
z = 4.2, p-value = 2.669e-05
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
0.6120018 0.8079982
sample estimates:
Cluster-weighted proportion
0.71

R> prop.test(sum(x), length(x))

1-sample proportions test with continuity correction

data:  sum(x) out of length(x), null probability 0.5
X-squared = 16.81, df = 1, p-value = 4.132e-05
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
0.6093752 0.7942336
sample estimates:
p
0.71
```

### Test of independence

The second question suggests a test of independence of extracurricular activity and gender. We start by producing cluster-weighted estimates of the proportion of students participating in each activity within each gender.

```
R> tab <- table(screen8$gender, screen8$activity, screen8$sch.id)
R> ptab <- prop.table(tab, c(1,3))
R> apply(ptab, c(1,2), mean)
academic     other    sports
F  0.3952102 0.2968473 0.3079425
M  0.3790267 0.3186699 0.3023035
```

The cluster-weighted proportions appear roughly similar, and we can test using `chisqtestClust`. Here, the default method of variance estimation is method of moments (`variance = "MoM"`), demonstrated to be best for the test of independence (Gregg et al., 2020).

```
R> chisqtestClust(screen8$gender, screen8$activity, screen8$sch.id)
Cluster-weighted Chi-squared test of independence with variance est:
MoM

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

Before proceeding to the next analysis, we note that further evidence of ICS in the `screen8` data can be demonstrated by implementing the standard chi-squared test for this question, which suggests that females were more likely to participate in academic extracurricular activities and males in sports.

```
R> prop.table(table(screen8$gender, screen8$activity), 1)
academic     other    sports
F  0.3891323 0.2979842 0.3128834
M  0.3370268 0.3120960 0.3508772

R> chisq.test(screen8$gender, screen8$activity)
```

```
Pearson's Chi-squared test

data:  screen8$gender and screen8$activity
X-squared = 6.9303, df = 2, p-value = 0.03127
```

### Tests of quantitative variables for two or more groups

We compare math test scores between males and females using the `ttestClust` function. We conclude that mean standardized test scores are equivalent between males and females, a departure from the conclusion reached by the standard t test ($p < .001$, results not shown).

```
R> 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
```

Even though this test does not make use of the $t$ distribution, we have named it as such to parallel the standard $t$-test means (`t.test` in R). Multi-group tests of quantitative parameters in **htestClust** implement jackknife variance estimation, so specification of variance estimation method is not necessary. In addition to the formula implementation used above, we note that `ttestClust` can also accept vectors of data and cluster identifiers for each of the two groups.

An alternative approach to this comparison, particularly if test scores were skewed in any way, would be a rank-based test. `wilcoxtestClust` implements the group-weighted analogue of the Wilcoxon test, which we use as an alternative method for the comparison of math test scores between males and females.

```
R> wilcoxtestClust(math ~ gender, id = sch.id, data = screen8, method = "group")
Group-weighted rank sum test

data:  math by gender, M = 73
z = -1.3799, p-value = 0.1676
alternative hypothesis: true location shift is not equal to 0
```

Our conclusion is the same as with the reweighted test of means. We note that this test requires estimation of the cluster-weighted empirical cumulative distribution (Dutta and Datta, 2016a) as well as jackknife variance estimation, so there is an added measure of computational expense in using `wilcoxtestClust`.

Finally, we compare reading test scores among the three groups defined by extracurricular activity, using `onewaytestClust`. Mean standardized reading test scores are not appreciably different among extracurricular activity groups.

```
R> 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
```

We have not shown the full functionality of the above-demonstrated functions, nor the **htestClust** functions for testing correlation, marginal homogeneity, and variance listed in Table 1. Their syntax and usage is similar and fully documented with examples in the help files.

## 5 Discussion

Standard model-based inference of clustered data can be biased when cluster or group size is informative. Reweighting methods that correct for this bias have been established and a number of authors have applied such weighting to develop direct hypothesis tests of marginal parameters in clustered data. Such tests can be interpreted as clustered analogues to common classical statistical tests, and include methods related to ranks, correlation, proportions, means and variances. While these methods are effective and intuitive, all but a few of these tests have remained inaccessible to many researchers due to an absence of convenient software.

In this paper we introduced **htestClust**, which is the first R package designed as a comprehensive library of inferential methods appropriate for clustered data with ICS/IWCGS. Most functions in **htestClust** perform hypothesis tests for clustered data that have an analogous classical form, and the interface of the package has been designed to reflect this relationship. Function syntax has been purposefully structured to resemble that of functions available in the native R environment that perform the analogous classical tests. Many functions have been designed with a secondary interface that operates through table or formula input, allowing flexibility in data structure. In addition to the hypothesis tests of marginal parameters, **htestClust** also includes functions to visualize potential informativeness and test for ICS. These tools allow analysts to explore the effect and degree of informativeness in their data.

With the exception of the test for ICS, the hypothesis tests performed by **htestClust** are derived through the asymptotic normality of reweighted parameters, and their asymptotic convergence is indexed by the number of clusters. As such, their use should only be considered when the number of clusters is sufficiently large (at least 30). Additionally, these methods retain a cluster-based marginal interpretation, making them appropriate when clusters, rather than intra-cluster observations, are the unit of interest. The marginal nature of these tests provides researchers with an analysis corresponding to a snapshot in time. If analysis of temporal aspects or effects of additional covariates is desired, readers might instead consider reweighted model-based methods such as those by Bible et al. (2016), Neuhaus and McCulloch (2011), and Wang et al. (2011). Future research will also be devoted to developing tests adjusting for informativeness due to quantitative covariates measured at the individual-within-cluster level.

**htestClust** is a tool to facilitate the analysis of clustered data, and we have designed its use to be accessible and intuitive. While the inferential methods performed by this package have been developed to correct for the biasing effects of ICS/IWCGS, they remain applicable when fluctuations of cluster or group size are unrelated to the outcome of interest. As such, this package is an effective resource for researchers addressing marginal analyses in clustered data with any variation in the cluster and/or group sizes.

## Computational details

The results in this paper were obtained using R 4.0.3 with the **MASS** 7.3.51 package.

## Acknowledgments

The authors would like to thank Lucas Koepke and an anonymous reviewer for their thorough review and insightful comments that improved the quality of this manuscript.

## Bibliography

J. Bible, J. D. Beck, and S. Datta. Cluster adjusted regression for displaced subject data (cards): Marginal inference under potentially informative temporal cluster size profiles. *Biometrics*, 72(2):441–451, 2016. URL https://doi.org/10.1111/biom.12456. [p54, 64]

V. J. Carey, T. Lumley, and B. Ripley. *gee: Generalized Estimation Equation Solver*, 2019. URL https://CRAN.R-project.org/package=gee. R package version 4.13-20. [p54]

A. Chaurasia, D. Liu, and P. S. Albert. Pattern-mixture models with incomplete informative cluster size: Application to a repeated pregnancy study. *Journal of the Royal Statistical Society C*, 67(1):255, 2018. URL https://doi.org/10.1111/rssc.12226. [p54]

S. Datta and G. A. Satten. Rank-sum tests for clustered data. *Journal of the American Statistical Association*, 100(471):908–915, 2005. URL https://doi.org/10.1198/016214504000001583. [p54, 55, 56]

S. Datta and G. A. Satten. A signed-rank test for clustered data. *Biometrics*, 64(2):501–507, 2008. URL https://doi.org/10.1111/j.1541-0420.2007.00923.x. [p54, 55, 56]

S. Dutta and S. 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, 2016a. URL https://doi.org/10.1111/biom.12447. [p54, 55, 56, 63]

S. Dutta and S. Datta. **ClusterRankTest***: Rank Tests for Clustered Data*, 2016b. URL https://CRAN.R-project.org/package=ClusterRankTest. R package version 1.0. [p57]

M. Gregg. *Marginal Methods and Software for Clustered Data With Cluster- and Group-Size Informativeness*. PhD thesis, UL (University of Louisville), 2020. [p54, 55, 56]

M. Gregg, S. Datta, and D. Lorenz. Variance estimation in tests of clustered categorical data with informative cluster size. *Statistical Methods in Medical Research*, 29(11):3396–3408, 2020. URL https://doi.org/10.1177/0962280220928572. [p54, 55, 56, 61, 62]

U. Halekoh, S. Højsgaard, J. Yan, et al. The r package **geepack** for generalized estimating equations. *Journal of Statistical Software*, 15(2):1–11, 2006. [p54]

E. B. Hoffman, P. K. Sen, and C. R. Weinberg. Within-cluster resampling. *Biometrika*, 88(4):1121–1134, 2001. URL https://doi.org/10.1093/biomet/88.4.1121. [p55]

Y. Huang and B. Leroux. Informative cluster sizes for subcluster-level covariates and weighted generalized estimating equations. *Biometrics*, 67(3):843–851, 2011. URL https://doi.org/10.1111/j.1541-0420.2010.01542.x. [p54, 56]

A.-M. Iosif and A. R. Sampson. A model for repeated clustered data with informative cluster sizes. *Statistics in Medicine*, 33(5):738–759, 2014. URL https://doi.org/10.1002/sim.5988. [p54]

Y. Jiang. *clusrank: Wilcoxon Rank Sum Test for Clustered Data*, 2018. URL https://CRAN.R-project.org/package=clusrank. R package version 0.6-2. [p57]

D. J. Lorenz, S. Datta, and S. J. Harkema. Marginal association measures for clustered data. *Statistics in Medicine*, 30(27):3181–3191, 2011. URL https://doi.org/10.1002/sim.4368. [p54, 55, 56]

D. J. Lorenz, S. Levy, and S. Datta. Inferring marginal association with paired and unpaired clustered data. *Statistical Methods in Medical Research*, 27(6):1806–1817, 2018. URL https://doi.org/10.1177/0962280216669184. [p54]

A. Mitani, E. Kaye, and K. Nelson. Marginal analysis of multiple outcomes with informative cluster size. *Biometrics*, 2020. URL https://doi.org/10.1111/biom.13241. [p54]

A. A. Mitani, E. K. Kaye, and K. P. Nelson. Marginal analysis of ordinal clustered longitudinal data with informative cluster size. *Biometrics*, 75(3):938–949, 2019. URL https://doi.org/10.1111/biom.13050. [p54]

J. M. Neuhaus and C. E. McCulloch. Estimation of covariate effects in generalized linear mixed models with informative cluster sizes. *Biometrika*, 98(1):147–162, 2011. URL https://doi.org/10.1093/biomet/asq066. [p64]

J. Nevalainen, H. Oja, and S. Datta. Tests for informative cluster size using a novel balanced bootstrap scheme. *Statistics in Medicine*, 36(16):2630–2640, 2017. URL https://doi.org/10.1002/sim.7288. [p55, 57]

M. Pavlou. *Analysis of Clustered Data When the Cluster Size is Informative*. PhD thesis, UCL (University College London), 2012. [p54]

M. Wang, M. Kong, and S. 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. URL https://doi.org/10.1177/0962280209347043. [p64]

J. M. Williamson, S. Datta, and G. A. Satten. Marginal analyses of clustered data when cluster size is informative. *Biometrics*, 59(1):36–42, 2003. URL https://doi.org/10.1111/1541-0420.00005. [p54, 55, 56]

*Mary Gregg*
*Department of Bioinformatics and Biostatistics*
*University of Louisville*
*485 East Gray Street*
*Louisville, KY 40202, USA*
*ORCiD: 0000-0003-2991-6939*
mary.gregg@louisville.edu

*Somnath Datta*
*Department of Biostatistics*
*University of Florida*
*2004 Mowry Rd*
*P.O. Box 117450*
*Gainesville, FL 32611 USA*
*ORCiD: 0000-0003-4381-1842*
somnath.datta@ufl.edu

*Douglas Lorenz*
*Department of Bioinformatics and Biostatistics*
*University of Louisville*
*485 East Gray Street*
*Louisville, KY 40202, USA*
*ORCiD: 0000-0001-8114-0926*
djlore01@louisville.edu