TESTING EQUALITY OF MEANS IN PARTIALLY MATCHED DATA

**Introduction**

It is common in biostatistics to test the equality of two means. In study designs with two independent, normally distributed samples, the two-sample *t*-test is appropriate. In study designs with two *paired* samples, at least two approaches are available. First, the one-sample *t*-test can be used on the unit-level differences between the first measurement and the second measurement. Second, although less routine, the two-sample *t*-test modified for correlated data can be used, assuming the correlation coefficient is known or estimated (Zimmerman 2012).

Some study designs may include a hybrid of independent samples and repeated measures. For example, a subset of participants in a drug trial may be lost to follow-up, precluding the calculation of difference scores, but the remainder of participants may have pre- and post-intervention measurements available. Such designs have been called partially paired (Guo and Yuan 2017), or partially correlated (Samawi and Vogel 2014).

Recently, methods have been proposed to evaluate the equality of means in study designs with partially paired data (Lin and Stivers 1974, Samawi and Vogel 2014, Guo and Yuan 2017). Such methods involve tests based on modified maximum likelihood, multiple imputation, or pooling statistics across the paired and unpaired subsets of data.

A separate type of disconnected data called *unmatched data* arise in paired or repeated measures study designs when unit-level identifiers have been lost or withheld, sometimes in the interest of anonymity. For example, an online survey is administered anonymously to an employee pool, with each employee submitting a response before and after an intervention. The feature of this dataset that would qualify it as unmatched is the absence of any identifier that could link one employee’s response from pre-intervention to the same employee’s response post-intervention. For convenience, we refer to such study designs as *paired but unmatched.* Study designs with paired but unmatched observations have received relatively less attention.

Data from paired but unmatched studies present a challenge for testing the equality of means. The one-sample *t*-test cannot be used, due to the inability to match samples and calculate difference scores. Using the two-sample ­*t*-test is equivalent to assuming samples are uncorrelated, which is often an unrealistic assumption of repeated measures data in biostatistics. An alternative approach involves calculating the minimum possible correlation coefficient given the observed data; this value can be used as the estimated correlation in the two-sample *t*-test modified for correlated data. However, this approach yields a maximally conservative test and is poorly powered.

In cases where even a small number of observations can be matched, an opportunity exists to construct a test for the equality of means that uses the matched pairs to inform an estimate of the correlation between all pairs. We refer to the study design in which observations are paired but only some observations are matched as *paired but partially matched,* or more concisely, *partially matched*. In this work, we show that several methods can powerfully test for the equality of means in partially matched data, while controlling Type I error rates at a nominal level.

**Methods**

*Notation / Definitions*

Let **X** and **Y** denote random variables for which we wish to compare means. Let denote the true, unobserved correlation between **X** and **Y**. We draw *n* paired samples , each of which represents a unit-level observation. Because data are paired but partially matched, we only observe matched pairs for *m* samples, and the remaining *n – m* paired samples are unmatched. For unmatched samples we cannot identify which observation in corresponds to its pair in . Our aim is to evaluate the difference in means, , between **X** and **Y**.

Assuming **X** and **Y** are normally distributed, and the variance of **X** and **Y** are equal, we have the following formula for the two-sample *t*-test modified for correlated data (Zimmerman 2012):

Note that the above statistic is equal to where *t* is the statistic from the two-sample *t*-test for independent variables with equally sized samples. Smaller values of result in a test statistic that is more conservative, while larger values of result in a test statistic that is more anti-conservative. Also note that ignoring correlation in unmatched data is equivalent to setting in the above equation and thus, assuming independence of **X** and **Y**.

In most applied scenarios this test requires an estimate of , which we will denote *r*. This additional requirement may explain the relative unpopularity of the test compared to the paired-samples *t*-test (which we previously referred to as the one-sample *t*-test on differences). However, Zimmerman (2012) demonstrated that compared to the paired-samples *t*-test, the modified *t*-test can improve power in modestly sized datasets (e.g., *n* = 25) while controlling the Type I error rate near the desired level of 0.05. This improvement in power is due to the greater degrees of freedom in the modified test statistic, *t’*, compared to the paired-samples statistic (2*n* – 2 degrees of freedom versus *n* – 1, respectively).

In partially matched data, the challenge in designing an appropriate test for the equality of means lies in finding a reasonable estimator of the correlation, *r*, despite observing only a subset of matched pairs in the sample. We considered several candidate estimators described in the next section.

*Estimators of correlation*

*1. Maximally conservative estimator:*

One can calculate the minimum possible correlation coefficient given the observed data, which would yield a maximally conservative test for the equality of means (i.e., it would result in the fewest number of rejected null hypotheses).

Let *j* denote the index of the samples sorted in ascending order. Then the following formula defines the maximally conservative estimate of the correlation:

Note that calculation of the numerator data in the above equation requires sorting observed samples from **X** in ascending order while sorting observed samples from **Y** in descending order.

Although expected to have considerable bias towards underestimating the true correlation, the above estimator is calculable even when data are unmatched. That is, when there are no matched samples with which to estimate correlation. The remaining estimators of correlation require at least one matched sample to be calculated.

*2. Pearson correlation of the matched samples:*

Perhaps the most intuitive correlation estimate in the presence of partially matched data would simply use the *m* matched samples and ignore the *n - m* matched samples. The following formula gives the Pearson correlation coefficient for matched samples:

The above estimator requires at least two matched samples to be calculable, and when only two matched samples are available, the estimator will return a value of either 1 or -1. However, with three or more matched samples we expect the above estimator to be substantially less biased than the maximally conservative estimator.

*3. Quantile estimator of the matched samples:*

Recognizing that the property of conservatism in estimating might be desirable, yet the degree of conservativism in might be too extreme, we hypothesized that a quantile estimate of rho might yield additional power while maintaining appropriate Type I error control.

The Fisher transformation of a Pearson correlation can be used to obtain an approximately normally distributed quantity and thus, a confidence interval for . Our estimator is constructed from the lower bound of an 80% confidence interval for the one-sided hypothesis test of a correlation between matched samples. Let be the standard error of the Fisher-transformed correlation coefficient. Then the following formula gives the quantile estimator of the correlation of *m* matched samples:

Note in the above that is the critical value chosen based on a desired confidence level. Since our desired confidence level was 80%, the critical value of was 0.842, and our estimator was a 20th quantile estimator of the correlation of the matched samples. We refer to this specific quantile estimator as .

Our rationale behind the quantile estimator was to account for the uncertainty in when computing the two-sample *t*-test modified for correlated data. This uncertainty exists even when data are entirely matched. Zimmerman (2012) noted that in simulation, the modified *t*-test using the sample correlation did not achieve the nominal significance level that the paired-samples *t*-test achieved until sample sizes of 100 or greater. An estimator of correlation that is biased toward underestimating the true correlation will result in a more conservative test. In theory, there exists some for which the Type I error is controlled at the desired level. In practice, we selected the 20th quantile as a semi-conservative estimator that may generalize well to a variety of datasets.

*Simulation study*

Data were simulated from bivariate normal distributions and from bivariate ordinal distributions, obtained by ‘binning’ values of the normal distribution to derive a seven-level ordinal value. The bins were selected to yield a right-skewed ordinal distribution, to mimic the data in our application.

We allowed the true correlation to vary over the following values: -0.9, -0.5, -0.25, 0, 0.25, 0.5, and 0.9. Note that for the ordinal distribution, the ‘true’ correlation was not preserved due to the transformation from continuous to ordinal.

We allowed the sample size in each group to vary between 10 and 200, we allowed the proportion of matched samples to vary from 0 to 1, and we allowed the true mean difference between groups to vary between 0 and 0.5. Group variances were held fixed at 1.

For every combination of the above parameters, we simulated 10,000 datasets. For all estimators of correlation under consideration, we monitored the bias and variance. We then derived the two-sample *t*-test modified for correlated data using the estimates obtained in each dataset. Finally, we computed the rate rejecting null hypotheses to examine Type I error control and power.

*Application to dataset*

[To-do]

**Results**

Not all estimators yielded valid correlation, particularly when the number of matched samples was small (i.e., less than four) or when the generating distribution was ordinal.

For normally distributed data with ten samples and two to five matched samples, the EM algorithm estimator and the Pearson estimator yielded similar bias; both tended to overestimate correlation when it was negative, and both tended to underestimate correlation when it was positive. However, the EM algorithm estimator was consistently less variable than the Pearson estimator (see below).

Chart, line chart

Description automatically generated

Under the same conditions (n=10, m=2-5), the EM algorithm yielded a *t*-test that was more conservative and controlled Type I error closer to the nominal level compared to the Pearson estimator. When the number of matched samples was two and the true correlation was positive, the EM algorithm yielded a slightly anti-conservative test (Type I error between 10 and 20%), whereas the Pearson estimator yielded an extremely anti-conservative test (Type I error rates over 50%). In fact, under this scenario, the assumption of independence yielded a test with better Type I error control compared to the Pearson estimator, for all values of correlation.

Chart, line chart

Description automatically generated

In generally small samples, the Pearson estimator yielded a test with superior power compared to the EM algorithm estimator. However, when the true correlation was non-negative, the EM algorithm estimator yielded a test that was generally better powered than the assumption of independence (see below).

Chart, line chart

Description automatically generated

Bayes estimator achieves lower MSE than EM (n=10, m=2).

Chart, line chart

Description automatically generated

Bayes estimator yields more conservative test, compared to EM.

Chart, line chart

Description automatically generated

Bayes estimator yields underpowered test, compared to EM. But both are better than available approaches (*t*-test and maximally conservative test).

**Chart

Description automatically generated with medium confidence**

**Discussion**

One of the practical questions for applications in partially matched datasets is whether a test for the equality of means can improve upon the unrealistic assumption that samples are independent (e.g., the two-sample *t*-test). We have shown that in small samples with only two matched pairs, the EM algorithm approach yields a test that is better powered, and consistent in controlling Type I error (albeit not at the nominal level), compared to the two-sample *t*-test when correlation is non-negative.

We have also shown that in such small sample cases, the test based on the EM algorithm estimator is consistently better than one based on the Pearson estimator. This is particularly relevant, for one’s first choice of an estimator may be the Pearson correlation, even in extremely small samples. But the EM algorithm estimator is substantially less variable than Pearson in partially matched data, since the EM approach considers information beyond the matched samples alone.

In partially matched data we have shown that better methods exist for testing the equality of means, compared to assuming independence or using the Pearson correlation of the matched samples.