`PartialPairs` Package Vignette

Sam van der Poel | Stony Brook University

July 15, 2021

The `PartialPairs` package provides five core functions for comparing paired samples `x` and `y` when both samples have missing data. The methods implemented here are motivated by the overview given in (Kuan & Huang, 2013). The following are the five methods:

- Liptak's weighted Z-test (Lipták, 1958)
- Kim et al.'s modified t-statistic (Kim et al., 2005)
- Looney and Jones's corrected Z-test (Looney & Jones, 2003)
- Lin and Stivers's MLE based test under heteroscedasticity (Lin & Stivers, 1974)
- Ekbohm's MLE-based test under homoscedasticity (Ekbohm, 1976)

In the documentation below, we will use the following common notation:

- n_1 denotes the number of complete paired entries in `x` and `y`. In R terms, n_1 is the quantity `sum(!is.na(x) & !is.na(y))`.
- n_2 denotes the number of entries where only `x`-data is present. In R terms, n_2 is the quantity `sum(!is.na(x) & is.na(y))`.
- n_3 denotes the number of entries where only `y`-data is present. In R terms, n_3 is the quantity `sum(!is.na(y) & is.na(x))`.

Each of the functions in `PartialPairs` checks whether appropriate sample size conditions are met. The sample sizes should satisfy $n_1 \geq 4$ and $n_2 + n_3 \geq 5$ in order for the desired functions to execute properly. If either of these inequalities is not satisfied, then depending on how much data is missing or present, the functions may instead exit, perform a paired sample t-test (when there is sufficient paired data), or perform a two-sample t-test (when there is sufficient unpaired data).

Where necessary, the functions in `PartialPairs` check whether the variance of input data is close to zero. The user is notified of this through an error message.

Sample data is provided and automatically loaded when `library(PartialPairs)` is called; it is stored in a data frame called `sampledata` with columns `x` and `y`. (If the data is not loaded initially, `data("sampledata")` will load it.) The data consist of 1500 observations (including 400 `NA`s in each of `x` and `y`) generated from two similar but unequal normal distributions. the following line runs Liptak's weighted Z-test on `sampledata`:

```
liptak.ztest(sampledata['x'], sampledata['y'],
             alternative = 'less')
```

When copying code with quotations ' from this vignette, be sure to replace the quotations with new quotations in your code as they are different Unicode characters.

Liptak's weighted Z-test

Usage: `liptak.ztest(...)`

Description: Liptak's weighted Z-test computes the p-values of a paired sample t-test on the n_1 paired entries and of a two-sample t-test on the $n_2 + n_3$ unpaired entries. The two p-values are then weighted and combined as detailed in (Kuan & Huang, 2013).

Input:

`x`	a non-empty numeric vector containing some `NA`s
`y`	a non-empty numeric vector containing some `NA`s (must match length of `x`)
`alternative`	specification of the alternative hypothesis. Takes values: `"two.sided"`, `"greater"`, or `"less"`

Output:

p-value associated with the test of the null hypothesis that the means are equal

Example:
```
# a case where true means are not equal
x = rnorm(400, 0, 1)
x[sample(1:400, size=75, replace=FALSE)] = NA
y = rnorm(400, 0.4, 4)
y[sample(1:400, size=75, replace=FALSE)] = NA
liptak.ztest(x, y, alternative = 'two.sided')
```

Kim et al.'s modified t-statistic

Usage: `modified.tstat(...)`

Description: Kim et al.'s modified t-statistic, as given by Equation (3) in (Kuan & Huang, 2013), follows an approximately standard Gaussian distribution under the null hypothesis.

Input:

`x`	a non-empty numeric vector containing some `NA`s
`y`	a non-empty numeric vector containing some `NA`s (must match length of `x`)
`alternative`	specification of the alternative hypothesis. Takes values: `"two.sided"`, `"greater"`, or `"less"`

Output:

p-value associated with the test of the null hypothesis that the means are equal

Example:
```
# a case where true means are not equal
x = rnorm(400, 0, 1)
x[sample(1:400, size=75, replace=FALSE)] = NA
y = rnorm(400, 0.4, 4)
```

```
y[sample(1:400, size=75, replace=FALSE)] = NA
modified.tstat(x, y, alternative = 'two.sided')
```

Looney and Jones's corrected Z-test

Usage: `corrected.ztest(...)`

Description: Looney and Jones's corrected Z-test is "corrected" in the sense that it adjusts the standard error of the difference of the two samples by accounting for the correlation between the n_1 paired observations. Under the null hypothesis, the resulting test statistic Z_{corr} has an asymptotic $N(0, 1)$ distribution.

Input:

`x`	a non-empty numeric vector containing some `NA`s
`y`	a non-empty numeric vector containing some `NA`s (must match length of `x`)
`alternative`	specification of the alternative hypothesis. Takes values: `"two.sided"`, `"greater"`, or `"less"`

Output:

p-value associated with the test of the null hypothesis that the means are equal

Example:
```
# a case where true means are not equal
x = rnorm(400, 0, 1)
x[sample(1:400, size=75, replace=FALSE)] = NA
y = rnorm(400, 0.4, 4)
y[sample(1:400, size=75, replace=FALSE)] = NA
corrected.ztest(x, y, alternative = 'two.sided')
```

Lin and Stivers's MLE-based test under heteroscedasticity

Usage: `modified.tstat(...)`

Description: Lin and Stivers's test makes use of a modified maximum likelihood estimator and assumption of heteroscedasticity. Under the null hypothesis, the resulting test statistic Z_{LS}, follows an approximate t distribution with n_1 degrees of freedom.

Input:

`x`	a non-empty numeric vector containing some `NA`s
`y`	a non-empty numeric vector containing some `NA`s (must match length of `x`)
`alternative`	specification of the alternative hypothesis. Takes values: `"two.sided"`, `"greater"`, or `"less"`

Output:

p-value associated with the test of the null hypothesis that the means are equal

Example:
```
# a case where true means are not equal
x = rnorm(400, 0, 1)
x[sample(1:400, size=75, replace=FALSE)] = NA
y = rnorm(400, 0.4, 4)
y[sample(1:400, size=75, replace=FALSE)] = NA
lin.mle.test(x, y, alternative = 'two.sided')
```

Ekbohm's MLE-based test under homoscedasticity

Usage: `modified.tstat(...)`

Description: Ekbohm's test makes use of a modified maximum likelihood estimator and assumption of homoscedasticity. Under the null hypothesis, the resulting test statistic Z_E, follows an approximate t distribution with n_1 degrees of freedom.

Input:

`x`	a non-empty numeric vector containing some `NA`s
`y`	a non-empty numeric vector containing some `NA`s (must match length of `x`)
`alternative`	specification of the alternative hypothesis. Takes values: `"two.sided"`, `"greater"`, or `"less"`

Output:

p-value associated with the test of the null hypothesis that the means are equal

Example:
```
# a case where true means are not equal
x = rnorm(400, 0, 1)
x[sample(1:400, size=75, replace=FALSE)] = NA
y = rnorm(400, 0.4, 4)
y[sample(1:400, size=75, replace=FALSE)] = NA
ekbohm.mle.test(x, y, alternative = 'two.sided')
```

References

Ekbohm, G. (1976). On comparing means in the paired case with incomplete data on both responses. *Biometrika*, *63*(2), 299–304.

Kim, B. S., Kim, I., Lee, S., Kim, S., Rha, S. Y., & Chung, H. C. (2005). Statistical methods of translating microarray data into clinically relevant diagnostic information in colorectal cancer. *Bioinformatics*, *21*(4), 517–528.

Kuan, P. F., & Huang, B. (2013). A simple and robust method for partially matched samples using the p-values pooling approach. *Statistics in medicine*, *32*(19), 3247–3259.

Lin, P.-E., & Stivers, L. E. (1974). On difference of means with incomplete data. *Biometrika*, *61*(2), 325–334.

Lipták, T. (1958). On the combination of independent tests. *Magyar Tud Akad Mat Kutato Int Kozl*, *3*, 171–197.

Looney, S. W., & Jones, P. W. (2003). A method for comparing two normal means using combined samples of correlated and uncorrelated data. *Statistics in medicine*, *22*(9), 1601–1610.