Balance Test

Satoshi Koiso

07/11/2022

Balance Test

The aim of randomization is to produce a balance in characteristics in the treatment/control groups.

Example

The `balsakhi` dataset is provided with the package `pwrcalc`. We'll use this dataset to estimate the control mean:

```
library(pwrcalc)
data(balsakhi)

# Let us view the `balsakhi `data that we have just loaded:
View(balsakhi)
```

Data Exploration

Let's explore the balance in the `balsakhi` study by presenting the means and standard deviations of the baseline test scores (total, math, and verbal) and age, gender, division, research group, as well as whether they cheated, they were helped, and they took it later. This can be done very easily with the `compareGroups()` function.

```
#Generate summaries of the variables by treatment group and save results as baselines
baselines <- compareGroups(bal ~ age + gender + div + researchgroup + sessiond + pre_tot + pre_math + p
#Use the createTable function to display the results saved in baselines
baseline.table <- createTable(baselines, show.ratio = FALSE, show.p.overall=TRUE)
#Display the created summary table
baseline.table
```

```
##
## --------Summary descriptives table by 'Dummy for balsakhi treatment'---------
##
## ____________________________________________________________________________
##                                          0            1       p.overall
##                                        N=5208       N=4990
##
## Age                                      .            .           .
## String variable for gender:                                      0.056
##     F                                 2716 (52.2%) 2507 (50.2%)
##     M                                 2492 (47.8%) 2483 (49.8%)
## Division code:                                                   0.043
##     A                                 3249 (62.4%) 3224 (64.6%)
```

```
##     B                                 1730 (33.2%) 1542 (30.9%)
##     C                                 229 (4.40%)  224 (4.49%)
## Supervising group for evaluation      5.90 (2.99)  5.92 (2.93)    0.720
## School session (morning or afternoon):                            0.047
##                                        72 (1.38%)   42 (0.84%)
##     Afternoon                         2657 (51.0%) 2605 (52.2%)
##     Both(M/A)                          27 (0.52%)   30 (0.60%)
##     Morning                           2452 (47.1%) 2313 (46.4%)
## Absolute total score on pre-test      32.1 (22.0)  31.7 (22.0)    0.351
## Absolute math score on pre-test       17.8 (12.4)  17.4 (12.5)    0.105
## Absolute verbal score on pre-test     14.3 (10.8)  14.3 (10.8)    0.971
## Cheated on pre-test:                                             <0.001
##                                       206 (3.96%)   44 (0.88%)
##     No                                4657 (89.4%) 4481 (89.8%)
##     Yes                               345 (6.62%)  465 (9.32%)
## Helped on pre-test:                                              <0.001
##                                       207 (3.97%)   44 (0.88%)
##     No                                4974 (95.5%) 4905 (98.3%)
##     Yes                                27 (0.52%)   41 (0.82%)
## Took the pre-test at a later date:                               <0.001
##                                       210 (4.03%)   44 (0.88%)
##     No                                4802 (92.2%) 4824 (96.7%)
##     Yes                               196 (3.76%)  122 (2.44%)
## Normalized total score on pre-test    0.00 (1.00)  0.01 (1.02)    0.688
## Normalized math score on pre-test     0.00 (1.00)  -0.01 (1.02)   0.733
## Normalized verbal score on pre-test   0.00 (1.00)  0.02 (1.03)    0.223
##
```

```
# to show ratio in table,
createTable(baselines, show.ratio=TRUE)
```

```
##
## --------Summary descriptives table by 'Dummy for balsakhi treatment'---------
##
## ---------------------------------------------------------------------------------------------------
##                                            0            1              OR              p.overall
##                                         N=5208       N=4990
##
## Age                                         .            .              .           .          .
## String variable for gender:                                                                 0.056
##     F                                 2716 (52.2%) 2507 (50.2%)       Ref.        Ref.
##     M                                 2492 (47.8%) 2483 (49.8%) 1.08 [1.00;1.17] 0.054
## Division code:                                                                              0.043
##     A                                 3249 (62.4%) 3224 (64.6%)       Ref.        Ref.
##     B                                 1730 (33.2%) 1542 (30.9%) 0.90 [0.83;0.98] 0.012
##     C                                 229 (4.40%)  224 (4.49%)  0.99 [0.81;1.19] 0.883
## Supervising group for evaluation      5.90 (2.99)  5.92 (2.93)  1.00 [0.99;1.02] 0.720     0.720
## School session (morning or afternoon):                                                      0.047
##                                        72 (1.38%)   42 (0.84%)        Ref.        Ref.
##     Afternoon                         2657 (51.0%) 2605 (52.2%) 1.68 [1.15;2.48] 0.007
##     Both(M/A)                          27 (0.52%)   30 (0.60%)  1.90 [0.99;3.64] 0.052
##     Morning                           2452 (47.1%) 2313 (46.4%) 1.61 [1.10;2.39] 0.013
## Absolute total score on pre-test      32.1 (22.0)  31.7 (22.0)  1.00 [1.00;1.00] 0.351     0.351
## Absolute math score on pre-test       17.8 (12.4)  17.4 (12.5)  1.00 [0.99;1.00] 0.105     0.105
## Absolute verbal score on pre-test     14.3 (10.8)  14.3 (10.8)  1.00 [1.00;1.00] 0.971     0.971
```

```
## Cheated on pre-test:                                                                                    <0.001
##                                        206 (3.96%)   44 (0.88%)        Ref.        Ref.
##     No                                 4657 (89.4%) 4481 (89.8%) 4.49 [3.27;6.31] 0.000
##     Yes                                345 (6.62%)  465 (9.32%)  6.28 [4.45;9.05] 0.000
## Helped on pre-test:                                                                                     <0.001
##                                        207 (3.97%)   44 (0.88%)        Ref.        Ref.
##     No                                 4974 (95.5%) 4905 (98.3%) 4.63 [3.37;6.50] 0.000
##     Yes                                 27 (0.52%)   41 (0.82%)  7.07 [3.96;12.9] <0.001
## Took the pre-test at a later date:                                                                      <0.001
##                                        210 (4.03%)   44 (0.88%)        Ref.        Ref.
##     No                                 4802 (92.2%) 4824 (96.7%) 4.78 [3.48;6.72] 0.000
##     Yes                                196 (3.76%)  122 (2.44%)  2.96 [2.00;4.44] <0.001
## Normalized total score on pre-test     0.00 (1.00)  0.01 (1.02)  1.01 [0.97;1.05] 0.688    0.688
## Normalized math score on pre-test      0.00 (1.00)  -0.01 (1.02) 0.99 [0.96;1.03] 0.733    0.733
## Normalized verbal score on pre-test    0.00 (1.00)  0.02 (1.03)  1.02 [0.99;1.06] 0.223    0.223
##
```

```
# Odds Ratio (OR) is a measure of association between exposure and an outcome
```

It looks like some variables have missing data. Let's do data cleaning.

```
missing <- miss_var_summary(balsakhi)
print(missing, n=76)
```

```
## # A tibble: 76 x 3
##    variable          n_miss pct_miss
##    <chr>              <int>    <dbl>
##  1 age                10198   100
##  2 pretakehome        10198   100
##  3 midtakehome        10198   100
##  4 posttakehome       10198   100
##  5 post_tot            1772    17.4
##  6 postpapersubtotal   1772    17.4
##  7 post_verb           1772    17.4
##  8 post_math           1772    17.4
##  9 post_mathnorm       1772    17.4
## 10 post_verbnorm       1772    17.4
## 11 post_totnorm        1772    17.4
## 12 postmath1std        1772    17.4
## 13 postmath2std        1772    17.4
## 14 postmath3std        1772    17.4
## 15 postverb1std        1772    17.4
## 16 postverb2std        1772    17.4
## 17 postverb3std        1772    17.4
## 18 mid_verb            1198    11.7
## 19 mid_verbnorm        1198    11.7
## 20 midverb2std         1198    11.7
## 21 mid_tot             1162    11.4
## 22 midpapersubtotal    1162    11.4
## 23 mid_math            1162    11.4
## 24 mid_mathnorm        1162    11.4
## 25 mid_totnorm         1162    11.4
## 26 midmath1std         1162    11.4
## 27 midmath2std         1162    11.4
## 28 midmath3std         1162    11.4
```

```
## 29 midverb1std         1162    11.4
## 30 midverb3std         1162    11.4
## 31 researchgroup        114     1.12
## 32 studentid              0     0
## 33 std                    0     0
## 34 schoolid               0     0
## 35 div                    0     0
## 36 gender                 0     0
## 37 sessiond               0     0
## 38 std3tc                 0     0
## 39 std4tc                 0     0
## 40 caltc                  0     0
## 41 precheated             0     0
## 42 prehelped              0     0
## 43 prelater               0     0
## 44 pre_tot                0     0
## 45 prepapersubtotal       0     0
## 46 standardtemp           0     0
## 47 midcheated             0     0
## 48 midhelped              0     0
## 49 midlater               0     0
## 50 postcheated            0     0
## 51 posthelped             0     0
## 52 postlater              0     0
## 53 attritprepost          0     0
## 54 attritpremid           0     0
## 55 attritmidpost          0     0
## 56 pre_verb               0     0
## 57 pre_math               0     0
## 58 bal                    0     0
## 59 male                   0     0
## 60 rank                   0     0
## 61 bot20                  0     0
## 62 third                  0     0
## 63 divid                  0     0
## 64 numstud                0     0
## 65 thirdinschool          0     0
## 66 bigschool              0     0
## 67 numthird               0     0
## 68 pre_mathnorm           0     0
## 69 pre_verbnorm           0     0
## 70 pre_totnorm            0     0
## 71 premath1std            0     0
## 72 premath2std            0     0
## 73 premath3std            0     0
## 74 preverb1std            0     0
## 75 preverb2std            0     0
## 76 preverb3std            0     0
```

```r
# should not use age and takehome
# found empty cells ("") are not treated as NA

# Convert "" to NA
balsakhi[balsakhi == ""] <- NA
```

```r
# chack NA
missing <- miss_var_summary(balsakhi)
print(missing, n=76)
```

```
## # A tibble: 76 x 3
##    variable          n_miss pct_miss
##    <chr>              <int>    <dbl>
##  1 age                10198   100
##  2 pretakehome        10198   100
##  3 midtakehome        10198   100
##  4 posttakehome       10198   100
##  5 standardtemp       10172    99.7
##  6 posthelped          2021    19.8
##  7 postlater           2021    19.8
##  8 postcheated         2002    19.6
##  9 post_tot            1772    17.4
## 10 postpapersubtotal   1772    17.4
## 11 post_verb           1772    17.4
## 12 post_math           1772    17.4
## 13 post_mathnorm       1772    17.4
## 14 post_verbnorm       1772    17.4
## 15 post_totnorm        1772    17.4
## 16 postmath1std        1772    17.4
## 17 postmath2std        1772    17.4
## 18 postmath3std        1772    17.4
## 19 postverb1std        1772    17.4
## 20 postverb2std        1772    17.4
## 21 postverb3std        1772    17.4
## 22 midcheated          1500    14.7
## 23 midhelped           1449    14.2
## 24 midlater            1449    14.2
## 25 mid_verb            1198    11.7
## 26 mid_verbnorm        1198    11.7
## 27 midverb2std         1198    11.7
## 28 mid_tot             1162    11.4
## 29 midpapersubtotal    1162    11.4
## 30 mid_math            1162    11.4
## 31 mid_mathnorm        1162    11.4
## 32 mid_totnorm         1162    11.4
## 33 midmath1std         1162    11.4
## 34 midmath2std         1162    11.4
## 35 midmath3std         1162    11.4
## 36 midverb1std         1162    11.4
## 37 midverb3std         1162    11.4
## 38 prelater             254     2.49
## 39 prehelped            251     2.46
## 40 precheated           250     2.45
## 41 researchgroup        114     1.12
## 42 sessiond             114     1.12
## 43 caltc                114     1.12
## 44 studentid              0     0
## 45 std                    0     0
## 46 schoolid               0     0
```

```
## 47 div                    0     0
## 48 gender                 0     0
## 49 std3tc                 0     0
## 50 std4tc                 0     0
## 51 pre_tot                0     0
## 52 prepapersubtotal       0     0
## 53 attritprepost          0     0
## 54 attritpremid           0     0
## 55 attritmidpost          0     0
## 56 pre_verb               0     0
## 57 pre_math               0     0
## 58 bal                    0     0
## 59 male                   0     0
## 60 rank                   0     0
## 61 bot20                  0     0
## 62 third                  0     0
## 63 divid                  0     0
## 64 numstud                0     0
## 65 thirdinschool          0     0
## 66 bigschool              0     0
## 67 numthird               0     0
## 68 pre_mathnorm           0     0
## 69 pre_verbnorm           0     0
## 70 pre_totnorm            0     0
## 71 premath1std            0     0
## 72 premath2std            0     0
## 73 premath3std            0     0
## 74 preverb1std            0     0
## 75 preverb2std            0     0
## 76 preverb3std            0     0
```

```
# select columns
base_var <- c("bal","gender","div","researchgroup","sessiond","pre_tot","pre_math", "pre_verb","precheat
balsakhi_base <- balsakhi[,base_var]

# drop na
balsakhi_base_cleaned <- drop_na(balsakhi_base)
missing <- miss_var_summary(balsakhi_base_cleaned)
print(missing, n=76)
```

```
## # A tibble: 14 x 3
##    variable      n_miss pct_miss
##    <chr>          <int>    <dbl>
##  1 bal                0        0
##  2 gender             0        0
##  3 div                0        0
##  4 researchgroup      0        0
##  5 sessiond           0        0
##  6 pre_tot            0        0
##  7 pre_math           0        0
##  8 pre_verb           0        0
##  9 precheated         0        0
## 10 prehelped          0        0
## 11 prelater           0        0
## 12 pre_totnorm        0        0
```

```
## 13 pre_mathnorm       0        0
## 14 pre_verbnorm       0        0
```

Finally, create the balance table

```
#Generate summaries of the variables by treatment group and save results as baselines
baselines <- compareGroups(bal ~ gender + div + researchgroup + sessiond + pre_tot + pre_math + pre_verb
#Use the createTable function to display the results saved in baselines
baseline.table <- createTable(baselines, show.ratio = FALSE, show.p.overall=TRUE)
#Display the created summary table
baseline.table
```

```
##
## --------Summary descriptives table by 'Dummy for balsakhi treatment'---------
##
## ____________________________________________________________________________
##                                        0            1       p.overall
##                                     N=4926       N=4904
##
## String variable for gender:                                      0.003
##     F                               2605 (52.9%) 2445 (49.9%)
##     M                               2321 (47.1%) 2459 (50.1%)
## Division code:                                                   0.485
##     A                               3096 (62.9%) 3139 (64.0%)
##     B                               1601 (32.5%) 1541 (31.4%)
##     C                               229 (4.65%)  224 (4.57%)
## Supervising group for evaluation    5.86 (3.03)  5.95 (2.93)    0.127
## School session (morning or afternoon):                           0.594
##     Afternoon                       2530 (51.4%) 2563 (52.3%)
##     Both(M/A)                        27 (0.55%)   30 (0.61%)
##     Morning                         2369 (48.1%) 2311 (47.1%)
## Absolute total score on pre-test    32.1 (22.1)  31.6 (22.0)    0.242
## Absolute math score on pre-test     17.8 (12.5)  17.4 (12.4)    0.100
## Absolute verbal score on pre-test   14.3 (10.8)  14.2 (10.7)    0.624
## Cheated on pre-test:                                            <0.001
##     No                              4595 (93.3%) 4442 (90.6%)
##     Yes                             331 (6.72%)  462 (9.42%)
## Helped on pre-test:                                              0.061
##     No                              4901 (99.5%) 4863 (99.2%)
##     Yes                              25 (0.51%)   41 (0.84%)
## Took the pre-test at a later date:                              <0.001
##     No                              4732 (96.1%) 4782 (97.5%)
##     Yes                             194 (3.94%)  122 (2.49%)
## Normalized total score on pre-test  0.00 (1.00)  0.00 (1.02)    0.855
## Normalized math score on pre-test   0.00 (1.00)  -0.01 (1.02)   0.473
## Normalized verbal score on pre-test 0.00 (1.00)  0.01 (1.02)    0.635
##
```

```
# to show ratio in table,
createTable(baselines, show.ratio=TRUE)
```

```
##
## --------Summary descriptives table by 'Dummy for balsakhi treatment'---------
##
## ____________________________________________________________________________________________
##                                        0            1              OR        p.ratio p.overall
```

```
##                                        N=4926       N=4904
##
## String variable for gender:                                                          0.003
##     F                                2605 (52.9%) 2445 (49.9%)       Ref.        Ref.
##     M                                2321 (47.1%) 2459 (50.1%) 1.13 [1.04;1.22]  0.003
## Division code:                                                                       0.485
##     A                                3096 (62.9%) 3139 (64.0%)       Ref.        Ref.
##     B                                1601 (32.5%) 1541 (31.4%) 0.95 [0.87;1.03]  0.235
##     C                                 229 (4.65%)  224 (4.57%)  0.96 [0.80;1.17]  0.713
## Supervising group for evaluation      5.86 (3.03)  5.95 (2.93)  1.01 [1.00;1.02]  0.127    0.127
## School session (morning or afternoon):                                               0.594
##     Afternoon                        2530 (51.4%) 2563 (52.3%)       Ref.        Ref.
##     Both(M/A)                         27 (0.55%)   30 (0.61%)  1.10 [0.65;1.86]  0.732
##     Morning                          2369 (48.1%) 2311 (47.1%) 0.96 [0.89;1.04]  0.351
## Absolute total score on pre-test      32.1 (22.1)  31.6 (22.0)  1.00 [1.00;1.00]  0.242    0.242
## Absolute math score on pre-test       17.8 (12.5)  17.4 (12.4)  1.00 [0.99;1.00]  0.100    0.100
## Absolute verbal score on pre-test     14.3 (10.8)  14.2 (10.7)  1.00 [1.00;1.00]  0.624    0.624
## Cheated on pre-test:                                                                 <0.001
##     No                               4595 (93.3%) 4442 (90.6%)       Ref.        Ref.
##     Yes                               331 (6.72%)  462 (9.42%)  1.44 [1.25;1.67] <0.001
## Helped on pre-test:                                                                  0.061
##     No                               4901 (99.5%) 4863 (99.2%)       Ref.        Ref.
##     Yes                               25 (0.51%)   41 (0.84%)  1.65 [1.01;2.76]  0.047
## Took the pre-test at a later date:                                                   <0.001
##     No                               4732 (96.1%) 4782 (97.5%)       Ref.        Ref.
##     Yes                               194 (3.94%)  122 (2.49%)  0.62 [0.49;0.78] <0.001
## Normalized total score on pre-test    0.00 (1.00)  0.00 (1.02)  1.00 [0.96;1.04]  0.855    0.855
## Normalized math score on pre-test     0.00 (1.00) -0.01 (1.02)  0.99 [0.95;1.03]  0.473    0.473
## Normalized verbal score on pre-test   0.00 (1.00)  0.01 (1.02)  1.01 [0.97;1.05]  0.635    0.635
##
```

```
# Odds Ratio (OR) is a measure of association between exposure and an outcome
```

Some variables are not balanced at baseline.