Table 1 and the tableone Package

David M. Vock

PubH 7485/8485

Key Idea

- Emphasis on summarizing data (both in graphical and tabular formats)
- Frequently instructors do not do a sufficient job explaining relevant summary measures for a particular analysis
- Want some tabular and graphical summaries of potential confounders

Observational Studies

- Observational study: Individuals are not assigned to treatment intervention by an experimental design
- Often unethical or impractical to do randomized trial
- In general, A is not independent of $\{Y^1, Y^0\}$
- Thus, the distribution of $Y|A = 1$ or $Y^1|A = 1$ does not equal the distribution of Y^1 and similarly the distribution of $Y|A = 0$ does not equal the distribution of Y^0
- Why? Confounding
- Heuristically, those who receive treatment may be inherently different than those who do not. Consequently, the associational parameters may reflect such inherent differences as well as any effect of treatment
- In other words, there are common causes of both Y and A

Potential Confounders

- IF the treatment was randomly assigned, then the disribution of any baseline (i.e., pre-randomization) covariate should be the same in each level of the treatment
- Covariates which are related to (i.e., causes of) treatment will be differentially distributed between different levels of the treatment
- Want ways of summarizing the distribution of covariates between treatment levels
- Such a table is often referred to as Table 1

Example of Table 1

(a) Recipient characteristics							
	Deceased donor recipients						
	Maintenance prednisone			Rapid discontinuation of prednisone			
	Median or Frequency	(25th, 75th centile) or percent	Percent Missing	Median or Frequency	(25th, 75th centile) or percent	Percent Missing	*P*-value
Age (Years)	52.7	(41.2, 61.8)	0	54.8	(43.4, 63.2)	0	<0.001
Race							
Black or African American	26038	32.9	0	7047	28.8	0	<0.001
Hispanic/Latino	12114	15.3		3527	14.4		
Asian	4882	6.2		1568	6.4		
White	34645	43.8		11819	48.3		
Other or Multiracial	1497	1.9		491	2.0		
Gender							
Female	32108	40.6	0	9156	37.4	0	<0.001
Male	47068	59.4		15296	62.6		
Transplant number							

TableOne

- TableOne is an R package created by Kazuki Yoshida that eases the construction of "Table 1", i.e., patient baseline characteristics table commonly found in biomedical research papers.
- This package is also very helpful in providing an overview of your data since it can summarize both continuous and categorical variables mixed within one table.

```
CreateTableOne(data=iris)
```

```
##
##                             Overall
##   n                          150
##   Sepal.Length (mean (SD)) 5.84 (0.83)
##   Sepal.Width (mean (SD))  3.06 (0.44)
##   Petal.Length (mean (SD)) 3.76 (1.77)
##   Petal.Width (mean (SD))  1.20 (0.76)
##   Species (%)
##      setosa                  50 (33.3)
##      versicolor              50 (33.3)
##      virginica               50 (33.3)
```

}

Some relevant information sources

https://cran.r-project.org/web/packages/tableone/index.html

https://cran.r-project.org/web/packages/tableone/vignettes/introduction.html

https://cran.r-project.org/web/packages/tableone/vignettes/smd.html

http://rstudio-pubs-static.s3.amazonaws.com/13321_da314633db924dc78986a850813a50d5.html

A New Categorical Variable

Let's add in a new column for color (0 for white, 1 for red) and make the table again:

```
new.iris <- cbind(iris, Petal.Color=rbinom(n=150, size=1, prob=0.5))
CreateTableOne(data=new.iris)
```

```
##
##                             Overall
##   n                          150
##   Sepal.Length (mean (SD)) 5.84 (0.83)
##   Sepal.Width (mean (SD))  3.06 (0.44)
##   Petal.Length (mean (SD)) 3.76 (1.77)
##   Petal.Width (mean (SD))  1.20 (0.76)
##   Species (%)
##      setosa                  50 (33.3)
##      versicolor              50 (33.3)
##      virginica               50 (33.3)
##   Petal.Color (mean (SD))  0.47 (0.50)
```

We can see that in this case, the package is unable to determine that Petal.Color is a categorical variable. We will have to set this manually.

Continuous vs Categorical Variables

We can specify which variables we would like in the table, and which variables should be considered categorical:

```
all.vars <- c("Petal.Length", "Petal.Width", "Species", "Petal.Color")
cat.vars <- c("Species", "Petal.Color")
CreateTableOne(vars=all.vars, data=new.iris, factorVars=cat.vars)
```

```
##
##                             Overall
##   n                          150
##   Petal.Length (mean (SD)) 3.76 (1.77)
##   Petal.Width (mean (SD))  1.20 (0.76)
##   Species (%)
##      setosa                   50 (33.3)
##      versicolor               50 (33.3)
##      virginica                50 (33.3)
##   Petal.Color = 1 (%)         70 (46.7)
```

Notice that Petal.Color only shows one level here.

Print Option 1: Show all levels

The default when a categorical variable has 2 levels is to only show one of the levels. To show all the levels for the categorical variables:

```
t1 <- CreateTableOne(vars=all.vars, data=new.iris, factorVars=cat.vars)
print(t1, showAllLevels = TRUE)
```

```
##
##                              level      Overall
##    n                                     150
##    Petal.Length (mean (SD))            3.76 (1.77)
##    Petal.Width (mean (SD))             1.20 (0.76)
##    Species (%)               setosa      50 (33.3)
##                              versicolor  50 (33.3)
##                              virginica   50 (33.3)
##    Petal.Color (%)           0           80 (53.3)
##                              1           70 (46.7)
```

Resetting Levels for Categorical Variables

Instead of showing 0/1 for Petal.Color, we can also reset the level for this variable so it displays as "White"/"Red":

```
new.iris$Petal.Color <- as.factor(new.iris$Petal.Color)
levels(new.iris$Petal.Color) <- c("White", "Red")
t1 <- CreateTableOne(vars=all.vars, data=new.iris)
print(t1, showAllLevels = TRUE)
```

```
##
##                             level      Overall
##   n                                     150
##   Petal.Length (mean (SD))            3.76 (1.77)
##   Petal.Width (mean (SD))             1.20 (0.76)
##   Species (%)               setosa      50 (33.3)
##                             versicolor  50 (33.3)
##                             virginica   50 (33.3)
##   Petal.Color (%)           White       80 (53.3)
##                             Red         70 (46.7)
```

Notice that if the levels are changed to character values, R will automatically change this variable into a "factor", you won't need to manually set it as a categorical variable.

Print Option 2: Interquartile Range

For nonnormal (skewed) continuous variables, the interquartile range (IQR) can be displayed instead of the mean and standard deviation:

```
nn.vars <- c("Sepal.Length", "Sepal.Width")
t2 <- CreateTableOne(data=iris)
print(t2, nonnormal = nn.vars)
```

```
##
##                              Overall
##   n                           150
##   Sepal.Length (median [IQR]) 5.80 [5.10, 6.40]
##   Sepal.Width (median [IQR])  3.00 [2.80, 3.30]
##   Petal.Length (mean (SD))    3.76 (1.77)
##   Petal.Width (mean (SD))     1.20 (0.76)
##   Species (%)
##      setosa                     50 (33.3)
##      versicolor                 50 (33.3)
##      virginica                  50 (33.3)
```

Checking Missingness

The TableOne package can also help you check for missingness in your data. Here we set the first 10 Petal.Length values to NA and check for missingness:

```
na.iris <- iris; na.iris$Petal.Length[1:10] <- rep(NA, 10)
t3 <- CreateTableOne(data=na.iris)
summary(t3)
```

```
##
##      ### Summary of continuous variables ###
##
## strata: Overall
##                n miss p.miss mean  sd median p25 p75 min max skew kurt
## Sepal.Length 150    0      0    6 0.8      6 5.1   6 4.3   8  0.3 -0.6
## Sepal.Width  150    0      0    3 0.4      3 2.8   3 2.0   4  0.3  0.2
## Petal.Length 150   10      7    4 1.7      4 1.7   5 1.0   7 -0.4 -1.2
## Petal.Width  150    0      0    1 0.8      1 0.3   2 0.1   2 -0.1 -1.3
##
## ============================================================================
##
##      ### Summary of categorical variables ###
##
## strata: Overall
##      var   n miss p.miss      level freq percent cum.percent
##  Species 150    0    0.0     setosa   50    33.3        33.3
##                          versicolor   50    33.3        66.7
##                           virginica   50    33.3       100.0
##
```

Stratify by Group and Testing

Often times you want to see summaries of different strata, with the strata usually being the treatment variable. Going back to our Petal.Color example:

```
CreateTableOne(data = new.iris, strata = "Petal.Color")
```

```
##                            Stratified by Petal.Color
##                             White        Red          p      test
##   n                           80           70
##   Sepal.Length (mean (SD)) 5.92 (0.84)  5.76 (0.82)    0.234
##   Sepal.Width (mean (SD))  3.07 (0.43)  3.04 (0.44)    0.733
##   Petal.Length (mean (SD)) 3.87 (1.78)  3.63 (1.76)    0.403
##   Petal.Width (mean (SD))  1.23 (0.76)  1.16 (0.77)    0.585
##   Species (%)                                          0.512
##      setosa                  25 (31.2)    25 ( 35.7)
##      versicolor              30 (37.5)    20 ( 28.6)
##      virginica               25 (31.2)    25 ( 35.7)
##   Petal.Color = Red (%)       0 ( 0.0)    70 (100.0)  <0.001
```

Note the group comparison p-value that are printed in the table. The hypothesis test functions used by default are chisq.test() for categorical variables (with continuity correction) and oneway.test() for continuous variables (with equal variance assumption, i.e., regular ANOVA). Two-group ANOVA is equivalent of t-test.

Print Option 3: Standardized Mean Differences

The Standardized Mean Difference (SMD), also known as Cohen's D, is a measure of distance between two group means in terms of one or more variables. In practice it is often used as a balance measure of individual covariates before and after propensity score matching. (More on this later in the semester...) Using our example from the previous slide:

```
t3 <- CreateTableOne(data = new.iris, strata = "Petal.Color")
print(t3, smd = TRUE)
```

```
##                            Stratified by Petal.Color
##                             White        Red          p      test SMD
##   n                           80           70
##   Sepal.Length (mean (SD)) 5.92 (0.84)  5.76 (0.82)   0.234       0.196
##   Sepal.Width (mean (SD))  3.07 (0.43)  3.04 (0.44)   0.733       0.056
##   Petal.Length (mean (SD)) 3.87 (1.78)  3.63 (1.76)   0.403       0.137
##   Petal.Width (mean (SD))  1.23 (0.76)  1.16 (0.77)   0.585       0.089
##   Species (%)                                         0.512       0.191
##      setosa                  25 (31.2)    25 ( 35.7)
##      versicolor              30 (37.5)    20 ( 28.6)
##      virginica               25 (31.2)    25 ( 35.7)
##   Petal.Color = Red (%)       0 ( 0.0)    70 (100.0)  <0.001         NaN
```

Print Option 4: Variable Name Labels

Using the "labelled" package, a label can be assigned to each variable and displayed in the final printed table. This way, your variable names stay intact for further manipulations in your code but your tableone output will be presentation quality:

```
new.iris <- set_variable_labels(new.iris,
                                Sepal.Length = "Length of Sepal",
                                Sepal.Width = "Width of Sepal",
                                Petal.Length = "Length of Petal",
                                Petal.Width = "Width of Petal",
                                Petal.Color = "Color of Petal"
)
print(CreateTableOne(data=new.iris), varLabels = TRUE)
```

```
##
##                                Overall
##   n                             150
##   Length of Sepal (mean (SD)) 5.84 (0.83)
##   Width of Sepal (mean (SD))  3.06 (0.44)
##   Length of Petal (mean (SD)) 3.76 (1.77)
##   Width of Petal (mean (SD))  1.20 (0.76)
##   Species (%)
##      setosa                     50 (33.3)
##      versicolor                 50 (33.3)
##      virginica                  50 (33.3)
##   Color of Petal = Red (%)      70 (46.7)
```

Export to LATEX/CSV

In my opinion, a huge advantage of using the tableone package is how easily this R table can be exported. To turn this table into LATEX code, we use the "xtable" package here:

```
final.table <- print(CreateTableOne(data=new.iris), varLabels = TRUE)
print(xtable(final.table))
```

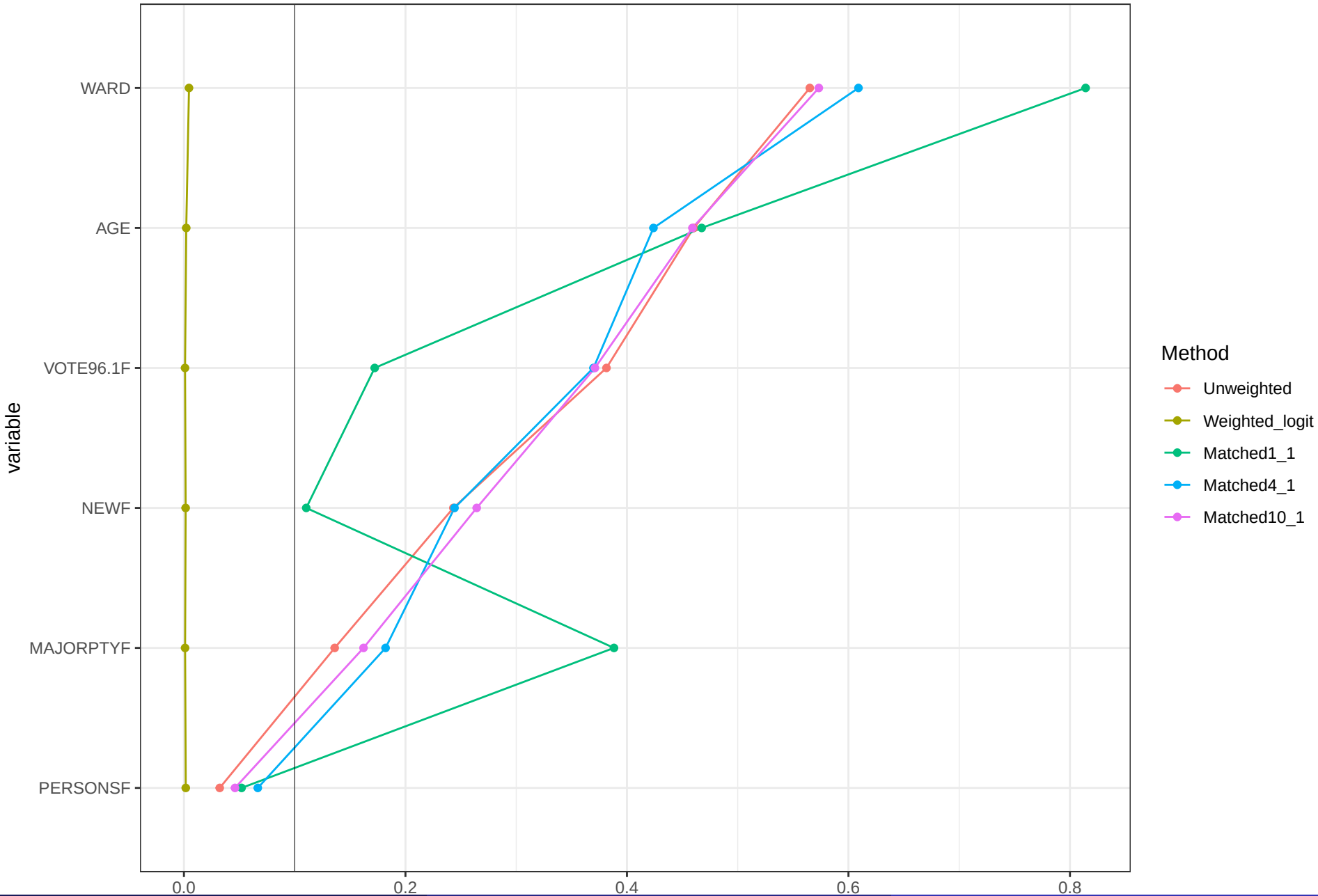
You can just as easily export this table to a CSV format (for Excel):

```
write.csv(final.table, file = "myTable.csv")
```

Plot of SMD

- If we have a lot of baseline covriates it is helpful to plot the absolute value of the standardized mean difference
- We will discuss methods in the future (e.g., weighting, matching, etc.) which help reduce the imbalance (i.e., absolute SMD) among groups. Plotting those on the same plot can show how we "fixed" the imbalance.
- Example given on the following slide of the SMD in the "original" dataset and then under various matching and weighting methods

Plot of SMD


WARD
AGE
VOTE96.1F
NEWF
MAJORPTYF
PERSONSF
variable
0.0
0.2
0.4
0.6
0.8
Method
Unweighted
Weighted_logit
Matched1_1
Matched4_1
Matched10_1

Other plots

- Can be useful to assess overlap in the distribution of covariates; will introduce later in the context of matching, weighting
- Should also graphically assess covariate outcome relationships; more on that in regression adjustment

Practice

Load the following dataset from the Mayo Clinic trial in primary biliary cirrhosis of the liver:

```
library(survival)
data(pbc)
```

Try answering the following questions using the pbc dataset:

- How many male patients are there in the dataset, and what is the percentage?
- Turn c("status", "trt", "ascites", "hepato","spiders", "edema", "stage") into categorical variables and show all levels. What percent of patients are stage 3?
- How many missing values are there for "copper" and "stage"? What percent of values are missing for "trig"?
- Display the IQR for c("bili","chol","copper","alk.phos","ast","trig","protime"). What is the median for "chol"? What is the 25th percentile value for "ast"?