

Descriptive Statistics

Juan Pablo Lewinger

6/14/2022

Descriptive Statistics

- First analytical task to understand the data and begin to see patterns
- Numerical and graphical description of the data
- Today we'll (mostly) focus on univariate summary statistics (i.e. one variable at time)
- We'll discuss bivariate summary statistics (i.e. two variables at a time) next week

Mayo Clinic's PBC data

- Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver
- Conducted between 1974 and 1984.
- Placebo controlled trial of the drug D-penicillamine.
- First 424 PBC patients
- First 312 cases in the data set participated in the randomized trial and contain largely complete data.
- Additional 106 subjects did not participate in the clinical trial but consented to have data collected

```
setwd("/Users/JP/Google Drive/Teaching/LAs BEST/2022")
pbc = read.csv('pbc.csv')
dim(pbc) #How many observations (rows) and how many variables (columns)

## [1] 418 20
```

Mayo Clinic's PBC data

Categorical variables:

- id
- status: 0=alive, 1=liver transplant, 2=dead
- trt: 1 = D-penicillamine, 2=placebo
- sex: 0=male, 1=female
- presence of ascites: 0=no 1=yes
- presence of hepatomegaly: 0=no 1=yes
- presence of spiders 0=no 1=yes
- presence of edema: 0=no edema and no diuretic therapy for edema; .5 = edema present without diuretics, or edema resolved by diuretics; 1 = edema despite diuretic therapy
- histologic stage of disease

Mayo Clinic's PBC data

Quantitative variables:

- number of days between registration and the earlier of death, transplantation, or study analysis time in July, 1986
- age in days
- serum bilirubin in mg/dl
- serum cholesterol in mg/dl
- albumin in gm/dl
- urine copper in ug/day
- alkaline phosphatase in U/liter
- SGOT (Serum glutamic oxaloacetic transaminase) in U/ml
- triglycerides in mg/dl
- platelets per cubic ml/1000
- prothrombin time in seconds

Mayo Clinic's PBC data

```
summary(pbc)
```

```
##           id           time           status           trt
## Min.      : 1.0    Min.      : 41    Min.      :0.0000    Min.      :1.000
## 1st Qu.:105.2    1st Qu.:1093    1st Qu.:0.0000    1st Qu.:1.000
## Median :209.5    Median :1730    Median :0.0000    Median :1.000
## Mean     :209.5    Mean     :1918    Mean     :0.8301    Mean     :1.494
## 3rd Qu.:313.8    3rd Qu.:2614    3rd Qu.:2.0000    3rd Qu.:2.000
## Max.     :418.0    Max.     :4795    Max.     :2.0000    Max.     :2.000
##                                     NA's      :106
##           age           sex           ascites           hepato
## Min.      :26.28    Length:418    Min.      :0.00000    Min.      :0.0000
## 1st Qu.:42.83    Class :character    1st Qu.:0.00000    1st Qu.:0.0000
## Median :51.00    Mode  :character    Median :0.00000    Median :1.0000
## Mean     :50.74                                Mean     :0.07692    Mean     :0.5128
## 3rd Qu.:58.24                                3rd Qu.:0.00000    3rd Qu.:1.0000
## Max.     :78.44                                Max.     :1.00000    Max.     :1.0000
##                                     NA's      :106    NA's      :106
##           spiders           edema           bili           chol
## Min.      :0.0000    Min.      :0.0000    Min.      : 0.300    Min.      : 120.0
## 1st Qu.:0.0000    1st Qu.:0.0000    1st Qu.: 0.800    1st Qu.: 249.5
## Median :0.0000    Median :0.0000    Median : 1.400    Median : 309.5
## Mean     :0.2885    Mean     :0.1005    Mean     : 3.221    Mean     : 369.5
## 3rd Qu.:1.0000    3rd Qu.:0.0000    3rd Qu.: 3.400    3rd Qu.: 400.0
## Max.     :1.0000    Max.     :1.0000    Max.     :28.000    Max.     :1775.0
## NA's      :106                                     NA's      :134
##           albumin           copper           alk.phos           ast
## Min.      :1.960    Min.      : 4.00    Min.      : 289.0    Min.      : 26.35
## 1st Qu.:3.243    1st Qu.: 41.25    1st Qu.: 871.5    1st Qu.: 80.60
## Median :3.530    Median : 73.00    Median :1259.0    Median :114.70
## Mean     :3.497    Mean     : 97.65    Mean     :1982.7    Mean     :122.56
## 3rd Qu.:3.770    3rd Qu.:123.00    3rd Qu.:1980.0    3rd Qu.:151.90
## Max.     :4.640    Max.     :588.00    Max.     :13862.4    Max.     :457.25
```

Mayo Clinic's PBC data

```
head(pbc)
```

```
##   id time status trt      age sex ascites hepato spiders edema bili chol
## 1  1  400      2   1 58.76523  f      1      1      1  1.0 14.5 261
## 2  2 4500      0   1 56.44627  f      0      1      1  0.0  1.1 302
## 3  3 1012      2   1 70.07255  m      0      0      0  0.5  1.4 176
## 4  4 1925      2   1 54.74059  f      0      1      1  0.5  1.8 244
## 5  5 1504      1   2 38.10541  f      0      1      1  0.0  3.4 279
## 6  6 2503      2   2 66.25873  f      0      1      0  0.0  0.8 248
##   albumin copper alk.phos      ast trig platelet protime stage
## 1    2.60    156   1718.0 137.95  172      190    12.2      4
## 2    4.14     54   7394.8 113.52   88      221    10.6      3
## 3    3.48    210    516.0  96.10   55      151    12.0      4
## 4    2.54     64   6121.8  60.63   92      183    10.3      4
## 5    3.53    143    671.0 113.15   72      136    10.9      3
## 6    3.98     50    944.0  93.00   63        NA    11.0      3
```

```
pbc$sex = factor(pbc$sex, levels = c('m', 'f'), labels = c('M', 'F'))
pbc$trt = factor(pbc$trt, levels = c(1,2), labels = c('D-penicillamine', 'placebo'))
pbc$status = factor(pbc$status, levels = c(0,1,2), labels = c('alive', 'liver transplant', 'dead'))
```

Mayo Clinic's PBC data

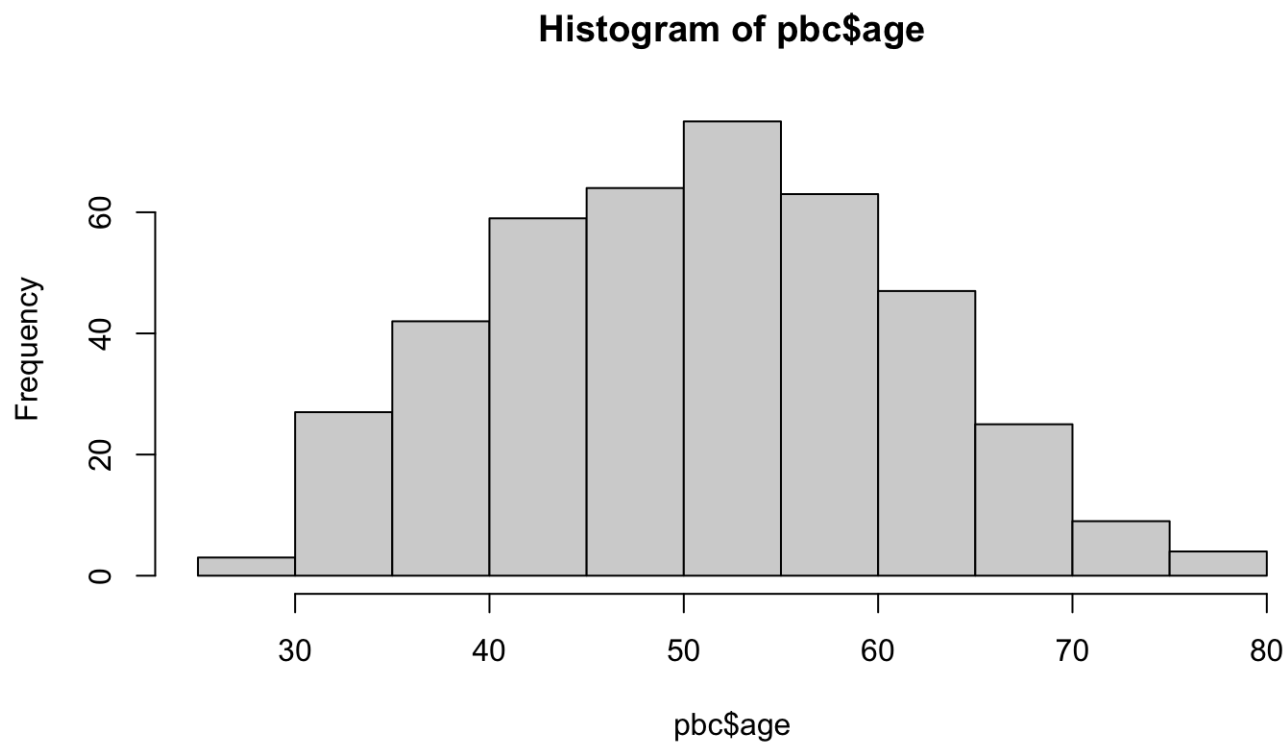
```
str(pbc)
```

```
## 'data.frame':    418 obs. of  20 variables:
## $ id      : int   1 2 3 4 5 6 7 8 9 10 ...
## $ time    : int  400 4500 1012 1925 1504 2503 1832 2466 2400 51 ...
## $ status  : Factor w/ 3 levels "alive","liver transplant",...: 3 1 3 3 2 3 1 3 3 3 ...
## $ trt     : Factor w/ 2 levels "D-penicillamine",...: 1 1 1 1 2 2 2 2 1 2 ...
## $ age     : num   58.8 56.4 70.1 54.7 38.1 ...
## $ sex     : Factor w/ 2 levels "M","F": 2 2 1 2 2 2 2 2 2 2 ...
## $ ascites : int    1 0 0 0 0 0 0 0 0 1 ...
## $ hepato  : int    1 1 0 1 1 1 1 0 0 0 ...
## $ spiders : int    1 1 0 1 1 0 0 0 1 1 ...
## $ edema   : num    1 0 0.5 0.5 0 0 0 0 0 1 ...
## $ bili    : num   14.5 1.1 1.4 1.8 3.4 0.8 1 0.3 3.2 12.6 ...
## $ chol    : int   261 302 176 244 279 248 322 280 562 200 ...
## $ albumin : num    2.6 4.14 3.48 2.54 3.53 3.98 4.09 4 3.08 2.74 ...
## $ copper   : int   156 54 210 64 143 50 52 52 79 140 ...
## $ alk.phos: num  1718 7395 516 6122 671 ...
## $ ast     : num   137.9 113.5 96.1 60.6 113.2 ...
## $ trig    : int   172 88 55 92 72 63 213 189 88 143 ...
## $ platelet: int   190 221 151 183 136 NA 204 373 251 302 ...
## $ protime : num   12.2 10.6 12 10.3 10.9 11 9.7 11 11 11.5 ...
## $ stage   : int    4 3 4 4 3 3 3 3 2 4 ...
```


Histograms

Histogram: univariate graphical summary for a quantitative/continuous variable

```
hist(pbc$age)
```



Descriptive stats for quantitative variables

Measures of location: mean, mode, median, first quartile, third quartile, minimum, maximum

```
mean(pbc$age)
```

```
## [1] 50.74155
```

```
median(pbc$age)
```

```
## [1] 51.00068
```

```
min(pbc[, 'age'])
```

```
## [1] 26.27789
```

```
max(pbc[, 5])    # Not recommended to subset by column number
```

```
## [1] 78.43943
```

Descriptive stats for quantitative variables

Measures of dispersion: variance, standard deviation, interquantile range (IQR)

```
var(pbc$age)
```

```
## [1] 109.1443
```

```
sd(pbc$age)
```

```
## [1] 10.44721
```

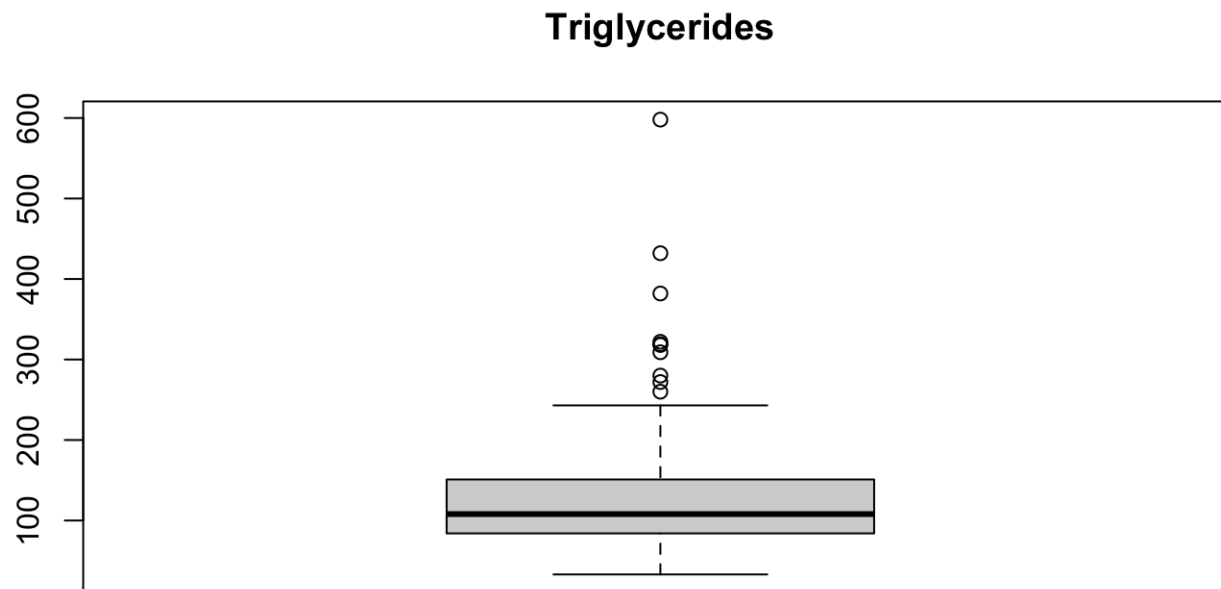
```
IQR(pbc[, 'age'])
```

```
## [1] 15.40862
```

Boxplots

Boxplots combines key summary stats (median, quartiles, IQR) into a summary plot

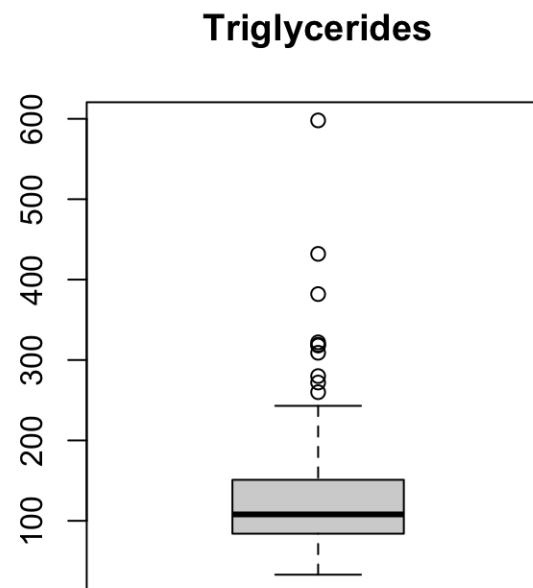
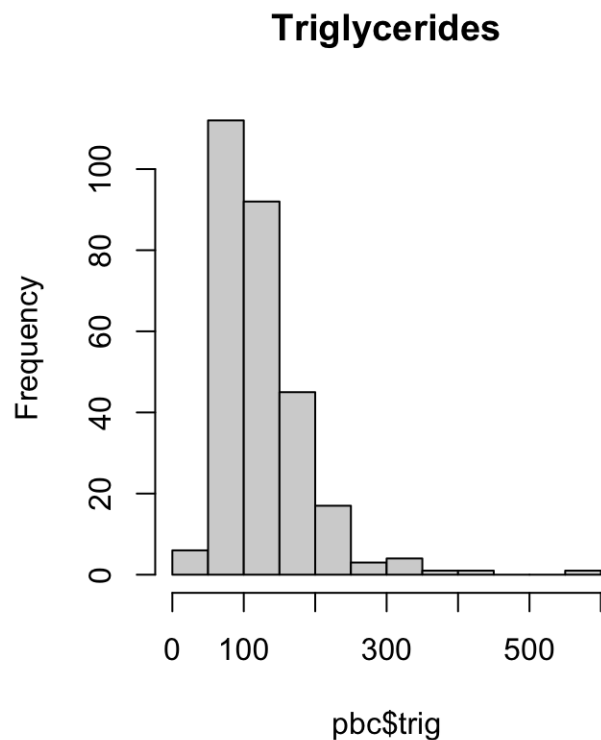
```
boxplot(pbc$trig, main = 'Triglycerides')
```



Is Triglycerides skewed left or right?

Histogram and boxplot on same plot

```
par(mfrow = c(1, 2))  
hist(pbc$trig, main='Triglycerides')  
boxplot(pbc$trig, main = 'Triglycerides')
```

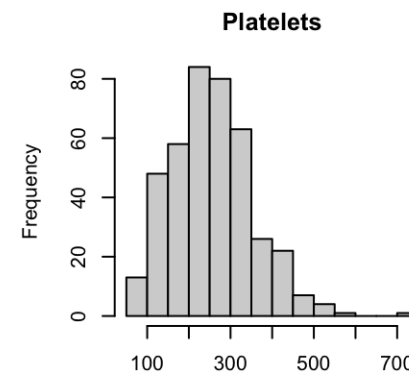
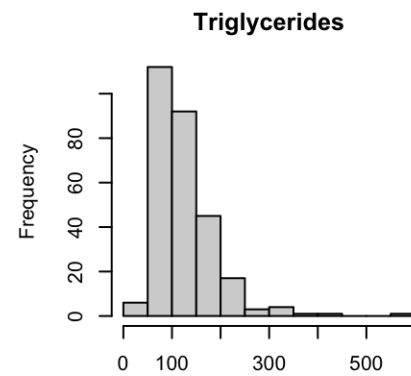
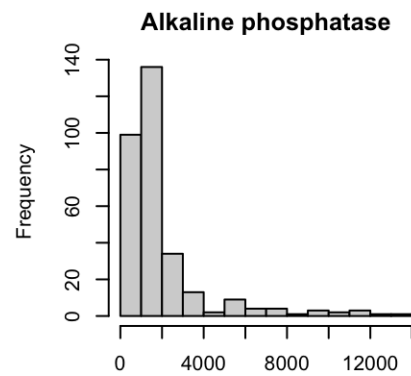
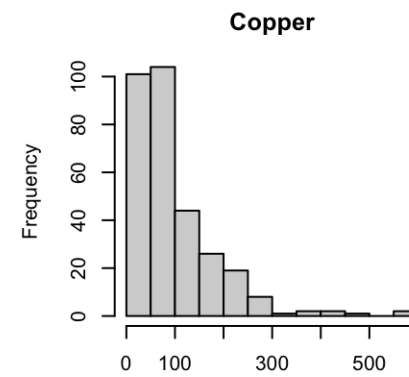
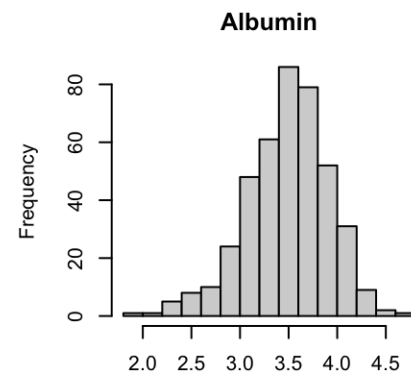
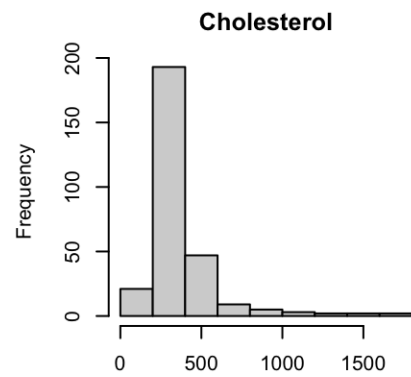


Multiple Histograms

```
par(mfrow = c(2, 3))

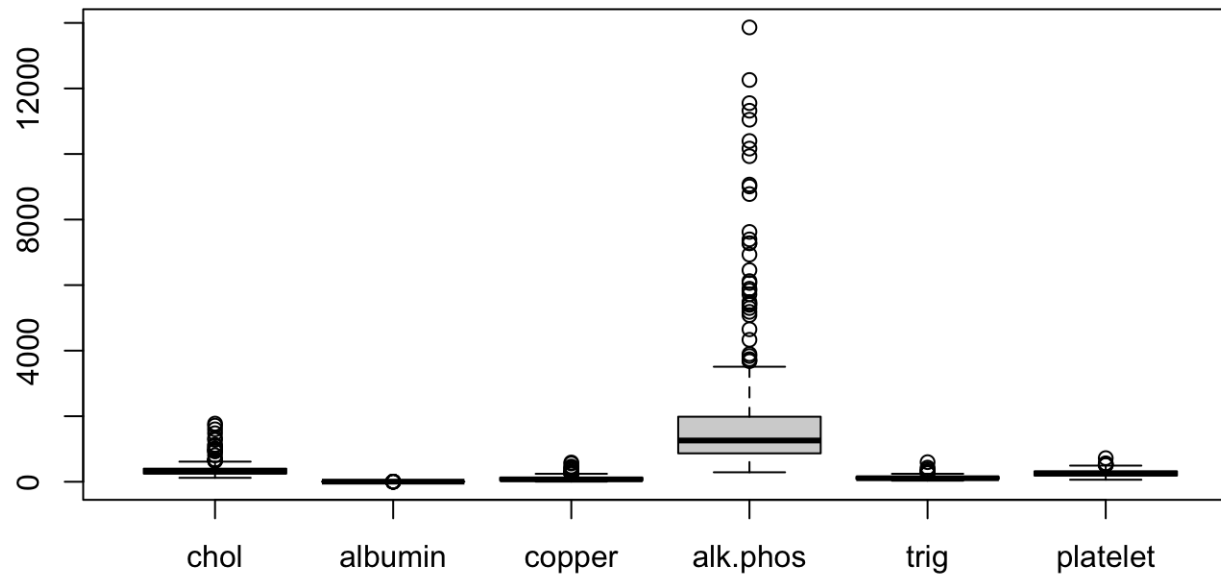
hist(pbc$chol, main='Cholesterol')
hist(pbc$albumin, main='Albumin')
hist(pbc$copper, main='Copper')
hist(pbc$alk.phos, main='Alkaline phosphatase')
hist(pbc$trig, main='Triglycerides')
hist(pbc$platelet, main='Platelets')
```

Multiple Histograms

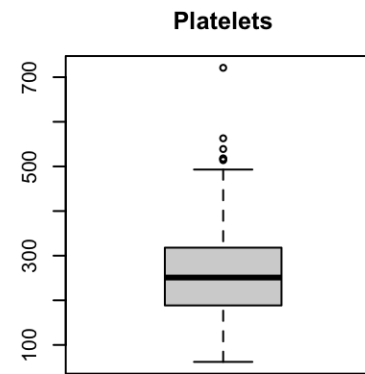
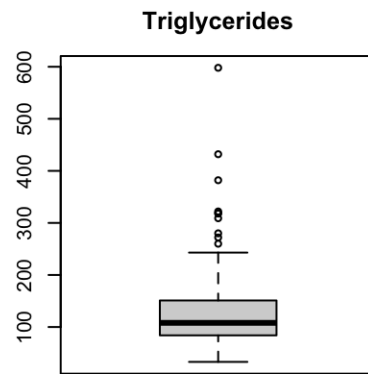
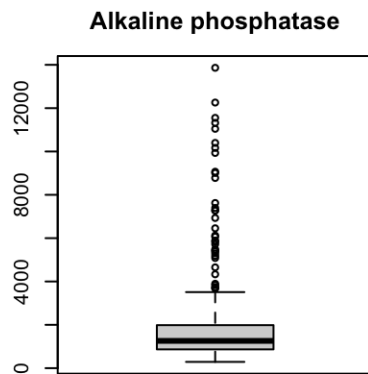
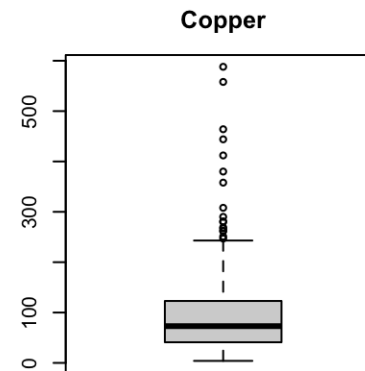
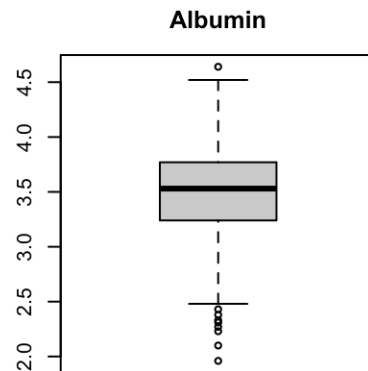
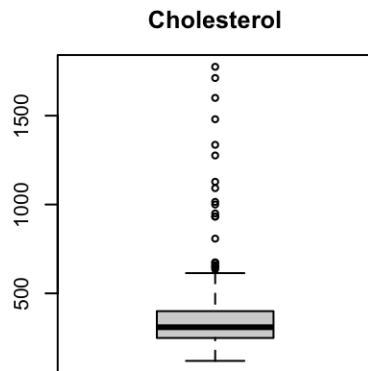


Multiple boxplots (Version 1)

```
boxplot(pbc[, c("chol", "albumin", "copper", "alk.phos", "trig", "platelet")])
```



Multiple boxplots (Version 2)



Descriptive stats for categorical variables

Numerical summary: Counts and proportion/percent per category

```
table(pbc$stage)
```

```
##  
##    1    2    3    4  
##  21   92  155  144
```

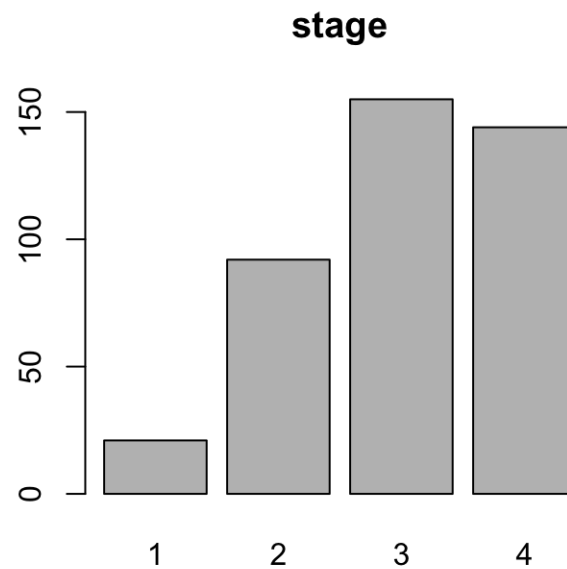
```
prop.table(table(pbc$stage))
```

```
##  
##           1           2           3           4  
## 0.05097087 0.22330097 0.37621359 0.34951456
```

Descriptive stats for categorical variables

Graphical summary: Barplot

```
barplot(table(pbc$stage), main = 'stage')
```



Package tableone

Adapted from tableone Vignette by Kazuki Yoshida

```
library(tableone)
```

- Makes it easy to construct table of baseline characteristics
- Commonly found in biomedical research papers as 'Table 1'.
- Can summarize continuous and categorical variables on same table.
- Categorical variables can be summarized as counts and/or percentages.
- Continuous variables can be summarized by means and standard deviations
- or by medians and interquartile ranges.

Table 1 example

Table 1. Baseline Characteristics of the Patients.*		
Variable	Rhythm-Control Group (N = 682)	Rate-Control Group (N = 694)
Male sex (%)	78	85
Age (yr)	66±11	67±11
Body-mass index†	27.8±5.4	28.0±5.1
Nonwhite race (%)‡	16	13
NYHA class III or IV (%)		
At baseline	32	31
During previous 6 mo	76	76
Predominant cardiac diagnosis (%)§		
Coronary artery disease	48	48
Valvular heart disease	5	5
Nonischemic cardiomyopathy	36	39
Congenital heart disease	1	1
Hypertensive heart disease	10	7
Coexisting conditions (%)		
Hypertension	49	46
Diabetes	22	20
Previous stroke or transient ischemic attack	11	8

From Roy et al. Rhythm Control versus Rate Control for Atrial Fibrillation and Heart Failure. NEJM (2008)

Single group summary

- Simplest use is for summarizing whole dataset.
- You can just feed in the data frame to the function `CreateTableOne()`.

```
CreateTableOne(data = pbc)
```

Single group summary

```
##
##                               Overall
##  n                               418
##  id (mean (SD))                209.50 (120.81)
##  time (mean (SD))              1917.78 (1104.67)
##  status (%)
##    alive                        232 (55.5)
##    liver transplant              25 ( 6.0)
##    dead                         161 (38.5)
##  trt = placebo (%)              154 (49.4)
##  age (mean (SD))                50.74 (10.45)
##  sex = F (%)                    374 (89.5)
##  ascites (mean (SD))            0.08 (0.27)
##  hepato (mean (SD))             0.51 (0.50)
##  spiders (mean (SD))            0.29 (0.45)
##  edema (mean (SD))              0.10 (0.25)
##  bili (mean (SD))               3.22 (4.41)
##  chol (mean (SD))               369.51 (231.94)
##  albumin (mean (SD))            3.50 (0.42)
##  copper (mean (SD))              97.65 (85.61)
##  alk.phos (mean (SD)) 1982.66 (2140.39)
##  ast (mean (SD))                122.56 (56.70)
##  trig (mean (SD))               124.70 (65.15)
##  platelet (mean (SD))           257.02 (98.33)
##  protime (mean (SD))            10.73 (1.02)
##  stage (mean (SD))              3.02 (0.88)
```

Subset of variables

Can specify a subset of variables to summarize using the vars argument (e.g. exclude id variable)

```
## Get variables' names
```

```
names(pbc)
```

```
## [1] "id"      "time"    "status"  "trt"     "age"     "sex"
## [7] "ascites" "hepato"  "spiders" "edema"   "bili"    "chol"
## [13] "albumin" "copper"  "alk.phos" "ast"     "trig"    "platelet"
## [19] "protime" "stage"
```

```
## Vector of variables to summarize
```

```
myVars <- c("status", "trt", "age", "sex", "ascites",  
            "albumin", "copper", "stage")
```

```
CreateTableOne(data = pbc, vars = myVars)
```


Subset of variables

```
##
##               Overall
##  n               418
##  status (%)
##    alive          232 (55.5)
##    liver transplant  25 ( 6.0)
##    dead            161 (38.5)
##  trt = placebo (%)  154 (49.4)
##  age (mean (SD))   50.74 (10.45)
##  sex = F (%)       374 (89.5)
##  ascites (mean (SD)) 0.08 (0.27)
##  albumin (mean (SD)) 3.50 (0.42)
##  copper (mean (SD)) 97.65 (85.61)
##  stage (mean (SD))  3.02 (0.88)
```

Categorical variables

- Some categorical variables are coded numerically in the dataframe (e.g. ascites, hepato)
- Need to either transform to factors as we did above for trt, sex and status
- or use factorVars argument to transform them on-the-fly.

```
catVars <- c("ascites", "stage")
```

Categorical variables

- Binary categorical variables are summarized as counts and percentages of the second level
- For 3+ category variable all levels are summarized
- Percentages are calculated after excluding missing values.

```
tab2 = CreateTableOne(vars = myVars, data = pbc, factorVars = catVars)
```

```
tab2
```

Categorical variables

```
##
##                                Overall
##  n                                418
##  status (%)
##    alive                232 (55.5)
##    liver transplant      25 ( 6.0)
##    dead                 161 (38.5)
##  trt = placebo (%)       154 (49.4)
##  age (mean (SD))        50.74 (10.45)
##  sex = F (%)            374 (89.5)
##  ascites = 1 (%)         24 ( 7.7)
##  albumin (mean (SD))    3.50 (0.42)
##  copper (mean (SD))     97.65 (85.61)
##  stage (%)
##    1                    21 ( 5.1)
##    2                    92 (22.3)
##    3                   155 (37.6)
##    4                   144 (35.0)
```

Showing all categories

```
print(tab2, showAllLevels = TRUE)
```

```
##
##           level           Overall
##  n                418
##  status (%)      alive      232 (55.5)
##                liver transplant  25 ( 6.0)
##                dead        161 (38.5)
##  trt (%)         D-penicillamine  158 (50.6)
##                placebo        154 (49.4)
##  age (mean (SD))  50.74 (10.45)
##  sex (%)         M           44 (10.5)
##                F           374 (89.5)
##  ascites (%)      0           288 (92.3)
##                1           24 ( 7.7)
##  albumin (mean (SD))  3.50 (0.42)
##  copper (mean (SD))  97.65 (85.61)
##  stage (%)        1           21 ( 5.1)
##                2           92 (22.3)
##                3          155 (37.6)
##                4          144 (35.0)
```

Recoding ascites

```
pbcs$ascites = factor(pbc$ascites, levels = c(0,1), labels = c('No', 'Yes'))
tab2 = CreateTableOne(vars = myVars, data = pbc, factorVars = catVars)
print(tab2, showAllLevels = TRUE)
```

```
##
##           level           Overall
##  n                418
##  status (%)      alive        232 (55.5)
##                  liver transplant  25 ( 6.0)
##                  dead          161 (38.5)
##  trt (%)          D-penicillamine  158 (50.6)
##                  placebo         154 (49.4)
##  age (mean (SD))  50.74 (10.45)
##  sex (%)          M            44 (10.5)
##                  F            374 (89.5)
##  ascites (%)      No           288 (92.3)
##                  Yes           24 ( 7.7)
##  albumin (mean (SD))  3.50 (0.42)
##  copper (mean (SD))  97.65 (85.61)
##  stage (%)         1            21 ( 5.1)
##                  2            92 (22.3)
##                  3           155 (37.6)
##                  4           144 (35.0)
```

Detailed information on missing values

```
summary(tab2$ContTable)
```

```
## strata: Overall
```

##	n	miss	p.miss	mean	sd	median	p25	p75	min	max	skew	kurt
## age	418	0	0	50.7	10.45	51.0	42.8	58.2	26	78.4	0.087	-0.62
## albumin	418	0	0	3.5	0.42	3.5	3.2	3.8	2	4.6	-0.468	0.57
## copper	418	108	26	97.6	85.61	73.0	41.2	123.0	4	588.0	2.304	7.62

Detailed information on missing values

```
summary(tab2$CatTable)
```

```
## strata: Overall
##      var    n miss p.miss      level freq percent cum.percent
##    status 418    0   0.0      alive  232   55.5       55.5
##                                liver transplant  25    6.0       61.5
##                                dead    161   38.5       100.0
##
##      trt 418  106  25.4 D-penicillamine  158   50.6       50.6
##                                placebo  154   49.4       100.0
##
##      sex 418    0   0.0      M    44   10.5       10.5
##                                F   374   89.5       100.0
##
##    ascites 418  106  25.4      No  288   92.3       92.3
##                                Yes   24    7.7       100.0
##
##      stage 418    6   1.4      1   21    5.1        5.1
##                                2   92   22.3       27.4
##                                3  155   37.6       65.0
##                                4  144   35.0       100.0
##
```


Summarizing nonnormal variables

- albumin, and copper are negatively and positively skewed respectively
- Age is symmetric, bell-shaped, normal looking
- Skewed distributions are not well summarized by mean and sd
- Use median and IQR for skewed/nonnormal variables

```
skewed <- c("albumin", "copper")
```

```
print(tab2, nonnormal = skewed)
```

Summarizing skewed variables

```
##
##                                Overall
##  n                                418
##  status (%)
##    alive                        232 (55.5)
##    liver transplant              25 ( 6.0)
##    dead                         161 (38.5)
##  trt = placebo (%)              154 (49.4)
##  age (mean (SD))                50.74 (10.45)
##  sex = F (%)                    374 (89.5)
##  ascites = Yes (%)              24 ( 7.7)
##  albumin (median [IQR])         3.53 [3.24, 3.77]
##  copper (median [IQR])          73.00 [41.25, 123.00]
##  stage (%)
##    1                            21 ( 5.1)
##    2                            92 (22.3)
##    3                           155 (37.6)
##    4                           144 (35.0)
```

Fine tuning

Check out `?print.TableOne` for the list of options (partial list below):

`catDigits`: Number of digits to print for proportions. Default 1.

`contDigits`: Number of digits to print for continuous variables. Default 2.

`quote`:

Whether to show everything in quotes. The default is FALSE. If TRUE, everything including the row and column names are quoted so that you can copy it to Excel easily.

`missing`: Whether to show missing data information.

`explain`:

Whether to add explanation to the variable names, i.e., (%) is added to the variable names when percentage is shown.

Stratified table

We may want to summarize by levels of a categorical variable (e.g., treatment)

```
myVars = myVars <- c("status", "age", "sex", "ascites", "albumin", "copper", "stage")
tab3 <- CreateTableOne(vars = myVars, strata = "trt", data = pbc, factorVars = catVars)

print(tab3, nonnormal = skewed, formatOptions = list(big.mark = ","))
```

Stratified table

```
##                               Stratified by trt
##                               D-penicillamine      placebo      p
##   n                               158              154
##   status (%)                                0.894
##     alive                83 (52.5)              85 (55.2)
##     liver transplant     10 ( 6.3)              9 ( 5.8)
##     dead                 65 (41.1)              60 (39.0)
##   age (mean (SD))      51.42 (11.01)          48.58 (9.96)      0.018
##   sex = F (%)          137 (86.7)             139 (90.3)      0.421
##   ascites = Yes (%)    14 ( 8.9)              10 ( 6.5)      0.567
##   albumin (median [IQR]) 3.56 [3.21, 3.83]      3.54 [3.34, 3.78]      0.950
##   copper (median [IQR]) 73.00 [40.00, 121.00] 73.00 [43.00, 139.00] 0.717
##   stage (%)                                0.201
##     1                12 ( 7.6)              4 ( 2.6)
##     2                35 (22.2)              32 (20.8)
##     3                56 (35.4)              64 (41.6)
##     4                55 (34.8)              54 (35.1)
##                               Stratified by trt
##                               test
##   n
##   status (%)
##     alive
##     liver transplant
##     dead
##   age (mean (SD))
##   sex = F (%)
##   ascites = Yes (%)
##   albumin (median [IQR]) nonnorm
##   copper (median [IQR])  nonnorm
##   stage (%)
##     1
##     2
##     3
##     4
```

Exporting table to csv

```
tab3Mat <- print(tab3, nonnormal = skewed, exact = "stage", quote = FALSE,  
                 noSpaces = TRUE, printToggle = FALSE)  
  
write.csv(tab3Mat, file = "myTable1.csv")
```