Exercise with solutions

Alessandra Meddis

2023-08-15

Exercise A: Reshaping the data

For this exercise we will work with a subset of "follicle" data, collected from patients with cancer that had OTC (ovarian tissue cryopreservation). Follicles were cultured for 8 days and the diameter was collected every 2 days. The aim of the study was to compare the follicles growth among different treatment groups.

Information are collected in two data sets.

Data set *follicle* with:

- Patient: patient ID
- Number: follicle ID
- Day_x: follicle diameter at Day x

Data set *patient* with:

- Patient ID
- Treatment group
- Type of Disease
- Age at Day 0

Question 1 Download both data sets and load them into R.

Consider Data set *patient*. Visualize the first lines of the data and print a summary of the data.

1. How many Patients?
2. what is the data type of each variable?

```
db_follicle<-read.csv("~/Desktop/KVN2021/Course/IntrotoR/data_exercise/follicle.csv")
db_pat<-read.csv("~/Desktop/KVN2021/Course/IntrotoR/data_exercise/patient.csv")
knitr::kable(head(db_pat))
```

Patient	Disease	Treatment	Age
1	Breast_cancer	FBS	31.6
2	Breast_cancer	FBS	29.4
3	Neurological_cancer	FBS	19.3
4	Rheumatoid_arthritis	FBS	20.3
5	Mb_Hodgkin	hPL	29.8
6	Sarcoma	HSA	19.7

```
dim(db_pat)
```

```
## [1] 14  4
```

```
summary(db_pat)
```

```
##     Patient        Disease           Treatment              Age
##  Min.   : 1.00   Length:14          Length:14          Min.   :19.30
##  1st Qu.: 4.25   Class :character   Class :character   1st Qu.:29.50
##  Median : 7.50   Mode  :character   Mode  :character   Median :31.40
##  Mean   : 7.50                                         Mean   :29.74
##  3rd Qu.:10.75                                         3rd Qu.:32.25
##  Max.   :14.00                                         Max.   :37.10
```

We have a total of 14 Patients. Patient and Age are quantitative and are either numerical or integer, while Disease and Treatment are categorical variables and are listed as characters.

Question 2

Treatment and Disease are *characters*, can we understand from the summary how many different diseases are in the data?

1. Would it be better if Treatment and Disease were of a different type? If yes, which one?
2. Transform them into factor. You can use the function **factor()** in **R**.
3. Print the summary of the data, can you see any difference?

We discussed that for easier representation it is better to have factor for categorical variables, thus we can transform Disease and Treatment into factors.

```
db_pat$Treatment<-factor(db_pat$Treatment)
db_pat$Disease<-factor(db_pat$Disease)

summary(db_pat)
```

```
##     Patient                          Disease  Treatment      Age
##  Min.   : 1.00   Brain_cancer            :1   FBS:11    Min.   :19.30
##  1st Qu.: 4.25   Breast_cancer           :8   hPL: 1    1st Qu.:29.50
##  Median : 7.50   Chronic_myeloid_leukemia:1   HSA: 2    Median :31.40
##  Mean   : 7.50   Mb_Hodgkin              :1             Mean   :29.74
##  3rd Qu.:10.75   Neurological_cancer     :1             3rd Qu.:32.25
##  Max.   :14.00   Rheumatoid_arthritis    :1             Max.   :37.10
##                  Sarcoma                 :1
```

The summary now shows the different levels of Treatment and Disease with the number of observations in that category.

Question 3

Age is a continuous covariate, we want to group people into two categories:

1. Show min,max and mean for age
2. Calculate the mean of Age for each disease group
3. Create a categorical variable for age with the median as threshold for the two categories **Hint: use the function *cut()***

```
summary(db_pat$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   19.30   29.50   31.40   29.74   32.25   37.10
```

```
min(db_pat$Age)
```

```
## [1] 19.3
```

```
max(db_pat$Age)
```

```
## [1] 37.1
```

```
mean(db_pat$Age)
```

```
## [1] 29.74286
```

```
aggregate(Age~Disease, db_pat, mean)
```

```
##                    Disease     Age
## 1             Brain_cancer 32.1000
## 2            Breast_cancer 32.9125
## 3 Chronic_myeloid_leukemia 31.9000
## 4               Mb_Hodgkin 29.8000
## 5      Neurological_cancer 19.3000
## 6     Rheumatoid_arthritis 20.3000
## 7                  Sarcoma 19.7000
```

```
db_pat$age.cat<-cut(db_pat$Age, breaks=c(15,32,40), labels=c("<32",">32"))

table(db_pat$age.cat,db_pat$Disease)
```

```
##
##       Brain_cancer Breast_cancer Chronic_myeloid_leukemia Mb_Hodgkin
##   <32            0             4                        1          1
##   >32            1             4                        0          0
##
##       Neurological_cancer Rheumatoid_arthritis Sarcoma
##   <32                   1                    1       1
##   >32                   0                    0       0
```

Question 4

1. Show the number of patients in each treatment group
2. Show the proportion of patients in each disease group
3. Show the number of patients by treatment group and disease (two-ways table)
4. Use the commands: $db\_patCancer < -ifelse(db_patDisease == "Breast\_cancer", "Breast\_cancer", "Others"))$ 3a. What do we obtain with this command? 3b. Can we do it in another way? **Hint: check combining groups in the slides**
5. Calculate the mean of Age for each group respect to *Cancer*

```
table(db_pat$Treatment)
```

```
##
## FBS hPL HSA
##  11   1   2
```

```
prop.table(table(db_pat$Disease))
```

```
##
##             Brain_cancer            Breast_cancer Chronic_myeloid_leukemia
##               0.07142857               0.57142857               0.07142857
##                Mb_Hodgkin      Neurological_cancer     Rheumatoid_arthritis
##               0.07142857               0.07142857               0.07142857
##                   Sarcoma
```

```
##               0.07142857
```

```
table(db_pat$Disease,db_pat$Treatment)
```

```
##
##                            FBS hPL HSA
##   Brain_cancer               0   0   1
##   Breast_cancer              8   0   0
##   Chronic_myeloid_leukemia   1   0   0
##   Mb_Hodgkin                 0   1   0
##   Neurological_cancer        1   0   0
##   Rheumatoid_arthritis       1   0   0
##   Sarcoma                    0   0   1
```

```
db_pat$Cancer<-ifelse(db_pat$Disease=="Breast_cancer","Breast_cancer","Others")

db_pat$Cancer2<-as.character(db_pat$Disease)
db_pat$Cancer2[db_pat$Cancer2!="Breast_cancer"]<-"Others"
db_pat$Cancer2<-factor(db_pat$Cancer2)

tapply(db_pat$Age, db_pat$Cancer, mean)
```

```
## Breast_cancer        Others
##      32.91250      25.51667
```

Question 5

Consider only patients with Breast cancer.

1. Subset data for Breast cancer patients
2. Show the number of patients for each Age gruop
3. Calculate mean and standard deviation for Age

```
db_BC<-subset(db_pat, Cancer=="Breast_cancer")

table(db_BC$age.cat)
```

```
##
## <32 >32
##   4   4
```

```
mean(db_BC$Age)
```

```
## [1] 32.9125
```

```
sd(db_BC$Age)
```

```
## [1] 2.700496
```

Exercise B: Reshaping data (Part II)

Consider Data set *follicle*. Visualize the first lines of the data and print a summary of the data.

```
head(db_follicle)
```

```
##   Number Patient     Day0     Day2     Day4     Day6    Day8
## 1      1       1 101.4590 112.1605 135.1980 160.2315 161.500
## 2      2       1  89.8315 141.3770 165.4925       NA      NA
## 3      3       1  90.2835 116.9870 122.6500 127.7305 129.447
```

```
## 4      4       1 120.3145 148.8840 166.9970 170.0245 170.740
## 5      5       1  93.0085 112.3135 120.8550 120.9000 120.940
## 6      6       2  83.9085 100.3520 112.6875       NA      NA
```

```
summary(db_follicle)
```

```
##      Number         Patient          Day0             Day2
##  Min.   : 1.00   Min.   : 1.0   Min.   : 48.32   Min.   : 59.87
##  1st Qu.:18.25   1st Qu.: 4.0   1st Qu.: 62.74   1st Qu.: 81.50
##  Median :35.50   Median : 7.5   Median : 83.91   Median :105.03
##  Mean   :35.50   Mean   : 7.5   Mean   : 84.70   Mean   :106.85
##  3rd Qu.:52.75   3rd Qu.:11.0   3rd Qu.: 95.56   3rd Qu.:118.05
##  Max.   :70.00   Max.   :14.0   Max.   :194.47   Max.   :263.01
##                                 NA's   :3        NA's   :5
##       Day4             Day6             Day8
##  Min.   : 62.97   Min.   : 71.34   Min.   : 72.48
##  1st Qu.: 96.96   1st Qu.:102.64   1st Qu.:103.78
##  Median :120.51   Median :126.42   Median :129.45
##  Mean   :127.01   Mean   :139.46   Mean   :146.24
##  3rd Qu.:147.24   3rd Qu.:172.66   3rd Qu.:176.93
##  Max.   :299.65   Max.   :304.94   Max.   :318.88
##  NA's   :11       NA's   :24       NA's   :29
```

Question 1

1. Calculate mean and standard deviation of the diameter at Day0 (Be carfeul, there are some missing!)
2. When we encounter into missing, we are often interested in the *complete case analysis* where we exclude patients with missing observations:

2a. Use the *na.omit* function (excludes all rows that have one missing values) (Run the command *db.CC<-na.omit(NameofDataFrame)**) 2b. check the dimension of the new data.frame 2c. Calculate mean and standard deviation of the diameter at Day0 from db.CC . Compare results with the ones in point 1.

```
mean(db_follicle$Day0, na.rm=TRUE)
```

```
## [1] 84.70384
```

```
sd(db_follicle$Day0, na.rm=TRUE)
```

```
## [1] 27.6819
```

```
db.CC<-na.omit(db_follicle)
dim(db.CC)
```

```
## [1] 41  7
```

```
mean(db.CC$Day0)
```

```
## [1] 91.08129
```

```
sd(db.CC$Day0)
```

```
## [1] 31.47671
```

Question 2

For each follicle the diameter was measured at day 0,2,4,6,8.

1. Are the data in a wide or long format?
2. Convert data from wide to long (or viceversa).

3. How many rows would we expect for each patient? Is it correct? (You can use the command *table(db$Patient)*)

Data are in a wide format because we have one row for each follicle and several columns to indicate the measurement at different timepoints. We can transform it in a long format using the function ***reshape****.*

```
db_long<-reshape(db_follicle, direction="long",
                 idvar="Number",
                 varying=c("Day0","Day2","Day4","Day6","Day8"),
                 timevar="Day",v.names=c("diameter"))

db_long$Day<-factor(db_long$Day, levels=1:5, labels=c("0","2","4","6","8"))


table(db_long$Patient)
```

```
##
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14
## 25 25 25 25 25 25 25 25 25 25 25 25 25 25
```

Question 3

We are interested in the follicle growth over time. We can calculate the diameter difference from time 0 at each time point:

1. Create a data.frame with Number (follicle ID) and the diameter at Day 0.
2. Rename the variable of diameter into *diameter0*
3. Merge this data.frame and the long format of your data set by *Number*
4. Create a new variable for the difference of diameter at each time point.

```
day0<-subset(db_long, Day==0)
day0<-day0[, c("Number","diameter")]
colnames(day0)<-c("Number","diameter0")

db_join=merge(db_long,day0, by="Number")

db_join$diam.change<-db_join$diameter-db_join$diameter0

head(db_join)
```

```
##   Number Patient Day diameter diameter0 diam.change
## 1      1       1   0 101.4590  101.4590      0.0000
## 2      1       1   8 161.5000  101.4590     60.0410
## 3      1       1   6 160.2315  101.4590     58.7725
## 4      1       1   4 135.1980  101.4590     33.7390
## 5      1       1   2 112.1605  101.4590     10.7015
## 6      2       1   2 141.3770   89.8315     51.5455
```

Question 4

Descriptive at baseline (Day0).

1. Merge the two data set: long version of *follicle* and *patient* to have baseline characteristics in one data.frame
2. Check if the number of observation for each Patient is correct (use *table()*)
3. Print min, max, median and standard deviation of diameter at Day0
4. Plot the histogram for the density of diameter at Day0

```
db_all<-merge(db_join, db_pat, by="Patient")
table(db_all$Patient)
```

```
##
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14
## 25 25 25 25 25 25 25 25 25 25 25 25 25 25
```

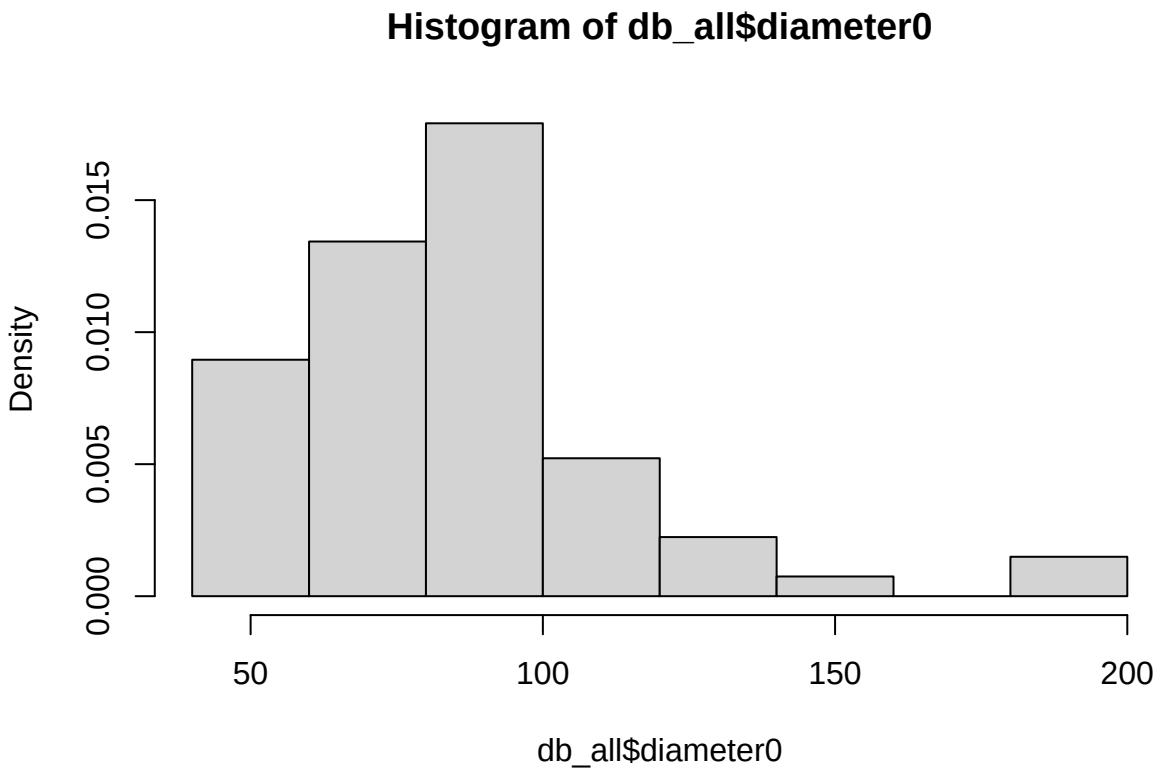
```
summary(db_all$diameter0)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   48.32   62.53   83.91   84.70   96.26  194.47      15
```

```
sd(db_all$diameter0, na.rm=TRUE)
```

```
## [1] 27.51564
```

```
hist(db_all$diameter0, prob=TRUE)
```



Question 5

Descriptive at Day 8

1. Calculate the average difference in diameter after 8 days.
2. Plot the histogram for the density of the difference in diameter after 8 days
3. Create the log-transformed variable for the difference in diameter at day 8.
4. Plot side-by-side the previous histogram and the histogram for the log-transformed variable.
5. Calculate mean and standard deviation for the difference in diameter after 8 days in each treatment group.
6. Create a boxplot for the difference in Diameter after 8 days by treatment group

```
tapply(db_all$diam.change,db_all$Day, mean, na.rm=TRUE)
```

```
##        0        2        4        6        8
```

```
##  0.00000 21.47157 40.04552 49.75852 55.16035
```

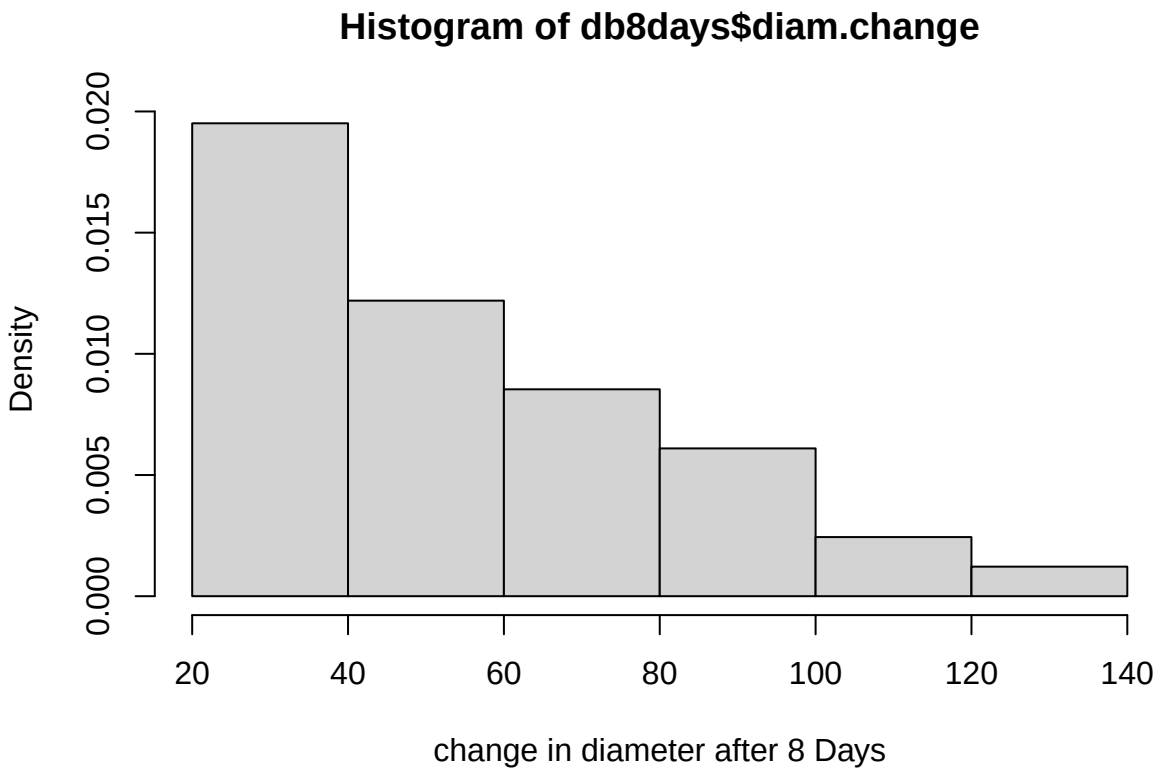
```
db8days<-subset(db_all,Day==8)
mean(db8days$diam.change, na.rm=TRUE)
```

```
## [1] 55.16035
```

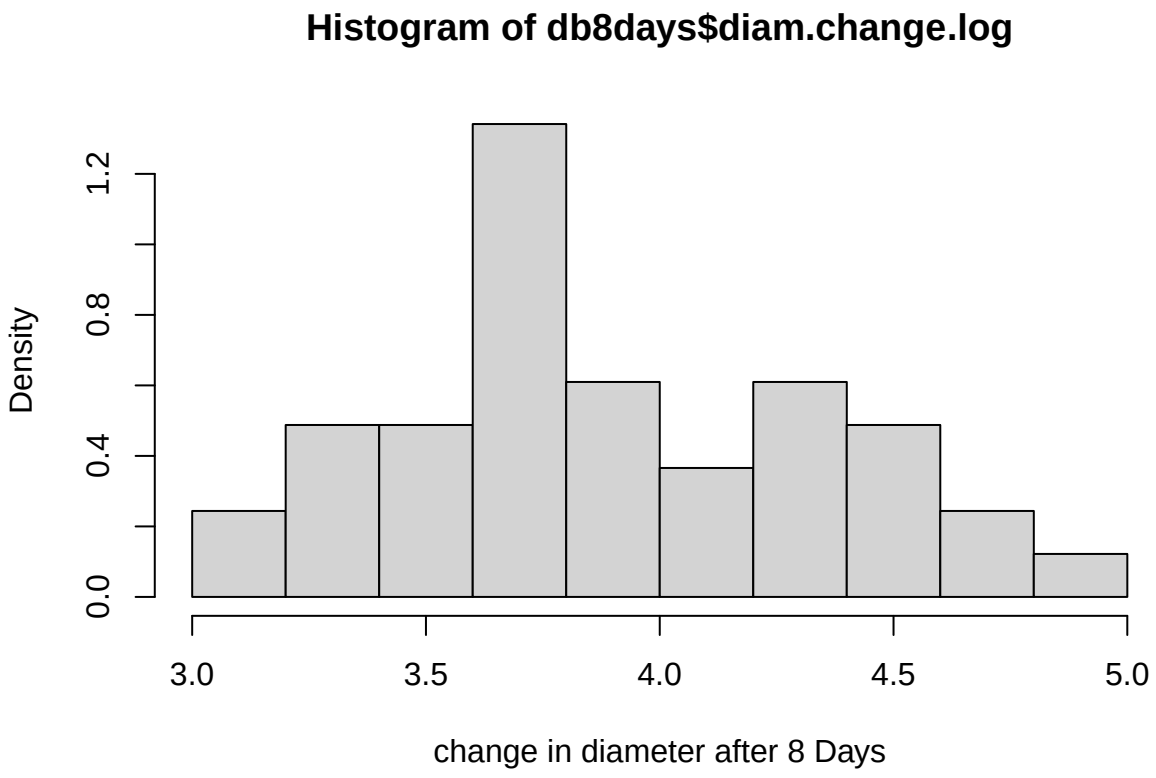
```
tapply(db8days$diam.change,db8days$Treatment, mean, na.rm=TRUE)
```

```
##      FBS      hPL      HSA
## 51.43953 84.88800 46.54333
```

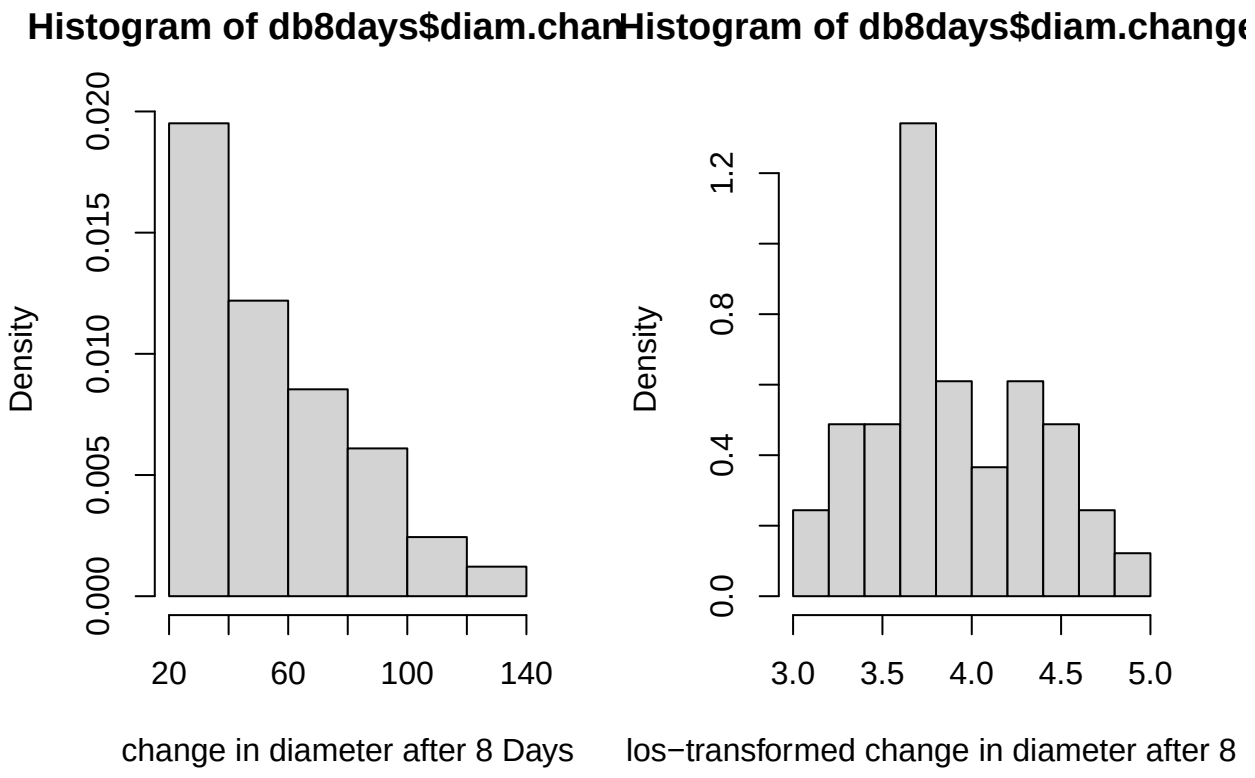
```
hist(db8days$diam.change, prob=TRUE, xlab="change in diameter after 8 Days")
```




```
db8days$diam.change.log<-log(db8days$diam.change)
hist(db8days$diam.change.log, prob=TRUE, xlab="change in diameter after 8 Days")
```




```
par(mfrow=c(1,2))
hist(db8days$diam.change, prob=TRUE, xlab="change in diameter after 8 Days")
hist(db8days$diam.change.log, prob=TRUE, xlab="los-transformed change in diameter after 8 Days")
```



```
boxplot(diam.change ~ Treatment, db8days,
        ylab="change in diameter after 8 Days")
```