Exercise 2

In [52]:

```
?ChickWeight
```

ChickWeight {datasets}   R Documentation

Weight versus age of chicks on different diets

Description

The `ChickWeight` data frame has 578 rows and 4 columns from an experiment on the effect of diet on early growth of chicks.

Usage

`ChickWeight`

Format

An object of class `c("nfnGroupedData", "nfGroupedData", "groupedData", "data.frame")` containing the following columns:

weight   a numeric vector giving the body weight of the chick (gm).

Time   a numeric vector giving the number of days since birth when the measurement was made.

Chick   an ordered factor with levels `18 < ... < 48` giving a unique identifier for the chick. The ordering of the levels groups chicks on the same diet together and orders them according to their final weight (lightest to heaviest) within diet.

Diet   a factor with levels 1, ..., 4 indicating which experimental diet the chick received.

Details

The body weights of the chicks were measured at birth and every second day thereafter until day 20. They were also measured on day 21. There were four groups on chicks on different protein diets.

This dataset was originally part of package `nlme`, and that has methods (including for `[`, `as.data.frame`, `plot` and `print`) for its grouped-data classes.

Source

Crowder, M. and Hand, D. (1990), *Analysis of Repeated Measures*, Chapman and Hall (example 5.3)

Hand, D. and Crowder, M. (1996), *Practical Longitudinal Data Analysis*, Chapman and Hall (table A.2)

Pinheiro, J. C. and Bates, D. M. (2000) *Mixed-effects Models in S and S-PLUS*, Springer.

See Also

`SSlogis` for models fitted to this dataset.

Examples

```
require(graphics)
coplot(weight ~ Time | Chick, data = ChickWeight,
       type = "b", show.given = FALSE)
```

[Package *datasets* version 4.0.3]

In [53]:
```
str(ChickWeight)
```

```
Classes 'nfnGroupedData', 'nfGroupedData', 'groupedData' and 'data.frame':
578 obs. of  4 variables:
 $ weight: num  42 51 59 64 76 93 106 125 149 171 ...
 $ Time  : num  0 2 4 6 8 10 12 14 16 18 ...
 $ Chick : Ord.factor w/ 50 levels "18"<"16"<"15"<..: 15 15 15 15 15 15 15 15 15
15 ...
 $ Diet  : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 1 ...
 - attr(*, "formula")=Class 'formula'  language weight ~ Time | Chick
  .. ..- attr(*, ".Environment")=<environment: R_EmptyEnv>
 - attr(*, "outer")=Class 'formula'  language ~Diet
  .. ..- attr(*, ".Environment")=<environment: R_EmptyEnv>
 - attr(*, "labels")=List of 2
  ..$ x: chr "Time"
  ..$ y: chr "Body weight"
 - attr(*, "units")=List of 2
  ..$ x: chr "(days)"
  ..$ y: chr "(gm)"
```

In [8]:
```
summary(ChickWeight)
```

```
     weight            Time          Chick        Diet
 Min.   : 35.0   Min.   : 0.00   13     : 12   1:220
 1st Qu.: 63.0   1st Qu.: 4.00   9      : 12   2:120
 Median :103.0   Median :10.00   20     : 12   3:120
 Mean   :121.8   Mean   :10.72   10     : 12   4:118
 3rd Qu.:163.8   3rd Qu.:16.00   17     : 12
 Max.   :373.0   Max.   :21.00   19     : 12
                                 (Other):506
```

The four variables / columns are weight, Time, Chick and Diet.

In [9]:
```
dim(ChickWeight)
```

578 · 4

The first number in the dim(ChickWeight) command referes to the number of rows in the ChickWeight dataframe. This means that there is a total of 578 records done in the experiment. There are 578 unique rows where each row has a unique combinations of a chick with a specific diet at one time interval.

Exercise 3

In [17]:
```
summary(ChickWeight$weight)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   35.0    63.0   103.0   121.8   163.8   373.0
```

The summary command gives us the high level statistics of the weight column in the ChickWeight dataframe.

In [18]:
```
head(ChickWeight$weight)
```

42 · 51 · 59 · 64 · 76 · 93

The head command gives us the first 5 values of the weight column in the ChickWeight dataframe.

In [19]:
```
mean(ChickWeight$weight)
```

121.818339100346

The mean command gives us the high average of the weight column in the ChickWeight dataframe.

In [20]:
```
myChkWts <- ChickWeight$weight
```

The command above saves the weight column from the ChickWeight into a variable called myChkWts

In [22]:
```
quantile(myChkWts, 0.50)
```

50%: 103

The quantile command above returns the 50% quantile (the median) of the myChkWts variable. I.e., it returns the median weight of the chicks dataframe.

Exercise 4

In [23]:
```
myChkWts <- ChickWeight$weight
```

```
In [35]:   quantile(myChkWts, .025)
           quantile(myChkWts, .975)
```
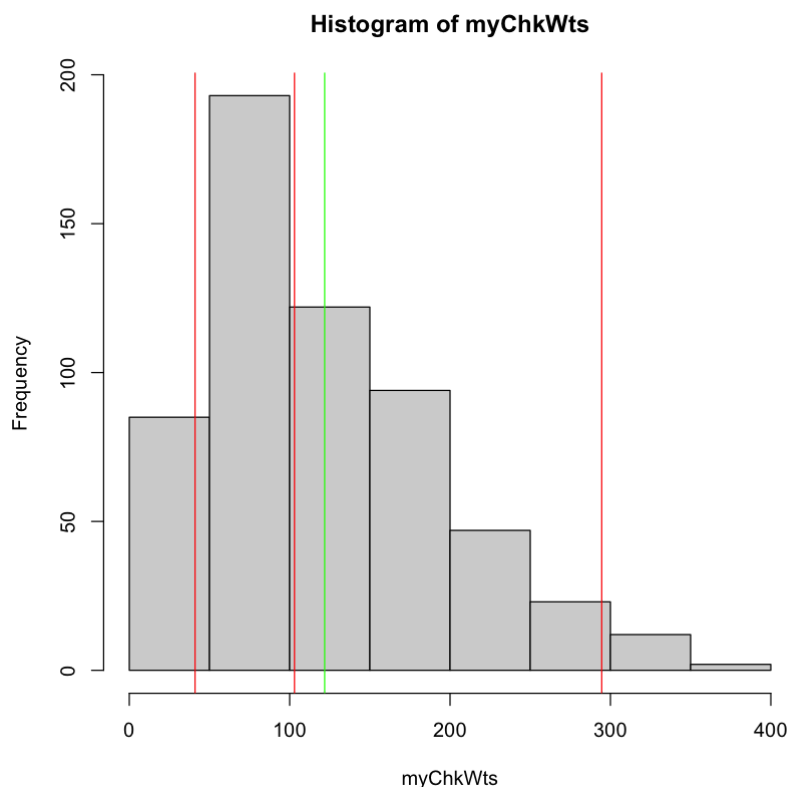
2.5%: 41

97.5%: 294.575

```
In [34]:   summary(myChkWts)
```

```
   Min.  1st Qu.  Median    Mean 3rd Qu.    Max.
   35.0     63.0   103.0   121.8   163.8   373.0
```

```
In [31]:   hist(myChkWts)
           abline(v=quantile(myChkWts, .975), col="red")
           abline(v=quantile(myChkWts, .025), col="red")
           abline(v=mean(myChkWts), col="green")
           abline(v=quantile(myChkWts, .5), col="red")
```


Histogram of myChkWts

The variable myChWts looks like a poisson distribution as most of the weights fall into the 50-100 bucket, and it has a much longer tail on the right side of the median. The median is 103 which means 50% of the weights lie before 103. The mean is 121.8 which means on average the weight of a chick is about 121.8. The 2.5% quantile is 41 and the 97.5% quatile is 294.575 which means that 95% of the data lies between these two points.

Exercise 5

```
In [48]:   chkWtsSamMeanDist <- replicate(n=1000, mean(sample(myChkWts, size=11, replace=TR
```

In [49]:
```
hist(chkWtsSamMeanDist)
abline(v=quantile(chkWtsSamMeanDist, .975), col="red")
abline(v=quantile(chkWtsSamMeanDist, .025), col="red")
```

Histogram of chkWtsSamMeanDist
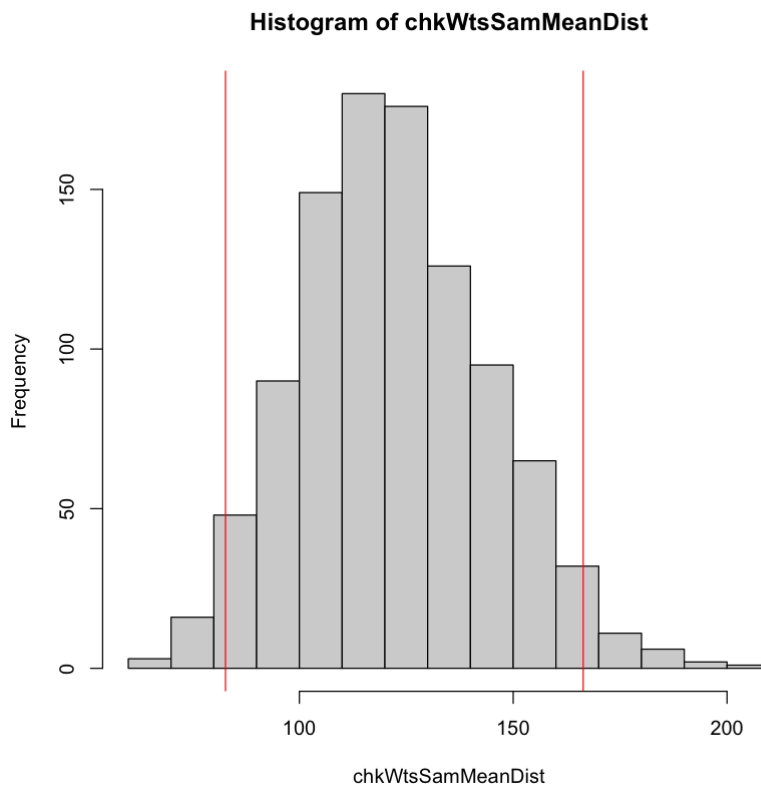
Exercise 6

The main difference between a distribution of raw data and the distribution of a sampling of means is that the distribution of raw data will not always be a normal distribution while the sampling of means will always be a normal distribution as the number of means increases. The sampling of means returns a description of how the mean of the raw data is distributed.

The 2.5% and the 97.5% quantiles are very different from each other as the sampling of means centers around the raw input mean while the raw data quantiles center around the median of the data.

Exercise 7

In [46]:
```
chkWtsSamMeanDist_n100 <- replicate(n=1000, mean(sample(myChkWts, size=100, repl
```

In [51]:
```
mean(chkWtsSamMeanDist)
```
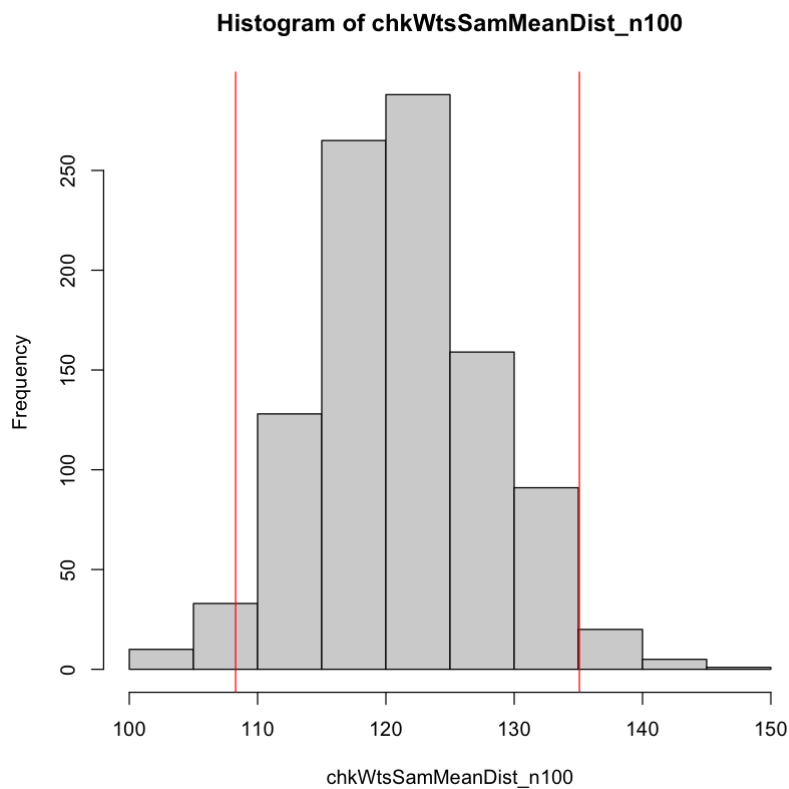
122.225272727273

In [50]:
```
mean(chkWtsSamMeanDist_n100)
```

121.30223

In [47]:
```r
hist(chkWtsSamMeanDist_n100)
abline(v=quantile(chkWtsSamMeanDist_n100, .975), col="red")
abline(v=quantile(chkWtsSamMeanDist_n100, .025), col="red")
```


Histogram of chkWtsSamMeanDist_n100

Even though both of the distributions have a center close to 120, the sampling distribution with n=100 has a higher confidence since the sample size is bigger. This means that the 2.5% and 97.5% quantiles will be more tightly bound.