Advanced Data Analysis
Haoyang Chen | hc2812 | Assignment 1

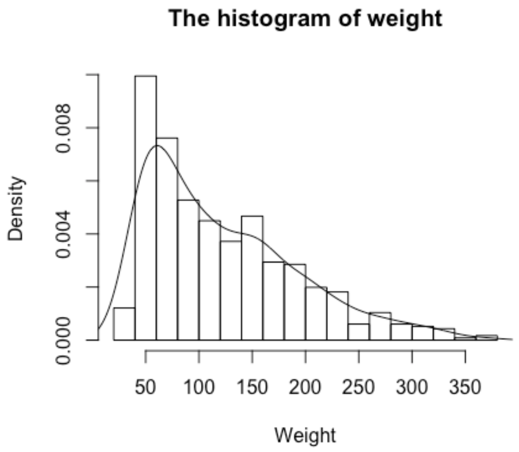
1.

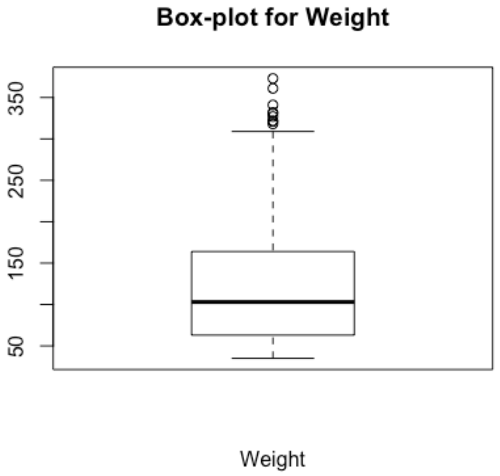
a). Perform EDA on the variable *weight*

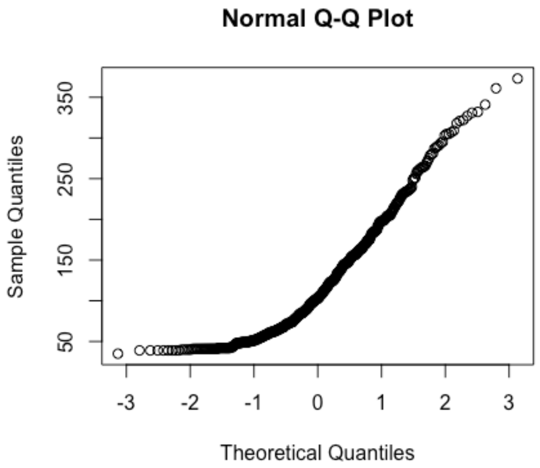
Variable *weight* does not have missing value.

Descriptive Statistics of *weight*

Mean	Median	Sample Variance	1st Quantile	3rd Quantile	IQR
121.82	103	5051.22	63.0	163.8	100.75







b) Estimate the bias associated with the sample standard deviation and median of *weight*

(1) Jackknife:
Sample Standard Deviation:

```
> jackknife(sd, weight)
[1] -0.03646702
```

(2) Median

```
> jackknife(median, weight)
[1] 0
```

(2) Bootstrap (100 times of sampling):
Sample Standard Deviation:

```
> bias.bootstrap(weight, sd)
[1] -0.1024876
```

(2) Median

```
> bias.bootstrap(weight, median)
[1] 0.85
```

2. Discuss the pros and cons of observational and controlled studies, giving illustrative examples.

Observational Studies:

Pros:

- Observational studies are easy to be conducted.
- Observation can help round out research by offering a real-world aspect to a hypothesis. It offers a better description and is less hypothetical than other methods.
- Observation allows you to create and observe actual situations and validates with actual result.

Cons:

- Observational studies can include a high degree of researcher bias.
- Observational studies could not control the possible confounding variables
- Observational studies rely on the interpretation of observation. And it can be difficult to create an accurate analysis from observation alone.

Example:
An observational research was conducted to study the association between the risk of lung cancer and smoking. We could get the behaviors. It offered the real-world picture. However, we could not make the conclusion of the association because there might exist other elements that would influence the result.

Controlled Studies:

Pros:

- Controlled studies can be very effective in attributing impact or causality to a program or a treatment.

Cons:

- Achieving true random controlled studies is extremely rare.
- It costs lots of money, resources and time to carry out an controlled studies.

Example:
In clinical trials, researchers can conduct controlled studies to find the efficacy of drugs for specific variables, for example, male or female, ethnic, etc. This method is easy to get rid of confounding variables than observational study.

Code:

```
# This is assignment 1
data("ChickWeight")
# a) EDA for variable weight
weight <- ChickWeight$weight

# Measures of Location: Mean, Median
mean(weight)
median(weight)

# Measures of dispersion: Sample Variance
var(weight)

# Histograms
hist(weight, probability = TRUE, breaks = 15, xlab = "Weight", ylab =
"Density", main = "The histogram of weight")
lines(density(weight))

# box-plot
boxplot(weight, main = "Box-plot for Weight", xlab = 'Weight')
# Q-Q plot
qqnorm(weight)

# b) Estimate the bias associated with  the sample standard deviation and
median of weight

# jackknife
jackknife <- function(func, data){
    estimator.sum <- 0
    for(i in 1:length(data)){
        estimator.sum <- estimator.sum + func(data[-i])
    }
    estimator.bias <- (length(data) - 1) * (estimator.sum / length(data) -
func(data))
    return(estimator.bias)
}

jackknife(sd, weight)
jackknife(median, weight)

# Bootstrap
library(bootstrap)
bias.bootstrap <- function(data, func){
    estimator.sample <- func(data)
    estimator.bootstrap <- bootstrap(data, nboot = 100, theta =
func)$thetastar
    return(mean(estimator.bootstrap - estimator.sample))
}

bias.bootstrap(weight, sd)
bias.bootstrap(weight, median)
```