Name: Huajie Chen
Matrikelnr.: 4199962

1. Descriptive Statistics

a)

$$\mathrm{E(X)} = \sum_{x \in \omega X} xPr(X = x)$$

---------->

$$\mathrm{E(X)} = \frac{n+1}{2}$$

b)

Normally distributed random vector of length 10

[1] 0.44528058 0.11423171 -0.61318183 -0.01041157 0.84263224 0.69371223 -0.40621217 -0.30956160 -1.10751399

[10] -0.30625571

empirical cumulative distribution function

-2.11442215 -1.37190620 -0.99484264 -0.69179349 -0.23813647 -0.17157899 -0.05853650 -0.01994341 0.87505357 0.90993548

the 0.75 quantile

-0.02959168

the 0.95 quantile

0.89423862

c)

mean = -0.3876171 median = -0.2048577

mean − median = -0.1827593

> y <- rnorm(100) > z <- rnorm(10000)

> mean(y)-median(y) > mean(z)-median(z)

[1] -0.07626176 [1] -0.001065337

It is clear that the longer/shorter the length of the vector is, the closer/further the distance between mean and median is. Because the longer the length of the vector is, the higher the density of the values in the vector is. Therefore, the mean is closer to median.

d)

the IQR for normally distributed data $\mathbf{N} = (\boldsymbol{\mu}, \boldsymbol{\sigma^2})$

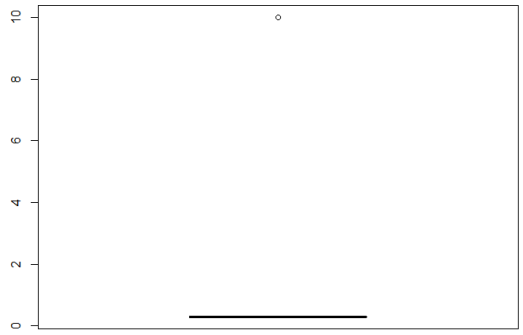
$$Q_1 = \mu - \sigma z = \mu - 0.67\sigma$$

$$Q_3 = \mu + \sigma z = \mu + 0.67\sigma$$

2. Boxplots and their interpretation

a)

In the first boxplot, there are only 2 whiskers. The upper whisker is at around 10, while the lower whisker and median line are combined together at around 0. And there is no outliers, Q1 or Q3. The values should be mostly the same at around 0 and there should be a single value at 10, like the plot below. (I cannot manage to plot a boxplot with only one single whisker)



In the second boxplot, there are no whisker. Q3 is at 3, Q1 is at 1, and median is at 2. Therefore, it should be a binomial distribution (discrete distribution).

b)

A data element x is called an outlier if $x \geq a \cdot IQR + Q3$ or $x \leq Q1 - a \cdot IQR$, where

IQR = interquartile range (Q3-Q1), measure of statistical dispersion

a = 3, a true outlier a = 1.5, a probable outlier.

The reason why x is called a true/probable outlier when a=3 or 1.5 is because John Tukey proposed this test, where k=1.5 indicates an "outlier", and k=3 indicates data that is "far out" (*Tukey, John W (1977). Exploratory Data Analysis. Addison-Wesley. ISBN 0-201-07616-0. OCLC 3058187.)