Assignment 0

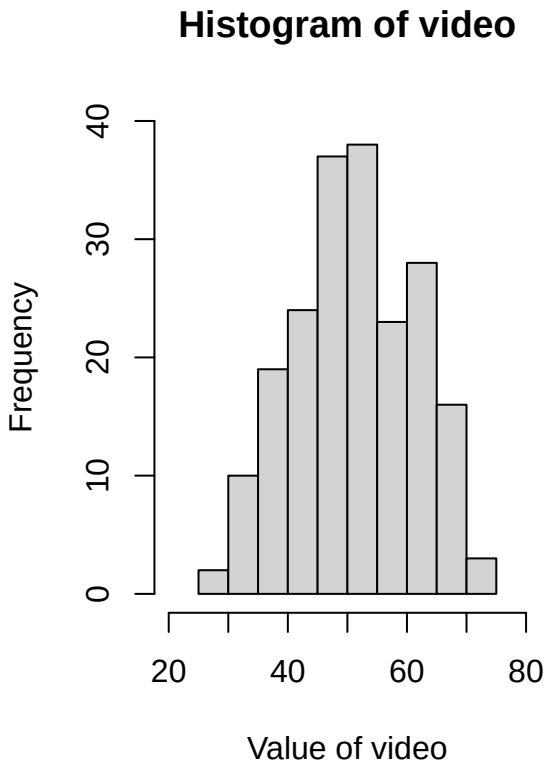
Example Author

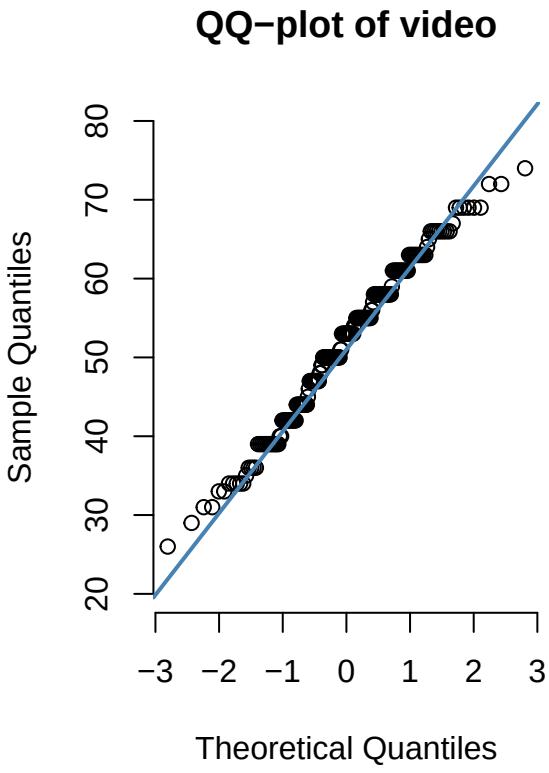
Example Date

Exercise 1

a)

From the given dataset *icecream.csv* the mean μ is 51.85





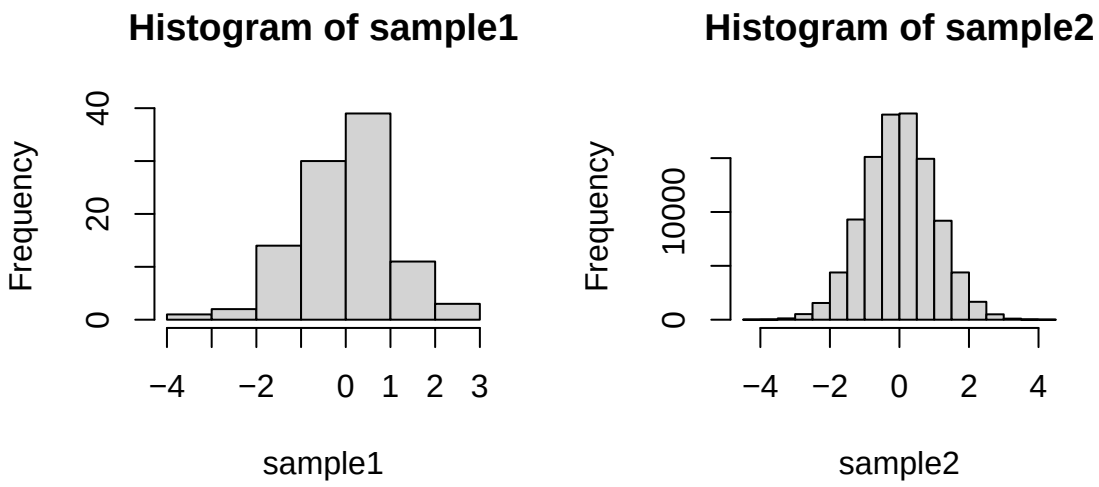
When analyzing the histogram and QQ-plot above the assumption of a normal distribution can be assumed since the histogram is bell shaped and the QQ-plot follows a straight line in comparison to the blue reference line.

The 97% Confidence Interval for the Mean μ is 51.85 has a lower bound of 50.3307225 and an upper bound of 53.3692775

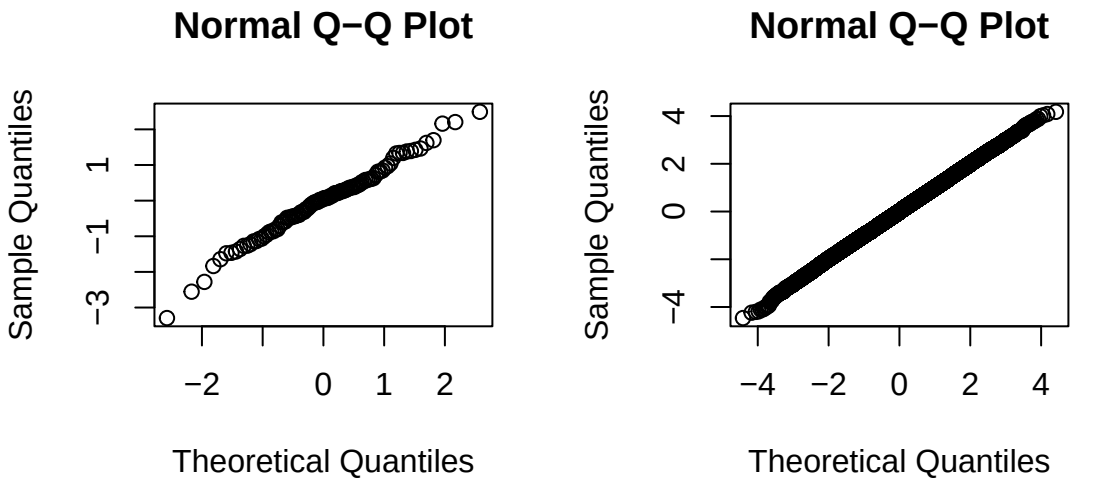
Required Sample Size: 206

We generate two samples of sizes 100 and 100000 from a standard normal distribution N(0,1) as follows: `sample1=rnorm(100)` , `sample2=rnorm(100000)`; then make histograms and QQ-plots for the both samples.

```
par(mfrow=c(1,2)); hist(sample1); hist(sample2) # two histograms next to each other
```




```
qqnorm(sample1); qqnorm(sample2) # two QQ-plots next to each other
```



For different samples, the figures are different. The quality of the histogram and QQ-plot depend on the sample size. If it is small, the histogram varies more and the QQ-plot varies more around a straight line whereas for large samples size the histogram is very stable and close to the true density, and the QQ-plot is straight in the middle with just some variation in the corners. The values of `mean` and `sd` only influence the scales on the axes, not the straightness of the line in the QQ-plot.

Now, we compute the means and standard deviations for the both samples, and summarize the results in the table.

Parameters	Est. for sample n=100	Est. for sample n=100000
mean=0	`mean(sample1)`=-0.0260713	`mean(sample2)`=-0.0014046
sd=1	`sd(sample1)`=1.0111301	`sd(sample2)`=1.0001411

The estimated mean and standard deviation are also clearly better for the second sample. This is not surprising as the second sample is of a much bigger size, i.e., containing much more data.

b) Given Z has a standard normal distribution, we need to compute the following probabilities: P(Z<2)=`pnorm(2)`=0.9772499, P(Z>-0.5)=`1-pnorm(-0.5)`=0.6914625, P(-1<Z<2)=`pnorm(2)-pnorm(-1)`=0.8185946.

c) For Z~N(0,1), the probabilities P(Z<2)=0.9772499, P(X>-0.5)=0.6914625 and P(-1 < Z < 2)=0.8185946 from b) can be estimated by using the data from a) as follows:

```
p1=sum(sample1<2)/length(sample1) # estimate of P(Z<2) for sample 1 with n=100
p2=sum(sample2<2)/length(sample2) # estimate of P(Z<2) for sample 2 with n=100000
p3=sum(sample1>-0.5)/length(sample1) # estimates of P(Z>-0.5) for sample 1
p4=sum(sample2>-0.5)/length(sample2) # estimates of P(Z>-0.5) for sample 2
p5=sum(sample1>-1&sample1<2)/length(sample1) # estimate of P(-1<Z<2) for sample 1
p6=sum(sample2>-1&sample2<2)/length(sample2) # estimate of P(-1<Z<2) for sample 2
c(p1,p2,p3,p4,p5,p6) # print all the estimates
## [1] 0.97000 0.97715 0.73000 0.68982 0.80000 0.81809
```

Summarize the results in the table. The 2nd and 3d columns in this table are the estimates of the corresponding theoretical probabilities from b).

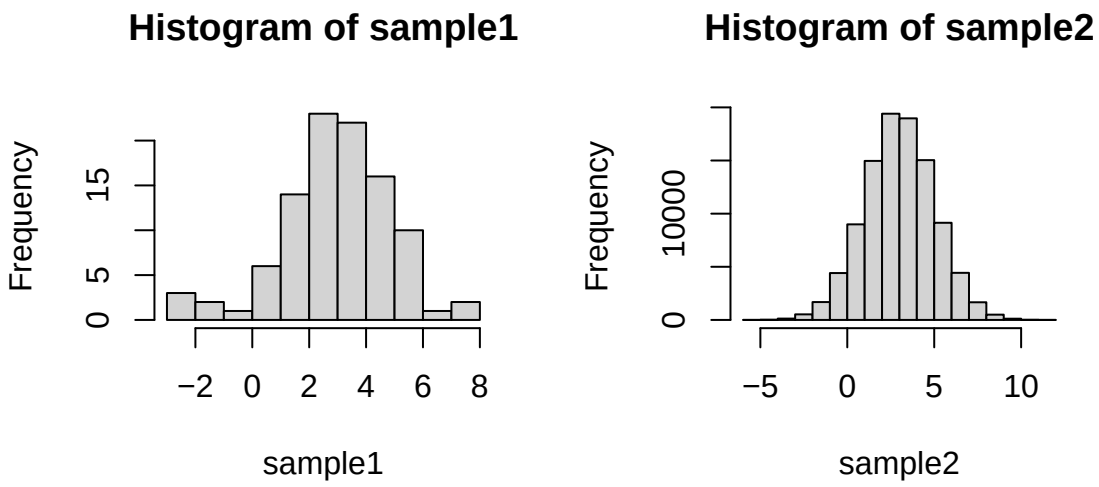
Probabilities from b)	Est. for sample n=100	Est. for sample n=100000
P(Z<2)=0.9772499	p1=0.97	p2=0.97715
P(Z>-0.5)=0.6914625	p3=0.73	p4=0.68982
P(-1<Z<2)=0.8185946	p5=0.8	p6=0.81809

The estimates based on the second sample are clearly better, because the second sample is larger.

d) As in a), we first generate the samples `sample1=rnorm(100,mean=3,sd=2)`, `sample2=rnorm(100000,3,2)`. Next, we estimate the parameters `mean` and `sd` and construct histrograms for the both samples.

Parameters	Est. for sample n=100	Est. for sample n=100000
mean=3	`mean(sample1)`=2.9636489	`mean(sample2)`=3.0002742
sd=2	`sd(sample1)`=1.9050996	`sd(sample2)`=2.0030053

```
par(mfrow=c(1,2)); hist(sample1); hist(sample2)
```



As before, the estimates and histrogram for the second sample are better as this sample is of a larger size.

For X~N(3,4), the probabilities are now found as follows: P(X<2)=`pnorm(2,mean=3,sd=2)`=0.3085375, P(X>-0.5)=`1-pnorm(-0.5,mean=3,sd=2)`=0.9599408, P(-1<X<2)=`pnorm(2,3,2)-pnorm(-1,3,2)`=0.2857874.

The value such that 95% of the outcomes is smaller than that value is nothing else but the 95%-quantile of the distribution N(3,4), which is `qnorm(0.95,mean=3,sd=2)`=6.2897073. Notice that it can also be found via the 95%-quantile `qnorm(0.95)` of the standard normal distribution as `3+2*qnorm(0.95)`=6.2897073.

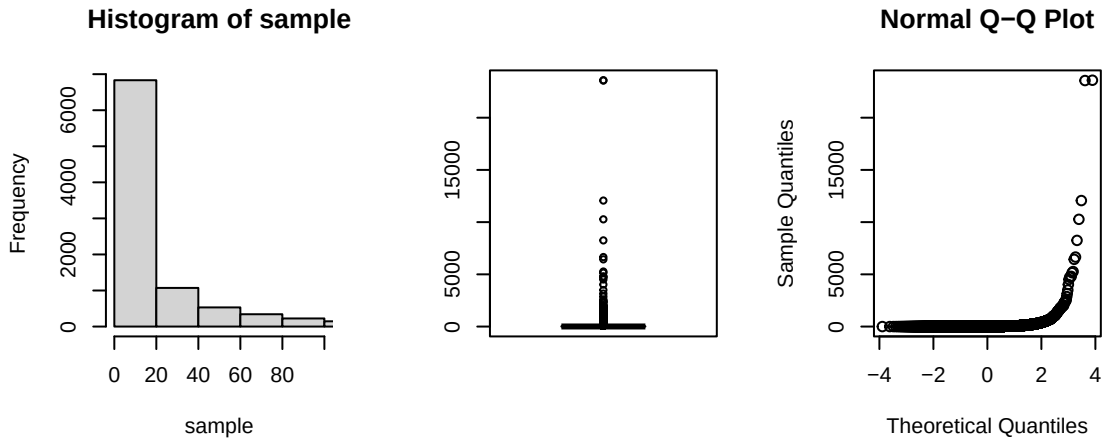
e) Any normal variable X~N(mu,sigma^2) can be generated from a standard normally distributed Z~N(0,1) as X=mu+sigma*Z. We generate in this way a sample of size 1000 from a normal distribution with `mean`=-10 and `sd`=5, and verify that the sample mean and sample standard deviation are close to the true values `mean`=-10 and `sd`=5.

```
sample=-10+5*rnorm(1000)
c(mean(sample),sd(sample)) # should be close to mean=-10 and sd=5
## [1] -10.309110   5.213198
```

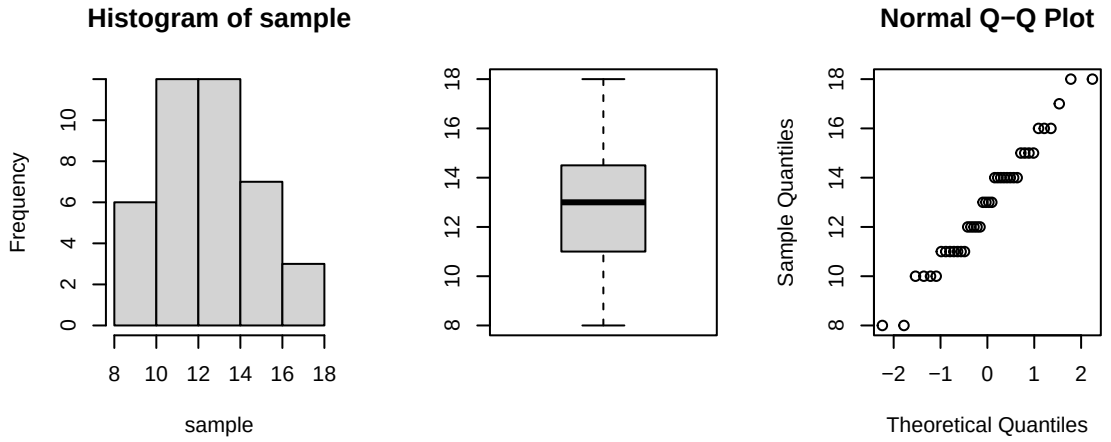
Exercise 2

We generate samples from the asked distributions and plot for each of the generated samples the histogram, boxplot and QQ-plot:

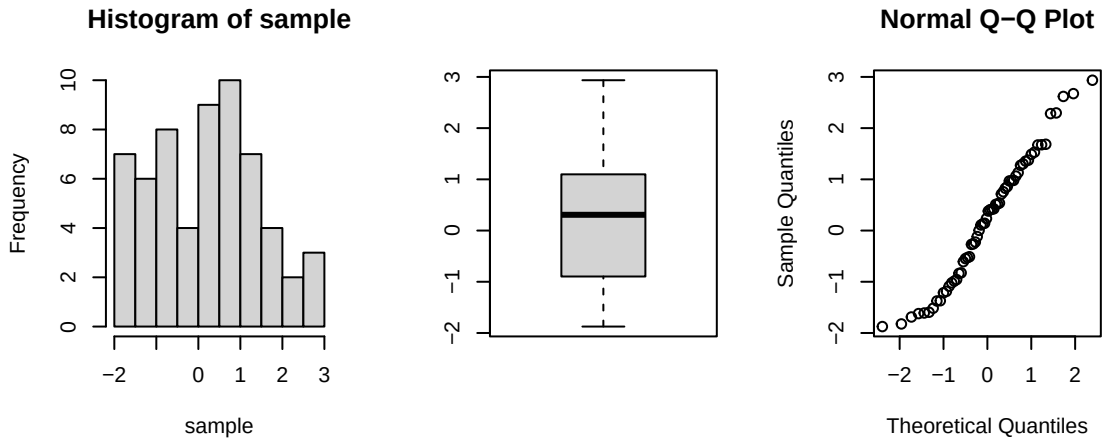
```
par(mfrow=c(1,3)) # two plots next each other
sample=rlnorm(10000,2,2) # from the lognormal distribution with mu=sigma=2
hist(sample,xlim=c(0,100),breaks=1000) # hist(sample) will not look good, why?
# to see the breaks: hist(sample,xlim=c(0,100),breaks=1000)$breaks
boxplot(sample) # a lot of outliers
qqnorm(sample) # of course, not normal
```



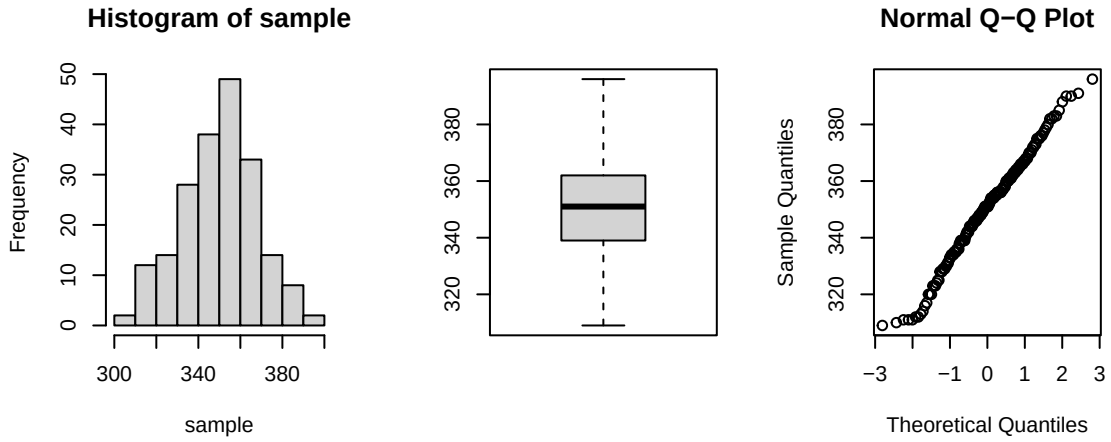

```
sample=rbinom(40,50,0.25) # from the binomial distribution with n=50 and p=0.25
hist(sample);boxplot(sample);qqnorm(sample) # looks like normal
```




```
sample=runif(60,-2,3) #from the uniform distribution on the interval [-2,3]
hist(sample);boxplot(sample);qqnorm(sample) # of course, not normal
```




```
sample=rpois(200,350) #from the Poisson distribution with lambda = 350
hist(sample);boxplot(sample);qqnorm(sample) # looks like normal
```



All but lognormal are symmetric (possibly not around zero), binomial and Poisson look like normal. Small sample sizes (10,40,60) show noise. Histograms are more stable and give better approximation of the true density for sufficiently large sample sizes.

Exercise 3

a)

b)

Exercise 4

a)-b) In these both cases the null hypothesis H0 holds because `mu=nu=180`.

a)

b)

c)

d) The null hypothesis H0 holds in a) and b) as `mu=nu`. Under H0, p-values are distributed uniformly on [0,1]. Hence the events {p<0.05} and {p<0.1} should occur approximately in 5% and 10% of cases respectively, and histograms of p-values should be close to uniform on [0,1]. In b) the approximations should be better because the variance is smaller.

In c) H0 does not hold (because `mu>nu`), so p-values are not uniformly distributed and `mean(p<0.05)` gives approximately the values of the power function at point `mu-nu=180-175=5`, which should approach 1 for a good test.

All these claims are confirmed by the simulations results in a), b) and c).