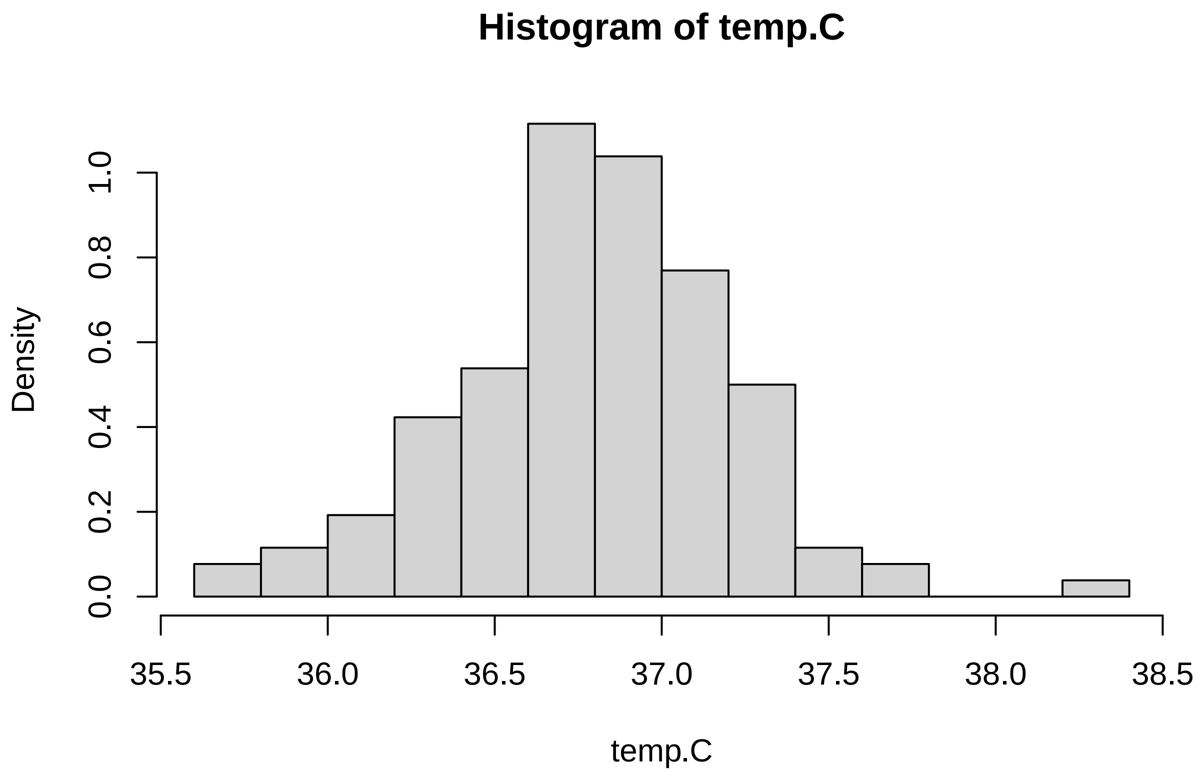
22112023_Stat_Learning

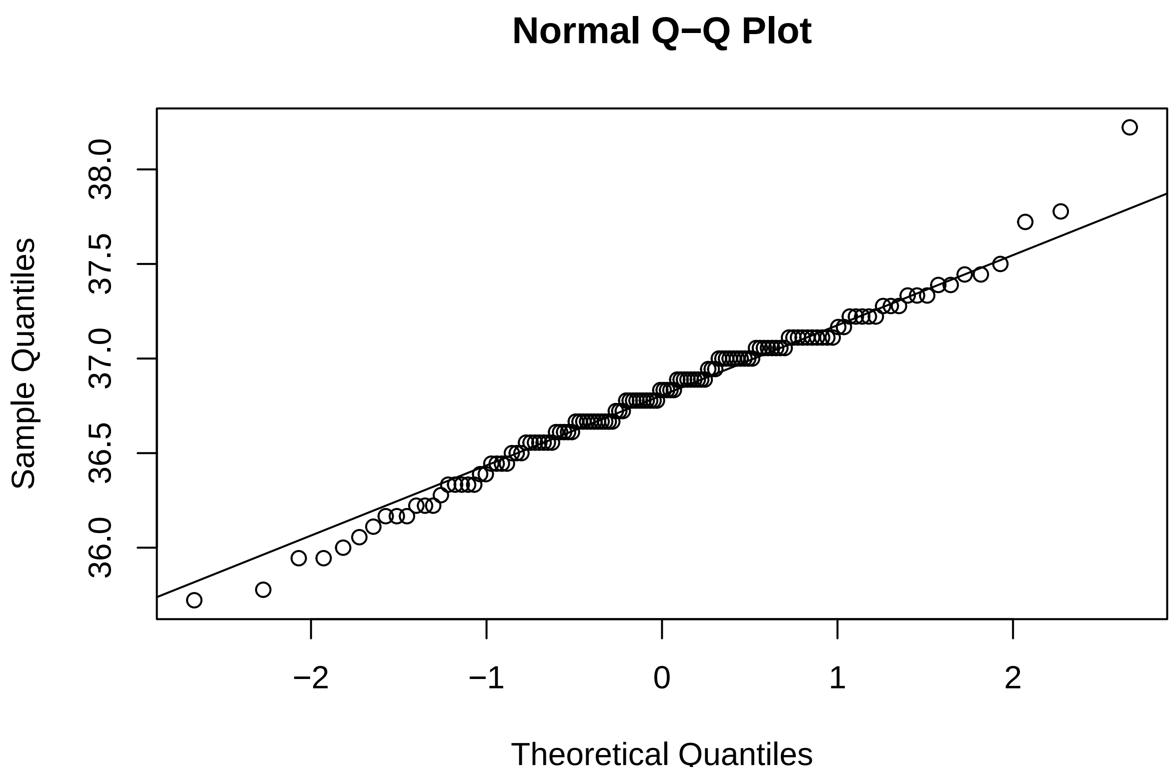
Mattia G.

2023-11-24

```r
###################################
# body temperature data
#################################
setwd("/Users/mattiagugole/Downloads/normtemp-data")
temp.data <- read.table("normtemp.txt", head=TRUE)
attach(temp.data)
temp.C <- (temperature-32)*5/9


# check normality

hist(temp.C, prob=TRUE, col="lightgray")
```




```r
qqnorm(temp.C)
qqline(temp.C)
```




```r
##################################
# Conf. Int. for mu with sigma known
#################################

# we assume X~N(mu, sigma^2) with sigma=0.45 known

x.bar <- mean(temp.C)
se <- 0.45/sqrt(length(temp.C))

# CI confidence level 95%

mu.lower <- x.bar - qnorm(0.975)*se
mu.upper <- x.bar + qnorm(0.975)*se

mu.lower
```

```
## [1] 36.72777
```

```r
mu.upper
```

```
## [1] 36.88248
```

```r
# more compact form

CI <- x.bar+c(-1, +1)*qnorm(0.975)*se
CI
```

```
## [1] 36.72777 36.88248
```

```r
# CI confidence level 1-alpha

alpha <- 0.05
# alpha <- 0.01
```

```
mu.lower <- x.bar - qnorm(1-alpha/2)*se
mu.upper <- x.bar + qnorm(1-alpha/2)*se

mu.lower
```

```
## [1] 36.72777
```

```
mu.upper
```

```
## [1] 36.88248
```

```
##################################
# Conf. Int. for mu with sigma unknown
#################################

# we assume X~N(mu, sigma^2)

n <- length(temp.C)
x.bar <- mean(temp.C)
se <- sd(temp.C)/sqrt(n)

alpha <- 0.05

mu.lower <- x.bar - qt(1-alpha/2, df=n-1)*se
mu.upper <- x.bar + qt(1-alpha/2, df=n-1)*se

mu.lower
```

```
## [1] 36.73445
```

```
mu.upper
```

```
## [1] 36.87581
```

```
# use of the function t.test()

t.test(temp.C, conf.level=0.95)
```

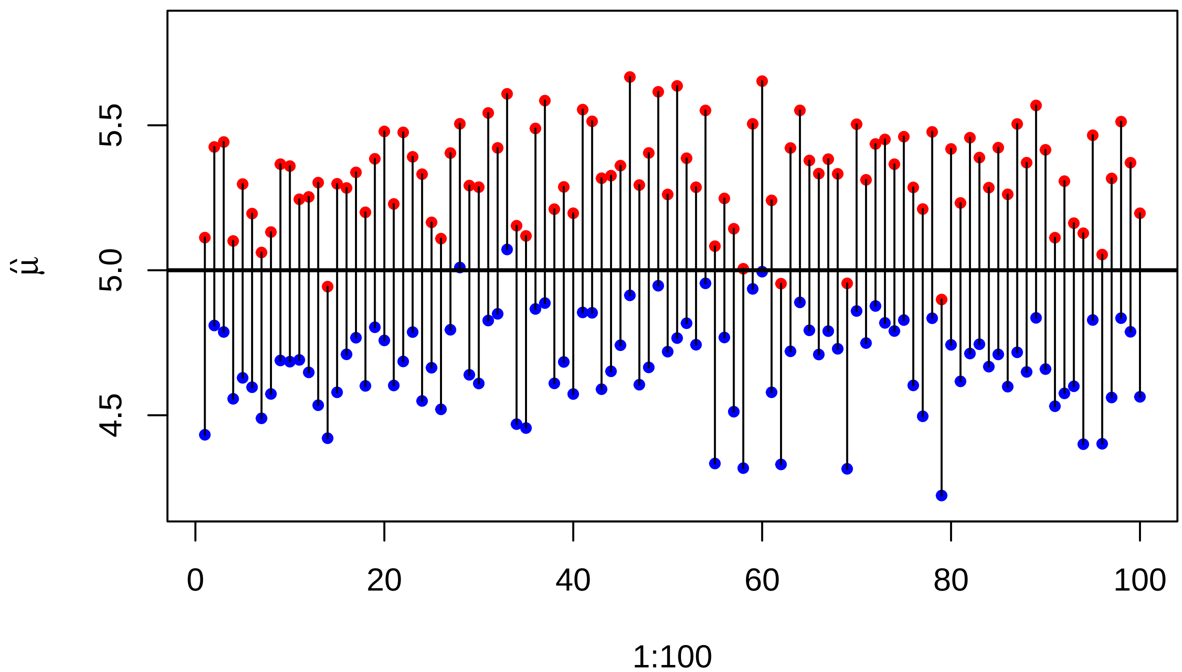
```
##
##  One Sample t-test
##
## data:  temp.C
## t = 1030.2, df = 129, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  36.73445 36.87581
## sample estimates:
## mean of x
##  36.80513
```

```
#############################################
# R script form slides: interpretation of CIs
#############################################

# simulate 1000 samples of size 40 from a normal distribution
# and for every sample compute a confidence interval for the mean (sigma unknown)
# the level of the interval is 0.95 (i.e. 95%)
```

```r
#
sampsize <- 40
mu.true <- 5
CI <- matrix(NA, ncol=2, nrow=1000)
for (i in 1:1000)
{
  x <- rnorm(sampsize, mu.true, 1)
  hat.mu <- mean(x)
  se <- sd(x)/sqrt(sampsize)
  mu.lower <- hat.mu - qt(0.975,sampsize-1)*se
  mu.upper <- hat.mu + qt(0.975,sampsize-1)*se
  CI[i,1] <- mu.lower
  CI[i,2] <- mu.upper
}

# compute the proportion of intervals that contain the true mean
# (should be close to the confidence level)
sum( (CI[,1] <= mu.true) & (CI[,2] >= mu.true) )/1000
```

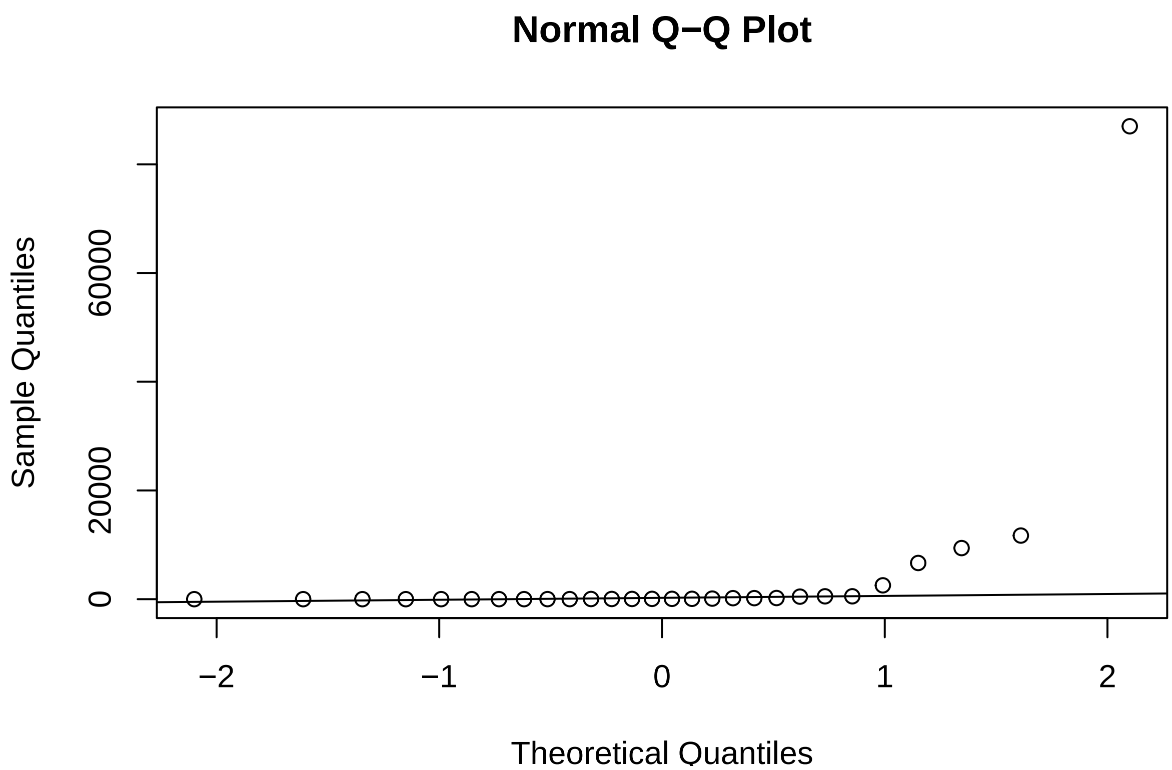
```
## [1] 0.955
```

```r
# visual representation for the fist 100 intervals
#
plot(1:100,CI[1:100,1],ylim=range(CI), ylab=expression(hat(mu)), pch=20, col="blue")
points(1:100,CI[1:100,2], pch=20, col="red")
segments(1:100,CI[1:100,2],1:100,CI[1:100,1])
abline(h=mu.true,col=1, lwd=2)
```




```r
###########################
# CI for body weight
###########################

library(MASS)
```

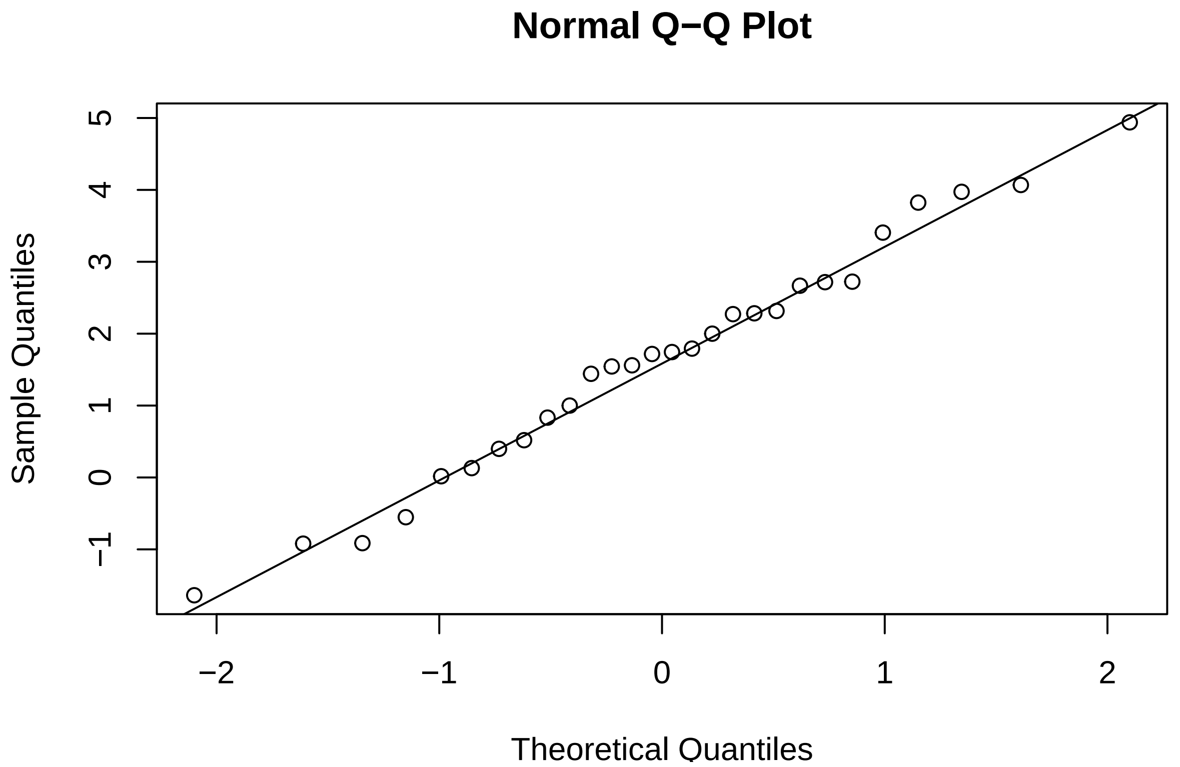
```r
data("Animals")
attach(Animals)

# body is not normally distributed
qqnorm(body)
qqline(body)
```




```r
t.test(body)
```

```
##
##  One Sample t-test
##
## data:  body
## t = 1.3737, df = 27, p-value = 0.1808
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -2112.028 10668.906
## sample estimates:
## mean of x
##  4278.439
```

```r
# log10(body) seems to be normally distributed
l.body <- log10(body)
qqnorm(l.body)
qqline(l.body)
```




```
# Conf. Int. for the mean of l.body

t.out <- t.test(l.body)
t.out$conf.int
```

```
## [1] 1.002872 2.272842
## attr(,"conf.level")
## [1] 0.95
```

```
# Conf. Int. for the geometric mean of body

10^t.out$conf.int
```

```
## [1]  10.06635 187.43142
## attr(,"conf.level")
## [1] 0.95
```

```
#####################################################################
# Exercises_01-Confidence intervals for the mean of a normal dist.pdf
#####################################################################


# Exercise 1
######################

#a
qt(0.95, df=11)
```

```
## [1] 1.795885
```

```
#b
qt(0.975, df=6)
```

```
## [1] 2.446912
```

```r
#c
qt(0.995, df=1)
```

```
## [1] 63.65674
```

```r
#d
qt(0.975, df=28)
```

```
## [1] 2.048407
```

```r
# Exercise 3
######################

# a
alpha <- (1-pt(2.776, 4))*2

round(1-alpha, 3)
```

```
## [1] 0.95
```

```r
# b
alpha <- (1-pt(2.718, 11))*2
round(1-alpha, 3)
```

```
## [1] 0.98
```

```r
# c
alpha <- (1-pt(5.841, 3))*2
round(1-alpha, 3)
```

```
## [1] 0.99
```

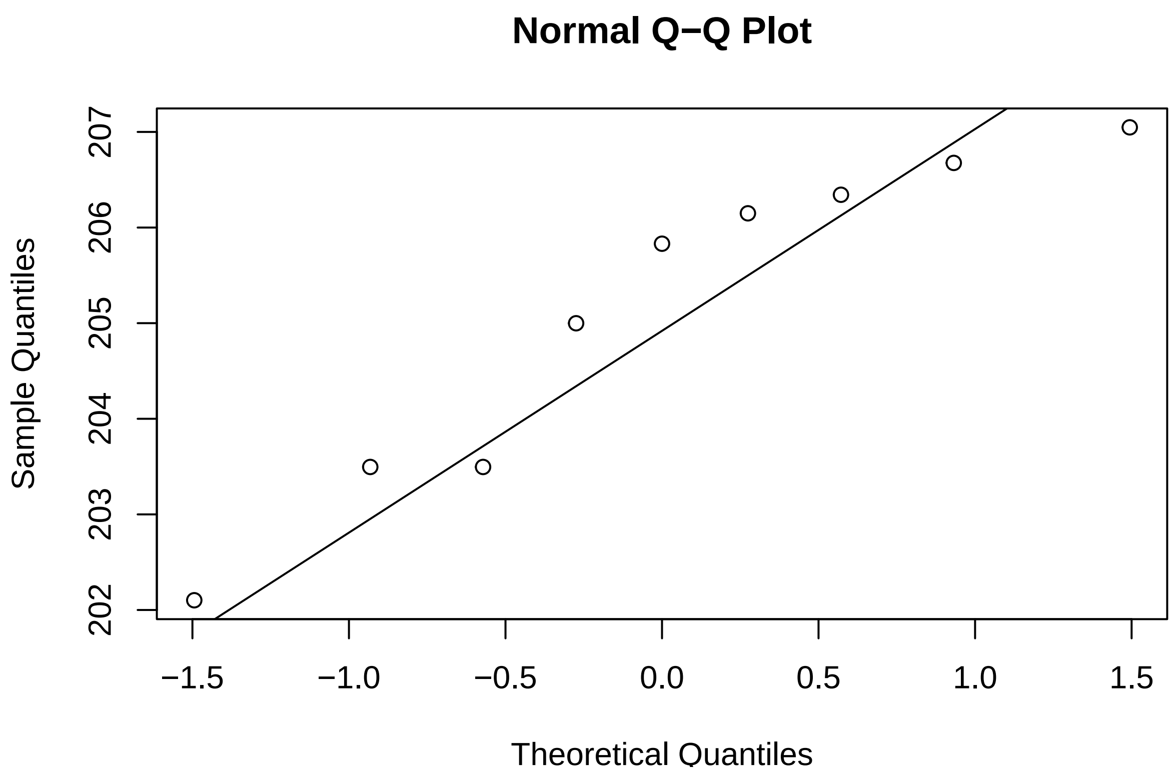
```r
# d
alpha <- (1-pt(1.325, 20))*2
round(1-alpha, 3)
```

```
## [1] 0.8
```

```r
# e
alpha <- (1-pt(1.746, 16))*2
round(1-alpha, 3)
```

```
## [1] 0.9
```

```r
# Exercise 7
######################


sample <- c(204.999, 206.149, 202.102, 207.048, 203.496, 206.343, 203.496, 206.676, 205.831)

qqnorm(sample)
qqline(sample)
```




```r
t.test(sample, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  sample
## t = 358.32, df = 8, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  203.8066 206.4468
## sample estimates:
## mean of x
##  205.1267
```

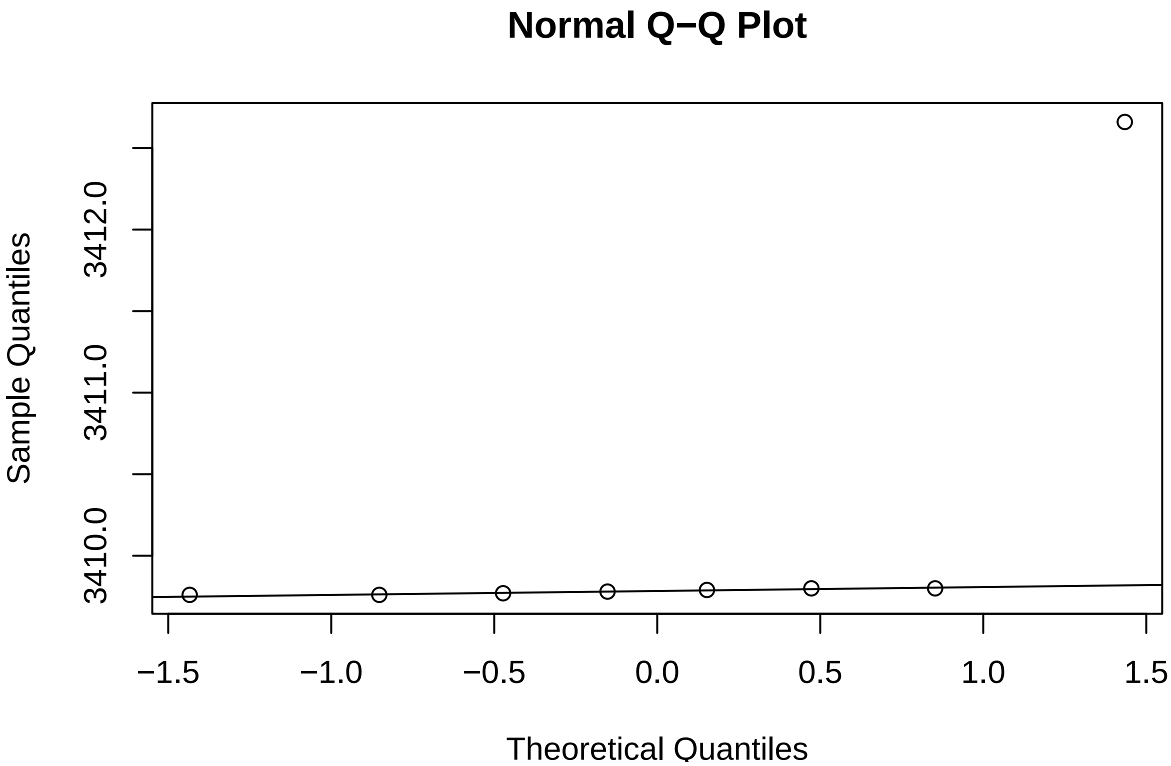
```r
# Exercise 8
######################

n     <- 8
x.bar <- 3410.14
s     <- 1.018

#a
x.bar+c(-1,1)*qt(0.975, 7)*s/sqrt(n)
```

```
## [1] 3409.289 3410.991
```

```r
#b
x.bar+c(-1,1)*qt(0.99, 7)*s/sqrt(n)
```

```
## [1] 3409.061 3411.219
```

```r
#c
sample <- c(3409.76, 3409.80, 3412.66, 3409.79, 3409.76, 3409.77, 3409.80, 3409.78)
qqnorm(sample)
qqline(sample)
```




```r
##########################################
# Hypotheses testing
##########################################

###############################
# test statistics and p-value
###############################


# null hypothesis         H_0: mu =  mu.0
# alternative hypothesis H_1: mu != mu.0

# value of mu.0 in the null hypothesis H_0
mu.0 <- 36.75


# the (unknown) true distribution is normal with
###################################################

# standard deviation
sigma <- 0.45

# and we consider two possible mean values

# a) case where H0 is true
```

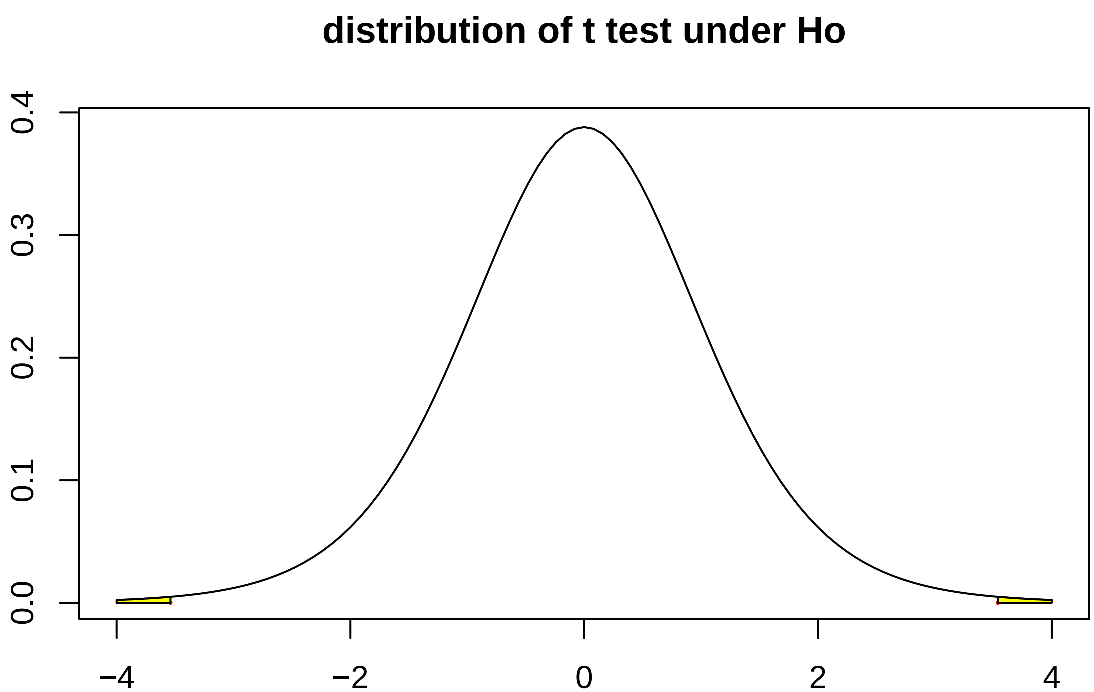
```r
mu.true <- 36.75

# case where H0 is not true
mu.true <- 36.30


# extract a sample
n <- 10
x <- rnorm(n, mean=mu.true, sd=sigma)

# compute the observed value of the test statistics
t.obs <- (mean(x)-mu.0)/(sd(x)/sqrt(n))
t.obs
```

```
## [1] -3.53939
```

```r
# represent the distribution of t statistics under H0
curve(dt(x, df=9), xlim=c(-4, 4), ylab="", xlab="", main="distribution of t test under Ho")

# represent the observed value of t statistics
lines(x=c(-abs(t.obs), -abs(t.obs)), y=c(0, dt(-abs(t.obs), n-1)), lty=3, lwd=2, col="red")
lines(x=c(abs(t.obs), abs(t.obs)), y=c(0, dt(abs(t.obs), n-1)), lty=3, lwd=2, col="red")

# color tails corresponding to p.value

# right
x <- seq(abs(t.obs), 4, length=100)
y <- dt(x, df=n-1)
polygon(c(x, max(x), abs(t.obs)), c(y, 0, 0), col="yellow")

# area under the right tail
1-pt(abs(t.obs), n-1)
```

```
## [1] 0.003160073
```

```r
# left
x <- seq(-4, -abs(t.obs),  length=100)
y <- dt(x, df=n-1)
polygon(c(x,-abs(t.obs), min(x)), c(y, 0, 0), col="yellow")
```




```
# area under the left tail
pt(-abs(t.obs), n-1)
```

```
## [1] 0.003160073
```

```
# compute p.value
p.value <- 2*pt(-abs(t.obs), df=n-1)
p.value
```

```
## [1] 0.006320146
```

```
#############################################
# test H_0:mu=36.75 in body temperature data
#############################################


temp.data <- read.table("normtemp.txt", head=T)
attach(temp.data)
```

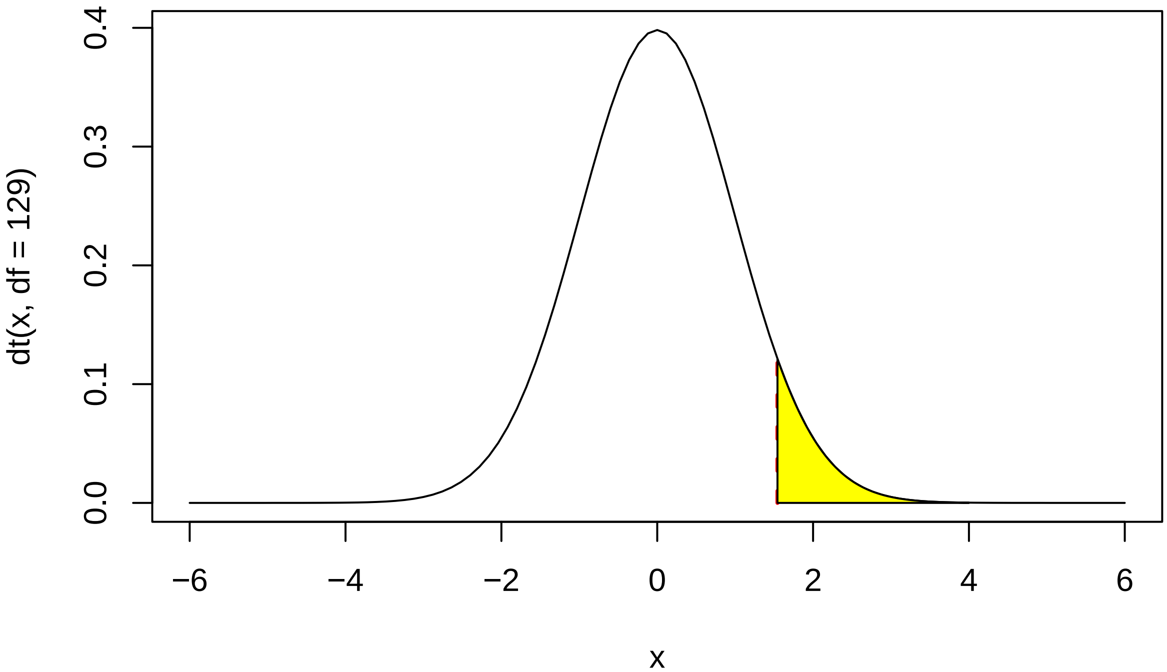
```
## The following objects are masked from temp.data (pos = 5):
##
##     gender, hr, temperature
```

```
temp.C <- (temperature-32)*5/9

x.bar <- mean(temp.C)
s <- sd(temp.C)
n <- length(temp.C)

# distribution under H_0

mu.0 <- 36.75

t.obs <- (x.bar-mu.0)/(s/sqrt(n))
```

```
t.obs
```

```
## [1] 1.543141
```

```
curve(dt(x, df=129), xlim=c(-6, 6))
lines(x=c(t.obs, t.obs), y=c(0, dt(t.obs, 129)), lty=2, lwd=2, col="red")
x <- seq(t.obs, 4, length=100)
y <- dt(x, 120)
polygon(c(x, max(x), t.obs), c(y, 0, 0), col="yellow")
```




```
p.value <- pt(-abs(t.obs), df=129)*2
p.value
```

```
## [1] 0.1252462
```

```
###########################################
# body temperature data - function t.test
###########################################

t.test(temp.C, mu=36.75)
```

```
##
##  One Sample t-test
##
## data:  temp.C
## t = 1.5431, df = 129, p-value = 0.1252
## alternative hypothesis: true mean is not equal to 36.75
## 95 percent confidence interval:
##  36.73445 36.87581
## sample estimates:
## mean of x
##  36.80513
```

```
help(t.test)
```

```r
#####################################################################
# Exercises_02-t-test for the mean of a normal distribution.pdf
#####################################################################


# Exercise 3
######################

# a)

# H0: mu <= 5
# H1: mu > 5

# b)

# define the decision rule

significance.level <- 0.05
significance.level
```

```
## [1] 0.05
```

```r
# right-sided alternative hypothesis
n <- 8
critical.value <- qt(0.95, df=n-1)
critical.value
```

```
## [1] 1.894579
```

```r
# compute the observed value of the
# test statistic

x.bar <- 6.5
s <- 1.9
mu.0 <- 5

t.obs <- (x.bar-mu.0)/(s/sqrt(n))
t.obs
```

```
## [1] 2.232969
```

```r
# is the empirical evidence in favor of H_1?
t.obs > critical.value
```

```
## [1] TRUE
```

```r
# approach based on the p.value

p.value <- 1-pt(t.obs, n-1)
p.value
```

```
## [1] 0.03035218
```

```r
# is the empirical evidence in favor of H_1?

p.value < significance.level
```

```
## [1] TRUE
```

```r
# When the exact p.value is not available,
# a decision can be made by using the
# approximate p.value obtained from
# statistical tables.

t.obs
```

```
## [1] 2.232969
```

```r
alpha <- c(0.80, 0.90, 0.95, 0.975, 0.99, 0.995)
qt(alpha, n-1)
```

```
## [1] 0.8960296 1.4149239 1.8945786 2.3646243 2.9979516 3.4994833
```

```r
# Exercise 4
######################




# a)

# H0: mu  = 23
# H1: mu != 23

# b)


# define the decision rule

significance.level <- 0.05
significance.level
```

```
## [1] 0.05
```

```r
# two-sided alternative hypothesis
n <- 10
critical.value <- qt(0.975, df=n-1)
critical.value
```

```
## [1] 2.262157
```

```r
# compute the observed value of the
# test statistic


x.bar <- 23.2
s <- 0.2
mu.0 <- 23

t.obs <- (x.bar-mu.0)/(s/sqrt(n))
t.obs
```

```
## [1] 3.162278
```

```r
# is the empirical evidence in favor of H_1?

abs(t.obs)> critical.value
```

```
## [1] TRUE
```

```r
# approach based on the p.value

p.value <- 2*pt(-abs(t.obs), n-1)
p.value
```

```
## [1] 0.01150799
```

```r
# is the empirical evidence in favor of H_1?

p.value < significance.level
```

```
## [1] TRUE
```

```r
# approximate p.value (if exact is not available)

t.obs
```

```
## [1] 3.162278
```

```r
alpha <- c(0.80, 0.90, 0.95, 0.975, 0.99, 0.995)
qt(alpha, n-1)
```

```
## [1] 0.8834039 1.3830287 1.8331129 2.2621572 2.8214379 3.2498355
```

```r
# Exercise 5
######################


# a.1)

# H0: mu >= 10
# H1: mu <  10


# a.2)

# define the decision rule

significance.level <- 0.05
significance.level
```

```
## [1] 0.05
```

```r
# left-sided alternative hypothesis
n <- 20
critical.value <- qt(0.05, df=n-1)
critical.value
```

```
## [1] -1.729133
```

```r
# compute the observed value of the
# test statistic

x.bar <- 6.7
s <- 3.9
mu.0 <- 10
```

```r
t.obs <- (x.bar-mu.0)/(s/sqrt(n))
t.obs
```

```
## [1] -3.784115
```

```r
# is the empirical evidence in favor of H_1?

t.obs< critical.value
```

```
## [1] TRUE
```

```r
# approach based on the p.value

p.value <- pt(t.obs, n-1)
p.value
```

```
## [1] 0.0006272178
```

```r
# is the empirical evidence in favor of H_1?

p.value < significance.level
```

```
## [1] TRUE
```

```r
# approximate p.value (if exact is not available)

t.obs
```

```
## [1] -3.784115
```

```r
alpha <- c(0.80, 0.90, 0.95, 0.975, 0.99, 0.995)
qt(alpha, n-1)
```

```
## [1] 0.8609506 1.3277282 1.7291328 2.0930241 2.5394832 2.8609346
```

```r
# b)

mu.0 <- 7.5

t.obs <- (x.bar-mu.0)/(s/sqrt(n))
t.obs
```

```
## [1] -0.9173612
```

```r
# is the empirical evidence in favor of H_1?

t.obs< critical.value
```

```
## [1] FALSE
```

```r
# approach based on the p.value

p.value <- pt(t.obs, n-1)
p.value
```

```
## [1] 0.185226
```

```r
# is the empirical evidence in favor of H_1?

p.value < significance.level
```

```
## [1] FALSE
```

```r
# approximate p.value (if exact is not available)

t.obs
```

```
## [1] -0.9173612
```

```r
alpha <- c(0.80, 0.90, 0.95, 0.975, 0.99, 0.995)
qt(alpha, n-1)
```

```
## [1] 0.8609506 1.3277282 1.7291328 2.0930241 2.5394832 2.8609346
```

```r
# Exercise 6
######################


# H_1:


# a)

# H0: mu  = 3.5
# H1: mu  > 3.5


# b)

# define the decision rule

significance.level <- 0.05
significance.level
```

```
## [1] 0.05
```

```r
# right-sided alternative hypothesis

n <- 6
critical.value <- qt(0.95, df=n-1)
critical.value
```

```
## [1] 2.015048
```

```r
# compute the observed value of the
# test statistic

sample <- c(3.45, 3.47, 3.57, 3.52, 3.40, 3.63)

n <- length(sample)
n
```

```
## [1] 6
```

```r
x.bar <- mean(sample)
x.bar
```

```
## [1] 3.506667
```

```r
s <- sd(sample)
s
```

```
## [1] 0.08406347
```

```r
mu.0 <- 3.5

t.obs <- (x.bar-mu.0)/(s/sqrt(n))
t.obs
```

```
## [1] 0.1942572
```

```r
p.value <- 1-pt(t.obs, n-1)
p.value
```

```
## [1] 0.4268102
```

```r
t.test(sample, mu=3.5, alternative = "g")
```

```
##
##  One Sample t-test
##
## data:  sample
## t = 0.19426, df = 5, p-value = 0.4268
## alternative hypothesis: true mean is greater than 3.5
## 95 percent confidence interval:
##  3.437513      Inf
## sample estimates:
## mean of x
##  3.506667
```

```r
# Exercise 8
######################

# a)

# H0: mu  <= 85
# H1: mu  > 85


# b)

# define the decision rule

significance.level <- 0.05
significance.level
```

```
## [1] 0.05
```

```r
# right-sided alternative hypothesis

n <- 6
critical.value <- qt(0.95, df=n-1)
critical.value
```

```
## [1] 2.015048
```

```r
# compute the observed value of the
# test statistic


sample <- c(93.2, 87.0, 92.1, 90.1, 87.3, 93.6)
```

```
n <- length(sample)
n
```

```
## [1] 6
```

```
x.bar <- mean(sample)
x.bar
```

```
## [1] 90.55
```

```
s <- sd(sample)
s
```

```
## [1] 2.901551
```

```
mu.0 <- 85


t.obs <- (x.bar-mu.0)/(s/sqrt(n))
t.obs
```

```
## [1] 4.68531
```

```
p.value <- 1-pt(t.obs, n-1)
p.value
```

```
## [1] 0.002703886
```

```
t.test(sample, mu=85, alternative = "g")
```

```
##
##  One Sample t-test
##
## data:  sample
## t = 4.6853, df = 5, p-value = 0.002704
## alternative hypothesis: true mean is greater than 85
## 95 percent confidence interval:
##  88.16307      Inf
## sample estimates:
## mean of x
##     90.55
```