28112023_Statistical_Learning

Mattia G.

2023-12-28

```
#####################################################
# inference for the variance of a normal distribution
######################################################

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

n <- length(temp.C)
n
```

```
## [1] 130
```

```
s2 <- var(temp.C)
s2
```

```
## [1] 0.1659128
```

```
s <- sd(temp.C)
s
```

```
## [1] 0.407324
```

```
ic.lower <- (n-1)*s2/qchisq(0.975, n-1)
ic.upper <- (n-1)*s2/qchisq(0.025, n-1)

# IC for the variance

round(c(lower=ic.lower, variance=s2, upper= ic.upper), 3)
```

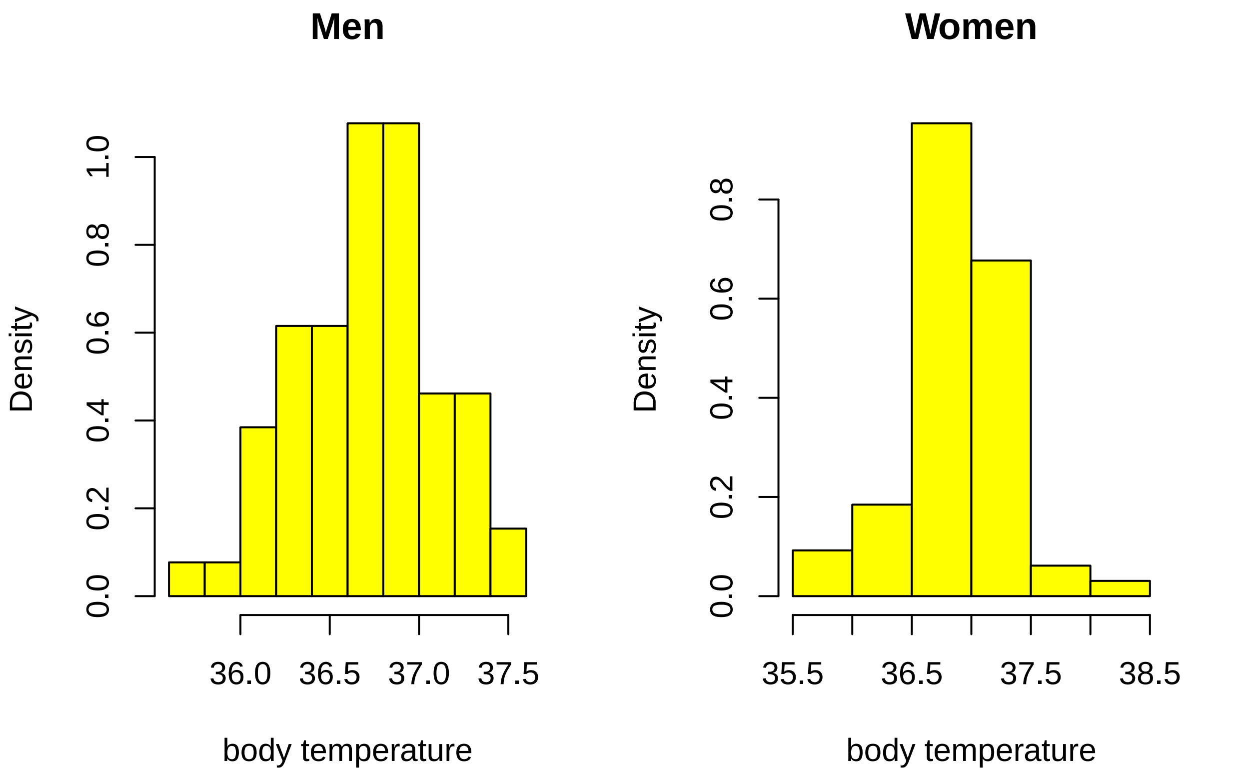
```
##    lower variance    upper
##    0.132    0.166    0.215
```

```
# IC for the sd

round(c(lower=sqrt(ic.lower), std.dev=sqrt(s2), upper= sqrt(ic.upper)), 3)
```

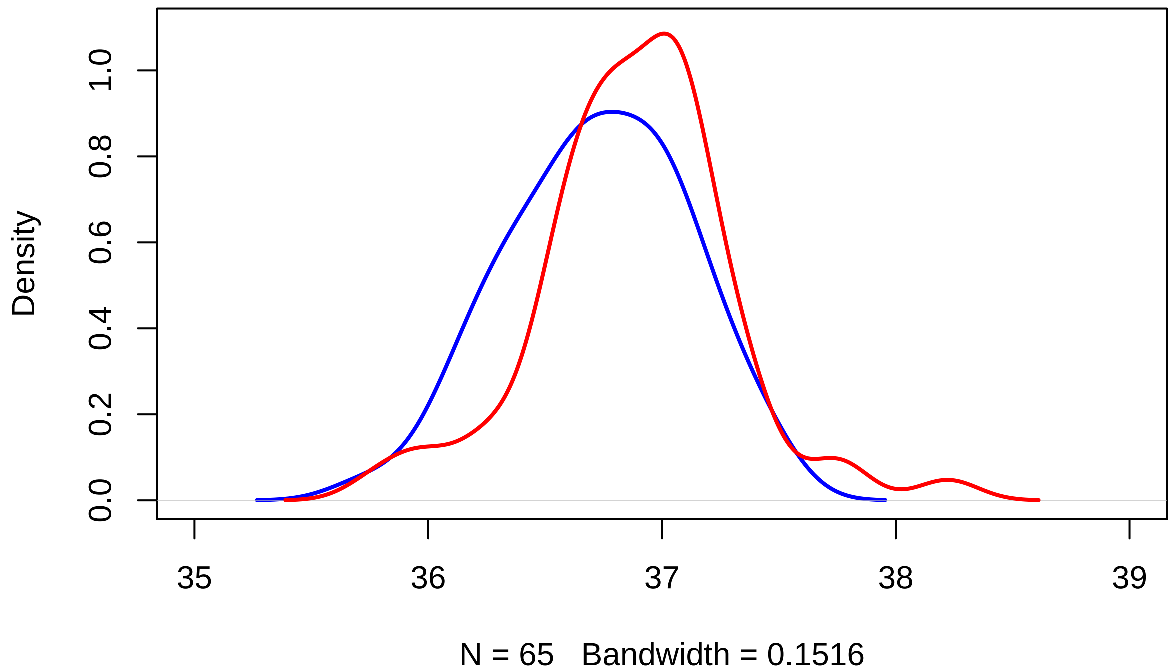
```
##   lower std.dev   upper
##   0.363   0.407   0.464
```

```
#########################################
# comparison of means: two populations
##########################################

# histograms comparison
par(mfrow=c(1,2))
```

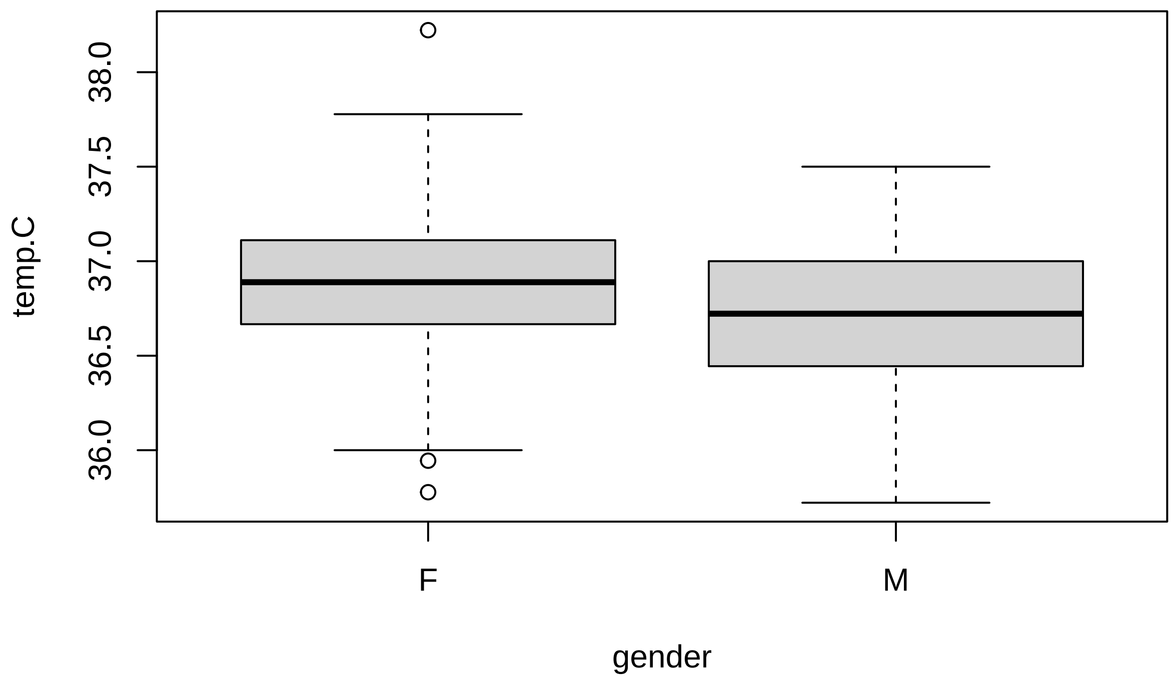
```r
hist(temp.C[gender=="M"], main="Men", prob=T, col="yellow", xlab="body temperature")
hist(temp.C[gender=="F"], main="Women", prob=T, col="yellow", xlab="body temperature")
```



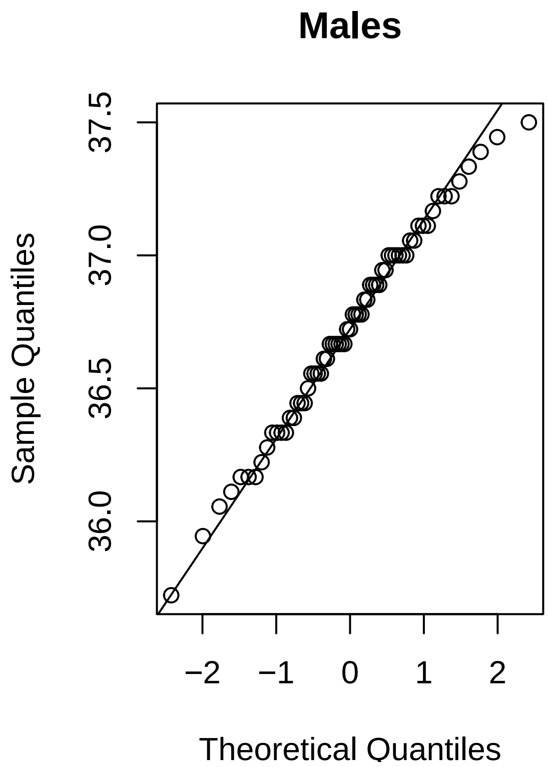

```r
par(mfrow=c(1,1))

d.m <- density(temp.C[gender=="M"])
d.f <- density(temp.C[gender=="F"])
plot(d.m, main="", lwd=2, col="blue", xlim=c(35, 39), ylim=c(0, 1.1))
lines(d.f, lwd=2, col="red")
```

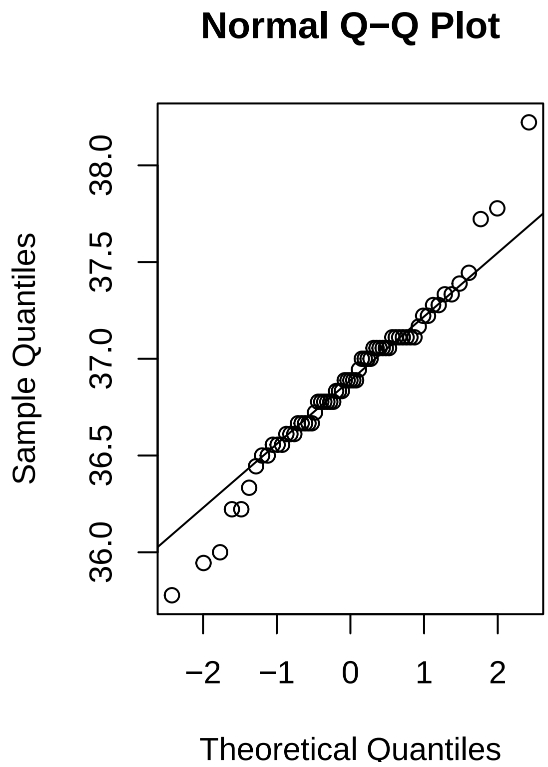



```
# side-by-side boxplots
boxplot(temp.C~gender, col="lightgray")
```




```
# normal quantile plots

par(mfrow=c(1,2))
qqnorm(temp.C[gender=="M"], main="Males")
qqline(temp.C[gender=="M"], main="Females")

qqnorm(temp.C[gender=="F"])
qqline(temp.C[gender=="F"])
```






```
par(mfrow=c(1,1))

# Males

n1 <- length(temp.C[gender=="M"])
n1
```

```
## [1] 65
```

```
m.1 <- mean(temp.C[gender=="M"])
m.1
```

```
## [1] 36.72479
```

```
s2.1 <- var(temp.C[gender=="M"])
s2.1
```

```
## [1] 0.1506974
```

```
# Females

n2 <- length(temp.C[gender=="F"])
n2
```

```
## [1] 65
```

```
m.2 <- mean(temp.C[gender=="F"])
m.2
```

```
## [1] 36.88547
```

```
s2.2 <- var(temp.C[gender=="F"])
s2.2
```

```
## [1] 0.1706093
```

```r
# pooled variance

pooled.var <- ((n1-1)*s2.1+(n2-1)*s2.2)/(n1+n2-2)

# test statistic

t.obs <- (m.1-m.2)/(sqrt(pooled.var*(1/n1 +1/n2)))
t.obs
```

```
## [1] -2.285435
```

```r
p.value <- 2*pt(-abs(t.obs), n1+n2-2)
p.value
```

```
## [1] 0.02393188
```

```r
# function t.test()

t.test(temp.C[gender=="M"], temp.C[gender=="F"], var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  temp.C[gender == "M"] and temp.C[gender == "F"]
## t = -2.2854, df = 128, p-value = 0.02393
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.29979966 -0.02156786
## sample estimates:
## mean of x mean of y
##  36.72479  36.88547
```

```r
t.test(temp.C~gender, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  temp.C by gender
## t = 2.2854, df = 128, p-value = 0.02393
## alternative hypothesis: true difference in means between group F and group M is not equal to 0
## 95 percent confidence interval:
##  0.02156786 0.29979966
## sample estimates:
## mean in group F mean in group M
##        36.88547        36.72479
```

```r
# comparing variances

t.obs <- s2.1/s2.2
t.obs
```

```
## [1] 0.8832897
```

```r
# critical values

qf(0.025, n1-1, n2-1)
```

```
## [1] 0.6099476
```

```r
qf(0.975, n1-1, n2-1)
```

```
## [1] 1.639485
```

```r
# compute the p.value
1/t.obs
```

```
## [1] 1.132131
```

```r
p.value <- pf(t.obs, n1-1, n2-1)+ (1-pf(1/t.obs, n2-1,n1-1))
p.value
```

```
## [1] 0.6210837
```

```r
var.test(temp.C[gender=="M"], temp.C[gender=="F"])
```

```
##
##  F test to compare two variances
##
## data:  temp.C[gender == "M"] and temp.C[gender == "F"]
## F = 0.88329, num df = 64, denom df = 64, p-value = 0.6211
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.5387604 1.4481404
## sample estimates:
## ratio of variances
##          0.8832897
```

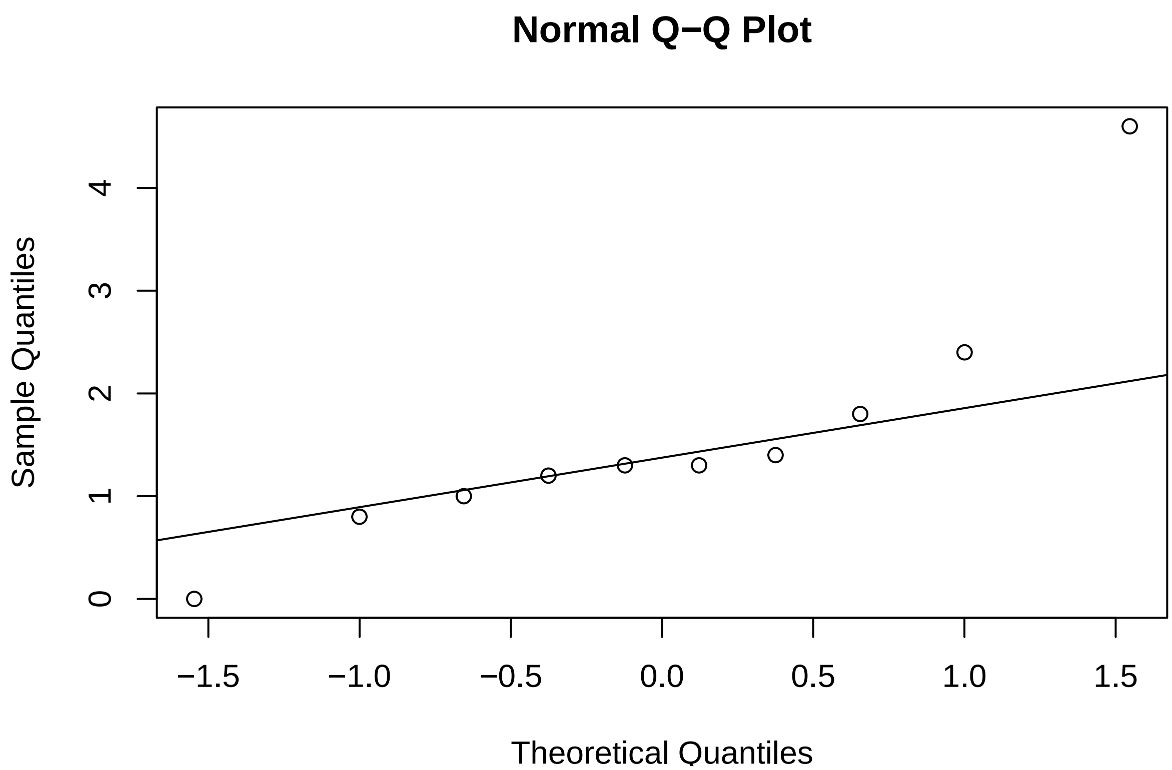
```r
var.test(temp.C~gender)
```

```
##
##  F test to compare two variances
##
## data:  temp.C by gender
## F = 1.1321, num df = 64, denom df = 64, p-value = 0.6211
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6905408 1.8561126
## sample estimates:
## ratio of variances
##           1.132131
```

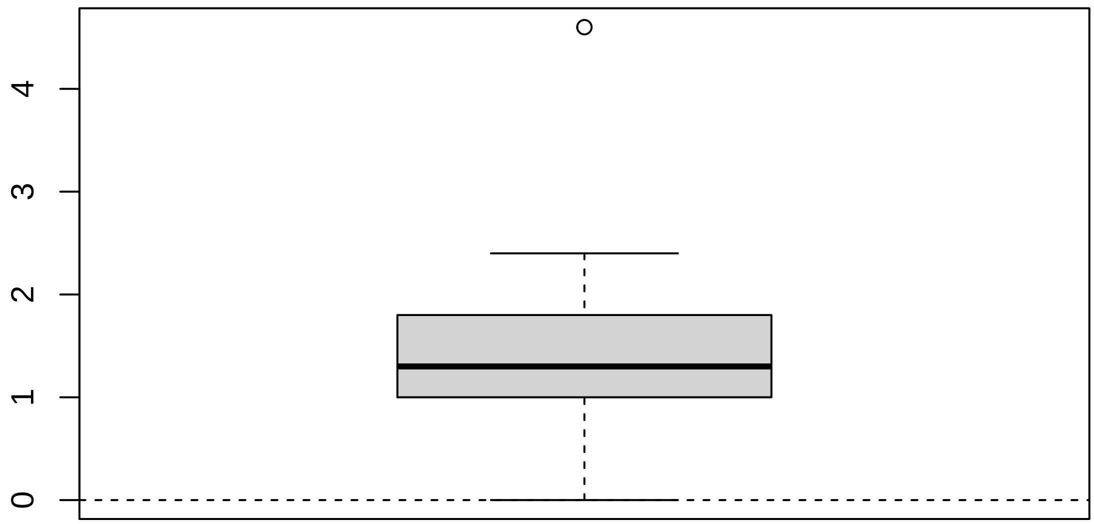
```r
# t.test non-equal variances

t.test(temp.C~gender,, var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  temp.C by gender
## t = 2.2854, df = 127.51, p-value = 0.02394
## alternative hypothesis: true difference in means between group F and group M is not equal to 0
## 95 percent confidence interval:
##  0.02156277 0.29980476
## sample estimates:
## mean in group F mean in group M
##        36.88547        36.72479
```

```
#####################################
# comparison of means for paired data
#####################################

data(sleep)
attach(sleep)

g1 <- extra[group==1]
g2 <- extra[group==2]

d <- g2-g1
qqnorm(d)
qqline(d)
```




```
boxplot(g2-g1, col="lightgray")
abline(h=0, lty=2)
```




```
t.test(d)
```

```
##
##  One Sample t-test
##
## data:  d
## t = 4.0621, df = 9, p-value = 0.002833
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.7001142 2.4598858
## sample estimates:
## mean of x
##      1.58
```

```
# wrong application of the t-test
t.test(g2, g1, paired=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  g2 and g1
## t = 1.8608, df = 17.776, p-value = 0.07939
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2054832  3.3654832
## sample estimates:
## mean of x mean of y
##      2.33      0.75
```

```
# t-test for paired data
t.test(g2, g1, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  g2 and g1
## t = 4.0621, df = 9, p-value = 0.002833
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  0.7001142 2.4598858
## sample estimates:
## mean of the differences
##                    1.58
```

```
t.test(d)
```

```
##
##  One Sample t-test
##
## data:  d
## t = 4.0621, df = 9, p-value = 0.002833
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.7001142 2.4598858
## sample estimates:
## mean of x
##      1.58
```

```
#########################################
# comparing proportions
######################################

M <- matrix(c(335, 75, 302, 105), ncol=2, byrow=TRUE)
M
```

```
##      [,1] [,2]
## [1,]  335   75
## [2,]  302  105
```

```
margin.table(M, margin=1)
```

```
## [1] 410 407
```

```
# Placebo

n1 <- 335+75
pp.hat <- 335/n1
pp.hat
```

```
## [1] 0.8170732
```

```
# Vitamin C

n2 <- 302+105
pc.hat <- 302/n2
pc.hat
```

```
## [1] 0.7420147
```

```
# pooled pi

p.hat <- (335+302)/(n1+n2)
p.hat
```

```
## [1] 0.7796818
```

```
se.hat <- sqrt(p.hat*(1-p.hat)*(1/n1+1/n2))
se.hat
```

```
## [1] 0.02900052
```

```r
t.obs <- (pp.hat-pc.hat)/se.hat
t.obs
```

```
## [1] 2.588175
```

```r
p.value <- 1-pnorm(t.obs)
p.value
```

```
## [1] 0.004824295
```

```r
# binom.test

prop.test(c(335, 302), c(n1, n2), alt="g")
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(335, 302) out of c(n1, n2)
## X-squared = 6.2688, df = 1, p-value = 0.006144
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.02508328 1.00000000
## sample estimates:
##    prop 1    prop 2
## 0.8170732 0.7420147
```