Tugas Besar IF2220 - Probabilitas dan Statistika

Part 4: One Sample Hypothesis

Anggota:
13521116 - Juan Christopher Santoso
13521162 - Antonio Natthan Krishna

All testing use significant of 5%

In [1]:
```
# Import Dataset
df <- read.csv("..\\test\\anggur.csv")

# Data Statistics
properties <- c("Rows", "Columns")
value <- c(nrow(df), ncol(df))
cbind(properties, value)

# List of Columns
columns_index <- c(1:ncol(df))
columns_name <- colnames(df)

# Display List
cbind(columns_index, columns_name)

# Significance
Significance <- 0.05
```

A matrix: 2 × 2 of type chr

properties	value
Rows	1000
Columns	12

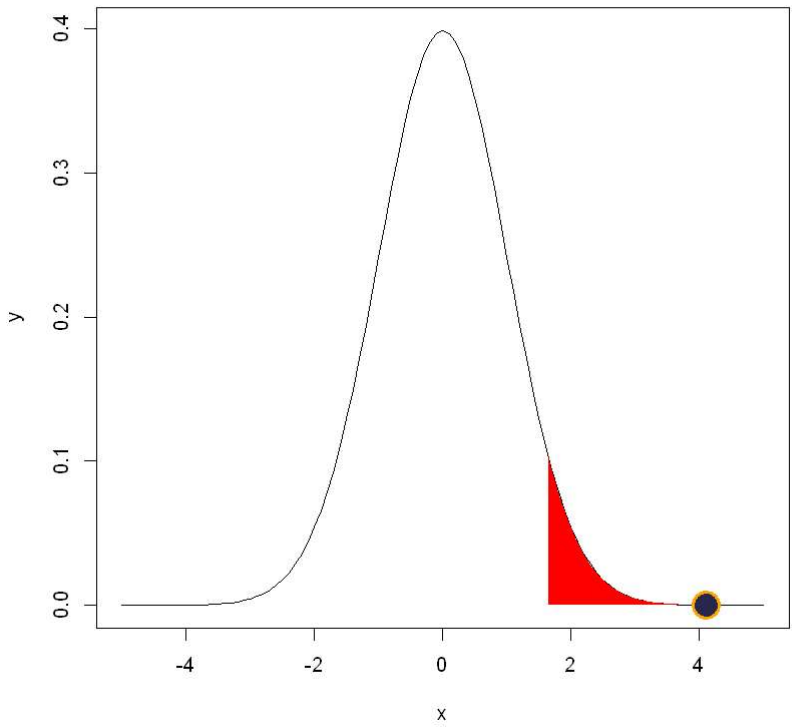
A matrix: 12 × 2 of type chr

columns_index	columns_name
1	fixed.acidity
2	volatile.acidity
3	citric.acid
4	residual.sugar
5	chlorides
6	free.sulfur.dioxide
7	total.sulfur.dioxide
8	density
9	pH
10	sulphates
11	alcohol
12	quality

1. Is the mean of `pH` greater than 3.29?

In [2]:
```
t <- (mean(df[,"pH"]) - 3.29) / (sd(df[,"pH"]) / sqrt(nrow(df)))
t0 <- qt(0.05, nrow(df)-1, lower.tail = FALSE)

cat("t  :", t, "\n")
cat("t0 :", t0, "\n")
cat("P-value:", 1- pt(t, nrow(df)-1))

# Plotting Critical Area
x <- seq(-5, 5, 0.1)
y <- dt(x, nrow(df)-1)
plot(x, y, type = "l")

x2 <- seq(t0,5,0.01)
y2 <- dt(x2, nrow(df)-1)
x2 = c(t0,x2,5)
y2 = c(0,y2,0)
polygon(x2,y2, col="red", border=NA)

lines(t, 0, type = "o", pch=21, bg="#28284d", cex=3, lwd=3, col="orange")
```

```
t  : 4.103781
t0 : 1.64638
P-value: 2.197958e-05
```



H0 = (mean pH == 3.29)
H1 = (mean pH > 3.29)

Use the significance 0.05
Using t distribution with degree 999 (1000-1),
Critical area : t > t(0.05),
Since t > t(0.05) (and p-Value < significance) which means t is located in critical area. Hence, we reject H0.

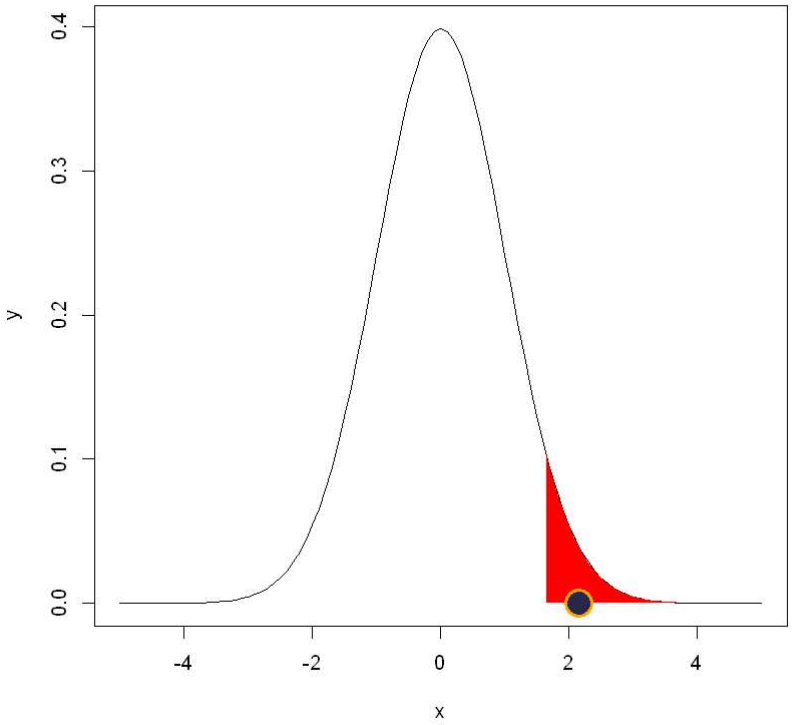
Conclusion: mean of population's pH greater than 3.29

2. Is the mean of `residual sugar` greater than 2.50?

In [3]:
```
t <- (mean(df[,"residual.sugar"])-2.5) / (sd(df[,"residual.sugar"]) / sqrt(nrow(df)
t0 <- qt(0.05, nrow(df)-1, lower.tail = FALSE)

cat("t  :", t, "\n")
cat("t0 :", t0, "\n")
cat("P-value:", 1 - pt(t, nrow(df)-1))

# Plotting Critical Area
x <- seq(-5, 5, 0.1)
y <- dt(x, nrow(df))
plot(x, y, type = "l")

x2 <- seq(t0,5,0.01)
y2 <- dt(x2, nrow(df)-1)
```

```
x2 = c(t0,x2,5)
y2 = c(0,y2,0)
polygon(x2,y2, col="red", border=NA)

lines(t, 0, type = "o", pch=21, bg="#28284d", cex=3, lwd=3, col="orange")
```

```
t  : 2.147962
t0 : 1.64638
P-value: 0.01597836
```



H0 = (mean residual sugar == 2.50)
H1 = (mean residual sugar > 2.50)

Use the significance 0.05
Using t distribution with degree 999 (1000-1),
Critical area : t > t(0.05),
Since t > t(0.05) (and p value < significance) which means t is located in critical area. Hence, we reject H0.

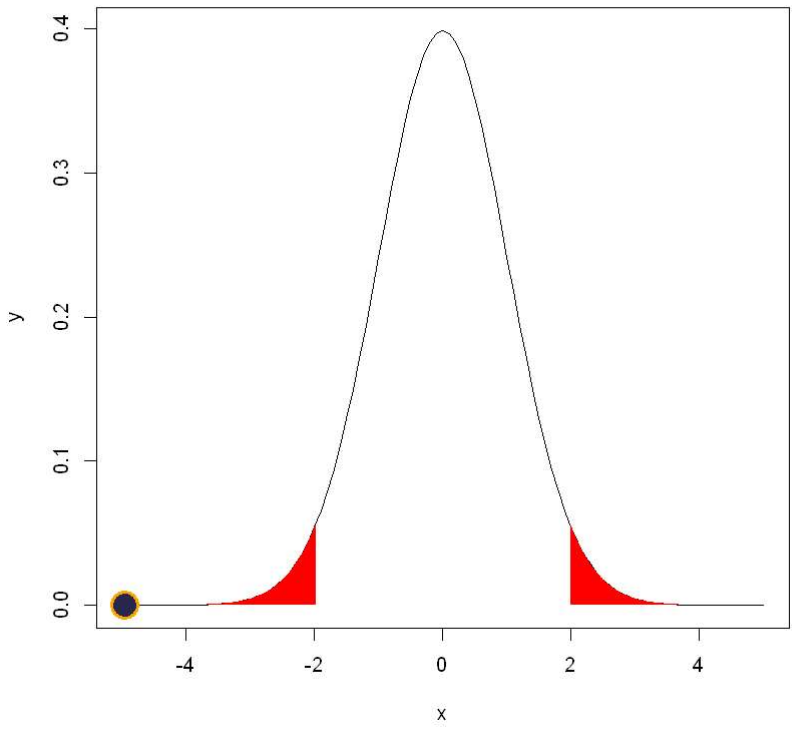
Conclusion: mean of population's residual sugar greater than 2.50

3. Is the mean of the first 150 row in column `sulphates` not 0.65?

In [4]:
```
t <- (mean(df[1:150,"sulphates"]) - 0.65) / (sd(df[1:150,"sulphates"]) / sqrt(150))
t0low <- qt(0.025, 150-1)
t0high <- qt(0.025, 150-1, lower.tail = FALSE)

cat("t  :", t, "\n")
```

```
cat("t0 low :", t0low, "\n")
cat("t0 high :", t0high, "\n")
cat("P-value:", pt(t, 150-1))

# Plotting Critical Area
x <- seq(-5, 5, 0.1)
y <- dt(x, nrow(df))
plot(x, y, type = "l")

x2 <- seq(-5,t0low,0.01)
y2 <- dt(x2, nrow(df)-1)
x2 = c(-5,x2,t0low)
y2 = c(0,y2,0)
polygon(x2,y2, col="red", border=NA)

x3 <- seq(t0high,5,0.01)
y3 <- dt(x3, nrow(df)-1)
x3 = c(x3,5,t0high)
y3 = c(0,y3,0)
polygon(x3,y3, col="red", border=NA)

lines(t, 0, type = "o", pch=21, bg="#28284d", cex=3, lwd=3, col="orange")
```

```
t  : -4.964843
t0 low : -1.976013
t0 high : 1.976013
P-value: 9.295076e-07
```



H0 = (mean first 150 sulphates == 2.50)

H1 = (mean first 150 sulphates != 2.50)

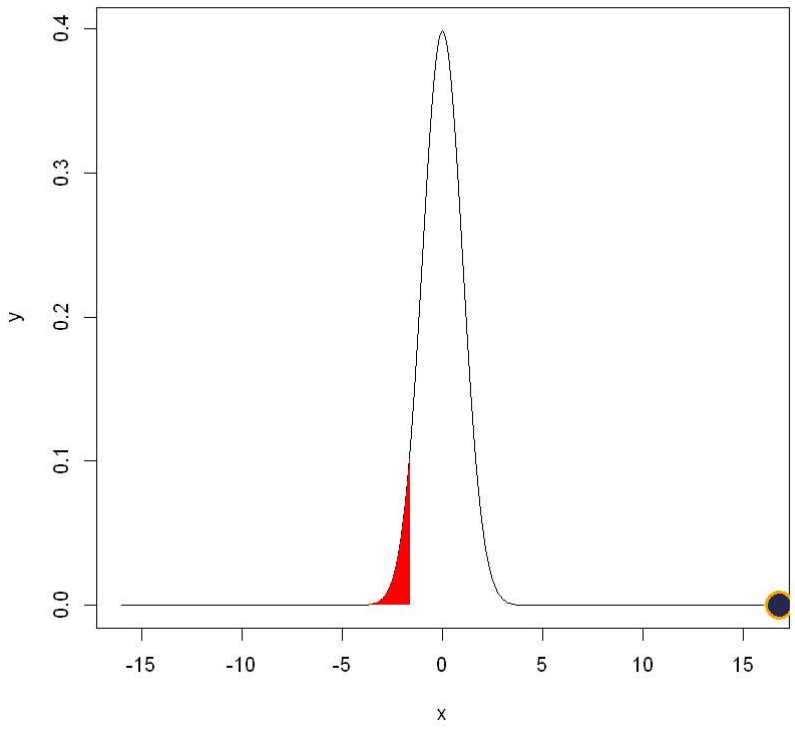
Use the significance 0.05
Using t distribution with degree 999 (1000-1),
Critical area : t < t(-0.025), t > t(0.025)
Since t < t0 (and p value < significance) which means t is located in critical area. Hence, we reject H0.

Conclusion: mean of population's sulphates is NOT 0.65

4. Is the mean of `total sulfur dioxide` lower than 35?

In [5]:
```
t <- (mean(df[,"total.sulfur.dioxide"]) - 35) / (sd(df[,"total.sulfur.dioxide"]) /
t0 <- qt(0.05, nrow(df)-1)

cat("t  :", t, "\n")
cat("t0 :", t0, "\n")
cat("P-value:", pt(t, nrow(df)-1))

# Plotting Critical Area
x <- seq(-16, 16, 0.1)
y <- dt(x, nrow(df))
plot(x, y, type = "l")

x2 <- seq(-16,t0,0.01)
y2 <- dt(x2, nrow(df)-1)
x2 = c(-16,x2,t0)
y2 = c(0,y2,0)
polygon(x2,y2, col="red", border=NA)

lines(t, 0, type = "o", pch=21, bg="#28284d", cex=3, lwd=3, col="orange")
```

```
t  : 16.78639
t0 : -1.64638
P-value: 1
```



H0 = (mean total sulfur dioxide == 35)
H1 = (mean total sulfur dioxide < 35)

Use the significance 0.05
Using t distribution with degree 999 (1000-1),
Critical area : t < t(0.05),
Since t > t(0.05) (and p value > significance) which means t is NOT located in critical area.
Hence, we accept H0.

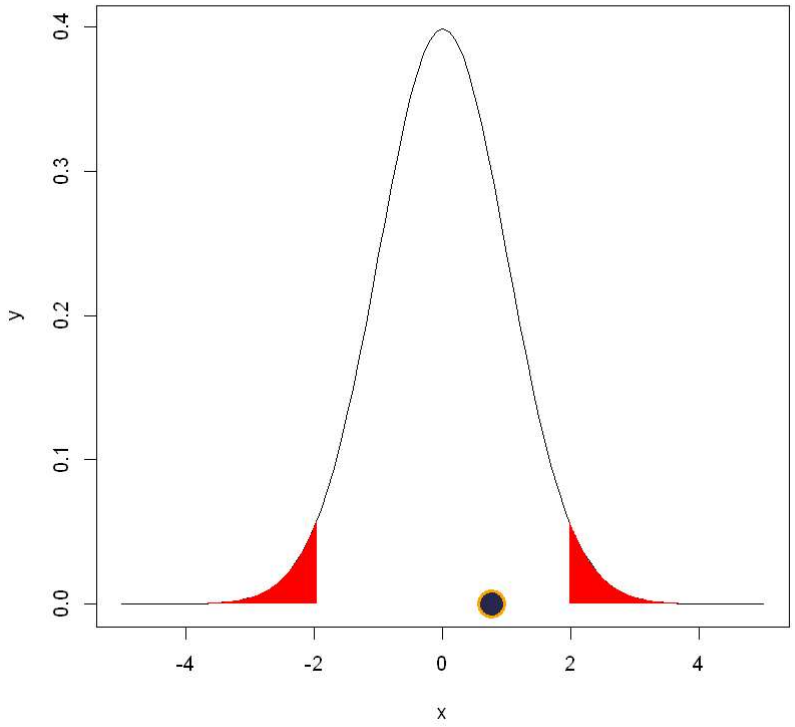
Conclusion: mean of population's total sulfur dioxide is NOT LOWER than 35

5. Is the proportion of the `total sulfur dioxide` which are more than 40 not 50%?

In [6]:
```
select <- df[df$"total.sulfur.dioxide" > 40,]
proportion <- nrow(select) / nrow(df)
z <- (proportion - 0.5) / sqrt(proportion*(1-proportion)/nrow(df))
z0 <- qnorm(0.025)

cat("z  :", z, "\n")
cat("z0 low :", z0, "\n")
cat("z0 high :", z0*-1, "\n")
cat("P-value:", 1-pnorm(z))

# Plotting Critical Area
x <- seq(-5, 5, 0.1)
y <- dnorm(x)
plot(x, y, type = "l")
```

```
x2 <- seq(-5,z0,0.01)
y2 <- dnorm(x2)
x2 = c(-5,x2,z0)
y2 = c(0,y2,0)
polygon(x2,y2, col="red", border=NA)

x3 <- seq(z0*-1,5,0.01)
y3 <- dnorm(x3)
x3 = c(x3,5,z0)
y3 = c(0,y3,0)
polygon(x3,y3, col="red", border=NA)

lines(z, 0, type = "o", pch=21, bg="#28284d", cex=3, lwd=3, col="orange")
```

```
z  : 0.7591653
z0 low : -1.959964
z0 high : 1.959964
P-value: 0.2238768
```



p = proportion of the `total sulfur dioxide` which are more than 40
H0 = (p == 0.5)
H1 = (p != 0.5)

Use the significance 0.05
Using normal distribution,
Critical area : z < z(-0.025), z > z(0.025)
Since z(-0.025) < z < z(0.025) (and p value > significance/2) which means z is NOT located in critical area. Hence, we accept H0.

Conclusion: mean of proportion of the total sulfur dioxide which are more than 40 is 50%