Chapter 7 - Hypothesis Testing

Dr. Alessandro Ruggieri

Contents

We take data on the average number of hours worked in a week across countries.

```
# import new data:hours worked rate across countries
data_hours <- read.csv("HOW_TEMP_SEX_ECO_GEO_NB_A-filtered-2021-01-12.csv",  header = TRUE)
# rename columns
names(data_hours) <- c("country", "label", "source",
                       "gender","occupations","area",
                       "year", "weeklyhours", "status",
                       "note1", "note2")
# subset matrix according to gender groups: Sex=all gender
data_hours<-data_hours[data_hours$gender == "Sex: Total",]
data_hours$gender <- NULL
# subset matrix according to occupations groups: occupations=Aggregate
data_hours<-data_hours[data_hours$occupations == "Economic activity (Aggregate): Total", ]
data_hours$occupations <- NULL
# subset matrix according to area groups: area=Aggregate
data_hours<-data_hours[data_hours$area == "Area type: National", ]
data_hours$area <- NULL
## Select 2018 data
data_hours_2018<-data_hours[data_hours$year == "2018",]
## Final data
x <- data_hours_2018$weeklyhours
```

One-Sample Two-sided T-testing

We can test whether the mean of the samples is equal to a specified value against the two-side alternative using the function t.test( )

Suppose we want to test the hypothesis that the average number of hours worked is equal to 35 ($H_0 : \mu = 35$) against the alternative that the the average number of hours is different than 35 ($H_0 : \mu \neq 35$). This T-test can be implemented as follows:

```
# Test whether the average weekly number of hours worked is 35 against a two-sided alternative
t.test(x, mu = 35)
```

```
##
##  One Sample t-test
##
## data:  x
```

```
## t = 10.436, df = 69, p-value = 7.796e-16
## alternative hypothesis: true mean is not equal to 35
## 95 percent confidence interval:
##  38.52083 40.18502
## sample estimates:
## mean of x
##  39.35293
```

One-Sample One-sided T-testing

We can test if the sample mean is lower or greater than a specific value by switching to a one-sided test. This is possibility by including the option alternative= "two.sided", "less", or "greater". Suppose in fact that we want to test the hypothesis that the average number of hours worked is equal to 35 ($H_0 : \mu = 35$) against the alternative that the the average number of hours is greater than 35 ($H_0 : \mu > 35$). Then we can write:

```r
# Test whether the average weekly number of hours worked is 35 against a one-sided alternative
t.test(x, mu = 35, alternative = 'greater')
```

```
##
##  One Sample t-test
##
## data:  x
## t = 10.436, df = 69, p-value = 3.898e-16
## alternative hypothesis: true mean is greater than 35
## 95 percent confidence interval:
##  38.65752      Inf
## sample estimates:
## mean of x
##  39.35293
```

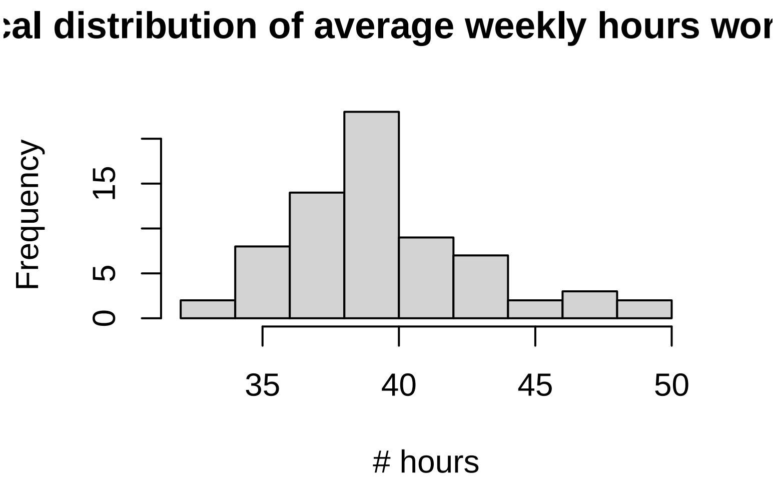
Suppose now that we want to test the hypothesis that the average number of hours worked is equal to 35 ($H_0 : \mu = 35$) against the alternative that the the average number of hours is lower than 35 ($H_0 : \mu < 35$). Then we can write:

```r
# Test whether the average weekly number of hours worked is 35 against a one-sided alternative
t.test(x, mu = 35, alternative = 'less')
```

```
##
##  One Sample t-test
##
## data:  x
## t = 10.436, df = 69, p-value = 1
## alternative hypothesis: true mean is less than 35
## 95 percent confidence interval:
##      -Inf 40.04834
## sample estimates:
## mean of x
##  39.35293
```

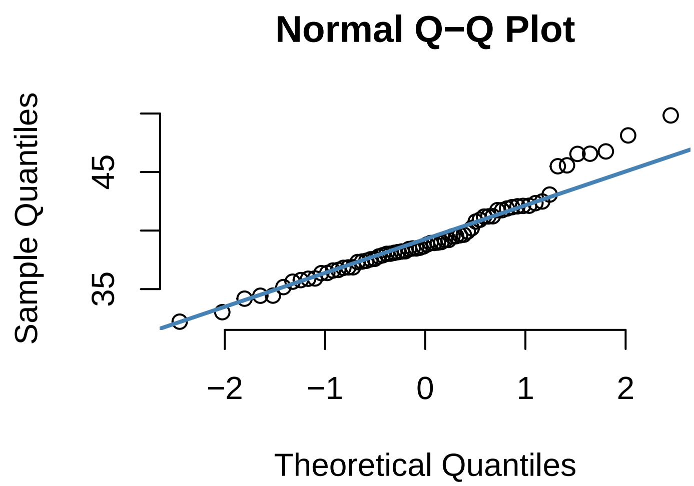
Remember that - when performing this test - we assumed that the sample data is extracted from a normally distributed random variable. We can check whether this assumption is satisfied by looking at the data histogram:

```r
# histogram of data
hist(x, xlab="# hours", main="Empirical distribution of average weekly hours worked in 2019")
```



We can also use the QQ plot:

```
# Sample quantile-to-quantile plot
qqnorm(x, pch = 1, frame = FALSE)
qqline(x, col = "steelblue", lwd = 2)
```



Two-Sample T-testing

Suppose now we want to compare the average weekly hours worked in UK against Portugal

We first extract the data as follows:

```
## Select and extract UK and Portugal data
data_hours_uk<-data_hours[data_hours$country == "United Kingdom",]
x <- data_hours_uk$weeklyhours
data_hours_pt<-data_hours[data_hours$country == "Portugal",]
y <- data_hours_pt$weeklyhours
```

Check if data have the same length

```
## Check if data have same length
length(x)==length(y)
```

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

```r
n<-length(x)
```

We then combine the data for UK and PT into a data frame

```r
# Create a data frame
z <- data.frame(
                group = rep(c("UK", "PT"), each = n),
                weeklyhours = c(x,  y)
                )
```

Then we can compute summary statistics by groups

```r
## Summarize statistics for hours worked separately by groups (UK and PT)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
group_by(z, group) %>%
  summarise(
    count = n(),
    mean = mean(weeklyhours, na.rm = TRUE),
    sd = sd(weeklyhours, na.rm = TRUE),
    min = min(weeklyhours),
    max = max(weeklyhours)
  )
```

```
## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 2 x 6
##   group count  mean    sd   min   max
##   <chr> <int> <dbl> <dbl> <dbl> <dbl>
## 1 PT       10  38.5 0.161  38.3  38.8
## 2 UK       10  35.8 0.161  35.6  36.0
```

We can test whether the average number of weekly hours worked in UK is the same as in Portugal ($H_0 : \mu_{UK} = \mu_{PT}$) against the alternative hypothesis that they are different ($H_0 : \mu_{UK} \neq \mu_{PT}$) using the function t.test( ) and specifying the samples we are comparing:

```r
t.test(x, y, alternative = "two.sided", var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = -37.571, df = 18, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.853196 -2.551004
## sample estimates:
## mean of x mean of y
```

```
##   35.8051   38.5072
```

We can also test whether the average number of weekly hours worked in UK is the same as in Portugal ($H_0 : \mu_{UK} = \mu_{PT}$) against the alternative hypothesis that hours worked are on average higher in UK ($H_0 : \mu_{UK} > \mu_{PT}$) using the function t.test( ) and specifying the samples we are comparing:

```r
t.test(x, y, alternative = "greater", var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = -37.571, df = 18, p-value = 1
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -2.826812       Inf
## sample estimates:
## mean of x mean of y
##   35.8051   38.5072
```

Remember that we imposed three major assumptions when performing this test:

- the two samples (UK and PT) are independent
- data in both samples follow a normal distribution
- data in both sample are drawn from a random variable with same variance