Week 3 R Practice

Anubhav Saha

06/10/2021

Introduction

Hypothesis testing is the use of statistics to determine the probability that a given hypothesis is true. The usual process of hypothesis testing consists of four steps:

1. Formulate the null hypothesis H0 (commonly, that the observations are the result of pure chance) and the alternative hypothesis Ha (commonly, that the observations show a real effect combined with a component of chance variation).
2. Identify a test statistic that can be used to assess the truth of the null hypothesis.
3. Compute the P-value, which is the probability that a test statistic at least as significant as the one observed would be obtained assuming that the null hypothesis were true. The smaller the P-value, the stronger the evidence against the null hypothesis.
4. Compare the p-value to an acceptable significance value alpha (sometimes called an alpha value). If p<=alpha, that the observed effect is statistically significant, the null hypothesis is ruled out, and the alternative hypothesis is valid.

```
knitr::opts_chunk$set(echo = TRUE)
library(MASS) #Loading the package MASS
library(datasets)
library(dplyr)
```

Load the MASS, datasets and other required packages in R using the library function

```
## Warning: package 'dplyr' was built under R version 3.6.3
```

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

```
## The following object is masked from 'package:MASS':
##
##     select
```

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

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

```
library(ggpubr)
```

```
## Warning: package 'ggpubr' was built under R version 3.6.3
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.6.3
```

```
## Loading required package: magrittr
```

One-sample t-test

Use the "chem" dataset to answer the question, "is the flour production company producing whole meal flour with greater than 1 part per million copper in it?"

```
chem #To glance our checm dataset
```

code:

```
##  [1]  2.90  3.10  3.40  3.40  3.70  3.70  2.80  2.50  2.40  2.40  2.70  2.20
## [13]  5.28  3.37  3.03  3.03 28.95  3.77  3.40  2.20  3.50  3.60  3.70  3.70
```

```
str(chem) #To get its structure
```

```
##  num [1:24] 2.9 3.1 3.4 3.4 3.7 3.7 2.8 2.5 2.4 2.4 ...
```

```
View(chem) #To view our dataset on a new window
MASS::chem #same as chem
```

```
##  [1]  2.90  3.10  3.40  3.40  3.70  3.70  2.80  2.50  2.40  2.40  2.70  2.20
## [13]  5.28  3.37  3.03  3.03 28.95  3.77  3.40  2.20  3.50  3.60  3.70  3.70
```

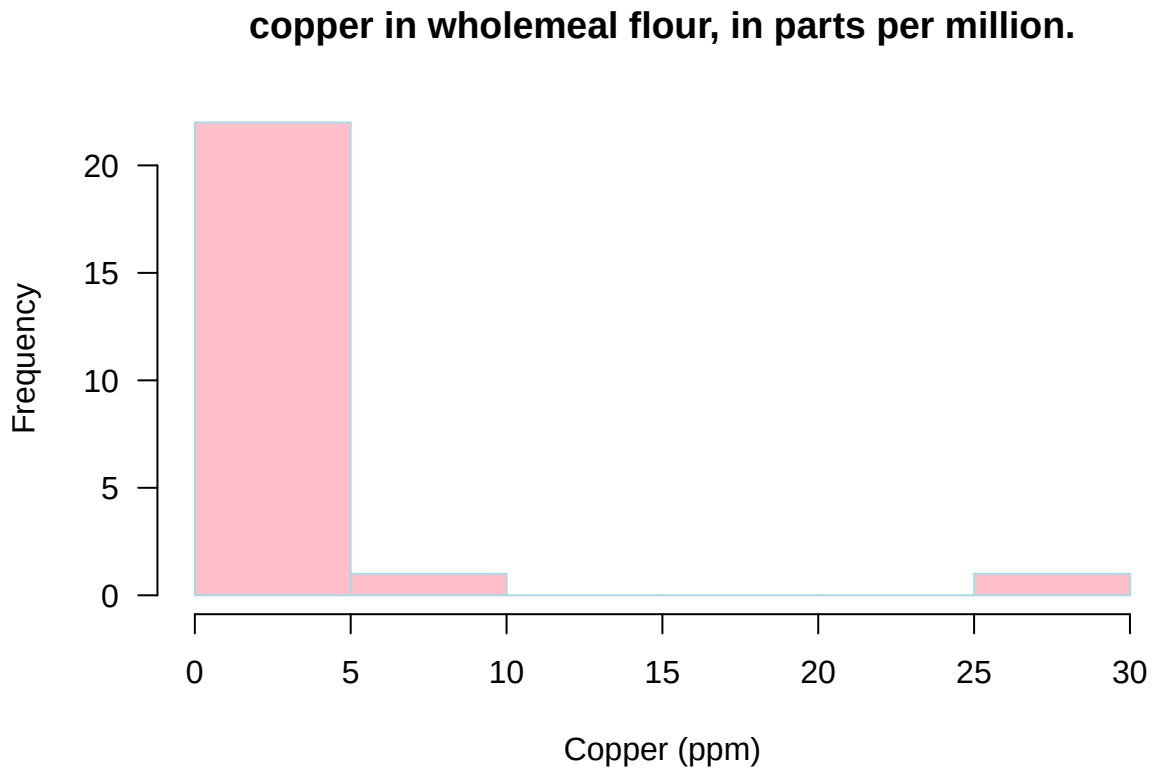
```
?chem #To get help on our dataset
```

```
## starting httpd help server ... done
```

```
summary(chem) #Gives the five number summary of our dataset
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.200   2.775   3.385   4.280   3.700  28.950
```

```r
hist(chem,
     main="copper in wholemeal flour, in parts per million.",
     xlab="Copper (ppm)",
     border="light blue",
     col="pink",
     las=1,
     breaks=5) #plots the histogram with given parameters
```



Step 1: Formulate Null and alternative Hypothesis As we need to find out if there is > 1ppm of copper present, we'll use a one tailed t test

Null hypothesis (H0) : The mean value of copper present is less than or equal to one, i.e. mu<=1

Alternative hypothesis (H1) : The mean value of copper present is greater than one, i.e. mu>1 (one tailed test)

Step 2: Identify the test statistic Since n < 30, we will use the t statistic instead of the z statistic.

```r
?t.test
```

Performs one and two sample t-tests on vectors of data.

Default arguments: t.test(x, y = NULL, alternative = c("two.sided", "less", "greater"), mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95, ...)

```
ttest_chem <- t.test(chem,
                     alternative="greater",
                     mu = 1,
                     conf.level=0.99)
ttest_chem
```

Step 3: Compute the P-value

```
##
##  One Sample t-test
##
## data:  chem
## t = 3.0337, df = 23, p-value = 0.002952
## alternative hypothesis: true mean is greater than 1
## 99 percent confidence interval:
##  1.577245      Inf
## sample estimates:
## mean of x
##  4.280417
```

```
attributes(ttest_chem)
```

```
## $names
##  [1] "statistic"   "parameter"   "p.value"     "conf.int"    "estimate"
##  [6] "null.value"  "stderr"      "alternative" "method"      "data.name"
##
## $class
## [1] "htest"
```

```
ttest_chem$p.value
```

```
## [1] 0.002951761
```

```
ttest_chem$conf.int
```

```
## [1] 1.577245      Inf
## attr(,"conf.level")
## [1] 0.99
```

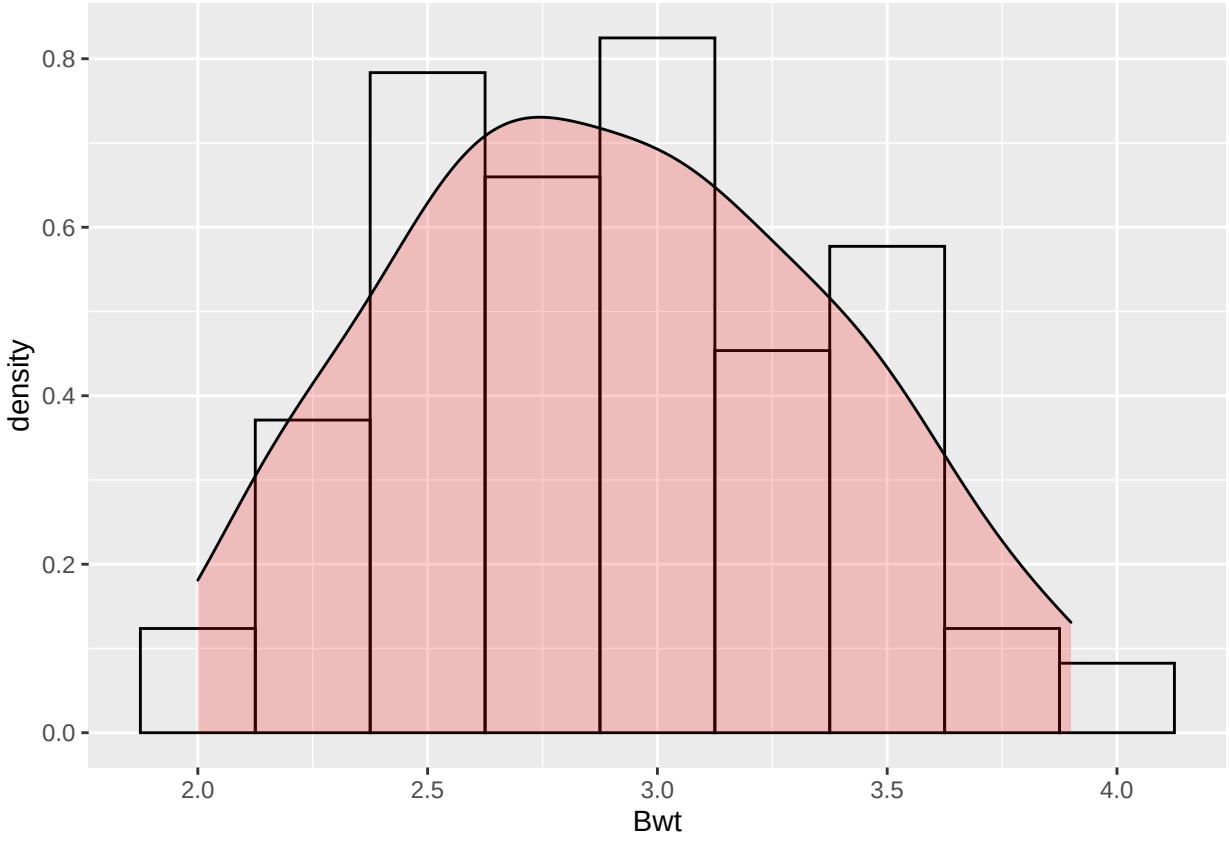
Step 4: Compare the p-value to alpha Since p-value < alpha (0.01), we can reject the NULL hypothesis and adopt our alternative hypothesis that mu > 1, i.e., the flour production company is producing whole meal flour with greater than 1 part per million copper in it.

Two-sample t-test

Use the "cats" dataset to answer the question, "do male and female cat samples have the same body weight?"

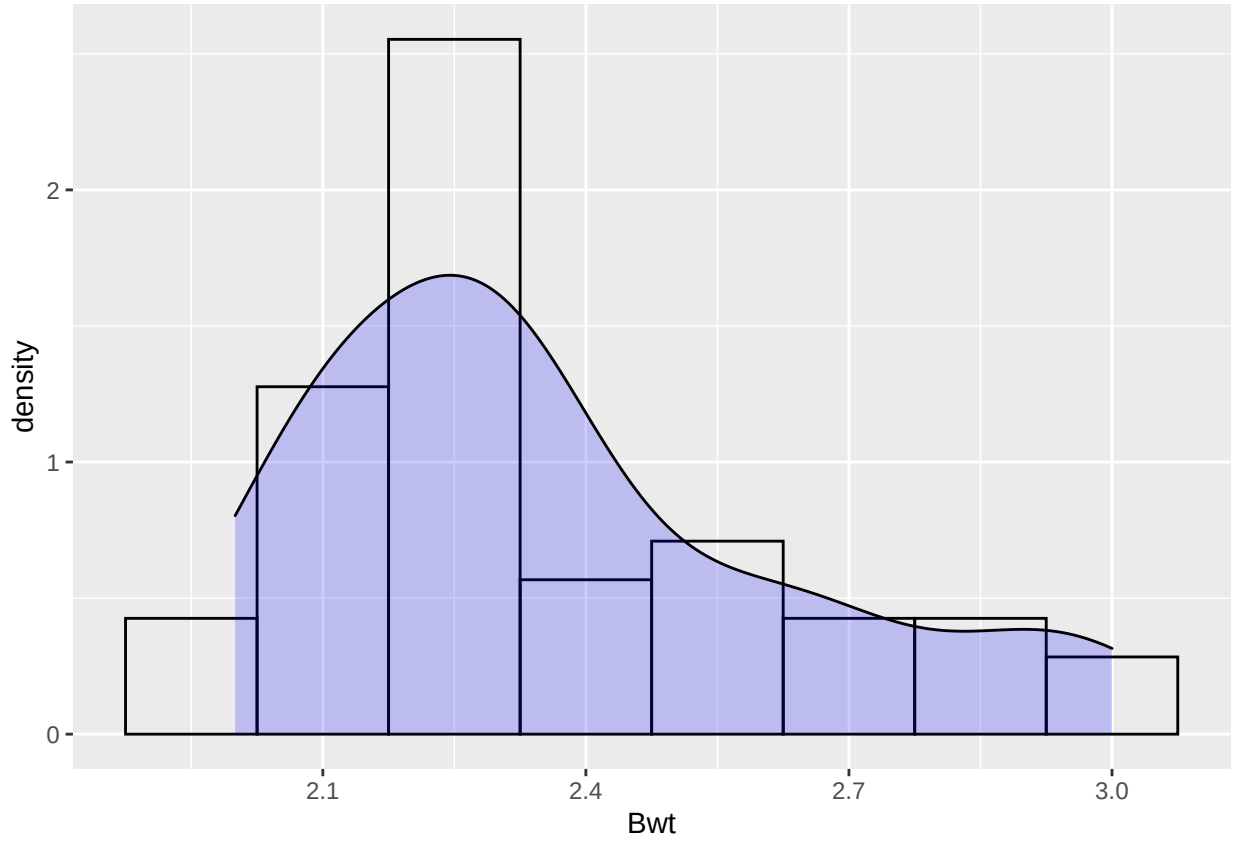
```
?cats
View(cats)
male <- subset(cats, subset=(cats$Sex=="M")) # used subset function to get separate vectors for male an
female <- subset(cats, subset=(cats$Sex=="F"))

ggplot(male,aes(x=Bwt)) +
    geom_histogram(aes(y=..density..),binwidth=0.25,colour="black", fill="light grey") +
        geom_density(alpha=.2, fill="red")
```



code:

```
ggplot(female,aes(x=Bwt)) +
    geom_histogram(aes(y=..density..),binwidth=0.15,colour="black", fill="light grey") +
        geom_density(alpha=.2, fill="blue")
```




```
male_sample <- sample(male, size = 29, replace = TRUE) #Taking sample of 29 values so that n<30 to sati
female_sample <- sample(female, size = 29, replace = TRUE)
```

As we need to find out if the bwt of male cats and female cats are same or not, we'll use a two tailed t test

Null hypothesis (H0) is as follows: Male and female cat samples have the same body weight, i.e difference of mean (mud) = 0

Alternative hypothesis (H1) is as follows: Male and female cat samples have unequal body weight, i.e difference of mean (mud) != 0

```
ttest_unpaired = t.test(male$Bwt, female$Bwt,
                        var.equal = FALSE,
                        paired = FALSE)
ttest_unpaired
```

Two sample unpaired t test with unequal variances

```
##
##  Welch Two Sample t-test
##
## data:  male$Bwt and female$Bwt
## t = 8.7095, df = 136.84, p-value = 8.831e-15
```

```
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.4177242 0.6631268
## sample estimates:
## mean of x mean of y
##  2.900000  2.359574
```

Intepretation of the results

Since p-value < alpha (0.05), we can reject the NULL hypothesis and adopt our alternative hypothesis that mud!=0, i.e., male and female cats do not have same body weights.

Paired t-test

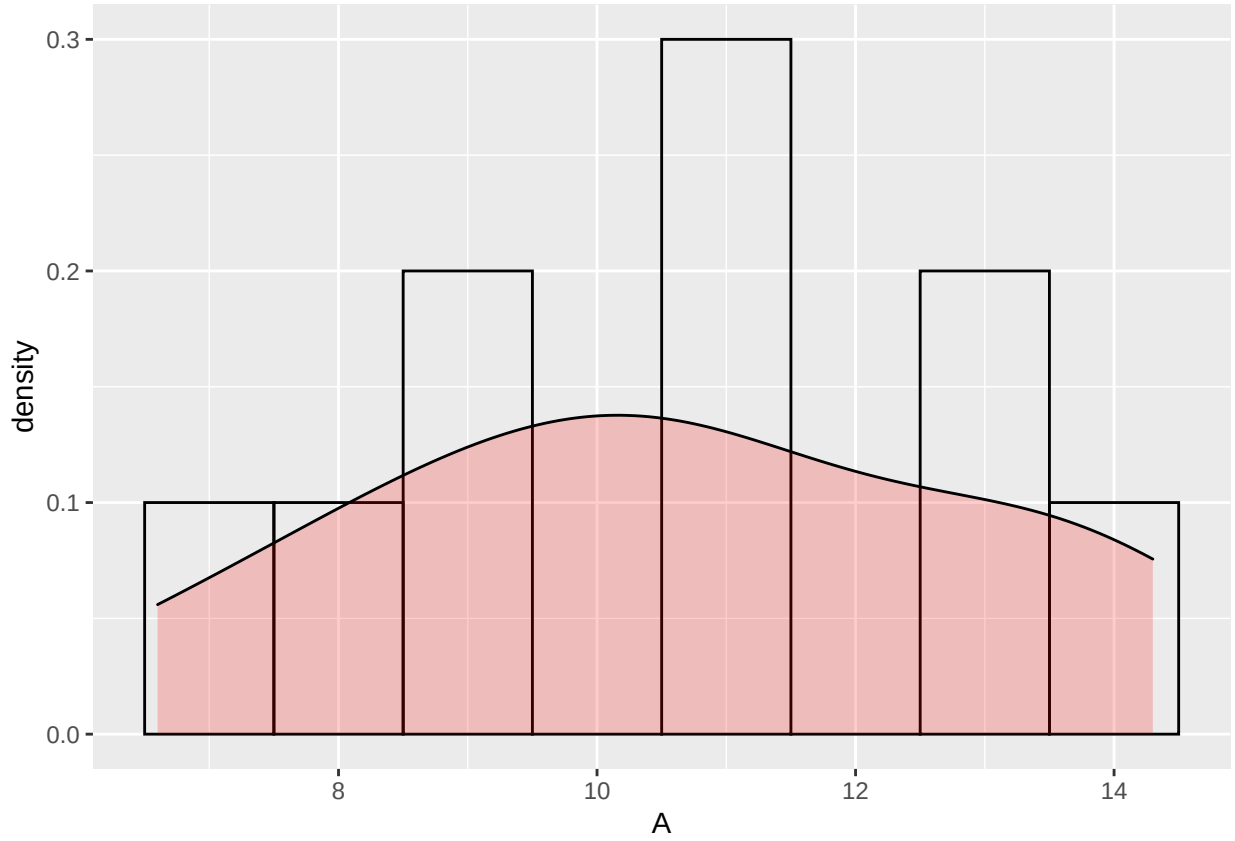
Use the "shoes" dataset to answer the question, "did material A wear better than material B?"

```
?shoes
shoes
```

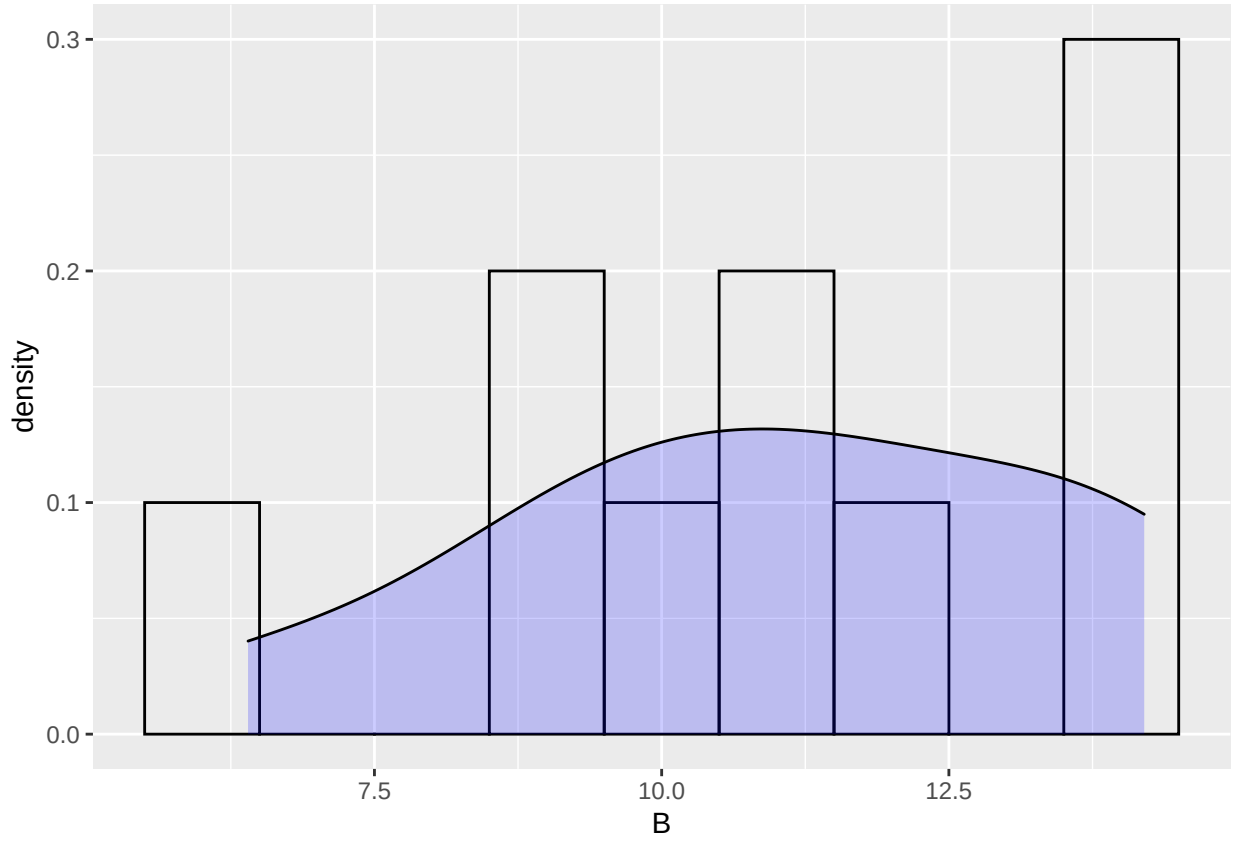
code

```
## $A
##  [1] 13.2  8.2 10.9 14.3 10.7  6.6  9.5 10.8  8.8 13.3
##
## $B
##  [1] 14.0  8.8 11.2 14.2 11.8  6.4  9.8 11.3  9.3 13.6
```

```
shoe <- data.frame(shoes)
ggplot(shoe,aes(x=A)) +
    geom_histogram(aes(y=..density..),binwidth=1,colour="black", fill="light grey") +
        geom_density(alpha=.2, fill="red")
```




```
ggplot(shoe,aes(x=B)) +
    geom_histogram(aes(y=..density..),binwidth=1,colour="black", fill="light grey") +
        geom_density(alpha=.2, fill="blue")
```



As we need to find out if the material A wears better than material B, we'll use a one - tailed two sampled paired t test

Null hypothesis (H0) is as follows: Shoe of materal A wears worse or equal to the shoe of material B
Alternative hypothesis (H1) is as follows: Shoe of materal A wears better than shoe of material B.

```
paired = t.test(shoes$A, shoes$B,
                paired=TRUE,
                alternative="greater")
paired
```

two sample paired t test

```
##
##  Paired t-test
##
## data:  shoes$A and shoes$B
## t = -3.3489, df = 9, p-value = 0.9957
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.6344264        Inf
## sample estimates:
## mean of the differences
##                   -0.41
```

Intepretation of the results

Since p-value > alpha (0.05), we fail to reject the NULL hypothesis i.e.at 95% confidence level, there is not enough evidence to deny that Shoe of materal A wears worse or equal to the shoe of material B

Independent two sample t-test

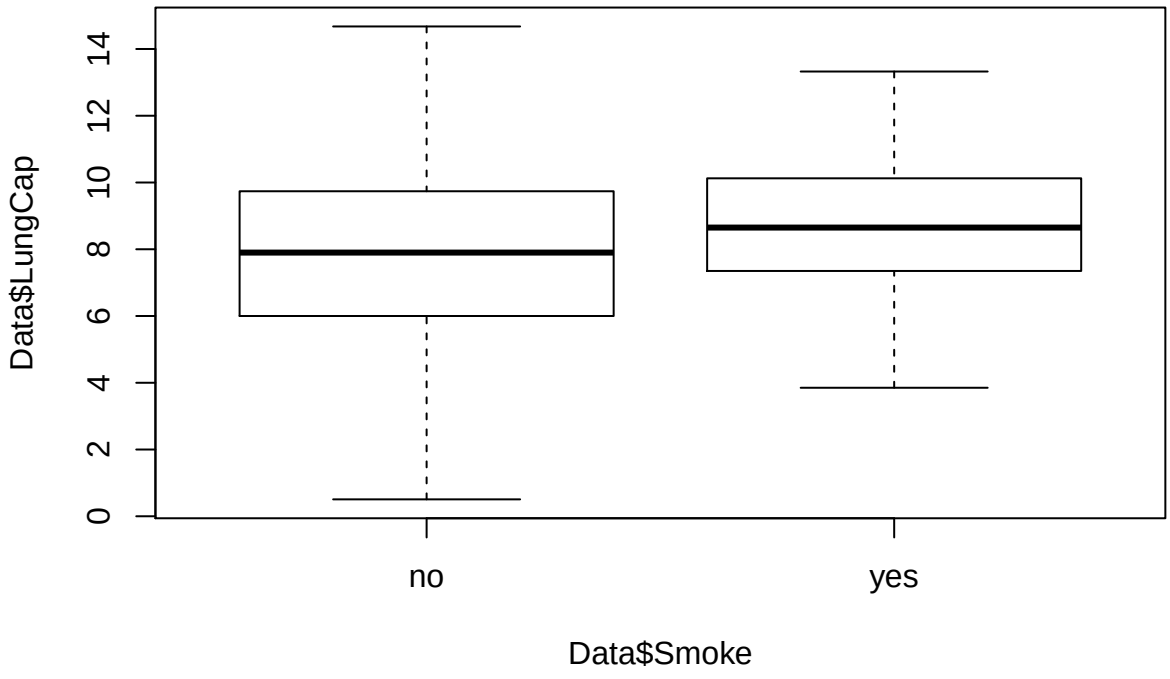
Link to the dataset lung capacity https://drive.google.com/file/d/0BxQfpNgXuWoIWUdZV1ZTc2ZscnM/edit?resourcekey=0-gqXT7Re2eUS2JGt_w1y4vA

```
Data <- read.csv(file.choose())
View(Data)
```

```
names(Data)
```

```
## [1] "LungCap"   "Age"       "Height"    "Smoke"     "Gender"    "Caesarean"
```

```
boxplot(Data$LungCap ~ Data$Smoke)
```



Ho : Mean Lung capacity of smokers = that of non smokers i.e. difference in mean = 0 Ha : Mean Lung capacity of smokers != that of non smokers i.e. difference in mean != 0

Based on the boxplot, non smokers seem to have a higher variance.

Let's confirm that by getting the variance

```r
var(Data$LungCap[Data$Smoke=="yes"])
```

```
## [1] 3.545292
```

```r
var(Data$LungCap[Data$Smoke=="no"])
```

```
## [1] 7.431694
```

```r
t.test(Data$LungCap ~ Data$Smoke)
```

Two sided t test

```
##
##  Welch Two Sample t-test
##
## data:  Data$LungCap by Data$Smoke
## t = -3.6498, df = 117.72, p-value = 0.0003927
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.3501778 -0.4003548
## sample estimates:
##  mean in group no mean in group yes
##          7.770188          8.645455
```

If we wanna test for difference in mean more than 0, we can provide that argument during the t test