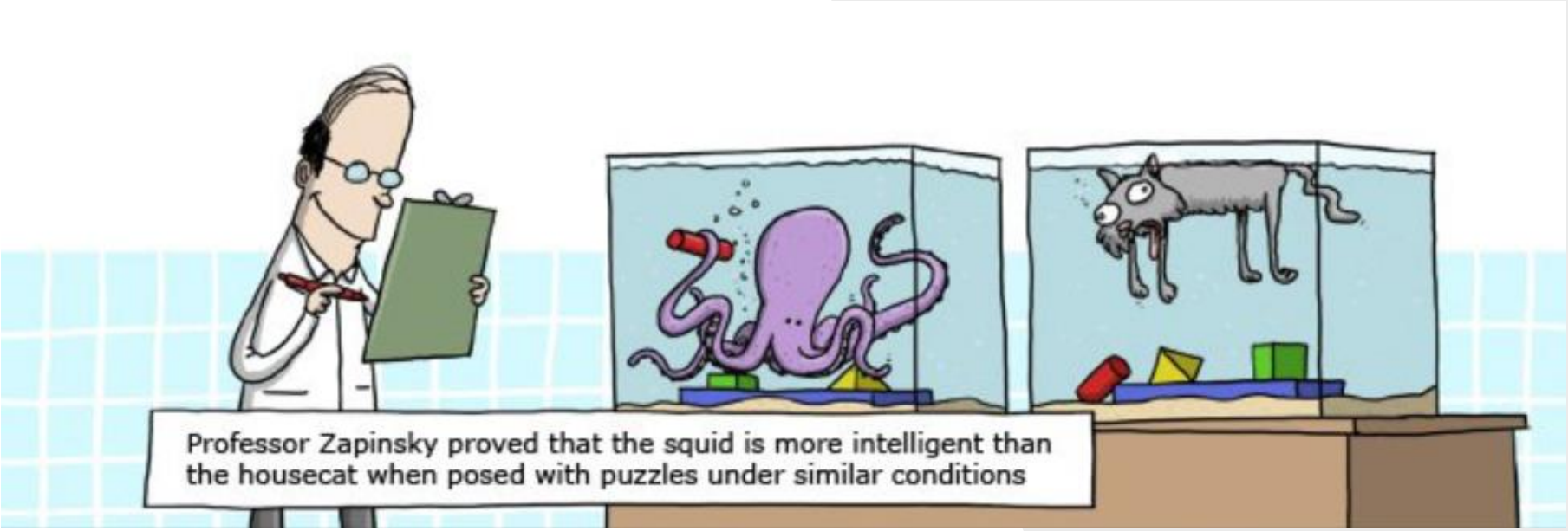
Tutorial 2

Statistics and its implementation in R



Exploratory data analysis (EDA)

- We would like to start out data analysis from exploratory analysis. This technique allows us to get the impression how does our data looks like and to act towards tiding the data and monitoring the accuracy of the analysis.
- The possible data visualization tools that allows us to perform EDA all through the analysis process include histogram, boxplots, log transforms etc.

Missing values

Scripts_and_markdowns_Bioinformatics_2021 - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Scripts_and_markdowns_Bioinformatics_2021

Tutorial_2.Rmd | R data sets

```
Data sets in package 'datasets':

AirPassengers           Monthly Airline Passenger Numbers
                        1949-1960
BJsales                 Sales Data with Leading Indicator
BJsales.lead (BJsales)
                        Sales Data with Leading Indicator
BOD                     Biochemical Oxygen Demand
CO2                     Carbon Dioxide Uptake in Grass Plants
ChickWeight             Weight versus age of chicks on different
                        diets
DNase                   Elisa assay of DNase
EuStockMarkets          Daily Closing Prices of Major European
                        Stock Indices, 1991-1998
Formaldehyde            Determination of Formaldehyde
```

Console | Terminal | Jobs

~/Bioinformatics/New_tutorials/Scripts_and_markdowns_Bioinformatics_2021/

```
> data()
>
```

Environment | History | Connections | Tutorial

Import Dataset | List

R | Global Environment

Data

iris	150 obs. of 5 variables

Files | Plots | Packages | Help | Viewer

Install | Update

Name	Description	Version
User Library		
askpass	Safe Password Entry for R, Git, and SSH	1.1
assertthat	Easy Pre and Post Assertions	0.2.1
backports	Reimplementations of Functions Introduced Since R-3.0.0	1.2.1
base64enc	Tools for base64 encoding	0.1-3
BH	Boost C++ Header Files	1.75.0-0
blob	A Simple S3 Class for Representing Vectors of Binary Data ('BLOBS')	1.2.1
broom	Convert Statistical Objects into Tidy Tibbles	0.7.5
bslib	Custom 'Bootstrap' 'Sass' Themes	0.2.4

Missing values

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

Tutorial_2_Markdown.Rmd | Untitled1*

Source on Save | Run | Source

```
1  head(airquality)
```

1:17 (Top Level) R Script

Console | Terminal | Jobs

~/Bioinformatics/New_tutorials/Tutorial 2/

```
> head(airquality)
  Ozone Solar.R Wind Temp Month Day
1    41     190  7.4   67     5   1
2    36     118  8.0   72     5   2
3    12     149 12.6   74     5   3
4    18     313 11.5   62     5   4
5    NA      NA 14.3   56     5   5
6    28      NA 14.9   66     5   6
>
```

Environment | History | Connections | Tutorial

Import Dataset | List

R | Global Environment

Files | Plots | Packages | Help | Viewer

airquality | Refresh Help Topic

R: New York Air Quality Measurements | Find in Topic

Description

Daily air quality measurements in New York, May to September 1973.

Usage

```
airquality
```

Format

A data frame with 153 observations on 6 variables.

[,1] Ozone	numeric Ozone (ppb)
[,2] Solar.R	numeric Solar R (lang)
[,3] Wind	numeric Wind (mph)
[,4] Temp	numeric Temperature (degrees F)
[,5] Month	numeric Month (1--12)
[,6] Day	numeric Day of month (1--31)

Missing values

- Missing values in a vector are denoted by the letters NA. The way different functions handle missing values varies from function to function.
- Its better to pre-process your data and select the rows you want to analyze before you proceed with the data analysis.

Missing values

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

Tutorial_2_Markdown.Rmd | Untitled1*

Source on Save | Run | Source

```
library(dplyr)
head(airquality)
#head of the airquality is piped into the filter function
#that outputs only the rows in which Solar.R is NOT NA.
#The final output is assigned into set1 variable.
set1<-head(airquality) %>% filter(!is.na(Solar.R))
set1
```

1:1 (Top Level) R Script

Console | Terminal | Jobs

~/Bioinformatics/New_tutorials/Tutorial 2/

```
  Ozone Solar.R Wind Temp Month Day
1    41     190  7.4   67     5   1
2    36     118  8.0   72     5   2
3    12     149 12.6   74     5   3
4    18     313 11.5   62     5   4
5    NA      NA 14.3   56     5   5
6    28      NA 14.9   66     5   6
> #head of the airquality is piped into the filter function
> #that outputs only the rows in which Solar.R is NOT NA.
> #The final output is assigned into set1 variable.
> set1<-head(airquality) %>% filter(!is.na(Solar.R))
> set1
  Ozone Solar.R Wind Temp Month Day
1    41     190  7.4   67     5   1
2    36     118  8.0   72     5   2
3    12     149 12.6   74     5   3
4    18     313 11.5   62     5   4
```

Environment | History | Connections | Tutorial

Import Dataset | List

R | Global Environment

Files | Plots | Packages | Help | Viewer

airquality | Refresh Help Topic

R: New York Air Quality Measurements | Find in Topic

Usage

```
airquality
```

Format

A data frame with 153 observations on 6 variables.

[,1] Ozone	numeric Ozone (ppb)
[,2] Solar.R	numeric Solar R (lang)
[,3] Wind	numeric Wind (mph)
[,4] Temp	numeric Temperature (degrees F)
[,5] Month	numeric Month (1--12)
[,6] Day	numeric Day of month (1--31)

Details

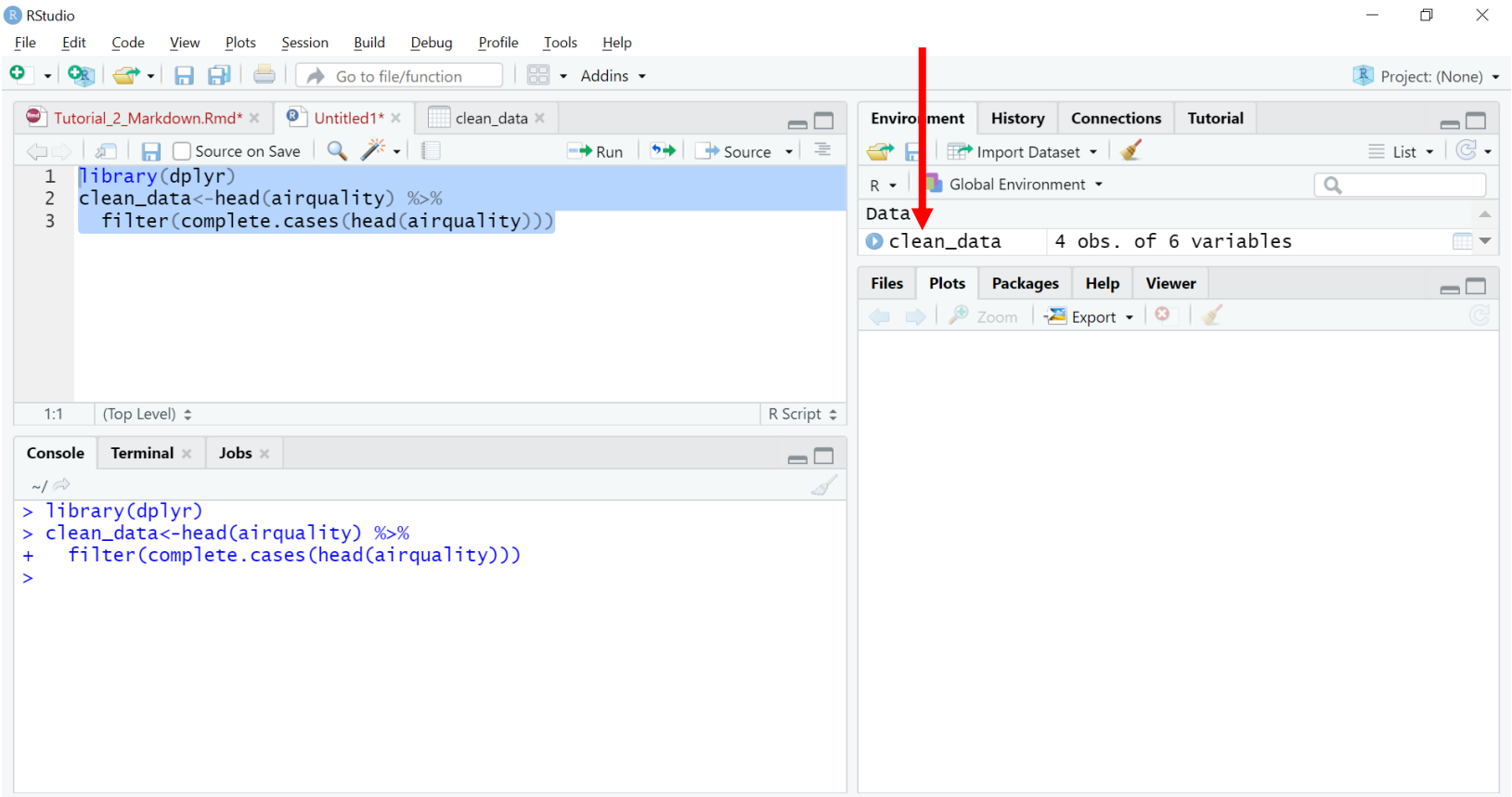
Daily readings of the following air quality values for May 1, 1973

%>% operator is defined at the dplyr package. The use of this operator is similar to the pipe use in Unix.

Missing values

- Sometimes we want to select only rows which have no missing values, so called complete cases.
- The complete.cases() function accepts a dataframe (or matrix) and tests whether each row is complete. It returns a vector with a TRUE/FALSE result for each row.

Missing values



Missing values

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

Tutorial_2_Markdown.Rmd* | Untitled1* | clean_data

Filter

	Ozone	Solar.R	Wind	Temp	Month	Day
1	41	190	7.4	67	5	1
2	36	118	8.0	72	5	2
3	12	149	12.6	74	5	3
4	18	313	11.5	62	5	4

Showing 1 to 4 of 4 entries, 6 total columns

Console | Terminal | Jobs

~/

```
> library(dplyr)
> clean_data<-head(airquality) %>%
+   filter(complete.cases(head(airquality)))
>
```

Environment | History | Connections | Tutorial

Import Dataset | List

R | Global Environment

Data

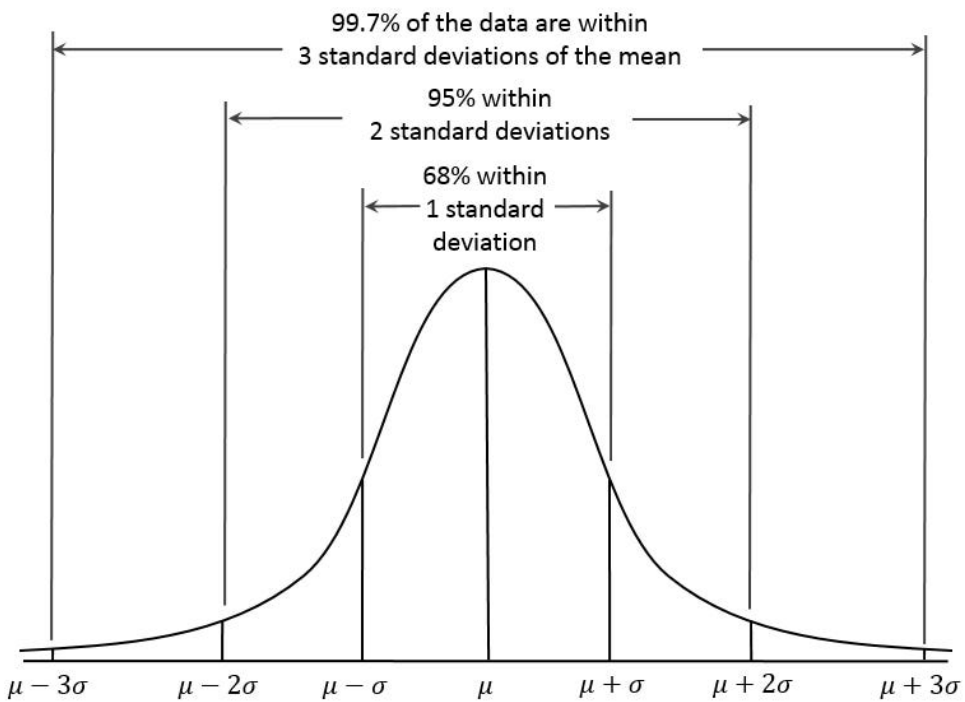
clean_data | 4 obs. of 6 variables

Files | Plots | Packages | Help | Viewer

Zoom | Export

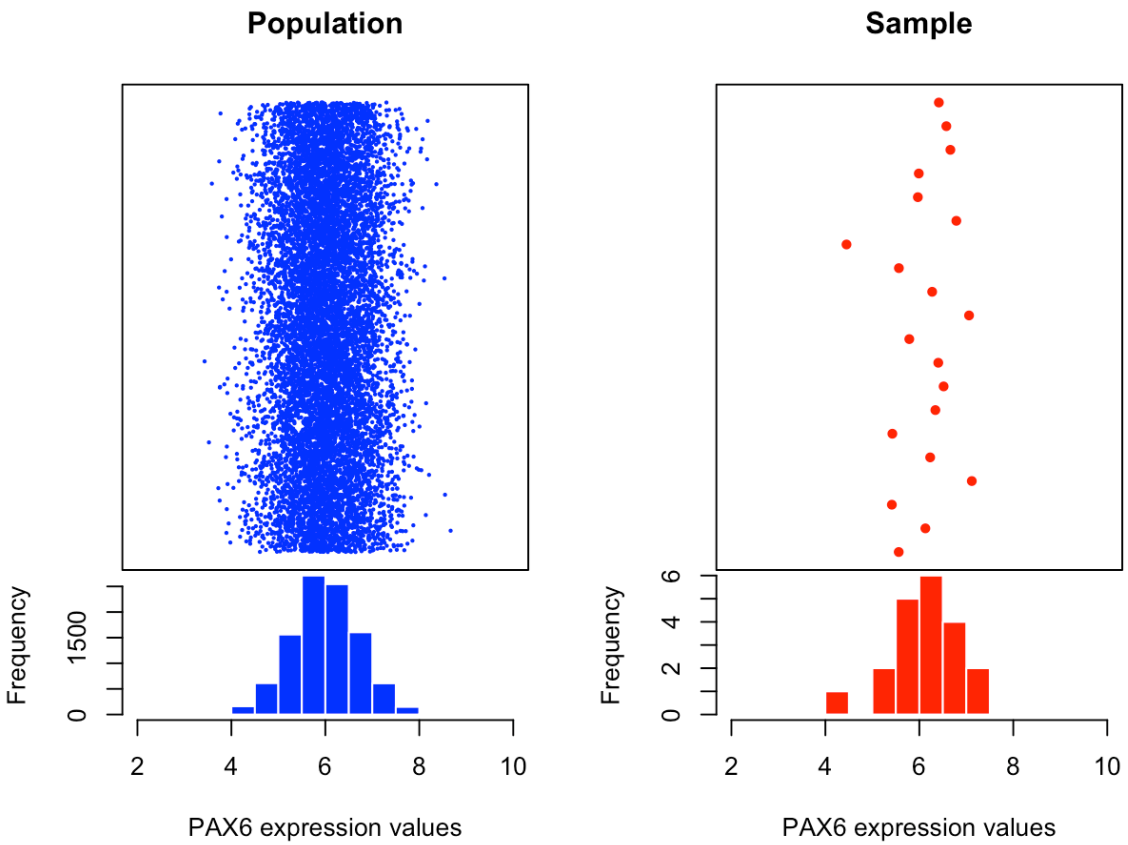
Normal (Gaussian) distribution

- Normal distribution of data can be observed in situations where data is randomly collected from independent sources.



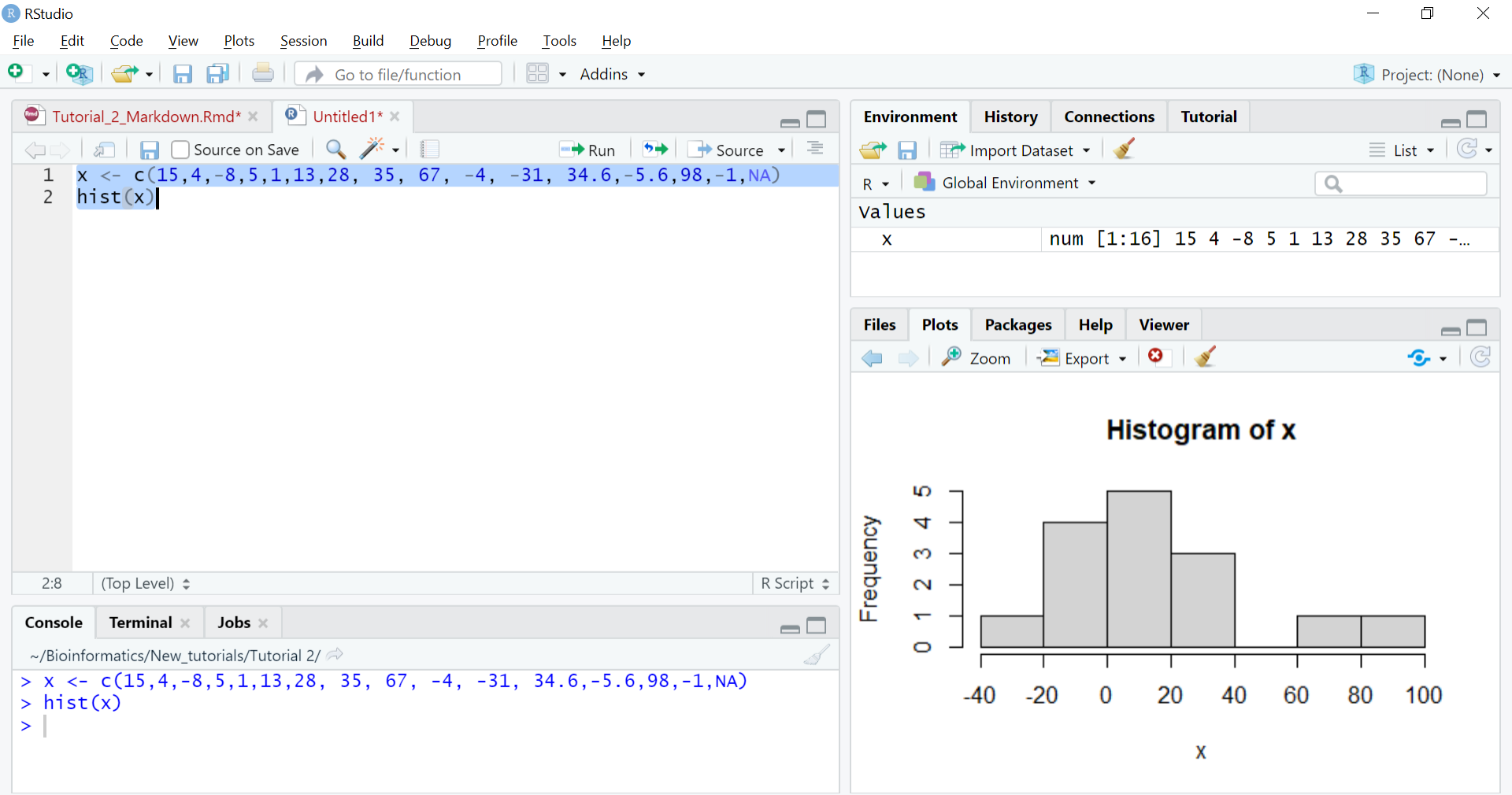
Dan Kernler, via Wikimedia Commons

Distribution of values



Histogram

The simplest display for the shape of a distribution of data can be done using a histogram- a count of how many observations fall within specified divisions ("bins") of the x-axis.



Let’s beautify the histogram a bit …

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

Tutorial_2_Markdown.Rmd* | Untitled1*

Source on Save | Run | Source

```
#Lets beautify the histogram a bit:

hist(x,)
```

3:8 (Top Level) R Script

Console | Terminal | Jobs

~/Bioinformatics/New_tutorials/Tutorial 2/

```
>
```

Environment | History | Connections | Tutorial

Import Dataset | List

R | Global Environment

Values

x	num [1:11] 15 4 -8 5 1 13 34.6 -5.6 …

Files | Plots | Packages | Help | Viewer

his | Refresh Help Topic

R: Histograms | Find in Topic

hist
hist.Date
hist.default
hist.POSIXt
history

hist {graphics} R Documentation

Histograms

Description

The generic function `hist` computes a histogram of the given data values. If `plot = TRUE`, the resulting object of class `"histogram"` is plotted by `plot.histogram`, before it is returned.

Usage

```
hist(x, ...)
```

Histogram

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

Tutorial_2_Markdown.Rmd* | Untitled1*

Source on Save | Run | Source

```
#Lets beautify the histogram a bit:

hist(x,)
```

xlim =
ylim =
xlab =
ylab =
axes =
plot =
labels =
nclass =

xlab

main title and axis labels: these arguments to `title()` get "smart" defaults here, e.g., the default `ylab` is `"Frequency"` iff `freq` is true.

Press F1 for additional help

3:8 (Top Level) R Script

Console | Terminal | Jobs

~/Bioinformatics/New_tutorials/Tutorial 2/

>

Environment | History | Connections | Tutorial

Import Dataset | List

R | Global Environment

Values

num [1:11] 15 4 -8 5 1 13 34.6 -5.6 ...

Help | Viewer

his | Refresh Help Topic

R: Histograms | Find in Topic

hist {graphics} R Documentation

Histograms

Description

The generic function `hist` computes a histogram of the given data values. If `plot = TRUE`, the resulting object of class `"histogram"` is plotted by `plot.histogram`, before it is returned.

Usage

```
hist(x, ...)
```

Histogram

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Tutorial_2_Markdown.Rmd | Untitled1*

```r
hist(x,
     col = "blue",
     breaks=20,
     xlim = c(-40,100),
     ylim =c(0,5),
     main = "Histogram of observations",
     xlab="Observation value",
     ylab = "Frequency of observations")
```

1:1 (Top Level) R Script

Console | Terminal | Jobs

~/Bioinformatics/New_tutorials/Tutorial 2/

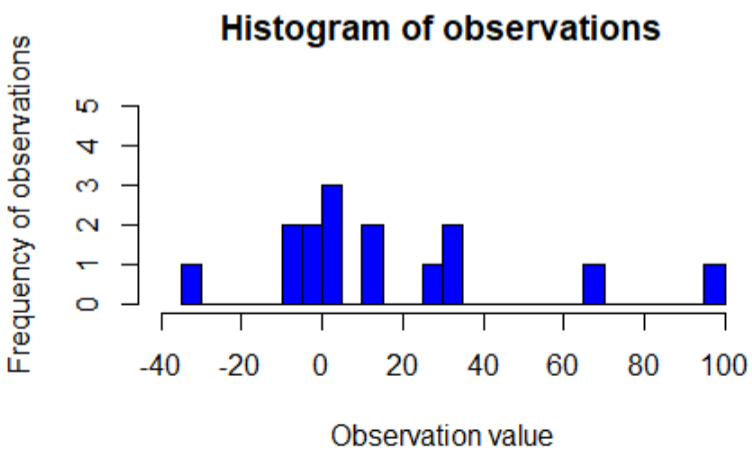
```
> hist(x,
+      col = "blue",
+      breaks=20,
+      xlim = c(-40,100),
+      ylim =c(0,5),
+      main = "Histogram of observations",
+      xlab="Observation value",
+      ylab = "Frequency of observations")
>
```

Environment | History | Connections | Tutorial

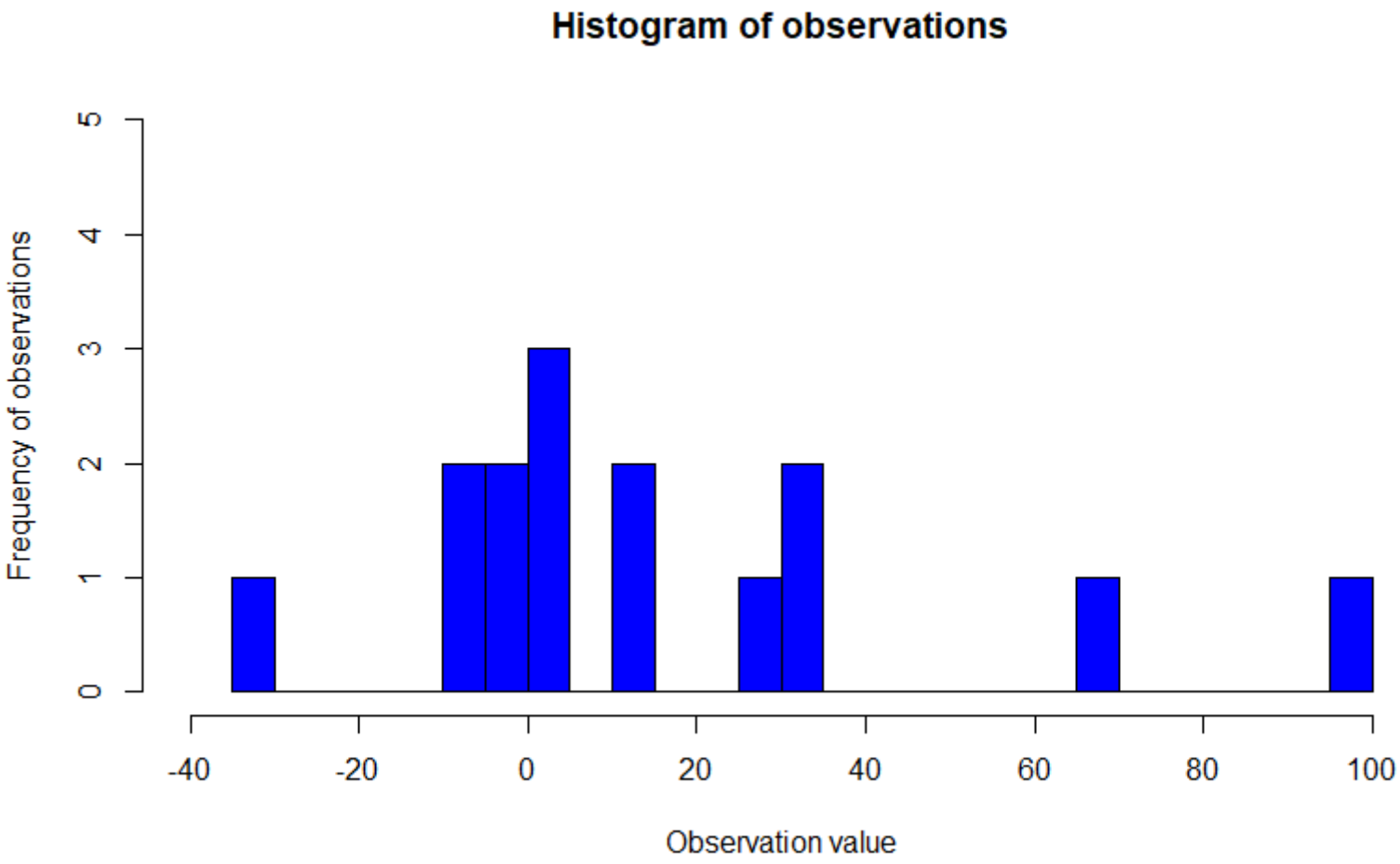
Values

x	num [1:16] 15 4 -8 5 1 13 28 35 67 -...

Files | Plots | Packages | Help | Viewer



Histogram



The distributions have parameters (such as mean and variance) that summarize them.

Mean and median are used to describe the central tendency of measurements.

- The **mean() function** is used to calculate the mean of a provided vector of numbers.
- The mean is easily affected by outliers . If certain values are very high or low compared to the bulk of the sample, this will shift mean toward those outliers.
- The **median()** function will calculate the median.
- The median is much less affected by outliers then mean as it is the value in a distribution where half of the values are above it and the other half are below.

Mean and median are used to describe the central tendency of measurements.

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

Tutorial_2_Markdown.Rmd* | Untitled1*

Source on Save | Run | Source

```r
# calculate mean from the vector with NA values
x <- c(15,4,-8,5,1,13,28, 35, 67, -4, -31, 34.6,-5.6,98,-1,NA)
mean(x)
# Drop NA values to get a result
mean(x,na.rm=T)
# calculate median
median(x,na.rm=T)
```

1:1 (Top Level) R Script

Console | Terminal | R Markdown | Jobs

~/Bioinformatics/New_tutorials/Tutorial 2/

```
> # calculate mean from the vector with NA values
> x <- c(15,4,-8,5,1,13,28, 35, 67, -4, -31, 34.6,-5.6,98,-1,NA)
> mean(x)
[1] NA
> # Drop NA values to get a result
> mean(x,na.rm=T)
[1] 16.73333
> # calculate median
> median(x,na.rm=T)
[1] 5
>
```

Environment | History | Connections | Tutorial

Import Dataset | List

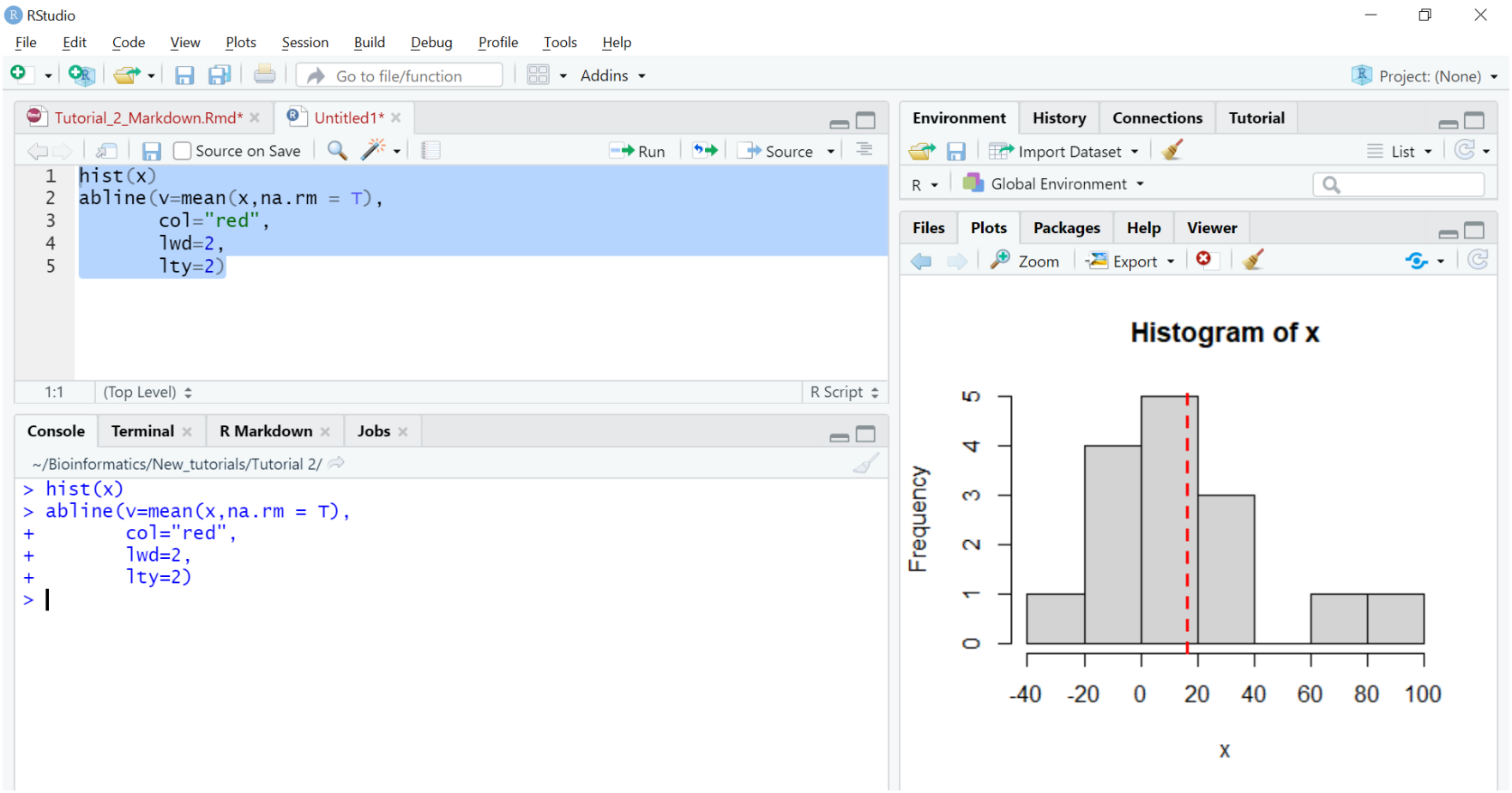
R | Global Environment

Values

x	num [1:16] 15 4 -8 5 1 13 28 35 ...

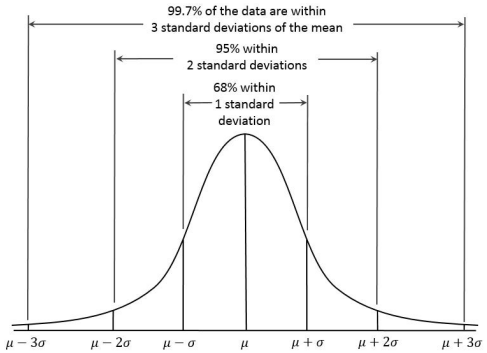
Files | Plots | Packages | Help | Viewer

Mean and median are used to describe the central tendency of measurements.

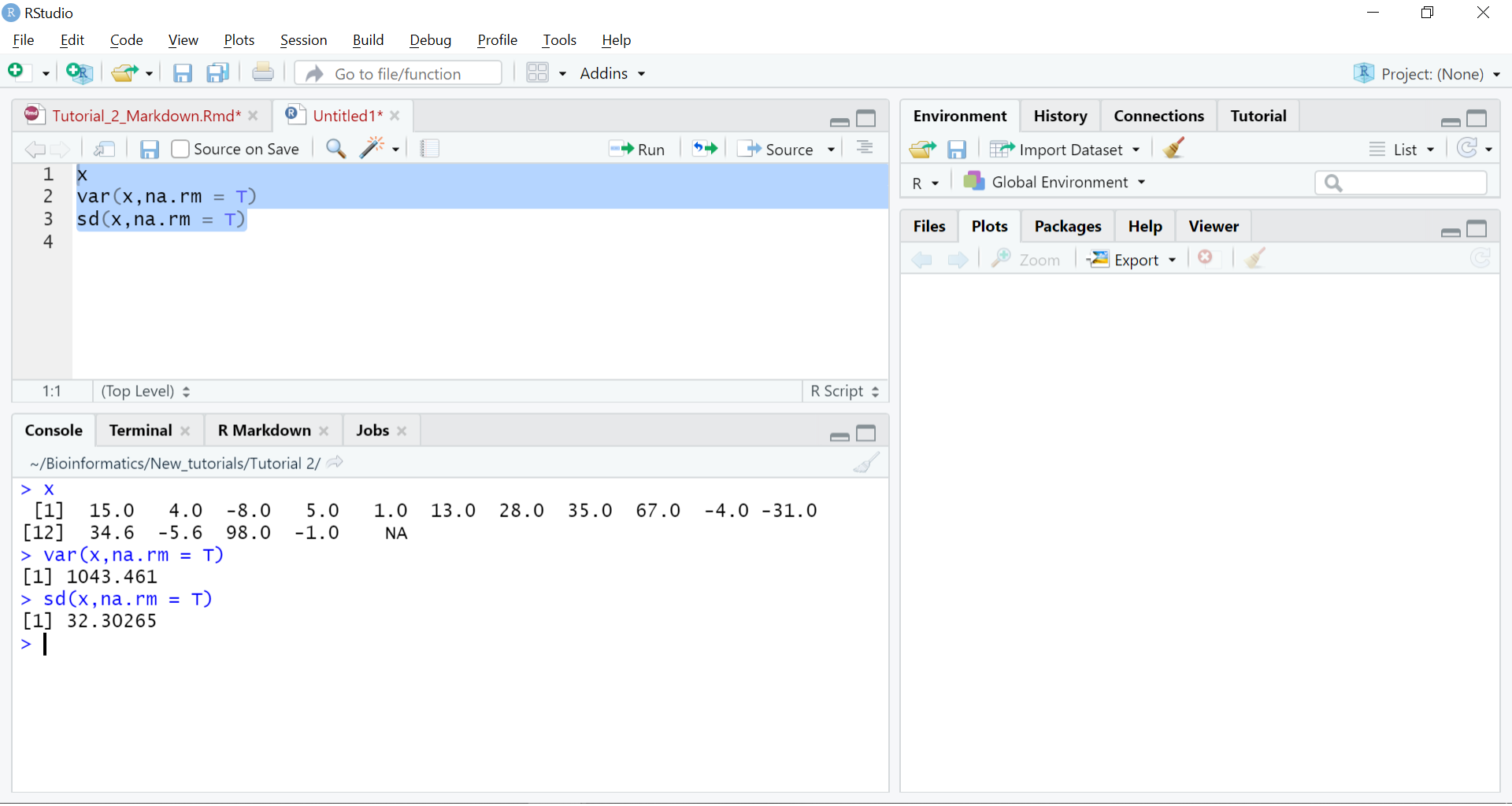


Measures that reflect variability in a distribution.

- **Variance** - squared distance of data points from the mean.
- **Standard deviation** - square root of the variance. It is measured in the same units as mean.
- A value around zero indicates there is not much variation in the values of the data points, and a high value indicates high variation in the values.



Measures that reflect variability in a distribution.



How can we test for difference between groups ?

Hypothesis testing

- Hypothesis testing is used to draw inferences about the overall data behavior performing statistical tests on a sample from the population.

How can we test for difference between groups ?

- Decide on a hypothesis to test, often called the “null hypothesis” (H_0). The null hypothesis assume that there is no anomaly in the tested population (ex. there is no difference between sets of samples).
- Decide on the “alternative hypothesis” (H_1). The alternative hypothesis that observation is due to real phenomena.
- Decide on a statistic test that will test the truth of the null hypothesis.
- Compare the result of the statistic test it to the observed value to establish significance, **the P-value**. Based on that, either reject or not reject the null hypothesis, H_0.

P-value

- To evaluate the significance of the difference between measurements in two groups we calculate the **p-value.**
- **P value can be described as** level of significance for the null hypothesis. By default, we will choose the level of significance as 5%.
- P-value is probability of the observed results assuming that there is no difference between the observations between the two samples (Ho).

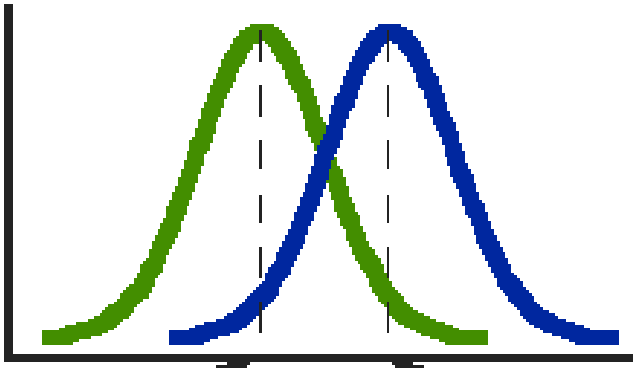
Significant p-value < 0.05**

If p-value < 0.05 ,
there is a statistically significant difference between the groups.

How can we decide whether the difference is significant ?

P-value is probability of the observed results assuming that there <u>is no difference</u> between the observations between the two samples (Ho).

If p-value < 0.05 ,
there is a statistically significant difference between the groups.



The null hypothesis is that the means (X1, X2) of two populations are equal

Which statistic test can we apply to find a p-value ?

- Multiple tests can be applied according to the type of question you are asking.
- T-tests (t.test() function) assume that a sampled population has a normal distribution and t-test generally can tolerate deviations from normality.
- Wilcoxon signed-rank test (wilcox.text() function) is used to compare paired data and as an alternative to paired t-test when the data is not normally distributed.

Let’s compare between 2 groups

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Tutorial_2_Markdown.Rmd* | Untitled1*

```r
set.seed(200)
treatment<-rnorm(30,mean = 4,sd=3)
control<-rnorm(30,mean = 3, sd=3)
obs<-mean(treatment)-mean(control)
```

2:35 (Top Level) R Script

Console | Terminal | Jobs

```
> treatment
 [1]  4.2542690  4.6793810  5.2976695  5.6741957  4.1792658  3.6560774
 [7]  0.9382650  3.1088461  4.5044501  8.2596170  3.7014248  1.5451091
[13]  2.5920933  5.7251349 -1.6152354  2.1045067  3.8726854  8.3263208
[19]  1.2373197  3.9531742  4.6583994  5.5017945  9.0080469  0.9893727
[25]  5.1825466 -0.4912178  7.7177690  4.1128666  0.8811493  4.4610025
> control
 [1] -0.1429148  5.1023438  5.2734351  0.6658121 -0.9396358  1.8861852
 [7]  2.7459195  2.2239633  3.6535719  4.1096699  5.7102985  4.1904406
[13]  2.5647460  2.6937124  2.0158888  2.8072973 -2.5943802  3.8322805
[19]  1.4882698  2.6094099  4.4589414 -0.6312660  0.3902501  4.0464665
[25]  3.0147758  6.9253995  0.1576763  2.9598471  6.5981406  2.7214632
> obs
[1] 1.249276
>
```

Environment | History | Connections | Tutorial

Import Dataset | List

R | Global Environment | treatment

Values

treatment	num [1:30] 4.25 4.68 5.3 5.67 4....

Files | Plots | Packages | Help | Viewer

Zoom | Export

Let’s compare between 2 groups

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Tutorial_2_Markdown.Rmd* | Untitled1*

```r
#build a histogram for treatment and for control measurements
hist(treatment,
     col= rgb(0,0,1,0.3),
     main = "Histogram of treatment and control values",
     xlab= "Measurement value")
hist(control,col=rgb(1,0,0,0.5),add=T)
```

7:1 (Top Level) R Script

Console | Terminal | Jobs

```
> hist(treatment,
+      col= rgb(0,0,1,0.3),
+      main = "Histogram of treatment and control values",
+      xlab= "Measurement value")
> hist(control,col=rgb(1,0,0,0.5),add=T)
>
```

Environment | History | Connections | Tutorial

Import Dataset | List

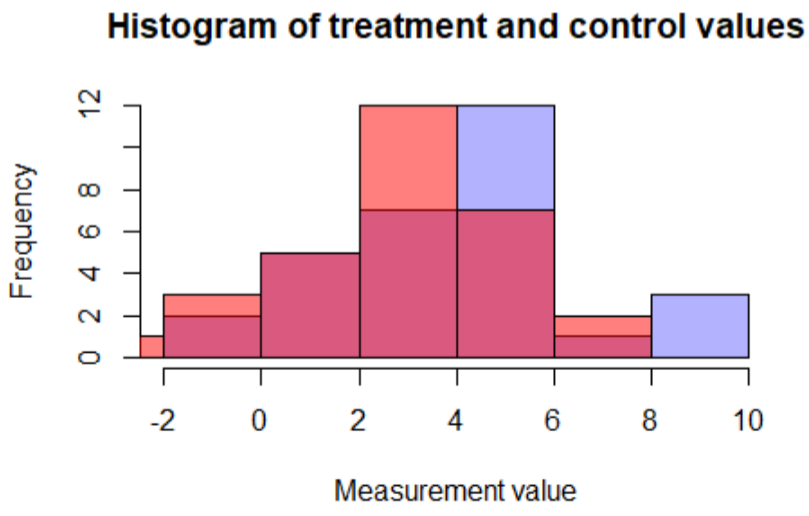
R | Global Environment | treatment

Values

treatment	num [1:30] 4.25 4.68 5.3 5.67 4.18 ...

Files | Plots | Packages | Help | Viewer

Zoom | Export



Let’s compare between 2 groups

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

Tutorial_2_Markdown.Rmd* | Untitled1*

Source on Save | Run | Source

```
#build a boxplot
study<-cbind(treatment,control)
colnames(study)<-c("treatment","control")
boxplot(study)
```

4:15 (Top Level) R Script

Console Terminal Jobs

~/

```
> study
       treatment    control
 [1,]  4.2542690 -0.1429148
 [2,]  4.6793810  5.1023438
 [3,]  5.2976695  5.2734351
 [4,]  5.6741957  0.6658121
 [5,]  4.1792658 -0.9396358
 [6,]  3.6560774  1.8861852
 [7,]  0.9382650  2.7459195
 [8,]  3.1088461  2.2239633
 [9,]  4.5044501  3.6535719
[10,]  8.2596170  4.1096699
[11,]  3.7014248  5.7102985
```

Environment History Connections Tutorial

Import Dataset | List

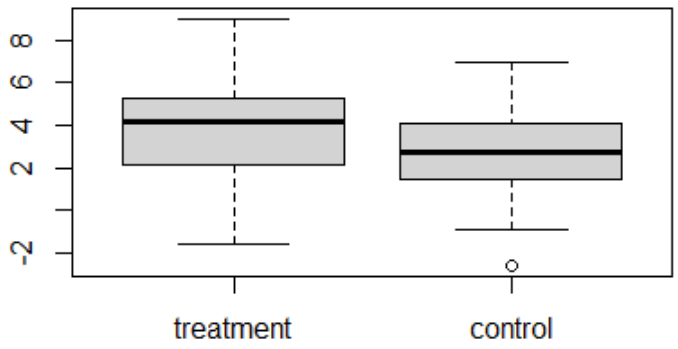
R | Global Environment | study

Data

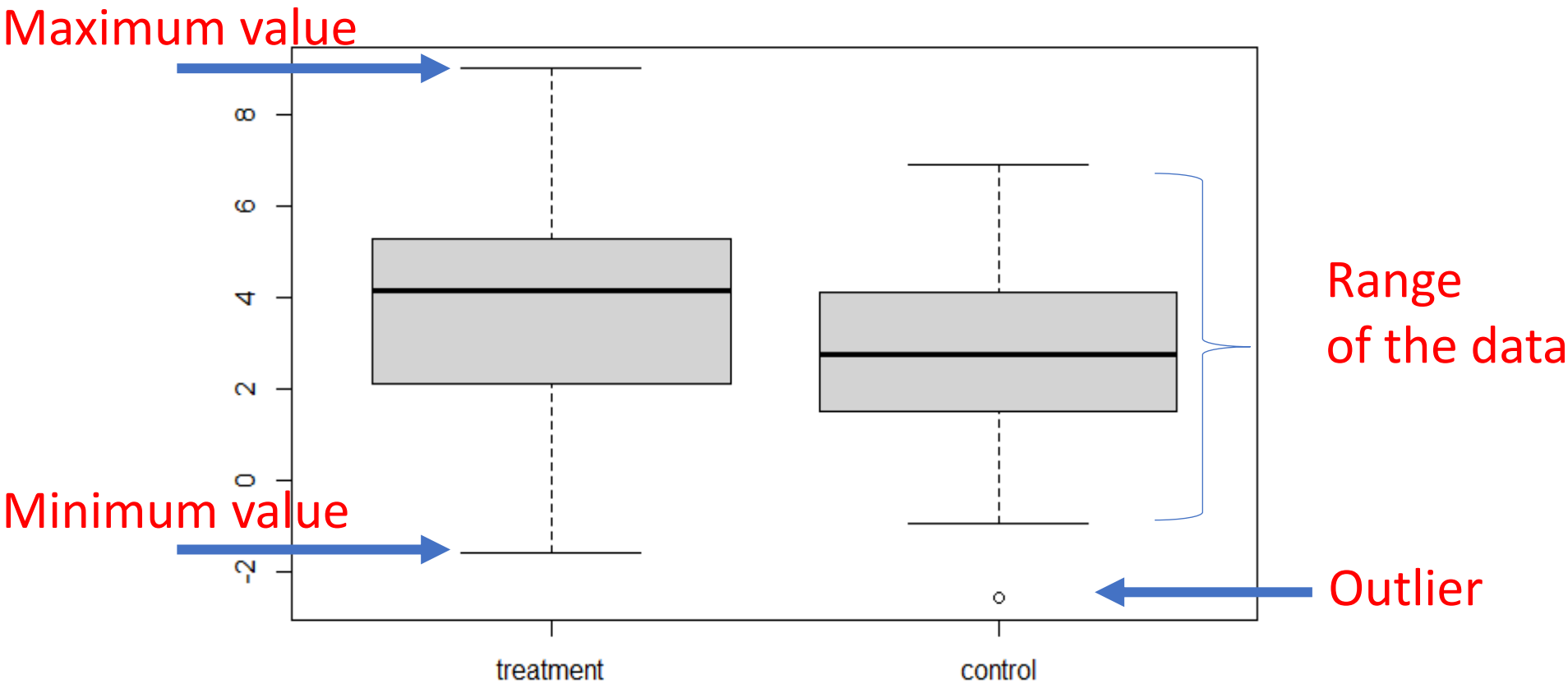
study	num [1:30, 1:2] 4.25 4.68 5.3 5.6...

Files Plots Packages Help Viewer

Zoom | Export

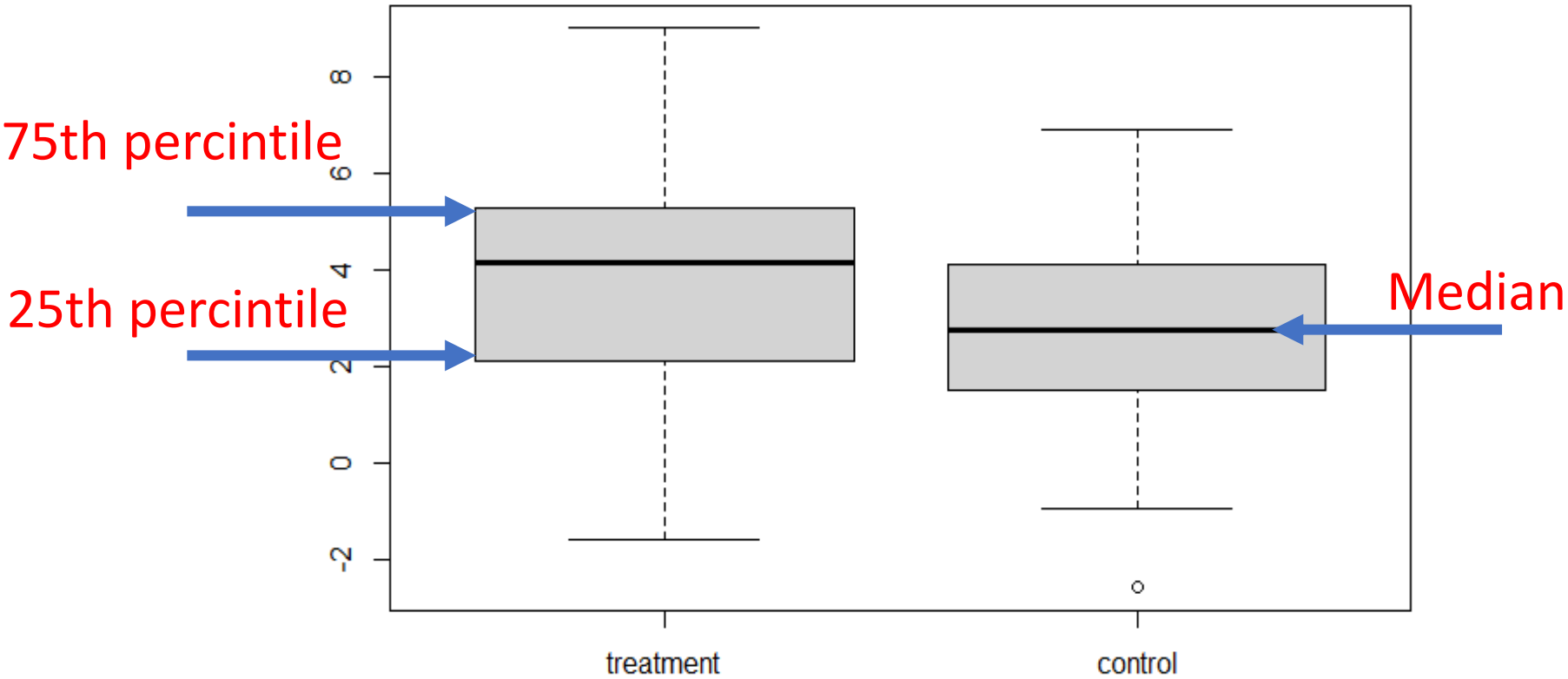


Let’s compare between 2 groups



The box represents 50% of the data

Let’s compare between 2 groups



The box represents 50% of the data.

x% of data will be less than x^{th} percentile and (100% – x%) of data will be more than x^{th} percentile.

Let’s compare between 2 groups

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Tutorial_2_Markdown.Rmd* | Untitled1* | clean_data

```r
#test whether the difference between two groups is significant
#(is it true phenomena or the difference is due to random variation)
t.test(treatment,control)
wilcox.test(treatment,control)
```

4:31 (Top Level) R Script

Console | Terminal | Jobs

```
        Welch Two Sample t-test

data:  treatment and control
t = 2.0175, df = 57.047, p-value = 0.04835
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.009358766 2.489193997
sample estimates:
mean of x mean of y
 3.933877  2.684600

> wilcox.test(treatment,control)

        Wilcoxon rank sum exact test

data:  treatment and control
W = 588, p-value = 0.04146
alternative hypothesis: true location shift is not equal to 0
```

Environment | History | Connections | Tutorial

Data

clean_data 4 obs. of 6 variables

Files | Plots | Packages | Help | Viewer

The distribution of one population is shifted to the left or right of the other, there is a difference in medians

Let’s compare between 2 groups

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Tutorial_2_Markdown.Rmd* | Untitled1* | clean_data

```r
#test whether the difference between two groups is significant
#(is it true phenomena or the difference is due to random variation)
t.test(treatment,control)
wilcox.test(treatment,control)
seed(200)
alt.treatment<-rnorm(50,mean = 4,sd=3)
t.test(alt.treatment,control)
wilcox.test(alt.treatment,control,paired = T)
```

5:1 (Top Level) R Script

Console | Terminal | Jobs

```
        Welch Two Sample t-test

data:  alt.treatment and control
t = 2.8873, df = 75.158, p-value = 0.005071
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.5359905 2.9210986
sample estimates:
mean of x mean of y
 4.413145  2.684600

> wilcox.test(alt.treatment,control,paired = T)
Error in wilcox.test.default(alt.treatment, control, paired = T) :
  'x' and 'y' must have the same length
>
```

Environment | History | Connections | Tutorial

Import Dataset | List

R | Global Environment

alt.treatme...	num [1:50] 5.459 0.369 1.39 5.0...
before	num [1:7] 200 191 193 213 241 ...

Files | Plots | Packages | Help | Viewer

Zoom | Export

Multiple testing correction

Significant p-value < 0.05 ***

***** For real data we must correct the p-value for multiple testing.**

Why?

Because the hypothesis testing is not error-free method of making decision. As more testing you are doing the probability to get a significant p-value by chance grows.

Multiple testing correction

- False positives (Type I error) – are “false discoveries” due to fact that we accept H_1 although we shouldn’t.
- False negatives (Type II error) – are cases in which can fail to accept the H_1 hypothesis when we should, meaning that we miss “true discoveries” by accepting Ho.

We expect to make more type I errors as the number of tests increase, which means we will reject the null hypothesis and accept H_1 by mistake.

Multiple testing correction

- We perform the p-value adjustment to multiple testing by p.adjust() function. Given a set of p-values, returns p-values adjusted using one of several methods.
- Which method to choose ?

Multiple testing correction

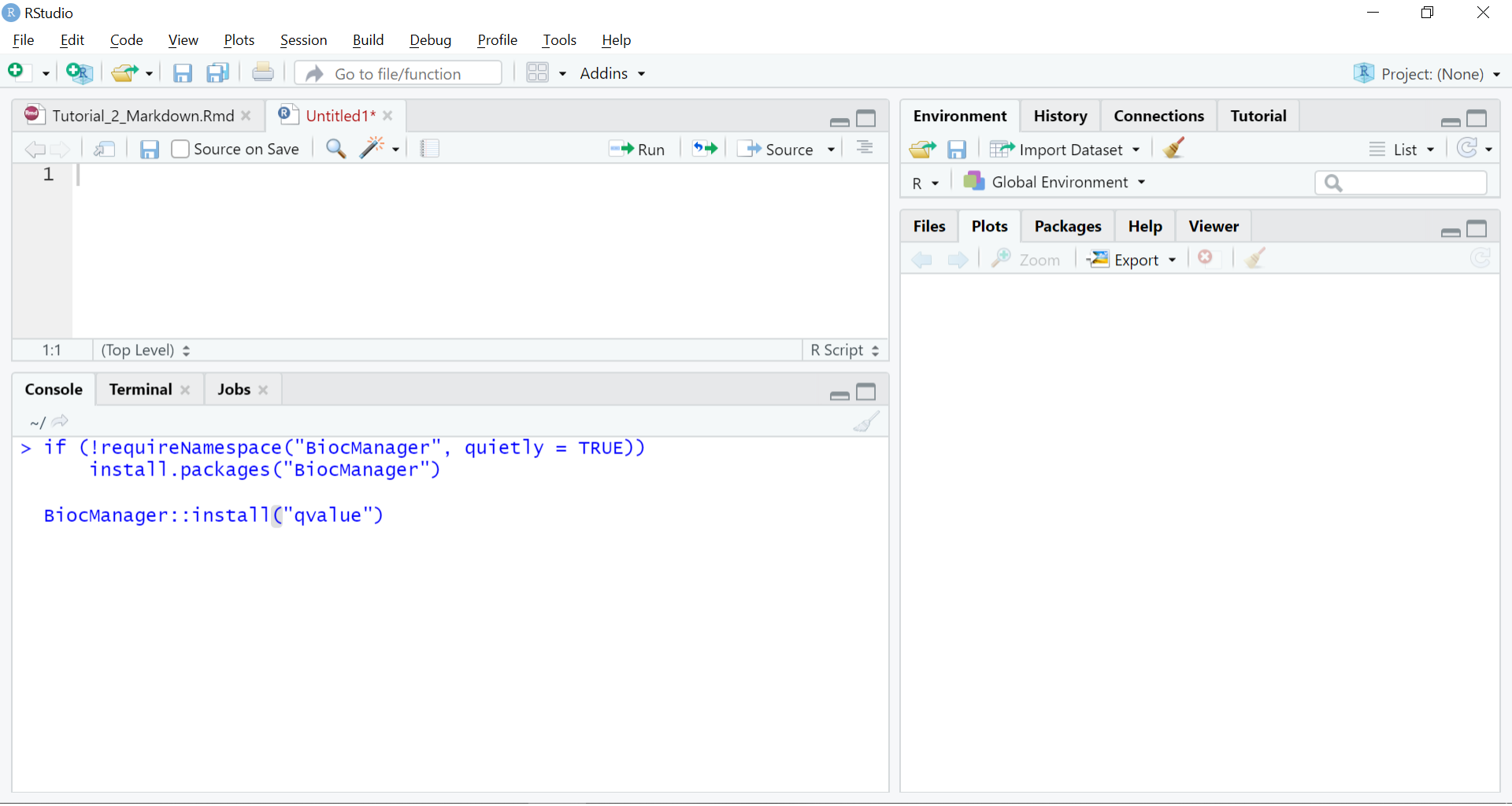
- If you concern about making even a single type 1 error, you can use **Bonferroni correction**. Using this correction you increase the type 2 error rate: you will falsely accept the null hypothesis more frequently, and so neglect findings that might be relevant.

- If your main concern is about the false positives you make, and you think it's important to you to avoid rejecting true H_1 hypothesis, you should apply **Benjamini and Hochberg (BH)** method which is also called **FDR correction**. This method allows you to control the proportion of the false discoveries.

Multiple testing correction

- Another method of multiple testing correction is calculation of **q-value**. Q-value is the proportion of significant results that turn out to be false . A q-value 0.01 would mean 1% of the tests called significant at this level will be truly null. Although they can be calculated differently the q-value and FDR adjusted P-value are synonymous within the genomics community.
- q-value is calculated using the qvalue package from Bioconductor.

qvalue package installation



https://bioconductor.org/packages/release/bioc/html/qvalue.html

Multiple testing correction

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Tutorial_2_Markdown.Rmd | Untitled1* | hedenfalk | hedenfalk

```r
library(qvalue)
#P-values and test-statistics from the Hedenfalk et al. (2001)
#gene expression dataset
data(hedenfalk)
p<-head(hedenfalk$p,10)
p
#calculate corrected p-values
bonferroni<-p.adjust(p,method = "bonferroni")
FDR<-p.adjust(p, method = "fdr")
q<-qvalue(p)
```

Console

```
> library(qvalue)
> #P-values and test-statistics from the Hedenfalk et al. (2001)
> #gene expression dataset
> data(hedenfalk)
> p<-head(hedenfalk$p,10)
> p
 [1] 0.0121261830 0.0750252366 0.9949211356 0.0417854890 0.8458138801
 [6] 0.2519242902 0.6586561514 0.0656813880 0.1232681388 0.0007129338
>
```

Environment | History | Connections | Tutorial

Data

hedenfalk	Large list (3 elements, 2.6 M...

Multiple testing correction

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Tutorial_2_Markdown.Rmd | Untitled1* | hedenfalk | hedenfalk

```r
library(qvalue)
#P-values and test-statistics from the Hedenfalk et al. (2001)
#gene expression dataset
data(hedenfalk)
p<-head(hedenfalk$p,10)
p
#calculate corrected p-values
bonferroni<-p.adjust(p,method = "bonferroni")
FDR<-p.adjust(p, method = "fdr")
q<-qvalue(p)
```

8:1 (Top Level) R Script

Console | Terminal | Jobs

```
> bonferroni<-p.adjust(p,method = "bonferroni")
> FDR<-p.adjust(p, method = "fdr")
> q<-qvalue(p)
> p
 [1] 0.0121261830 0.0750252366 0.9949211356 0.0417854890 0.8458138801
 [6] 0.2519242902 0.6586561514 0.0656813880 0.1232681388 0.0007129338
> bonferroni
 [1] 0.121261830 0.750252366 1.000000000 0.417854890 1.000000000
 [6] 1.000000000 1.000000000 0.656813880 1.000000000 0.007129338
> FDR
 [1] 0.060630915 0.150050473 0.994921136 0.139284963 0.939793200
 [6] 0.359891843 0.823320189 0.150050473 0.205446898 0.007129338
>
```

Environment | History | Connections | Tutorial

Import Dataset | List

R | Global Environment | hedenfalk

Data

hedenfalk	Large list (3 elements, 2.6 M...

Files | Plots | Packages | Help | Viewer

Zoom | Export

Multiple testing correction

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Tutorial_2_Markdown.Rmd | Untitled1* | hedenfalk | hedenfalk

```r
library(qvalue)
#P-values and test-statistics from the Hedenfalk et al. (2001)
#gene expression dataset
data(hedenfalk)
p<-head(hedenfalk$p,10)
p
#calculate corrected p-values
bonferroni<-p.adjust(p,method = "bonferroni")
FDR<-p.adjust(p, method = "fdr")
q<-qvalue(p)
```

8:1 (Top Level) R Script

Environment | History | Connections | Tutorial

Data

hedenfalk Large list (3 elements, 2.6 M...

Console | Terminal | Jobs

```
> p
 [1] 0.0121261830 0.0750252366 0.9949211356 0.0417854890 0.8458138801
 [6] 0.2519242902 0.6586561514 0.0656813880 0.1232681388 0.0007129338
> q
$call
qvalue(p = p)

$pi0
[1] 1

$qvalues
 [1] 0.060630915 0.150050473 0.994921136 0.139284963 0.939793200
 [6] 0.359891843 0.823320189 0.150050473 0.205446898 0.007129338
```

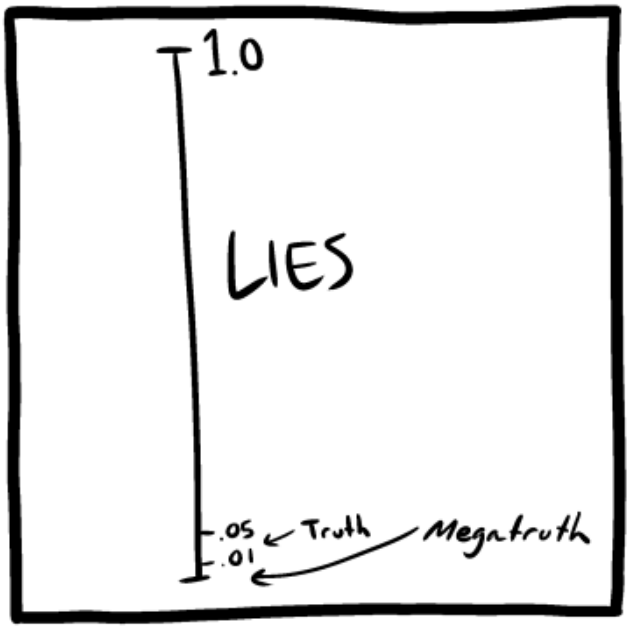
Files | Plots | Packages | Help | Viewer

Mutiple testing correction

- You should choose a method familiar in your field of study. You should apply common sense and choose how you balance the probability of making a type I error relative to a type II error.
- In a preliminary study, you'll probably want to keep as many significant values as possible to not exclude potentially significant factors from future studies.
- In a clinical study that can decide the human lives fate you'd want to have a very high level of certainty before concluding that one treatment is better than another and you will use more stringent multiple testing correction, even in the price of false negatives.

Beware of p-haking !

- *"P-haking is the misuse of data analysis to find patterns in data that can be presented as statistically significant when in fact there is no real underlying effect" — [Wikipedia]*

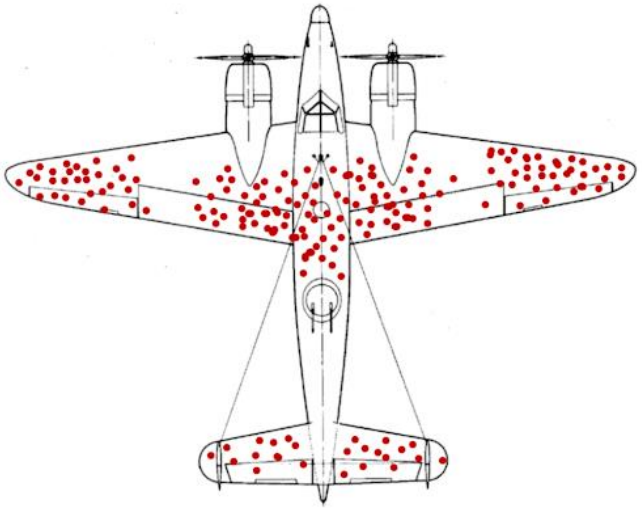


https://www.smbc-comics.com/index.php

“If you’ll torture the data long enough, it will confess to anything” (C. Ronald H. Coase)

Cherry picking

Survivorship bias



You are cordially invited to read more about statistical fallacies you must avoid at: https://www.geckoboard.com/best-practice/statistical-fallacies/

https://www.pinterest.com/pin/240238961352821718/

A kind reminder:
Use your resources wisely.

- https://cran.r-project.org/doc/contrib/Short-refcard.pdf
- https://stackoverflow.com/
- https://stats.stackexchange.com/
- https://community.rstudio.com/

- What statistical analysis should I use?
- https://stats.idre.ucla.edu/r/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-r/

Whats next?

- Next lesson we will dive into the DNA sequence analysis.