Statistical Methodology for Software Engineers

Hadas Lapid, PhD

Resources

- All course materials on Moodle
- “An Introduction to Statistical Methods and Data Analysis” by R. Lyman Ott and Michael Longnecker (on moodle)
- R Software: https://rstudio.com/products/rstudio/download/#download
- Web-textbook: http://www.statsoft.com/Textbook

Contents Today

- **Software installation**
- **Software ide introduction**
- **Basic concepts – standard functions checkout**
- **Probability distribution plots**
- **Sample data frequency plots**

Software Installation

Go to https://rstudio.com/products/rstudio/download/

RStudio Desktop

Open Source License

Free

Choose the Rstudio Desktop option

DOWNLOAD

Learn more

1. Install R. RStudio requires R 3.0.1+.

Download R interpreter from link: https://cran.rstudio.com/

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux
- Download R for (Mac) OS X
- Download R for Windows

Which leads you here

../software installation

Which leads you here

R for Windows

Subdirectories:

base	Binaries for base distribution. This is what you want to **install R for the first time**.
contrib	Binaries of contributed CRAN packages (for R >= 2.13.x; managed by Uwe Ligges). There is also information on third party software available for CRAN Windows services and corresponding environment and make variables.
old contrib	Binaries of contributed CRAN packages for outdated versions of R (for R < 2.13.x; managed by Uwe Ligges).
Rtools	Tools to build R and R packages. This is what you want to build your own packages on Windows, or to build R itself.

R-3.6.2 for Windows (32/64 bit)

And finally here

Download R 3.6.2 for Windows (83 megabytes, 32/64 bit)

Installation and other instructions

New features in this version

After clicking download starts

R-3.6.2-win.exe

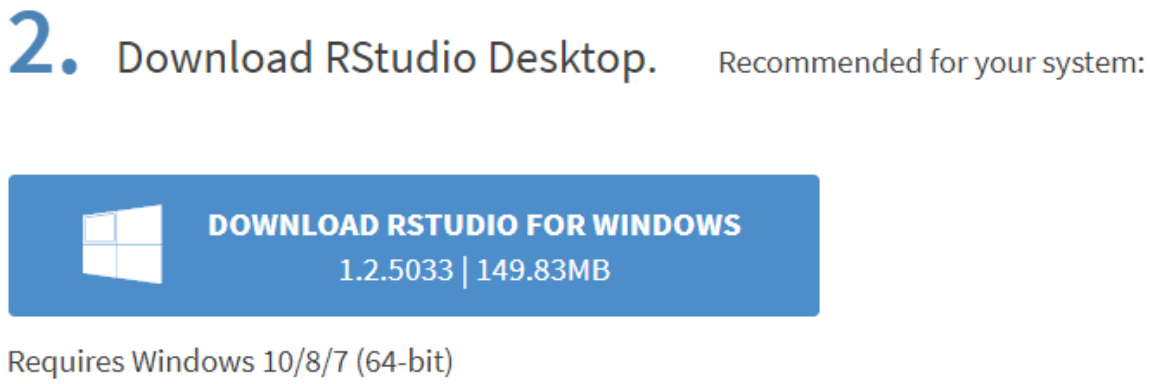
28.7/82.4 MB, 13 secs left

../software installation

After R installation (v3.6.2 as of Spring semester 2020)
Install the integrated development environment (ide)

Go back to
https://rstudio.com/products/rstudio/download/#download

and click on



After clicking download starts

Follow installation instructions till done

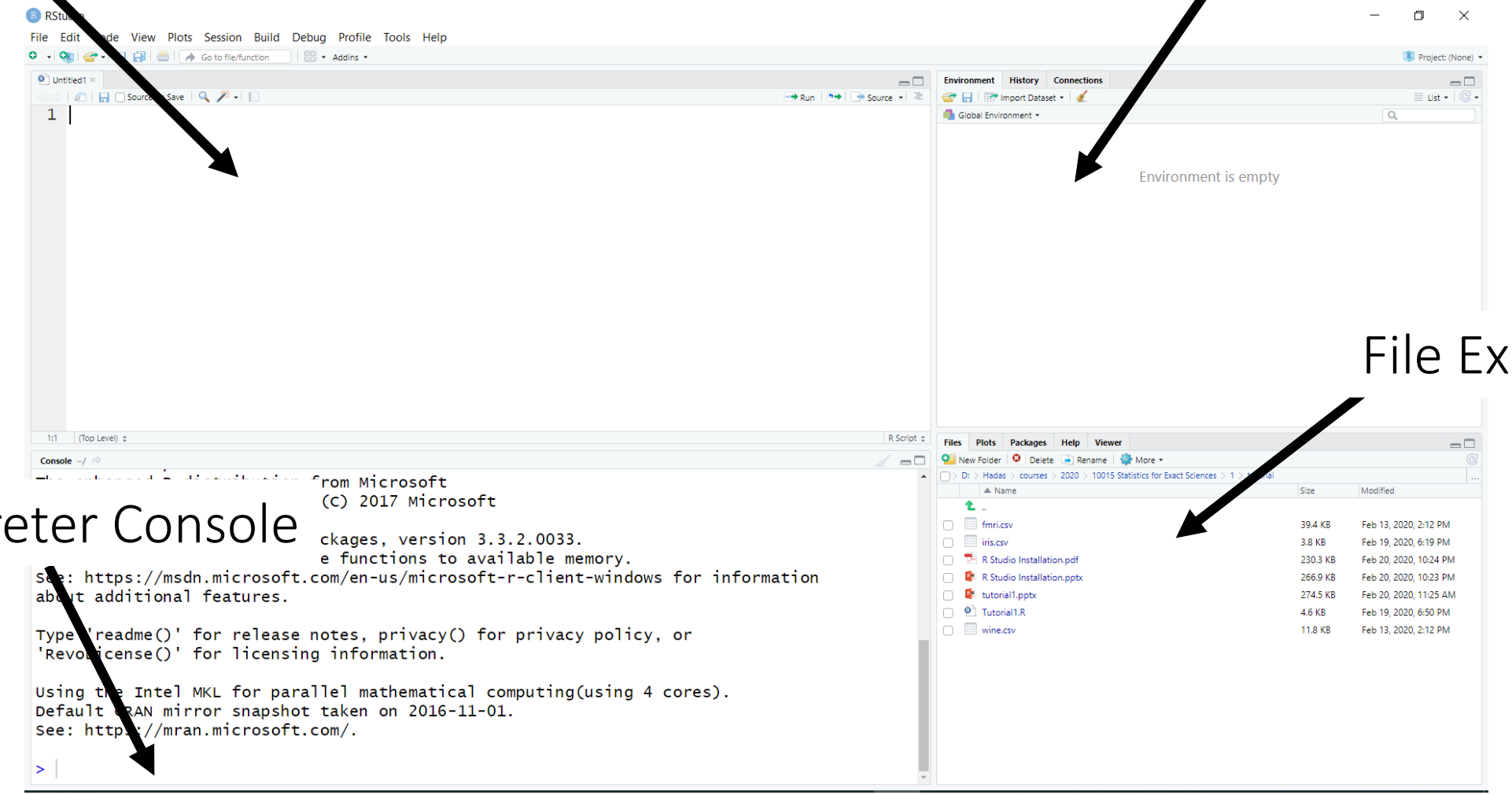
ide introduction

Editor

Environment Variables

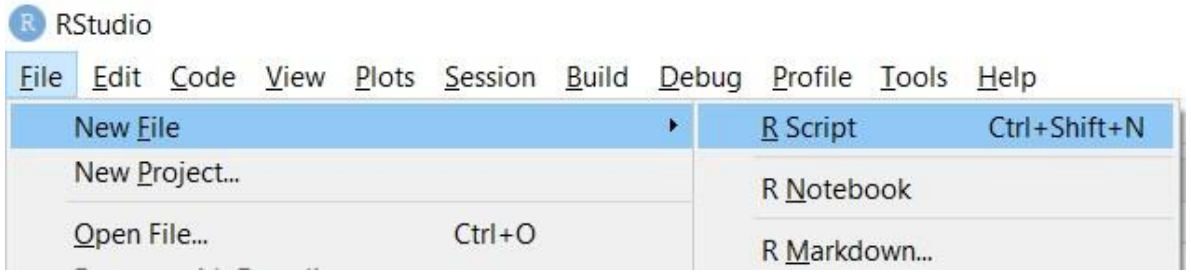
File Explorer

Interpreter Console



ide introduction

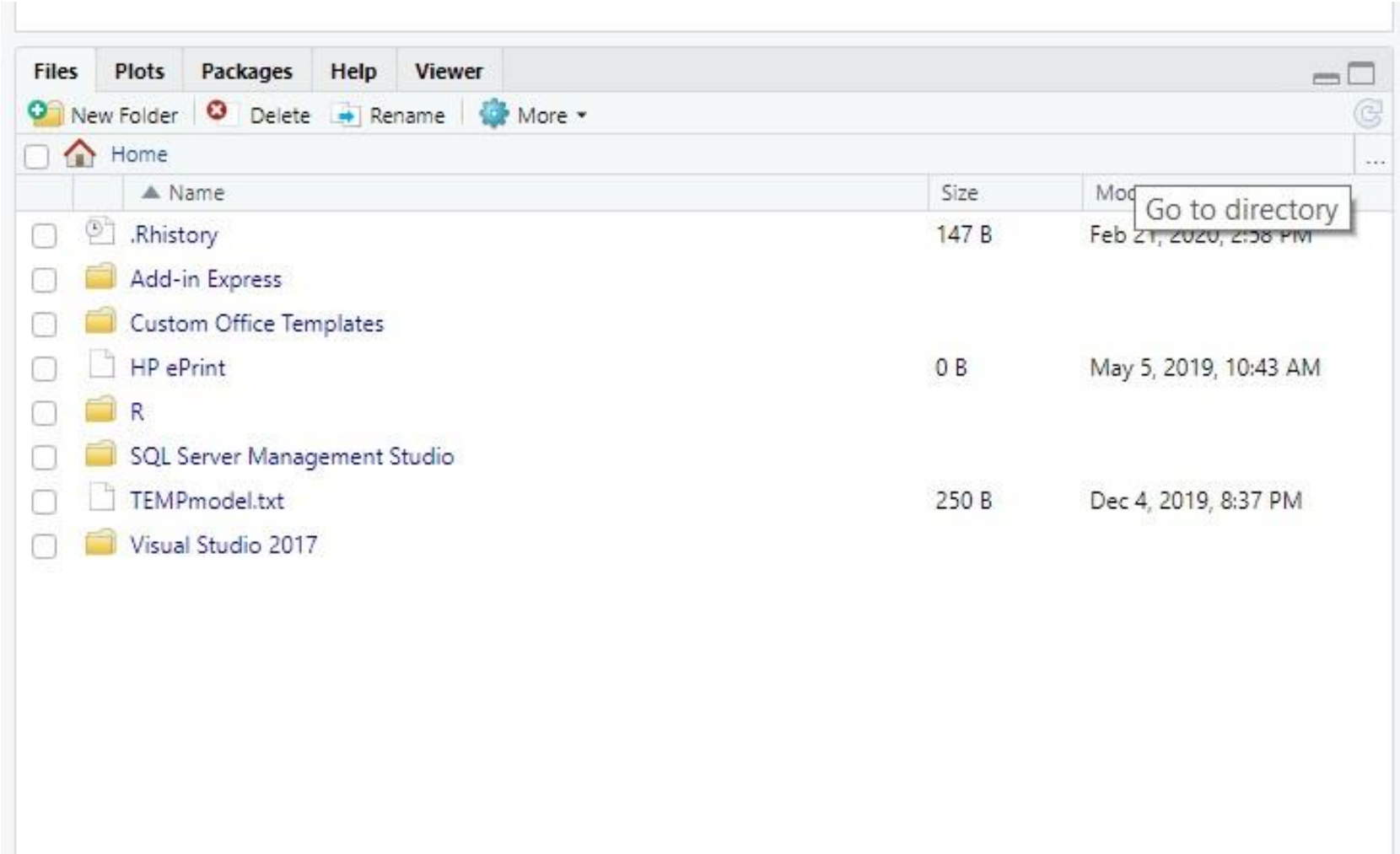
Opening a new script



Running a script or a line/section

ide introduction

Change current working directory



ide introduction

Getting formal help

```
> help("hist")
```

Files | Plots | Packages | Help | Viewer

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.

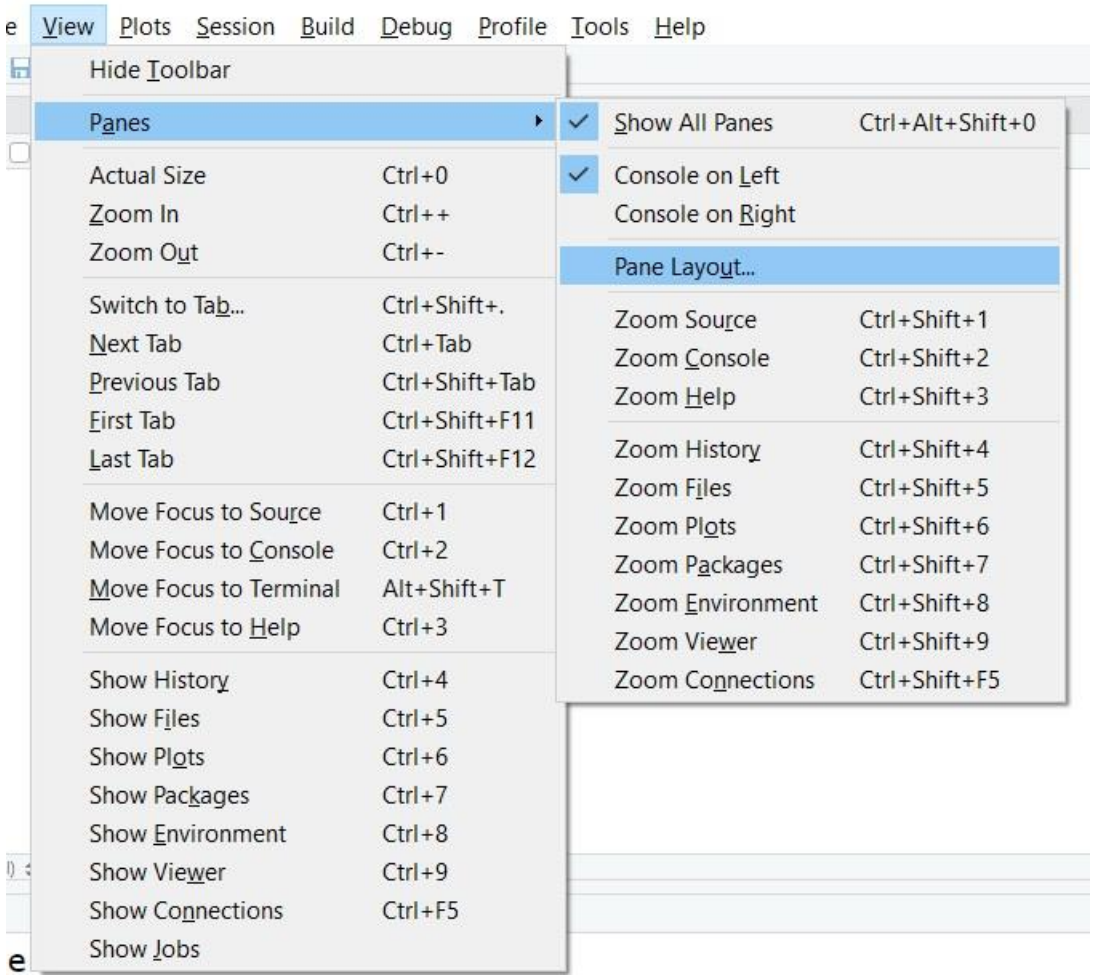
Usage

```
hist(x, ...)

## Default S3 method:
hist(x, breaks = "Sturges",
     freq = NULL, probability = !freq,
     include.lowest = TRUE, right = TRUE,
     density = NULL, angle = 45, col = NULL, border = NULL,
```

ide introduction

Viewing panes if closed/missing



ide introduction

Current Env Installed Packages

Files | Plots | Packages | Help | Viewer

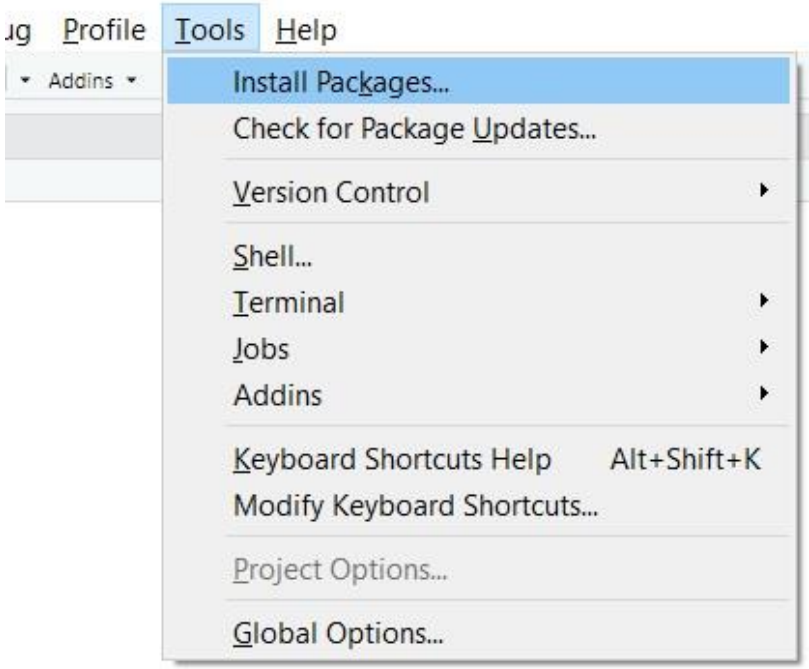
Install | Update

Name	Description	Version
chron	Chronological Objects which can Handle Dates and Times	2.3-47
colorspace	Color Space Manipulation	1.2-7
data.table	Extension of Data.frame	1.9.6
DBI	R Database Interface	0.5-1
dichromat	Color Schemes for Dichromats	2.0-0
digest	Create Compact Hash Digests of R Objects	0.6.10
dplyr	A Grammar of Data Manipulation	0.5.0
e1071	Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien	1.6-7
ggplot2	An Implementation of the Grammar of Graphics	2.1.0
ggpubr	'ggplot2' Based Publication Ready Plots	0.1.0
ggrepel	Repulsive Text and Label Geoms for 'ggplot2'	0.6.3
ggsci	Scientific Journal and Sci-Fi Themed Color Palettes for 'ggplot2'	1.5
gtable	Arrange 'Grobs' in Tables	0.2.0
labeling	Axis Labeling	0.3
lazyeval	Lazy (Non-Standard) Evaluation	0.2.0
magrittr	A Forward-Pipe Operator for R	1.5
moments	Moments, cumulants, skewness, kurtosis and related tests	0.14
munsell	Utilities for Using Munsell Colours	0.4.3

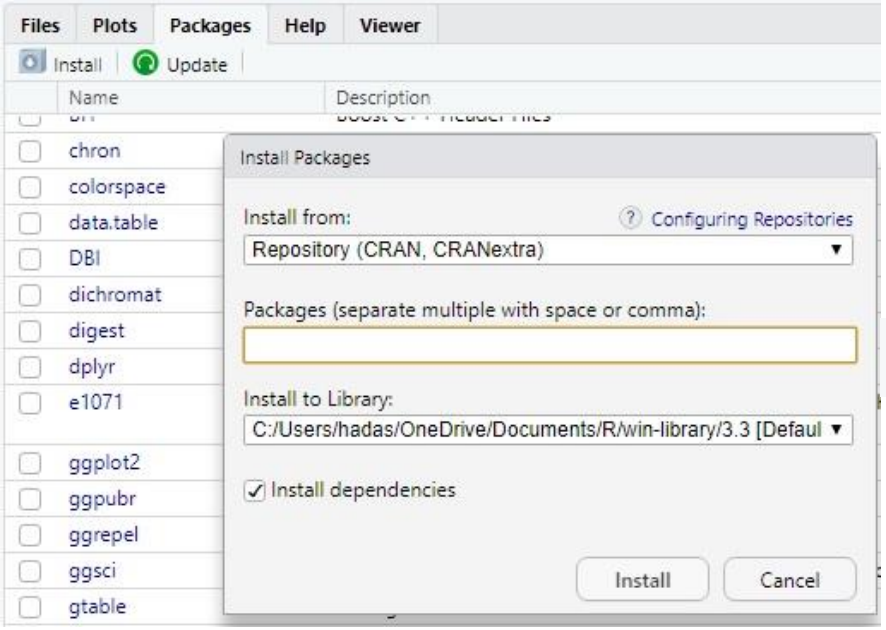
ide introduction

Package installation

From Toolbar



From Packages window



From Console

```
> install.packages("vctrs")
```

Basic concepts
Initiating a script

- Set working directory to current directory using setwd() function

```
setwd("D:/Hadas/courses/…")
```

- Clear the interpreter's working memory:

```
rm(list = ls())
```

- Import libraries you intend to use in the code, e.g.:

```
library(ggplot2)
library(tidyr)
library(data.table)
```

Basic concepts
standard functions checkout

1. Loading a csv file into an R dataframe:

```
> df = read.csv(FILEPATH)
```

Tutorial1.R × df ×

Filter

	sepal_length	sepal_width	petal_length	petal_width	species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3.0	1.4	0.1	setosa
14	4.3	3.0	1.1	0.1	setosa
15	5.8	4.0	1.2	0.2	setosa
16	5.7	4.4	1.5	0.4	setosa

2. Extracting a feature from a dataframe:

```
> y = df$sepal_width
```

```
y                num [1:150] 3.5 3 3.2 3.1 3.6 3.9 3.4...
```

Basic concepts
standard functions checkout

3. **Calculate mean** with standard R function ‘mean’
Control precision with R function ‘round’
> m = round(mean(y), 4)

m	3.0573

4. Now calculate mean analytically
Control precision with R function ‘round’

m_calc	3.0573

4. Compare between the standard R function and the analytic calculation:
> m_bool = m==m_calc

m_bool	TRUE

- Repeat this exercise for the **standard deviation** of y (use sd) and the **standard error of the mean** (no built-in function, use equation)

Probability distribution plots

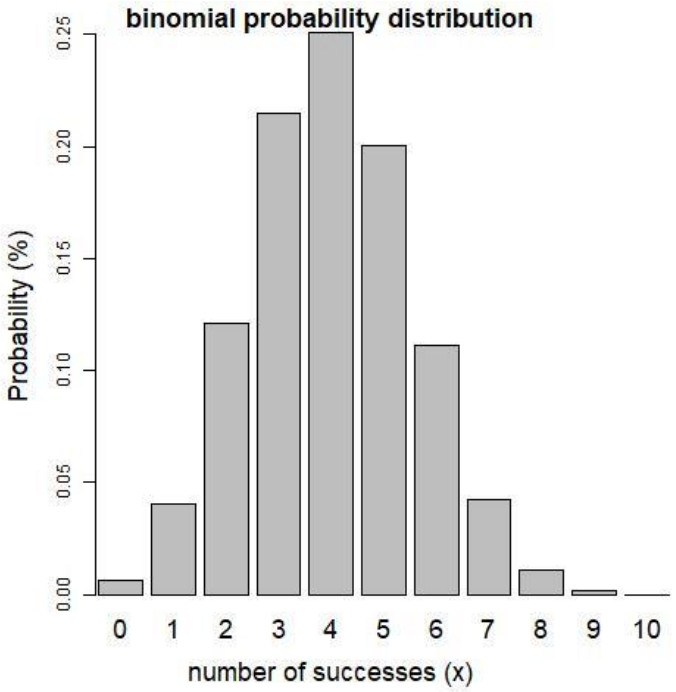
Q1. Plot a **binomial probability** distribution with 10 samples and p=0.4 chance of success

- Span x axis using seq() function
- Create binomial distribution using dbinom() function
- Plot binomial distribution as a function of trial using barplot() function

p.s.

- You can save the plot using:

jpeg(file=FILEPATH) before the barplot() and dev.off() function thereafter



Probability distribution plots

Q2. Given a binomial probability distribution with n=10 samples and p=0.4 chances of success

1. Calculate the probability for 5 successes (P(x=5))
2. Calculate the expectation value (μ)
3. Calculate the standard deviation (σ)

At home:

- Plot a Poisson probability distribution of up to 15 possible events per hour, with Poisson variable λ=3
- Calculate the probability for 5 events per hour (P(x=5))

frequency (histogram) plots

Aim: display the frequency distribution of Iris petal width

How: use ‘hist’ function to get variable distribution

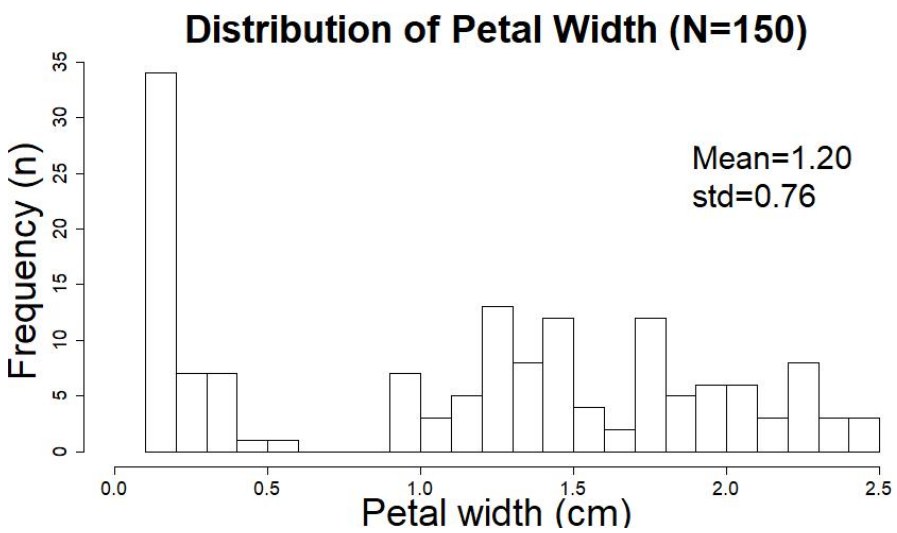
1. Get petal_width from df and assign it to y variable
2. Calculate petal_width mean using **mean()**
3. Calculate petal_width standard deviation using **sd()**
4. Plot frequency distribution using **hist()** function

Notice: ‘breaks’ argument defines number of bins

‘xlim’, ‘ylim’ define the graph axes scales

xlab, ylab, main define graph titles

5. Add legend with calculated sample mean and standard deviation



frequency (histogram) plots

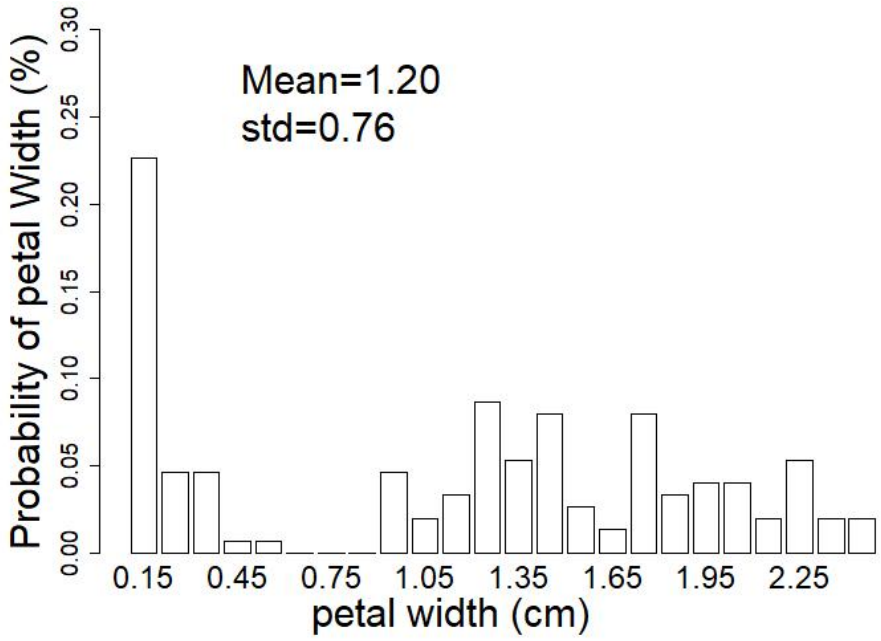
Bonus

Aim: display the probability distribution of Iris petal width

How: use 'hist' function output to plot the *normalized* distribution

Hist() returns the sample distribution parameters.
Plot the **probability distribution of petal width** by normalizing the counts by the overall count

- Assign the variable "counts" from histogram output to "freq_counts" variable and divide it by the vector sum (normalization)
- Assign "mids" from histogram output to "x" variable
- Plot the probability distribution of "freq_counts" with barplot() function



Quantile plots (boxplot)

- Plot quantile representation using boxplot() function Of sepal and petal width.
- Use ‘names=’ argument to control x axis labels and ‘cex.axis =‘ argument to control axis font size.

- Display vector quantiles using summary() function

```
> summary(df$sepal_width)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.800   3.000   3.057   3.300   4.400
```